TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Big Data - charakteristika a zpracování nestrukturovaných dat

Diplomová práce

Studijní program: N6209 – Systémové inženýrství a informatika
Studijní obor: 6209T021 – Manažerská informatika

Autor práce: **Bc. Bára Smolová**
Vedoucí práce: Ing. Dana Nejedlová, Ph.D.

Liberec 2018

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Akademický rok 2017 / 2018

Zadání diplomové práce

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bára Smolová, Bc.**
Osobní číslo:	E16000391
Studijní program:	N6209 Systémové inženýrství a informatika
Studijní obor:	N6209T021 – Manažerská informatika
Zadávající katedra:	katedra informatiky
Vedoucí práce:	Ing. Dana Nejedlová, Ph.D.
Konzultant práce:	Ing. Štěpán Aubrecht Trask solutions a. s., BI Architekt

Název práce: **Big Data - charakteristika a zpracování nestrukturovaných dat**

Zásady pro vypracování:

1. Big Data a jejich využití.
2. Možnosti zpracování nestrukturovaných dat současnými nástroji.
3. Zpracování nestrukturovaných dat vybranými nástroji.
4. Zhodnocení řešení.

Seznam odborné literatury:

HOFMAN, Markus a Andrew CHISHOLM. 2015. *Text Mining and Vizualization: Case Studies Using Open-Source Tools*. Boca Raton: Chapman and Hall/CRC. ISBN 978-1482237573.

AGGARWAL, Charu C. 2015. *Data Mining: The Textbook*. Berlin: Springer. ISBN 978-3-319-14141-1.

MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. 2014. *Big Data*. Brno: Computer Press. ISBN 978-80-251-4119-9.

HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. 2015. *Big Data a NoSQL databáze*. Praha: GRADA Publishing. ISBN 978-80-247-5466-6.

PROQUEST. 2017. *Databáze článků ProQuest* [online]. Ann Arbor, MI, USA: ProQuest. [cit. 2017-09-28]. Dostupné z: http://knihovna.tul.cz/

Rozsah práce:	65 normostran
Forma zpracování:	tištěná / elektronická
Datum zadání práce:	31. října 2017
Datum odevzdání práce:	31. srpna 2019

Technická univerzita v Liberci Ekonomická fakulta

L.S.

prof. Ing. Miroslav Žižka, Ph.D.
děkan Ekonomické fakulty

doc. Ing. Klára Antlová, Ph.D.
vedoucí katedry

V Liberci dne 31. října 2017

Prohlášení

Byla jsem seznámena s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědoma povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracovala samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum: 13. 12. 2018

Podpis: Smeluí

Poděkování

Ráda bych poděkovala, vedoucí mé diplomové práce Ing. Daně Nejedlové, Ph.D. především za ochotu, podnětné rady, připomínky a návrhy. Dále mé poděkování patří mé rodině a mým přátelům za obrovskou podporu, kterou mi byli po celou dobu mého studia.

Anotace

Diplomová práce se zabývá metodami zpracování Big dat s bližším zaměřením na data nestrukturovaná. Teoretická část práce je zaměřena na charakteristiku Big dat, specifikaci zdrojů nestrukturovaných dat a na popis dostupných metod pro zpracování dat z těchto zdrojů prostřednictvím v současné době existujících nástrojů, dále se teoretická část práce zaměřuje na popis aplikačních oblastí těchto metod a také na přínosy zpracování Big dat. V praktické části práce jsou nestrukturovaná data z různých zdrojů zpracována vybranými nástroji, které jsou popsány v teoretické části této práce. Na závěr je definován a zhodnocen výsledek zpracování nestrukturovaných dat.

Klíčová slova

Analýza sentimentu, Big data, detekce podvodů, nestrukturovaná data, datové proudy, strojové učení, umělá inteligence, analýza dat v reálném čase, zpracování přirozeného jazyka.

Annotation

Thesis name: Big Data - characteristics and processing of unstructured data

The diploma thesis focuses on methods of processing Big Data with a closer look at unstructured data. The theoretical part is concentrated on characterization of Big Data, specification of unstructured data sources and description of available methods for data processing from these sources through the currently available tools. Furthermore this thesis aims on the description of the scope in which these methods are used and also on the benefits of processing Big Data. In the practical part of the thesis unstructured data are processed by selected tools, which are described in the theoretical part of this thesis. Finally the result of unstructured data processing is defined and evaluated.

Key words

Sentiment Analysis, Big data, Fraud detection, Unstructured data, Streaming data, Machine Learning, Artificial Intelligence, Real Time Data Analysis, Natural Language Processing.

Obsah

Seznam ilustrací

Seznam tabulek

Seznam použitých zkratek

AAM	Active Appearance Model (Aktivní vzhledový model)
AI	Artificial Intelligence (Umělá inteligence)
ASM	Active Shape Model (Aktivní tvarový model)
ASR	Automatic Speech Recognition (Automatické rozpoznávání řeči)
ASU	Automatic Speech Understanding (Automatické pochopení řeči)
CAGR	Compound Annual Growth Rate (Ukazatel míry návratnosti investic po dobu trvání investice)
CEP	Complex Event Processing (Komplexní zpracování událostí)
CM	Condition monitoring (Monitorování dat dle stanovených podmínek)
CPB	Constant Percentage Bandwifth (Konstatní procentní šířka pásma)
DAST	Dynamic Application Security Testing (Dynamické bezpečnostní testování aplikací)
EDI	Electronic Data Interchange (Elektronická výměna dat)
ESP	Event Stream Processing (Zpracování událostí – proudů dat)
FACS	Facial action units detection (Detekce obličejových akčních jednotek)
FDS	Fraud Detection System (Systém pro detekci podvodů)
FFT	Fast Fourier Transform (Rychlá Fourierova transformace)
FPS	Frame per second (Obrázků na sekundu)
HDP	Hrubý domácí produkt (Ukazatel)
IBM	International Business Machines (Společnost)
ICT	Information and Communication Technology (Informační a komunikační technologie)
IoT	Internet of Things (Internet věcí)
IT	Information Technology (Informační technologie)
JSON	JavaScript Object Notation (Datový formát)
kb/s	Kilobit per second (Kilobitů za sekundu, 1024 bps)
kHz	(Kilohertz, 1 000 Hz)
LBP	Local Binary Pattern (Lokální binární vzor)
LSA	Latent semantic analysis (Latentní sémantická analýza)
MLP	Multi Layer Perception (Vícevrstvé neuronové sítě)
MM	Mathematical Morphology (Matematická morfologie)

NER	Named Entity Recognition (Rozpoznávání pojmenovaných entit)
NLP	Natural Language Processing (Zpracování přirozeného jazyka)
PB	Petabyte (Petabajt, 10^{15} bajtů)
PCA	Principal Component Analysis (Analýza hlavních komponent)
POS	Parts of Speech Tagging (Označování částí řeči)
RBF	Radial Basis Function (Radiální funkce báze)
SAST	Static Application Security Testing (Statické bezpečnostní testování aplikací)
SIEM	Security Information and Event Management (Management bezpečnostních informací a událostí)
SOM	Self Organizing Map (Samoorganizující se sítě)
SRS	Speech System Recognition (Systém pro rozpoznání řeči)
STT	Speech to Text (Převod řeči na text)
STFT	Short Time Fourier Transformation (Krátkodobá Fourierova transformace)
SVM	Support Vector Machines
TB	Terabyte (Terabajt, 10^{12} bajtů)
TTS	Text to Speech (Převod textu na řeč)
XML	eXtensible Markup Language (Rozšiřitelný značkovací jazyk)
ZB	Zettabyte (Zetabajt, 10^{21} bajtů)

Úvod

Diplomová práce se zabývá charakteristikou Big dat s bližším zaměřením na zpracování nestrukturovaných dat coby v současné době nepříliš využívaného zdroje dat pro podporu rozhodování. Tato práce navazuje na bakalářskou práci autorky a dále ji rozšiřuje.

Problematika zpracování nestrukturovaných dat je v současné době velmi aktuální a stále se vyvíjí nové metody a způsoby, jak lze získat z těchto dat kýžené výsledky.

Cílem práce je zmapovat současný stav problematiky zpracování nestrukturovaných dat, včetně popisu v současné době využívaných metod a postupů zpracování, a porovnat tento stav se stavem, který popisovala bakalářská práce autorky. Dílčím cílem práce je prakticky ukázat zpracování audiovizuálního záznamu vybranými prostředky s využitím analýzy sentimentu.

Diplomová práce je rozdělena na dvě stěžejní části. Teoretická část práce je zaměřena na charakteristiku Big dat s bližším zaměřením na data nestrukturovaná. Jsou zde popsány současné metody zpracování nestrukturovaných dat v závislosti na jejich formátu, dále jsou zde definovány přínosy zpracování nestrukturovaných dat.

Praktická část práce navazuje na teoretickou část práce. V této části jsou popsány nástroje pro zpracování nestrukturovaných dat a je zde názorně předvedeno praktické užití vybraných nástrojů pro zpracování audiovizuálního záznamu.

V závěru diplomové práce jsou vyhodnoceny přínosy využití nástrojů a metod pro zpracování Big dat a také jsou interpretovány výsledky provedené analýzy.

1 Analýza a zhodnocení současného stavu problematiky

Současná moderní společnost je bezesporu obklopena nepřeberným množstvím chytrých zařízení, která by nám měla umožnit zefektivnit náš čas či si zjednodušit současnou hektickou dobu. Tato zařízení jsou pro velkou část populace nepostradatelná.

Automobily sbírají informace o svých pasažérech a cestě, spotřebě paliva, telefonních hovorech či o stylu jízdy. Chytré hodinky své uživatele ráno probudí v závislosti na jejich spánkovém cyklu, celý den měří tep, počet kroků a trasu, kterou uživatel šel pěšky/běžel/ujel, a také počet kalorií, které danou aktivitou spálil. V současné době je běžné trávit několik hodin denně na internetu. Prostřednictvím sociálních sítí či emailů probíhá velká část lidské komunikace, dochází zde ke sdílení názorů a soukromých fotografií. Jak uvádí ve své práci Šmahaj (2014), lidé, kteří využívají tyto technologie, tak často i nevědomky sdílí své osobní a velice soukromé informace, které mohou být využity, a často i zneužity jinými subjekty (Kasík, 2017).

Vzhledem k růstu množství chytrých zařízení a vývoji moderních technologií dochází v posledních letech k exponenciálnímu nárůstu množství dat, jak uvádí Helms (2015) a jak popisuje bakalářská práce autorky (Smolová, 2016).

Společnost IDC předpovídá, že do roku 2025 se zvýší množství dat na 163 ZB (zetabajtů), což je desetinásobek vygenerovaných dat v roce 2016 (v tomto roce bylo vygenerováno 16,1 ZB). (Reinsel a kol., 2017)

Množství strukturovaných dat roste pozvolna, zatímco množství nestrukturovaných dat roste stále rychlejším tempem. Za poslední dva roky se změnila nejen rychlost nárůstu množství dat, ale i skladba nestrukturovaných dat. Vzrostlo především množství senzorových dat a dat z webu, především dat ze sociálních sítí. Společnosti se více soustředí na analýzu senzorových a uživatelských dat za účelem podrobné profilace nejen společnosti a vybraných skupin, ale i jednotlivců. Tato profilace poskytne podrobný pohled na současné či potencionální zákazníky, z hlediska např. náboženského či politického vyznání, preferencí a zájmů, vztahů mezi dalšími subjekty a zákazníky, zaměstnání, majetku, věku, pohlaví

apod. Vzhledem k tomuto faktu jsou stále vyvíjeny nové nástroje a metody, které zefektivní analýzu těchto dat, jak popisují například Stefan Stieglitz a kol. (2018), Jun Mi a kol. (2017) a další.

Zpracování Big dat využívají prakticky všechna odvětví – bankovnictví, výroba, vzdělávání, zdravotnictví, obchod a veřejný sektor.

Data, která poskytují uživatelé o své osobě i provozní data, představují obrovské bohatství. Tento fakt si společnosti stále více uvědomují a jsou ochotné investovat nemalé částky na sběr, zpracování a vizualizaci nejen uživatelských dat, ale také dat o svém provozu (výroba, obchod, atd.). Data lze v dnešní době považovat za velmi váženou komoditu. (Ishikawa, 2015)

Zpracování tzv. Big dat je v současné době poměrně běžné, pokud tedy smýšlíme o Big datech pouze jako o velkých objemech dat. Problém nastává v tom momentě, kdy považujeme pojem Big data za data nejen velká, ale i různorodá – v různých formátech a z různých zdrojů, jak popisuje Marr (2015).

Dle studia literatury provedeného autorkou této práce se zdá, že v současnosti neexistuje žádná aplikace/platforma, která by zvládla analyzovat všechny druhy dat – strukturovaná i nestrukturovaná data z různých zdrojů (sociální sítě, logové záznamy, JSON soubory atd.) a zároveň nad těmito daty provést v současné době všechny dostupné druhy analýzy (např. analýzu výrazu tváře apod.). Společnosti se musí prozatím spokojit s pestrou nabídkou aplikací, které se zaměřují na konkrétní druhy dat a na vybraný druh analýzy (analýza sentimentu, detekce podvodů, analýza vztahů a souvislostí subjektů apod.).

Zdá se, že se dnešní svět pohybuje každým dnem rychleji. Klade se stále větší a větší důraz na rychlost, s jakou jsou získány odpovědi na otázky. V určitých odvětvích, jako je například bankovnictví či bezpečnost, jsou odpovědi potřeba nejlépe ihned. Je kladen důraz nejen na rychlé zpracování dat, ale také i na zpracování dat v reálném čase. V tomto případě jsou data ihned po vytvoření zpracována a vizualizována tak, aby přinesla kýžené výsledky v reálném čase. Tato analýza je velmi náročná z hlediska technologie, ale předchází velkým škodám a bezpečnostním rizikům. Touto analýzou se zpracovávají data bankovních systémů

(transakční data) či bezpečnostních systémů – monitoring logových záznamů zařízení (např. kamer). (Kudyba, 2014)

Velmi často bývají špatně označována za Big data i data, která ve skutečnosti nesplňují podmínky (objem, různorodost apod.) a Big daty nejsou, jak popisuje ve svém článku i Marr (2015).

Trend zpracování Big dat se průběhu let mění. Tuto změnu způsobuje vývoj technologií využívaných pro zpracování Big dat, např. využívání cloudových služeb, rozvoj tzv. blockchain databází, ale také i rozvoj analytických metod založených na sofistikovaných statistických algoritmech, např. vývoj deep learning (metoda strojového učení), optimalizace dosavadních algoritmů apod. (Shaffer, 2017)

Trendem roku 2019 pro zpracování Big dat bude zpracování Dark dat, Quantum Computing (kvantové výpočty), Edge Computing (změna technologie síťového přenosu dat) a také větší zaměření na data z IoT (Internet věcí – Internet of Things), jejichž množství má vzrůst až na úroveň CAGR 28,5 % (Some, 2018). Zatímco trendem roku 2016 bylo zaměření na bezpečnost Big dat, rozvoj a užití strojového učení (především rozvoj hlubokého učení) a rozvoj technologií pro ukládání Big dat (všech formátů struktur dat), jako např. NoSql databáze, Hadoop apod. (Shah, 2015 a Pandit, 2016 a také Rijmenam, 2015)

Současná doba již umožňuje jednoduše zpracovávat data z různých aplikací, sociálních sítí apod. Jsou vyvíjeny specializované nástroje např. na klasifikaci tzv. youtuberů. (Del Vecchio a spol., 2017)

Vzhledem k nárůstu výpočetních kapacit současných strojů dochází i k nahrazení standardních pravděpodobnostních modelů z oblasti strojového učení s učitelem za tzv. hluboké učení (deep learning), které aplikuje strojové učení s učitelem i bez něj. Stále častěji jsou využívány umělé neuronové sítě, které umožňují distribuované paralelní zpracování dat. (Veselovská, 2017)

Z hlediska zabezpečení Big dat je současným trendem vznik technologií zabezpečujících data tak, aby je bylo možno analyzovat bez zneužití. (Sušický, Mikeška, 2015)

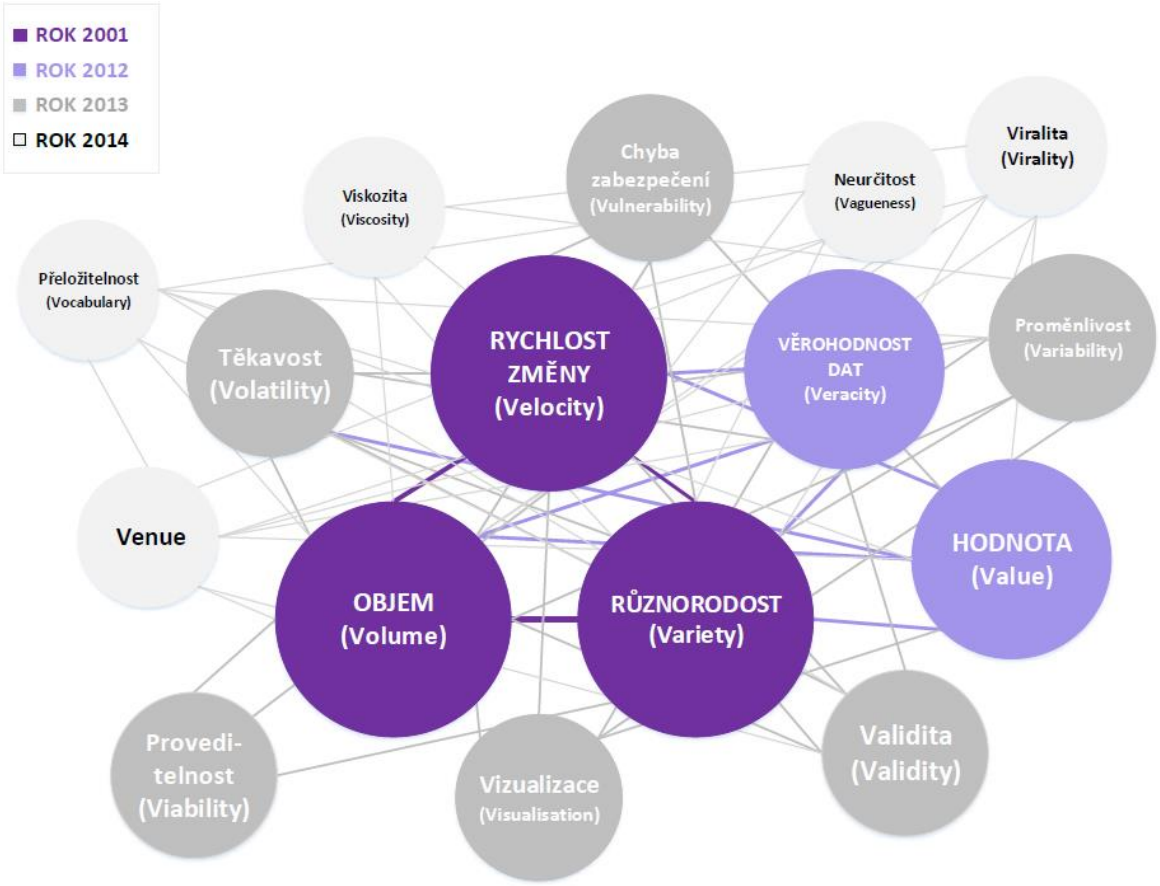
2 Big Data

Co to vlastně jsou Big data? Jedná se pouze o velké objemy dat? Či o různorodá data? Big data lze definovat prostřednictvím tzv. V-parametrů. Dle počátečních písmen parametrů, která by tato data měla splňovat např. **V**olume (objem), **V**elocity (rychlost změny), **V**ariety (různorodost) a další. V současné době stále probíhají dohady o počtu parametrů, které by měla data splňovat, aby se dala označit za Big data. Dosud neexistuje jednotná definice – obvykle se liší počet charakteristik (počet V-parametrů).

2.1 Definice Big dat a její vývoj

Definice se Big dat má za sebou velmi dlouhý vývoj, postupně byly přidávány nové charakteristiky.

V této práci jsou popsány pouze charakteristiky, které jsou uznávány a všeobecně přijímány. Existuje nespočet dalších charakteristik, které ovšem z nějakého důvodu nejsou příliš známé, či jsou rozporovány.



Obrázek 1: Vývoj definice Big dat v letech
Zdroj: Přeloženo a modifikováno z (Shafer, 2017, 15. 02. 2018)

3V – Volume, Velocity, Variety (2001)

První charakteristiky pro definici Big dat vytvořil již v roce 2001 Douglas Laney, datový specialista ze společnosti Meta Group (která byla později připojena ke společnosti Gartner), ve svém výzkumu, jak popisuje společnost Gartner ve svém článku (Gartner, 2005). Tato studie byla zaměřena na data management a autor v ní definuje Big data prostřednictvím 3V-parametrů (Volume, Velocity, Variety). (Laney, 2001)

Tyto stěžejní 3 parametry a další rozvoj definice Big dat znázorňuje Obrázek 1.

Volume (objem dat)

Objem je nejdůležitějším parametrem Big dat – dle svého velkého objemu jsou i Big data pojmenována, viz kapitola 2.3 Vznik pojmu Big data.

Za Big data lze považovat data o objemu v řádech PB (petabajtů), ale v České Republice lze hovořit i o řádech v TB (terabajtech). (Sušický, Mikeška, 2015)

+ Velocity (rychlost zpracování)

Dalším parametrem je i rychlost zpracování nově generovaných dat. Vzhledem ke stále rychlejšímu růstu množství dat je nutné data zpracovávat co nejrychleji.

Některá data vyžadují zpracování v reálném čase – jedná se tzv. o streamovaná data (transakce, mobilní zařízení apod.), která popisuje kapitola 3.4.1 Zpracování logových záznamů.

+ Variety (různorodost)

Různorodost je jedním ze základních parametrů pro Big data. Právě jejich různorodost činí z analýzy těchto dat velmi složitou disciplínu, která má stále velký prostor pro další rozvoj.

V základním dělení dle struktury rozdělujeme data na strukturovaná a nestrukturovaná. Často se také můžeme setkat s pojmem semistrukturovaná data. Tato problematika je podrobněji popsána v kapitole 2.2 Různorodost dat.

V dělení dle zdroje dat lze dělit data na externí a interní.

4V – Volume, Velocity, Variety, Veracity (2011)

Vědci ze společnosti IBM v roce 2011 přidali charakteristiku Veracity (důvěryhodnost dat) pro lepší a přesnější definování Big dat. (IBM, 2014)

Volume, Velocity a Variety – viz definice 3V

> **+ Veracity (věrohodnost dat) –** Big data pocházejí z různých zdrojů. Z tohoto důvodu je nutné počítat s určitou úrovní nevěrohodnosti dat. Data ze sociálních sítí, internetu apod. mohou být nekonzistentní, neúplná – nevěrohodná. Tento parametr způsobuje zkreslení výsledků analýz.

4V a C – Volume, Velocity, Variety, Veracity, Complexity

Společnost SAS věří, že kromě V-parametrů je nutné přidat i C-parametr – Complexity. (SAS Institute Inc., 2018)

Volume, Velocity, Variety, Veracity

> **+ Complexity (složitost, komplexnost)**
>
> Data pocházejí z různých nezávislých zdrojů. Úkolem při zpracování Big dat je tato data provázat, sloučit, vyčistit porovnat a transformovat. Je nutné spojit a korelovat vztahy mezi daty – získat vazby.

5V – Volume, Velocity, Variety, Veracity, Value (2012)

Další definice prostřednictvím 5V-parametrů. Největším propagátorem a také tvůrcem této definice je Bernard Marr (datový expert a autor knih o Big datech), který definoval pátý parametr – Value. (Marr, 2014)

Volume, Velocity, Variety a Veracity – viz definice 4V

+ Value (hodnota)

Big data by měla představovat určitou hodnotu pro daný byznys. Analýzou Big dat by společnosti měly získat kýžené výhody (obchodní výhody ve formě nových faktů apod.) nad svou konkurencí, která tyto analýzy neprovádí.

5V – Volume, Velocity, Variety, Value, Viability (2013)

Tuto definici uvádí Biehn (2013).

Volume, Velocity, Variety, Value

+ Viability (proveditelnost)

Výsledky Big data analýzy by se měly co nejvíce přibližovat realitě.

6V – Volume, Velocity, Variety, Veracity, Value, Vulnerability (2017)

Společnost Experian zastává názor, že Big data je možné definovat prostřednictvím 6V. (Experian, 2017)

Volume, Velocity, Variety, Veracity, Value

+ Vulnerability (chyba zabezpečení)

Tento parametr zohledňuje fakt, že stále více lidí si uvědomuje, že firmy jsou schopné své uživatele ovlivňovat právě díky znalosti jejich zvyklostí, preferencí apod. (Marr, 2016).

7V – Volume, Velocity, Variety, Variability, Veracity, Visualisation, Value (2013)

Ve svém příspěvku takto definuje Big data například McNulty (2014).

Volume, Velocity, Variety, Veracity, Value

+ Variability (proměnlivost, variabilnost)

Tato charakteristika je způsobená změnou významů a vývoje významu dat (např. slov v textu). Projevuje se především při analýze sentimentu (pozitivní/negativní emoce – ironie).

+ Visualisation (vizualizace dat)

Data, která jsou zpracována, ale nejsou vizualizována, neposkytnou potřebné výsledky. Obtížnost vizualizace je jedním z parametrů pro analýzu Big dat.

Současné nástroje čelí technickým problémům – omezená paměť, dlouhý čas odezvy, funkčnost a další, jak uvádí ve svém článku Firican (2017).

8V – Volume, Velocity, Variety, Veracity, Visualisation, Value, Viscosity, Virality (2014)

Společnost M-Brain definuje Big data prostřednictvím 8V-parametrů. (M-Brain, 2018)

Volume, Velocity, Variety, Veracity, Value, Visualisation

+ Viscosity (viskozita dat)

Velmi úzce spojené s Velocity (rychlost změny). Tento parametr odkazuje na fakt, jak moc je složité zpracovávat Big data – různé druhy dat, zpracování v reálném čase, složitost požadovaného zpracování.

+ Virality (viralita)

Tento parametr popisuje, jak rychle se data šíří mezi subjekty. M-Brain (2018)

10V – Volume, Velocity, Variety, Variability, Veracity, Visualisation, Value, Viscosity, Validity, Volatility (2014) a další

Tutu definici uvádí například Firican (2017).

Volume, Velocity, Variety, Variability, Veracity, Visualisation, Value, Viscosity

+ Validity (validita dat)

Parametr je velmi podobný Veracity (důvěryhodnosti dat). Tento parametr se zaměřuje na kvalitu podkladových dat, resp. zda jsou data přesná a kvalitní, jak popisuje ve svém článku Firican (2017).

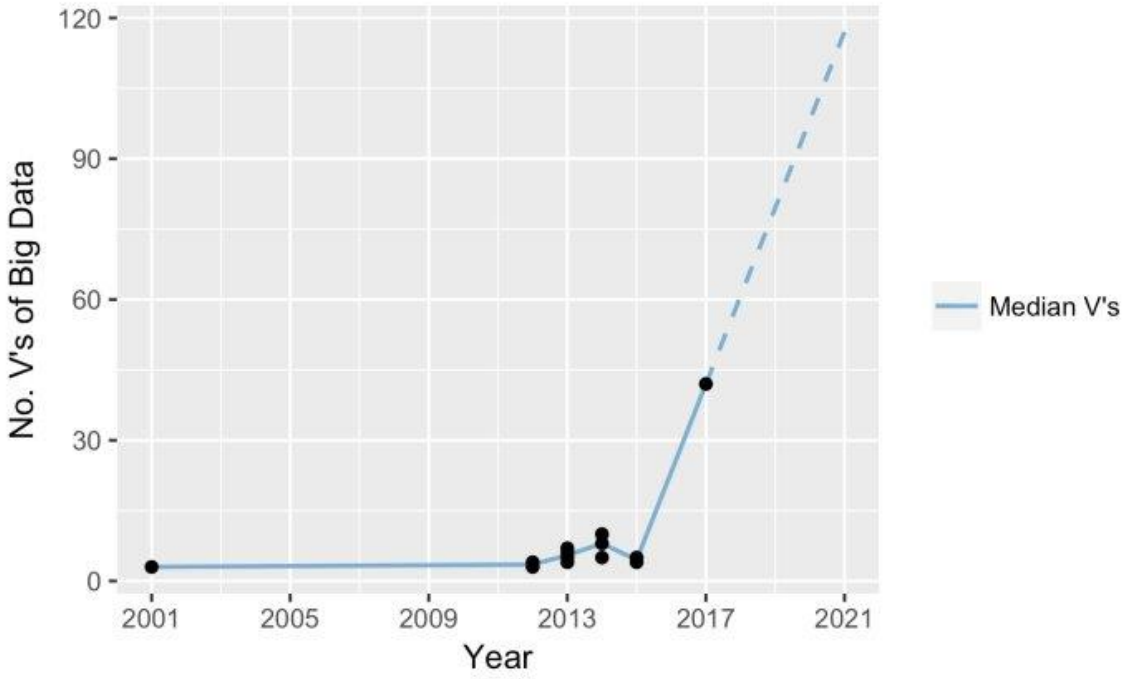
Datoví vědci stráví přípravou dat až 80 % času, z toho až 60 % času zabere čištění a organizování dat. (Gil Press, 2016)

+Volatility (těkavost)

Tento parametr řeší problematiku, jak dlouho data uchovávat.

S určitou dávkou humoru pojednává o růstu množství V-parametrů i článek (Shafer, 2017), kde autor uvádí celkem 42 V-parametrů, viz Obrázek 2 níže.

Vývoj počtu V-parametrů pro definici Big dat



Obrázek 2: Vývoj počtu V-parametrů pro definici Big dat
Zdroj: (Shafer, 2017, 15. 02. 2018)

V současnosti nelze určit, jaké charakteristiky jsou opravdu nutné pro správnou definici Big dat. V této práci budeme považovat za Big data taková data, která splňují základní definici 5V (Volume – objem, Velocity – rychlost změny, Variety – různorodost, Value – hodnota pro byznys), definovanou Bernardem Marrem. (Marr, 2014)

2.2 Různorodost dat

Jak je již popsáno výše, tak různorodost dat (Variety) je důležitým parametrem dat, která lze označit za Big data.

Data dělíme na data strukturovaná (tabulky, adresní údaje, rejstříky, relační databáze apod.), nestrukturovaná (volný text, audio, grafiku, video a další) a semistrukturovaná (data ve formátu XML, JSON, EDI a další). Semistrukturovaná data budou dále řazena pod data strukturovaná, jelikož proces zpracování je téměř identický.

V této práci nebude tato problematika dále rozebírána s odkazem na bakalářskou práci autorky (Smolová, 2016), kde je různorodost dat (kapitola 2.1 Různorodost dat) podrobně popsána.

V bakalářské práci autorky (Smolová, 2016) jsou popsány i zdroje těchto dat, ale pro snazší pochopení problematiky zpracování nestrukturovaných dat je vhodné podrobnější popis zdrojů dat, rozdělený do dvou základních kategorií.

Data vytvořená lidskou interakcí s lidmi nebo zařízeními

- *Sociální sítě (např. Facebook, Youtube, Twitter, blogy, Instagram apod.)*

  Uvedené sítě obsahují neuvěřitelné množství informací. Současná společnost zde sdílí své názory, zážitky, fotografie i osobní informace, to znamená velké datové bohatství.

  Nyní již existuji kvalitní analytické nástroje, které pomáhají tato data změnit v informace, které firmy požadují. Na sociálních sítích je velmi často využívána analýza sentimentu – rozpoznání emoce/postoje.

  Na fotografiích z těchto sítí je testována umělá inteligence.

- *Elektronická pošta (Messenger zprávy a další)*

  Elektronická komunikace mezi uživateli může být také zpracována za účelem identifikace vztahů mezi subjekty.

- *Záznamy o aktivitě uživatelů na internetu (Web Analytics)*

  Velké množství cenných dat je získáváno prostřednictvím analýzy chování/aktivity uživatelů na internetu, jak popsal ve svém výzkumu pro konferenci Big data i Benjamins (2014). Z chování uživatelů lze rozpoznat následující, např. o jaký výrobek má uživatel zájem, pohlaví, věk, kolikrát uživatel navštívil daný web. Získat informace lze i prostřednictvím cookies.

  Často využívána je webová analýza ve formě služby Google Analytics.

- *Data z mobilních zařízení*

  Mobilní zařízení mají přístup k biometrickým datům uživatele (jedinečné, trvalé a měřitelné znaky uživatele), a to jak k fyziologickým (obličej, otisk prstu, duhovka apod.), tak i k behaviorálním datům (gesta, tempo psaní, hlas). (Koong a kol., 2014)

  Tzv. „wearables“ (chytré hodinky, sportovní hrudní pásy apod.) měří zdravotní stav (např. srdeční tep) jedince a předávají tato data dalším zařízením či aplikacím.

- **Strojově vytvořená data**

  - *Strojová data (logové záznamy aplikace, zařízení)*

    Společnosti provádějí sběr, analýzu a kontrolu logových záznamů a kontrolu svých systémů, zařízení či aplikací. Tato data jsou analyzována za účelem monitoringu.

    Strojová data jsou často zpracovávána v reálném čase. V případě zachycení předem definovaných událostí lze okamžitě zasáhnout a předcházet možných škodám.

- *Senzorová data (data z chytrých zařízení)*

  Senzory jsou stále častěji využívány pro různá odvětví. Jak uvádí Jeyanthi (2016), senzorová data úzce souvisí s tzv. IoT. Jak je již popsáno výše, dnešní moderní společnost využívá nepřeberné množství nových technologií a zařízení – chytré hodinky, chytré mobilní telefony, chytré automobily, lednice, pračky a další. Tato zařízení snímají pomocí senzorů počínání svých uživatelů, data zpracovávají a vyhodnocují. Využitím daných dat lze získat téměř kompletní přehled o určitém uživateli (data o jeho zdraví, preferencích, životním stylu, zaměstnání i rodině).

  Senzorová data nám také monitorují např. dopravní situaci, stav pacienta v nemocnici, počasí apod. Mají velmi široké spektrum využití.

2.3 Vznik pojmu Big data

Poprvé tento termín použili vědci Michael Cox a David Ellsworth ve svém výzkumu prováděném pro společnost NASA. (Friedman, 2012)

Termín byl poprvé použit v roce 1997 v souvislosti s problematikou vizualizace velkého množství dat. Konkrétně ve své studii o problému napsali:

„Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data.“ (Cox a Ellsworth, 1997)

Volný překlad: „Vizualizace vytváří velkou výzvu pro počítačové systémy: datové soubory jsou příliš velké, přesahují kapacitu hlavní paměti, lokálního disku, a dokonce i vzdáleného disku. Nazvali jsme tento problém jako Big data.” (Cox a Ellsworth, 1997)

Tento termín ale proslavila až práce „Big-Data Computing – Creating revolutionary breakthroughs in commerce, science, and society“ publikovaná v prosinci roku 2008 prominenty v IT odvětví – Rendal E. Bryant (Carnegie Mellon University), Randy H. Katz (University of California, Berkeley), Edward D. Lazowska (University of Washington). Celé znění studie k nahlédnutí. (Bryant, Katz a Lazowska, 2008)

Tato studie předpovídala změnu chování firem, vědeckých výzkumníků, praktických lékařů apod. prostřednictvím zpracování Big dat.

V současnosti je termín Big data ve slovníku Oxford English Dictionary definován následovně:

***„big data** n. Computing (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.“* (Oxford University Press, 2018)

Volný překlad textu výše je následující: „Big data (výpočetní technika): Data dosahující velmi velkého objemu, obvykle tak velkého, že jejich manipulace a řízení způsobuje velké logistické problémy, (rovněž) v odvětví výpočetní techniky se vyskytuji takto objemná data.“ (Oxford University Press, 2018)

2.4 Analýza Big dat

Zpracování se skládá z několika důležitých procesů (Press, 2016 a Agrawal, Bernstein a kol., 2012):

- **Stanovení cíle**
- **Příprava dat**
    - *Extrakce dat* – definování zdrojů dat, jejich formátů a nastavení filtrů na nežádoucí data.

- *Anotace dat* – tvorba metadat (tzv. data o datech např. definice zdroje dat)

- *Čištění dat* – upravení dat do požadované podoby (dochází např. k odstranění interpunkčních znamének).

- *Integrace dat a agregace dat* – tvorba datové základy pro analýzu.

- **Analýza dat**

  - *Volba metod, algoritmu či technologií pro zpracování* – blíže popsáno v kapitole 3 **Metody zpracování nestrukturovaných dat**.

  - *Zpracování dat* – příklad zpracování nestrukturovaných dat viz kapitola 6 Zpracování nestrukturovaných dat vybranými prostředky .

  - *Interpretace dat* – např. tvorba reportů.

- **Vyhodnocení výsledku analýzy**

2.5 Bezpečnost

Čím větším množstvím dat firma disponuje, tím více by měla soustředit svou pozornost na jejich zabezpečení.

Je kladen velký důraz na zabezpečení údajů v databázích či v podnikových systémech.

Pro zajištění bezpečnosti se využívá nespočet různých ověření (certifikáty, autorizace, elektronický podpis apod.), aby nedocházelo k neoprávněnému přístupu do systémů. Bohužel se zdá, že hackeři jsou vždy o krok napřed.

Bezpečnostní hrozbu představuje také současný trend – využívání cloudových služeb. Tato uložiště nejsou vhodná pro citlivé firemní či osobní údaje, jelikož k těmto datům má přístup také druhá strana. Za únik citlivých informací z cloudových uložišť může obvykle selhání lidského faktoru.

V současné době existuji moduly, které umožnují zabezpečení dat uložených v cloudu.

Big data jako zdroj jsou využívána k analýze SIEM a také k detekci podvodů (Fraud Detection). Tyto technologie jsou popsány v následujících kapitolách 2.5.1.1 Fraud Detection (detekce podvodů) a 2.5.1.2 SIEM (Security Information and Event Management).

2.5.1 Technologie pro zajištění bezpečnosti

Tato kapitola je věnována popisu technologií pro zajištění bezpečnosti, které využívají logové záznamy, např. pro detekci podvodů či detekci bezpečnostních rizik.

2.5.1.1 Fraud Detection (detekce podvodů)

Cílem Fraud Detection je včasné odhalení podvodného jednání. Tato disciplína využívá tzv. Fraud Detection systému (FDS – Fraud Detection Systems), jedná se o systém pro detekci podvodů, nežádoucího či anomálního chování.

Jedná se o aplikace pro dolování volného textu, které mají udělena oprávnění k přístupu k firemním dokumentům – mají právo provádět analýzu interní komunikace, elektronické žádosti, objednávky přes internet, smlouvy apod., např. udělené právo procházet e-maily zaměstnanců podniku ve snaze detekovat slovo či slovní spojení, fráze, které by mohly naznačovat podvodné jednání. Pokud aplikace detekuje možný podvod, je daný dokument či komunikace označena za rizikovou a předána na příslušné řešitele.

Aplikace pro detekci podvodů třídí dané dokumenty do smysluplných shluků (shluková analýza) – např. do shluku – komunikace nepředstavující riziko, riziková komunikace, podezřelá komunikace apod.

Systém pro detekci podvodného chování se využívá především v odvětví bankovnictví, telekomunikací, e-commerce apod.

Pro detekci podvodného chování jsou využívány dva základní vědecké obory:

- Statistika

- Umělá inteligence – strojové učení

Pro detekci podvodného chování se využívají například následující techniky (ACL, 2018):

- *Výpočet statistických parametrů* (např. průměry, odchylky, či nápadně nízké či vysoké hodnoty) – identifikace odchylek, které by mohly představovat podvodné jednání.

- *Klasifikace* – hledání vzorů v datech (ve volném textu či v číselných datech).

- *Stratifikace čísel* – identifikace neobvyklých položek (příliš velké či malé položky).

- *Digitální analýza s využitím Benfordova zákona* – Benfordův zákon je matematický zákon, který říká, že ve skupině čísel, které představují reálné hodnoty čehokoliv, bude jedničkou začínat zhruba 30 % čísel, dvojkou bude začínat cca 17,6 % čísel, trojkou 12,5 % číselných hodnot a jen 4,57 % čísel devítkou. Tímto přírodním zákonem se řídí soubory jakýchkoliv přirozených dat bez ohledu na jejich podstatu.

  Při podvodném jednání mají lidé tendenci vymýšlet falešné výsledky tak, že začínají na všechna čísla se stejnou pravděpodobností. Benfordův zákon tedy umožňuje detekovat možný podvod nebo je využíván jako jednoduchý test regulérnosti hodnot.

- *Spojení dat z různých zdrojů* – spojení dat z různých zdrojů umožňuje nacházet a identifikovat souvislosti mezi daty (např. shoda jmen, adres, čísla účtů) v takových případech, kde by tyto souvislosti existovat neměly.

- *Testování na duplicity* – identifikace duplicitních transakcí (např. v případě, že je jediná faktura od dodavatele proplacena vícekrát apod.).

- *Testování mezer* – pro tento typ testování jsou využívány aplikace, které se specializují na monitoring změn a operací provedených s daným souborem (např. software IDEA, tyto aplikace využívají funkce, které detekují, zda nějaké položky chybí a v jakých polích. Tyto funkce pracují jak s číselnými hodnotami, tak i s datovými (čas) a znakovými (text) hodnotami.

- *Sčítání číselných hodnot* – identifikace kontrolních součtů, které mohly být neoprávněně upraveny (zfalšovány).

- *Ověřování vstupních dat* – identifikace podezřelých či nevhodných časů pro zadávání dat (např. zadání dat po půlnoci, kdy už nikdo nepracuje apod.).

Existuje také velké množství dalších technik, kterými lze předcházet či včas detekovat podvod. Touto problematikou se zabývá Amanda Nieweler (2015), jedná se např. o:

- *Větší množství reportingových mechanismů* – čím větší množství reportingových mechanismů, tím existuje větší pravděpodobnost nalezení pochybení či podvodu.

- *Proškolení zaměstnanci* – je nutná také spolupráce zaměstnanců. Zaměstnanci by měli být proškoleni na odhalování podvodného chování a také být připraveni na podvod reagovat a ohlásit ho příslušnému oddělení.

- *Minimalizace příležitostí k podvodnému* jednání – firemní politika by měla být nastavena tak, aby minimalizovala příležitost provést jakýkoliv podvod. Měly by být nastaveny kontrolní mechanismy.

2.5.1.2 SIEM (Security Information and Event Management)

Tato analýza slouží k monitoringu logových záznamů z různých heterogenních zařízení, z různých zdrojů za účelem identifikace bezpečnostních hrozeb, které mohou být bezpečnostními incidenty.

Tato technologie rozpozná a upozorní na bezpečnostní hrozby na základě definovaných pravidel zaměstnance podniku, kteří spravují zabezpečení podnikového systému. Tito pracovníci by měli být schopni na tuto hrozbu nebo tento incident zareagovat a v případě bezpečnostního incidentu minimalizovat škody. O této problematice pojednává i Chuvakin a kol. (2013) či Montesino a kol. (2012).

Bezpečnostní událost – je stav systému, služby nebo sítě, který může představovat možné porušení bezpečnostní politiky, nebo selhání bezpečnostního opatření. Mezi bezpečnostní události jsou řazeny veškeré dosud nenastálé situace, které mohou být důležité z pohledu bezpečnosti informací. (Miroslav Čermák, 2014)

Bezpečnostní incident – *„Jedná se o bezpečnostní událost, která představuje narušení bezpečnosti informací v informačních systémech nebo narušení bezpečnosti služeb a sítí elektronických komunikací“* dle zákona o kybernetické bezpečnosti. (Čermák, Miroslav, 2014)

SIEM je tzv. Log Management nástroj, který umožňuje zpracovávat velké množství logů z různých zdrojů.

Tyto systémy umožňují analyzovat v reálném čase jak samotné logy, tak i data z aplikací IPS/IDS, z firewallů atd. s využitím CEP (Complex Event Processing) technologie. Data z heterogenních zdrojů jsou agregována a vytváří komplexní přehled o připojených zařízeních a aplikacích. SIEM vytvoří přehled, který umožňuje dávat jednotlivé informace do logických souvislostí, tj. umožní vytvářet modely příslušných korelací. (Čermák, Miroslav, 2014)

Princip rozpoznávání útoků

Je velmi důležité mít správně nastavená korelační pravidla technologie SIEM. Takto nastavená pravidla lehce identifikují případné hrozby.

SIEM technologie zpracovává události v reálném čase (zpracovává proudy dat prostřednictvím CEP viz kapitola 3.4.1 Zpracování logových záznamů). V okamžiku, kdy technologie rozpozná bezpečnostní hrozbu, přiřadí jí důležitost a zalarmuje bezpečnostní týmy, které mohou na situaci pohotově reagovat.

Záznamy o útocích jsou ukládány a slouží jako vzory pro rychlejší identifikaci bezpečnostního incidentu v budoucnosti.

3 Metody zpracování nestrukturovaných dat

Jednoduché analýzy nestrukturovaných dat (např. Mrak slov, Strom slov) jsou realizovány za pomoci jednoduchých matematických a statistických výpočtů, zatímco na složitější analýzy (např. analýza sentimentu, analýza multimédií) je třeba použít umělou inteligenci (tzv. AI, Artificial Intelligence). Tato práce není zaměřena na principy fungování umělé inteligence. Z tohoto důvodu bude umělá inteligence a její disciplíny, které se používají pro zpracování nestrukturovaných dat popsány velmi stručně.

3.1 Umělá inteligence

Marvin Minsky (1967) definoval umělou inteligenci jako *"umělá inteligence je věda o vytváření strojů nebo systémů, které budou při řešení určitého úkolu užívat takového postupu, který – kdyby ho dělal člověk – bychom považovali za projev jeho inteligence."* Autorka práce se s touto definicí ztotožňuje.

Umělá inteligence využívá pro zpracování nestrukturovaných dat Strojové učení (včetně umělých neuronových sítí).

Umělé neuronové sítě zajišťují mimo jiné např. rozpoznání objektů obrázků, predikci vývoje. Učí se rozpoznávat, identifikovat a definovat výsledky za pomoci strojového učení.

3.1.1 Umělá neuronová síť

Základním prvkem neuronové sítě je jednoduchý procesor – percepton. Percepton (neboli neuron) je matematický model biologického neuronu.

Pro umělé neuronové sítě platí (Mendelu, 2018) a (Shrimphood, 2018):

- Percepton (neuron) může mít pouze jeden výstup, ale neomezené množství vstupů.
- Každý percepton (neboli neuron) má prahovou hodnotu (tzv. potenciál neuronu).
- Každý vstup má svou váhu určenou synapsí (spojením mezi jednotlivými neurony).

Typy neuronových sítí:

- **Vícevrstvé neuronové sítě (MLP – Multi Layer Perception)**
    - Použití – predikce (na základě časových řad, vývoje trendu), klasifikace, aproximace.
- **Hopfieldovy sítě**
    - Použití – asociativní paměť, klasifikátor (OCR), optimalizace (problém obchodního cestujícího).
- **Samoorganizující se sítě (SOM – Self Organizing Map, neboli Kohenenova síť)**
    - Použití – shlukování, klasifikace (např. zákazníků).
- **Radiální báze (neboli RBF sítě)**
    - Použití – klasifikace, regrese.

  a další (Mendelu, 2018)

Základní algoritmy neuronové sítě:

- **učení s učitelem** – srovnávání aktuálního výstupu s požadovaným výstupem. Cílem tohoto algoritmu je snížit rozdíl mezi těmito výstupy na minimum – nalézt chybu a minimalizovat ji. Chybu je možné minimalizovat přenastavením váhy a prahu neuronové sítě.
- **učení bez učitele** – není znám výstup. Síť se učí systémem třídění vstupu. Sadu vzorů, které síť obdrží, roztřídí do skupin. Cílem učení bez učitele je získat konzistentní výstup. Aby toho neuronová síť dosáhla, je nutné změnit např. topologii sítě, či reagovat na typického zástupce skupin.

Prapůvod neuronových sítí je v biologii.

Paralelismus je největší devízou umělých neuronových sítí (zajišťuje distribuované paralelní zpracování dat).

Strojové učení

Je oblast umělé inteligence (AI), která zajišťuje schopnost počítačového systému „učit se“. Strojové učení využívá oblast statistiky a data miningu.

Základní algoritmy strojového učení (Brownlee, 2016):

- **Učení s učitelem (Supervised Learning) –** počítačovému systému poskytneme pouze vstupní data, bez požadavků na výstup.

- **Učení bez učitele (Unsupervised Learning) –** počítači poskytneme vstupní data i požadovaný výstup.

- **Učení se zpětnou vazbou (Reinforcement Learning) –** učení počítačového systému na základě zpětné vazby (odměny a tresty na základě odvedené práce).

Kroky strojového učení (Jain, 2015):

1. **Sběr dat** – sběr dat z různých zdrojů a v různých formátech (Excel, OLAP kostky, Access, textové dokumenty). Vytvořený soubor dat je stěžejní pro strojové učení.
2. **Příprava dat** – soubor dat musí být kvalitní, aby se z něho daly vyčíst co nejpřesnější informace, tzn. je nutné odstranit nesrovnalosti v datech (např. chybějící údaje).
3. **Trénování modelu** – volba správného algoritmu a reprezentace dat. Data jsou rozdělena na dvě části, na část trénovací a část testovací.
4. **Ohodnocení modelu** – model, který byl otestován na trénovací množině dat, je vhodné aplikovat na testovací část, a zjistit tak přesnost modelu a výkon.
5. **Aplikace či přetrénování modelu** – Pokud přesnost a výkon modelu odpovídá stanovených požadavkům, je možné tento model aplikovat. Pokud současný model nesplňuje požadované parametry, je nutné model přetrénovat (návrat ke kroku 3. Trénování modelu ale využití jiných či optimalizovaných algoritmů).

Problémy řešené strojovým učením:

- **Klasifikace** – označení jednotlivých objektů, které určí, do jaké skupiny se řadí.
- **Regrese** – analýza existujících dat za účelem předpovědi dalšího chování dat (např. zkoumání změn trendu, odhad vývoje cen nemovitostí apod.).
- **Clusterování (Shlukování)** – shlukování podobných objektů (seskupování textů s podobným tématem, obrázků s podobnými objekty apod.).
- **Asociace** – identifikace pravidel v datech (vztahy mezi množinami dat). Toto řešení využívá např. odvětví Business Inteligence.

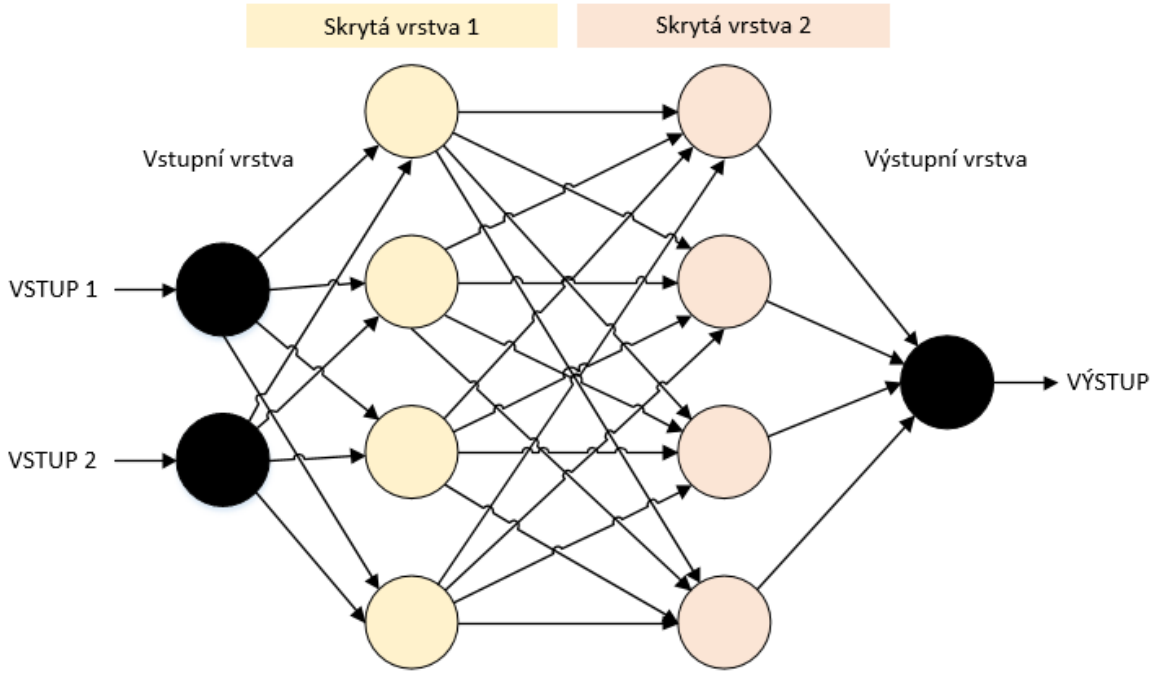
Strojové učení je používáno pro (Garetta, 2015):

- **Zpracování obrázků**
  - Označování objektů obrázku/videa
  - Optické rozeznání znaků (OCR)
- **Textovou analytiku**
  - Analýzu sentimentu
  - Extrakce informací
  - Filtrace spamu
- **Dobývání dat**
  - Předpověď dalšího vývoje
  - Asociační pravidla
  - Seskupování
  - Detekce anomálií
- **Robotiku**

Hluboké učení (tzv. Deep learning)

Hluboké učení využívá neuronové sítě, které jsou tvořeny několika vrstvami propojených umělých neuronů (perceptorů), viz Obrázek 3.

Hloubka modelu = počtu vrstev reprezentujících data.



Obrázek 3: Vícevrstvá neuronová síť (dopředná)
Zdroj: vlastní zpracování, inspirováno dle Holčíka a Komendy, 2015

Další podkapitoly kapitoly 3 jsou zaměřeny na použití metod umělé inteligence k dobývání znalostí, konkrétně na analýzu textu, multimédií a dat v reálném čase.

3.2 Analýza textu

Analýza volného textu patří mezi složité a velmi rozvíjené technologie.

3.2.1 Jednoduché zpracování nestrukturovaných dat

Mezi tyto metody řadíme postupy zpracování dat, které jsou založené na statistice či matematice. Vstupními soubory jsou textová data, mohou být strukturovaná i nestrukturovaná.

Touto problematikou se zabývá bakalářská práce autorky. (Smolová, 2016)

Příklady statistického a matematického zpracování nestrukturovaných dat lze vidět na Obrázcích č. 3 a č. 4.

Touto problematikou se zabývá také monografie Hofmanna a Chisholma (2015).

- **Mrak slov (Word Cloud)**

Obrázek 4: Word Cloud – Mrak slov

Zdroj: Text z kapitoly Analýza a zhodnocení současného stavu problematiky zpracovaný službou na stránce https://www.wordclouds.com/ dne 15. 02. 2018.

- **Strom slov (Word Tree)**

Obrázek 5: Strom slov – Word Tree

Zdroj: Text z kapitoly Analýza a zhodnocení současného stavu problematiky zpracovaný službou na stránce https://www.jasondavies.com/wordtree dne 15. 02. 2018.

Algoritmus pro zpracování textu do vybraného tvaru je založen na jednoduchých matematických a statistických výpočtech. Např. na Obrázku č. 3 jsou největší ta slova, která mají v analyzovaném textu nejvyšší četnost. Čím je menší četnost daného slova, tím se slovo nachází v obrázku v menší velikosti.

Zatímco na Obrázku č. 4 je uživatelem definováno slovo (v tomto případě „Data“). Slovo je vyhledáno v celém textu a následně je za toto slovo doplněn kontext, ve kterém se slovo nachází. V případě, že se za slovem „Data“ se nachází v kontextu jiné slovo či diakritika – v tomto případě čárka (,) – více než jednou, je opět znázorněno ve větší velikosti a také v kontextu. Na tomto principu je postavena celá analýza Stromu slov. Problematiku blíže popisují ve své práci Martin Wattenberg a Fernanda Viégas, 2007.

Mrak slov je využíván především pro grafické znázornění hrubého obsahu daného textu. Strom slov umožňuje vyhledávat v textu a vytvořit představu o jeho obsahu.

3.2.2 Zpracování přirozeného jazyka

Zpracování přirozeného jazyka (též Natural Language Processing (NLP), komputační lingvistika) počítačem je stále velmi aktuální a také velmi rozsáhlé téma. V této práci je tato problematika velmi stručně popsána.

Cílem tohoto zpracování je porozumění přirozenému textu strojem či počítačem (extrakce důležitých dat a pochopení textu jako celku). K tomu jsou využívány nástroje a algoritmy, které využívají znalosti formální lingvistiky, informatiky (využití umělé inteligence (AI)), akustiky a dalších vědních oborů. (Veselovská, 2017)

Pro správné pochopení této problematiky je nutné si definovat, co považujeme za přirozený jazyk. Dle Mluvnice současné češtiny I.:

„můžeme definovat přirozený jazyk jako systém verbálních znaků (morfémů, slov, vět), který slouží k mezilidské komunikaci.“ (Cvrček a kol., str. 18, 2015)

Při zpracování přirozeného textu se využívá několik disciplín pro zpracování nestrukturovaných dat, jako např. zpracování řeči, tokenizace, extrakce vztahů, kategorizace dokumentů, detekce vět, klasifikace a shrnutí textu atd., jak popisuje ve své knize. (Reese, 2015)

Zpracování přirozeného jazyka je řazeno mezi těžší disciplíny zpracování nestrukturovaných dat. Velké množství problémů, které se vyskytují v této disciplíně lze vyřešit relativně jednoduše, ale obsahuje i velké množství problémů, které je nutné řešit prostřednictvím sofistikovanějších technik (s využitím například hlubokého učení). (Reese, 2015 a Veselovská, 2017)

Zpracování přirozeného jazyka komplikuje velké množství faktorů, např. velké množství přirozených jazyků s rozdílnou syntaxí, sémantikou.

Procesy při zpracování textu a problémy, kterou mohou nastat:

- **Tokenizace** – rozdělení vět, resp. celého textu na jednotlivá slova. Nejmenší prvek textu se nazývá token.

  Nejdříve je nutné si stanovit, jakým způsobem, resp. dle jakého prvku textu budou jednotlivá slova oddělena od sebe. Velmi často bývá využito tzv. bílých znaků (mezera, tabulátor, odřádkování), ale ne vždy je to vhodné řešení, např. pro jazyky se speciálními znaky. Při využití pouze bílých znaků pro oddělení slov poté nastává problém s interpunkčními znaménky, jako je např. čárka (,), tečka (.), jelikož jsou spojena se svými slovy, např. „konec!“ je považováno za jiné slovo než „konec“.

  Výstupem tohoto procesu je stream tokenů.

- **Normalizace –** cílem tohoto procesu je převedení jednotlivých tokenů do normalizované podoby s využitím:

  - **Stematizace –** hledání kořene slova, jednotlivých slov (resp. tokenů). Stematizace je využívána například v internetových vyhledávačích.

    - **Morfém –** minimální, významově nedělitelná jednotka (předpona, přípona, vpona. Jak je popsáno na cestinaveslovniku.cz (2018).

  - **Lematizace –** proces, ve kterém je slovo převedeno do základního tvaru (např. běhání -> běhat)

- **Koreference slov –** určení vztahů mezi jednotlivými slovy v textu.

- **Význam slov –** zjištění významu daného slova (tokenu) – v případě homonym je identifikace významu velmi složitá. Je nutné pochopit kontext věty. (Smolová, 2016)

Možná využití zpracování přirozeného jazyka

Jak popisuje Reese (2017), tento druh analýzy je využíván k řešení různorodých problémů ve velkém množství disciplín, např.:

- **Strojový překlad** – překlad z jednoho přirozeného jazyka do druhého.

- **Sumarizace** – sumarizace odstavců, článků, souboru dokumentů – resp. krátké shrnutí celé vybrané části textu.

- **Rozpoznávání pojmenovaných entit** (NER – Named Entity Recognition) – tato metoda umožňuje identifikovat entity v textu a klasifikovat je do předdefinovaných kategorií.

- **Analýza sentimentu** – zjištění postoje autora k dané problematice. O této problematice blíže pojednává dále kapitola 3.2.4 Analýza sentimentu.

- **Označování částí řeči** (POS – Parts of Speech Tagging) – tato disciplína slouží k označení slov v textu – např. přiřazení slovního druhu k danému slovu.

- **Vyhledávání** – identifikace konkrétních prvků v textu, zjištění počtu výskytu daného slova.

- **Rozpoznávání řeči** – rozpoznávání lidské řeči, jazyka.

- **Generování přirozeného jazyka** – schopnost interpretovat data přirozeným jazykem.

- **Zodpovídání dotazů** – stroje (počítače) dokáží reagovat na lidský dotaz (na přirozený jazyk). Velký úspěch slaví společnost IBM se svým superpočítačem – IBM Watson, který vyhrál znalostní soutěž Jeopardy. (Gabbatt, 2011)

3.2.3 Analýza zdrojového kódu

Kromě strojového kódu (logových záznamů) je také možné analyzovat soubory/skripty, které obsahují zdrojový kód webové stránky/aplikace.

Dle typu analýzy se můžeme zaměřit na zjištění bezpečnostních hrozeb, slabých míst, vytíženost daných částí kódu, či pouze pochopit zdrojový kód v bližších souvislostech.

Existuje velké množství nástrojů pro analýzu zdrojového kódu. Tyto nástroje automatizovaně testují zdrojový kód a podávají o něm požadované informace.

V současné době jsou k dispozici např. tyto druhy analýzy zdrojového kódu:

- **Analýza bezpečnostních rizik zdrojového kódu**

  Tzv. SAST (Static Application Security Testing) – testování zdrojového kódu aplikací z hlediska bezpečnosti. Detekce existujících chyb v kódu, které by mohly ohrozit zabezpečení aplikace. Jedná se například o aplikace IBM Security Appscan, Veracode, AttackFlow a další.

  Opakem SAST je DAST (Dynamic Application Security Testing). DAST netestuje zdrojový kód aplikace, ale bezpečnostní rizika aplikace při jejím užívání.

- **Analýzu zdrojového kódu za účelem porozumění –** tyto nástroje pomáhají vývojářům pochopit, udržovat a dokumentovat zdrojový kód. Analýzou kódu dochází k vytěžení potřebných informací pro tvorbu diagramů vztahů, listu použitých proměnných a postupů či k vizualizaci architektury kódu.

  Mezi tyto nástroje řadíme např. Roslyn od firmy Microsoft, Understand a další.

Nástroje pro analýzu zdrojového kódu jsou závislé na programovacím jazyku, ve kterém je aplikace vytvořena, vyvíjena.

Na základě výsledků těchto analýz jsou tvořeny reporty.

3.2.4 Analýza sentimentu

Analýza sentimentu spadá pod NLP (Natural Language Processing). Daná analýza je aplikována na volný text či na snímek/fotografii (např. sentiment výrazu tváře). Tato problematika je v současné době velmi populární, přestože stále nejsou vyřešeny všechny problémy (např. rozpoznání ironie). (Godsay, 2015)

Analýza je často využívána pro zjištění sentimentu z komentářů u příspěvků na sociálních sítích, jako je např. Facebook, Twitter, Instagram, Youtube a další.

Cílem je určit, jaký postoj mají lidé k dané věci (k článku/informaci, kterou okomentovali), tzn. jak subjektivně vnímají daný příspěvek/informaci, jaké z toho mají pocity, emoce = sentiment. (Godsay, 2015)

Analýza sentimentu je považována za velmi složitou disciplínu pro zpracování přirozeného jazyka.

Postup pro rozpoznání sentimentu textu:

1. *Zjištění, zda zvolený text obsahuje sentiment* (subjektivní text) či neobsahuje (objektivní text).

2. *Určení citového zabarvení textu* – pozitivní, negativní, neutrální.

Jak již bylo zmíněno výše, analýza sentimentu se stále potýká s několika problémy. Tato disciplína se využívá již několik let, ale dosud není definováno, jak rozpoznat ironii, humor, sarkasmus, porozumění kontextu apod. Problémy při analýze textu, např. v anglickém jazyce, může způsobit např. i slovo v jiném jazyce – např. latinské. Cizojazyčná slova jsou až na výjimky považována za neutrální. Další problémy může způsobit i použití negace v českém jazyce. (Godsay, 2015)

Z těchto důvodů současné nástroje pracují s určitou mírou chybovosti.

Dalším problémem je jazyk. Existuje nepřeberné množství systémů pro analýzu sentimentu volného textu v anglickém jazyce, zatímco velmi málo aplikací pro jiné jazyky (např. čeština, slovenština apod.).

Analýze sentimentu využívá dvě základní metody, které jsou založené na:

1. **Lexikonu slov**

   Tato metoda používá k určení sentimentu lexikon, který obsahuje sémanticky orientovaná slova nebo fráze. V lexikonu se nachází slovo a jeho polarita (jestli se jedná o pozitivní, či negativní sentiment).

   Dle počtu a míry polarity výskytu těchto slov/frází v textu je určen sentiment textu.

   Lexikony mohou být vytvořeny ručně i automaticky (některým slovům je přiřazena polarita, následně jsou vyhledána slova podobného významu a těm je přiřazena polarita stejná). (Taboada a kol., 2011)

2. **Strojové učení**

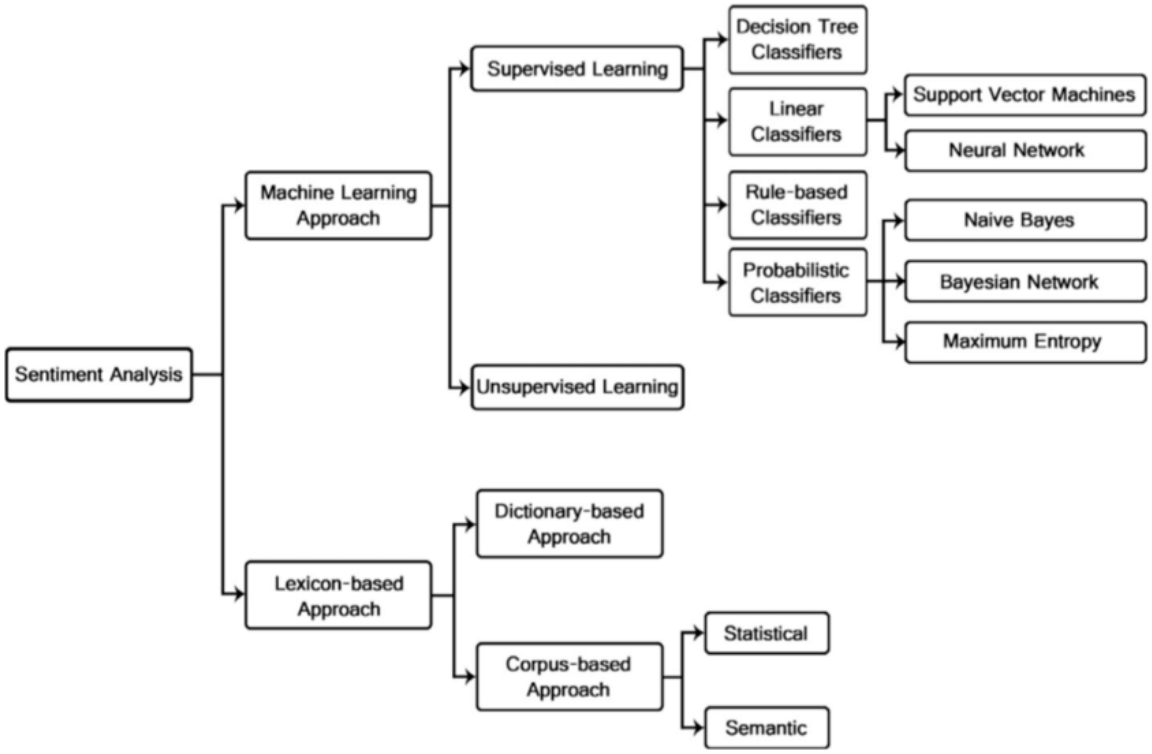
   Jak popisuje Pang a kol. (2002 a 2008), základem strojového učení pro analýzu sentimentu je vytrénování klasifikátoru na základě principu učení s učitelem ve sbírce anotovaných textů. Každý text je zde reprezentován vektorem slov, n-gramy (sled n po sobě následujících slov), skip-gramy v kombinaci s jinými typy sémantických rysů, které se pokoušejí modelovat syntaktickou strukturu vět, zintenzivněním, negací, subjektivitou či ironií.

Existuje několik technik strojového učení využívaných pro analýzu sentimentu, např.

- **SVM** (Support Vector Machine)

- **Hluboké strojové učení**
- **LSA** (Latent Semantic Analysis)
- **Naive Bayes**

a další (viz Obrázek 6 níže)



Obrázek 6: Metody analýzy sentimentu
Zdroj: převzato od Voight, Kieslinger a Schäfer, 2017.

Interpretace výsledků

Výsledek analýzy sentimentu je interpretován nejčastěji desetinnými čísly s polarizací – aby bylo možné určit typ sentimentu. Každý nástroj může mít nastavená jiná pravidla pro interpretaci výsledků (rozdílné intervaly, ve kterých je sentiment pozitivní, negativní či neutrální).

Např. výsledek je roven +0,28, tzn. plusová hodnota vyšší jak např. 0,2 je označena na pozitivní sentiment.

Prediktivní analýza

Tento druh analýzy využívá statistiku, Big data analýzu a strojové učení za účelem předpovědi pravděpodobného budoucího vývoje analyzované situace.

Prediktivní analýza využívá různé nástroje nebo algoritmy pro vytváření prediktivních modelů, které jsou využívány k charakteristice historických informací. Tyto modely jsou dále použity k předpovězení povahy a pravděpodobnosti budoucích událostí.

Jak popisuje ve své knize McCue (2007), lze tuto problematiku vysvětlit prostřednictvím hry pro děti, ve které se propojují jednotlivé body. Propojením těchto bodů vznikne vybraný obrázek.

Propojením určitých historických informací nám vznikne vybraný model a tento model lze dále využít k předpovězení budoucího vývoje.

Prediktivní analýza je využívána při detekci podvodů, v marketingu a v obchodu (např. nákupní chování uživatelů, předpověď obratu firem apod.). Často jsou pro prediktivní analýzu využívány specializované analytické nástroje, statistické aplikace či programovací jazyky (např. R, Python a další). (Kumar, 2016)

Prediktivní analýza využívá základní tři techniky:

- **Regresní analýza**
    - Lineární regresní analýza
    - Logistická regresní analýza – tento model je zobecněním lineárního regresního modelu.

- **Analýza časových řad**

- **Strojové učení**

Kvalita předpovědi

Přesnost předpovědi se odvíjí od kvality vstupních dat a také od toho, jak dalekou budoucnost má analýza předpovědět. Prediktivní analýza předpoví s vyšší přesností budoucí rok než budoucnost za několik let. (Bari, Chaouchi a Jung, 2014)

Kroky prediktivní analýzy

1. **Definice cílů prediktivní analýzy** – ve vztahu k podnikovým cílům, strategii apod.

2. **Sběr dat** – shromáždění dat v různých formátech a z různých zdrojů (interní i externí, data z aplikací apod.).

3. **Analýza –** prověření dostupných dat a návrh modelu.

4. **Zpracování dat –** definovaným modelem.

5. **Učení se z dat** – upravení modelu, aby odpovídal požadavkům.

6. **Nasazení prediktivních modelů do testovacího prostředí** – nasazení Business inteligence a správně vytvořených procedur modelu.

7. **Testování** – otestování správnosti výstupů vytvořeného modelu prediktivní analýzy.

8. **Nasazení prediktivních modelů do ostrého provozu** – nasazení prediktivní analýzy do ostrého provozu.

3.3 Analýza multimédií

Za multimédia jsou považovány soubory, které kombinují formy obsahu textu, audia a videa.

V této kapitole se podrobněji zaměříme na analýzu audiovizuálních záznamů. Dané problematice je věnována praktická část práce – 6. Zpracování nestrukturovaných dat vybranými prostředky .

3.3.1 Analýza zvuku

Zvuk je mechanické vlnění hmotných částic, které je charakterizováno parametry pohybu částic v prostředí. Zdrojem zvuku je kmitající těleso, jehož vzruchy se v prostoru šíří formou postupného podélného vlnění (zvukové vlny). (Bernat, 2010)

Výsledkem analýzy zvukového záznamu videa může být:

- **Přepis (Speech To Text - STT)** – této problematice se věnuje část Přepis řeči na text, viz níže.

- **Vizualizace –** lze interpretovat např.:

    - **Úrovně zvukových signálů** – tento proces zajišťuje reprezentaci zvukového signálu signálem obrazovým. Díky této vizualizaci je možné např. vyhledat a opravit nedostatky záznamu.

      Zvukový signál je charakterizován také intenzitou a akustickým tlakem.

      Vizualizace časového průběhu zvukového signálu je nejčastější vizualizací tohoto typu.

- **Frekvenční spektrum zvuku** – prostřednictvím analýzy frekvence. Tento druh analýzy poskytuje pohled na měřený časový průběh akustického, vibračního nebo obecného signálu. (Ekosoftware, 2018)

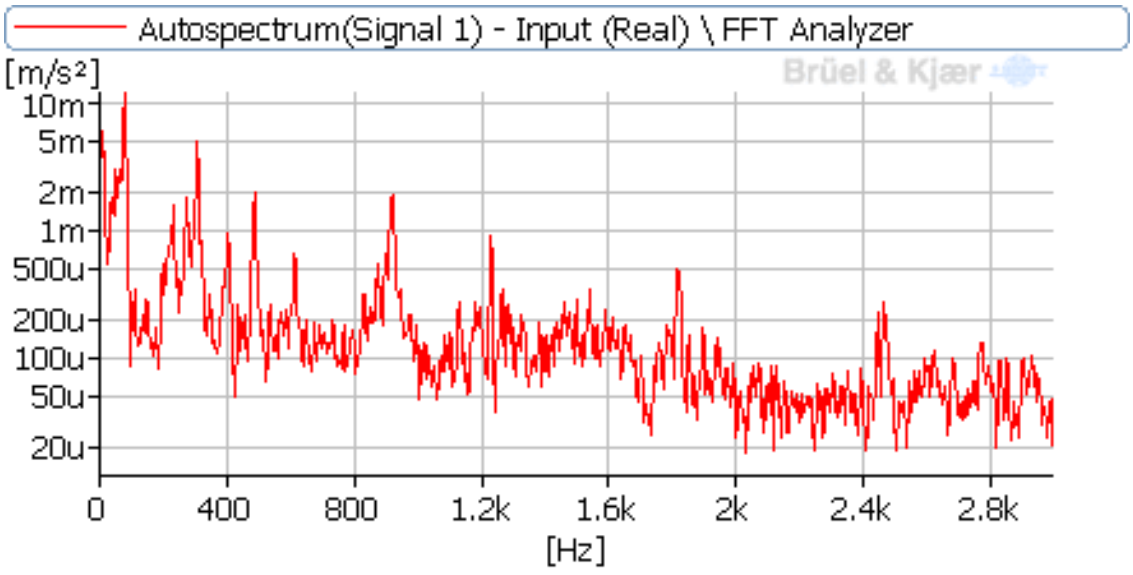
    - **Typy frekvenčních analýz**

        - **CPB (Constant Percentage Bandwifth)** – měření zvuku (hluku) za použití frekvenčních filtrů. (Ekosoftware, 2018)

        - **FFT (Fast Fourier Transform)** – používá se ke zpracování vibračních signálů.

        - **Řádová analýza (Order Analýza)** – využití FFT analýzy, jen jsou před zahájením analýzy provedeny úpravy v naměřeném časovém úseku. (Ekosoftware, 2018)

        - **STFT (Short Time Fourier Transformation) –** algoritmus pro výpočet časově frekvenčního zobrazení (jak popisuje Brejl a Šebesta, 1999).



Obrázek 7: Frekvenční spektrum zvuku (FFT analýza)

Zdroj: Ekosoftware, 2018.

Přepis řeči na text

Problematikou přepisu řeči na text se zabývají tzv. STT nástroje (Speech To Text nástroje). Naopak nástroje pro převod textu na řeč jsou označovány zkratkou TTS (Text to Speech). Jedná se o nástroje, které provádí převody mezi textem a řečí prostřednictvím výpočetní techniky.

Nástroje pro STT využívají tzv. technologie ASR (Automatic Speech Recognition, česky: automatické rozpoznání řeči, jinak také SRS – Speech Recognition System).

Nástroje pro rozpoznávání hlasu se stále učí. Přepis řeči na text je považován za jednu z nejsložitějších problematik zpracování nestrukturovaných dat. Především kvůli komplexnosti lidského jazyka, jak popisuje ve své studii Forsberg (2003) a jak je blíže popsáno níže.

Problémy spojené s automatickým rozpoznáváním řeči (ASR) (Forsberg, 2013):

- **Lidské porozumění řeči v porovnání s ASR** – ASR nikdy nebude rozumět sdělení na takové úrovni jako člověk, a to především kvůli znalostem, které člověk má o řečníkovi (např. věk, pohlaví, rozpoložení člověka apod.). Na základě těchto znalostí je člověk schopen i předpovídat, jakým směrem se bude řečník dále ubírat, co a jak řekne. ASR tuto možnost nemá, proto se může úrovni, na které je lidské porozumění, pouze přiblížit.

- **Řeč těla** – pokud ASR nekomunikuje s technologiemi pro rozpoznávání řeči lidského těla (v případě videa), chybí zcela informace sdělené tímto prostředkem.

- **Hluk** – vysoká míra hluku může způsobit velké problémy při automatickém rozpoznávání řeči. ASR nerozpozná slova přehlušená jiným zvukem, lidský mozek je schopen si tato slova domyslet. S tímto bodem souvisí i problém tzv. variability kanálu – kdy je obsah sdělení (jeho kontext) přerušen akustickou vlnou.

- **Rozdíly mezi řečnickým jazykem/psaným jazykem** – problém při využití například slangu. ASR nerozpoznaná slang správně.

- **Průběžná mluva** – pokud řečník mluví příliš rychle a nedělá žádné odmlky/pauzy mezi jednotlivými slovy, je pro ASR velmi obtížně rozpoznat hranice těchto slov.

- **Variabilita řečníka** – každý člověk má svůj vlastní způsob vyjadřování, lišící se zejména v těchto aspektech:

    - Hlas a jeho variabilita – každý řečník má jinou výšku, frekvenci apod.

    - Pohlaví

    - Rychlost řeči

    - Dialekt

    - Styl mluvčího

- **Dvojznačnost** – ASR není schopno rozpoznat zamýšlený význam dvojznačného slova.

Důležité prvky řeči pro správné ASR:

- **Hlas a jeho variabilita** – blíže popsáno v předchozím odstavci

- **Slovník** – je důležité, o jaký jazyk přepisu se jedná. Každý jazyk má svůj slovník. Pokud řečník použije slovo, které se nenachází ve slovníku ASR, není možné ho přepsat. Důležité také je stanovit jazyk, jakým řečník mluví, aby nedocházelo k záměně slov podobně znějících v různých jazycích.

- **Kontext**

- **Prostředí** – kvalita přepisu je také závislá na prostředí, ve kterém se řečník nachází. Pokud je řečník v hlučném prostředí, nebude přepis nikdy tak vysoké kvality jako v případě, kdy je prostředí kolem řečníka tiché.

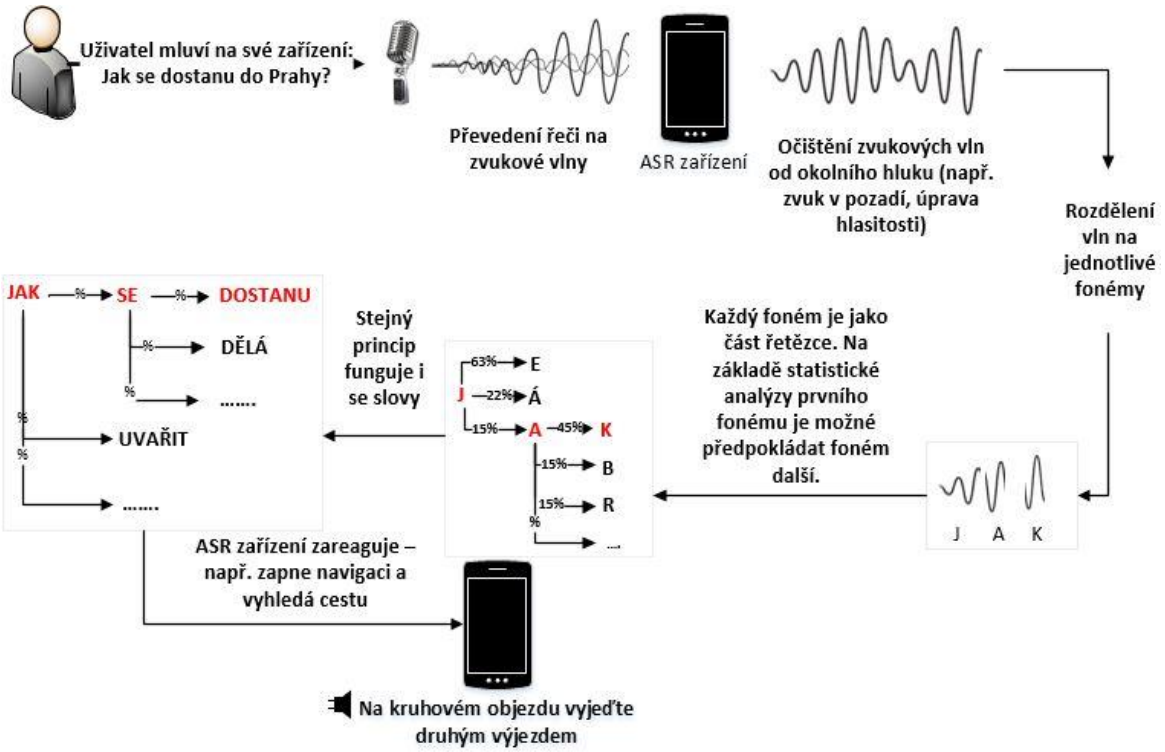
Výstupy ASR:

- **Text** – přepis řeči na text (STT – Speech to Text).

- **Akce zařízení zajišťující ASR, na podnět** – jedná se o tzv. ASU (Automatic Speech Understanding) – zařízení zajišťující ASR porozumí tomu, co po něm uživatel vyžaduje, co říká. Jedná se například o položení otázky, na kterou ASR zařízení vyhledá odpověď či jinak zareaguje (např. viz screen níže). Jedná se o efektivní cestu, jak lze komunikovat s chytrými zařízeními.

ASR je velmi často využíváno v aplikacích pro automobilový průmysl, zdravotnictví, armádu a v mnoha dalších oblastech.

Základní princip fungování ASR



Obrázek 8: Princip fungování ASR – automatického rozpoznávání řeči
Zdroj: Přepracován obrázek dostupný z https://www.noobpreneur.com/2014/10/21/automatic-speech-recognition-works-learns/

Existují dva způsoby komunikace uživatele s ASR zařízeními:

- **Řízený dialog** – ASR zařízení nabídne uživateli varianty a uživatel si z nich musí vybrat.

- **Přirozená konverzace** – zařízení je schopné zpracovat konverzaci s otevřeným koncem, porozumět jí (ASU) a zareagovat.

ASR má algoritmy pro:

- **Detekce řeč neřeč** – rozlišení mezi řečí a jinými zvuky/tichem.

- **Detekce mluvčích a jejich odlišení** – zjištění kolik lidí mluví a jestli jejich dialog na sebe navazuje, či si mluvčí skáčou do řeči. Odlišení mluvčích.

- **Detekce tzv. crosstalku** – pokud se mluvčí přeřvávají – je možné filtrovat pouze vybraný hlas. (Cetin a Shriberg, 2006)

- **Detekce jazyka** – např. detekce češtiny, angličtiny apod.

- **Detekce pohlaví** – zjištění, zda je mluvčí muž nebo žena.

- **Detekce věku** – dle výšky hlasu, použitých slov apod.

- **Detekce klíčových slov** – např. detekce názvu země, organizace v řeči. Využívá se pro prozkoumání, zda se v daném textu nemluví o dané značce (např. o Škoda Auto).

Přepis audio záznamů či diktované řeči na text je v současné době užíván ve všech odvětvích, např. v obchodu, kde umožňuje uspořit čas díky automatickému přepisu, identifikuje volajícího, zjišťuje demografická data o uživatelích a další.

3.3.2 Analýza digitálního snímku

Princip této analýzy spočívá v detekci objektů daného snímku. Tyto objekty lze dále zkoumat, např. pokud analýza identifikuje na snímku osobu, lze rozpoznat pohlaví, věk, rasu na základě analýzy detekované tváře (např. rasovou/etnickou příslušnost lze rozpoznat na základě odstínu pleti).

Základní kroky analýzy snímku

Segmentace – proces rozdělení snímku do oblastí s homogenními vlastnostmi. Metody segmentace:

- **Prahování** – zkoumání jasové hodnoty pixelů za účelem určení prahů. (Ličev a Sojka, 2009)

- **Zpracování binárních obrazů** – jedná se o obraz, jehož pixely nabývají vždy jedné ze dvou hodnot (např. hodnot 0 – popředí a 1- pozadí snímku) (Ličev a Sojka, 2009). Lze je zpracovat různými přístupy:

    - **Matematická morfologie (MM)** – zajišťuje techniku detekce geometrických struktur (např. obrysů objektů), založených na teorii množin, teorii uspořádání, topologii a na náhodných funkcích. (Owens, 1997)

    - **Ztenčování** – odstraňování krajních pixelů objektu tak, aby nedošlo k narušení jeho obrysu. (Ličev a Sojka, 2009)

- **Metody rozpoznávání objektu**

    - **Rozpoznávání objektu –** pro rozpoznání objektu je důležitý popis objektů, existují dvě základní metody popisu:

        - Příznaková – předpoklad znalosti snímku ve formě vektoru příznaků.

        - Syntaktická – metody popisu na základě posloupnosti a hierarchie.

      (Ličev a Sojka, 2009)

    - **Neuronové sítě**

        - Konvoluční neuronová síť (Charu C. Aggarwal, 2018)

        - Kompetitivní síť a Kohenenovo učení (dvouvrstvá neuronová síť)

          (Ličev a Sojka, 2009)

- **Fourierovy transformace průběhu křivosti** – zkoumání průběhu křivosti hranic zkoumaných objektů (dle Ličev a Sojka, 2009).

- **Jasové atributy** – rozpoznání objektů na základě výrazné změny jasů pixelu v dané oblasti.

Transformace – úprava snímku za účelem zvýraznění informací o např. poloze objektů v prostoru (Ličev a Sojka, 2009). Způsoby transformace obrázků jsou např.:

- **Filtrace**

- **Detekce hran** – založeno na matematických výpočtech. Hrany jsou identifikovány na základě výrazné změny jasu pixelů v dané oblasti. (Marr a Hildreth, 1980)

- **Úpravy barev** – např. za účelem zvýraznění objektů.

Extrakce vlastností – extrakce vlastností snímku či objektu.

Identifikace – porovnání extrahovaných vlastností snímku či objektů s vlastnostmi snímků či objektů v databázi.

Zhodnocení – zhodnocení analýzy na základě geometrických a fotometrických charakteristik (např. určení o jaký objekt se jedná, detekce barev apod.)

Analýza snímku za účelem:

- **Identifikace objektů snímku –** např. identifikace automobilu, domu apod.

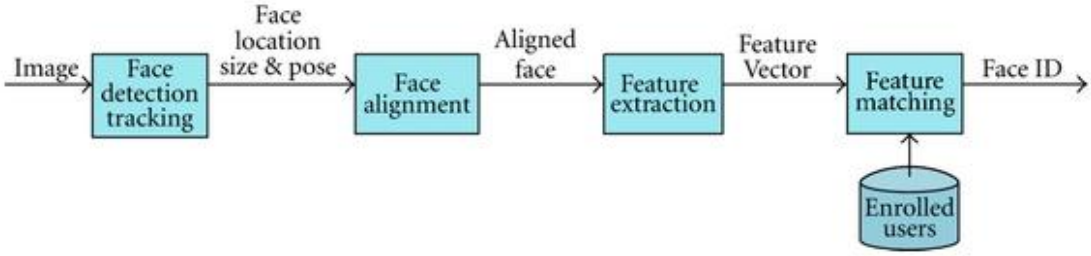
- **Zjištění vlastností snímku –** např. za účelem detekce barev snímku.

Analýza tváře

Prostřednictvím této analýzy lze z tváře vyčíst osobnostní, demografické, sociální a další charakteristiky. V současné době existuje nepřeberné množství nástrojů, které z tváře vyčtou pohlaví, věk, emoce, cílení pohledu a další.

Těmito nástroji lze zpracovat jak fotografie, tak i video (jsou zpracovány jednotlivé snímky v přesném pořadí).

Technologie, které tyto nástroje využívají, jsou na velmi vysoké úrovni a využívají hluboké neuronové sítě.



Obrázek 9: Segmenty identifikace tváře
Zdroj: Le, 2011

Identifikace tváře

V této práci popíšeme přibližný postup metody identifikace obličeje založené na strojovém učení.

Postup identifikace tváře:

1. **Detekce tváře** (Face detection tracking) – nejdříve je nutné na obrázku detekovat tvář mezi jinými objekty. Existuje několik různých přístupů:

    a. Znalostní metody

b. Invariantní metody založené na prvcích

c. Metody založené na porovnávání šablon

d. Metody založené na strojovém učení

2. **Ohraničení tváře** (Face alignment) – dochází k bližšímu určení tváře a k její normalizaci. Jsou detekovány základní prvky obličeje (kontury nosu, uší, úst, brady, obočí apod.) na základě významných bodů. Pro detekci významných bodů tváře je možné použít:

   - **Umělé neuronové sítě** – tuto síť je nutné nejdříve natrénovat nad databází s velkým počtem snímků (čím je vyšší počet snímku, na kterých neuronové sítě trénují, tím je vyšší přesnost detekce). V této problematice jsou využity dopředné vícevrstvé sítě/perceptony (tzv. MLP).

   - **Vyhledávání SVR (Support Vector Recognition)** – „*SVR je trénováno na predikci pozice bodu vhodného obrazového deskriptoru (obvykle založeno na LBP histogramech)*“ ,jak píše Gruber (2015).

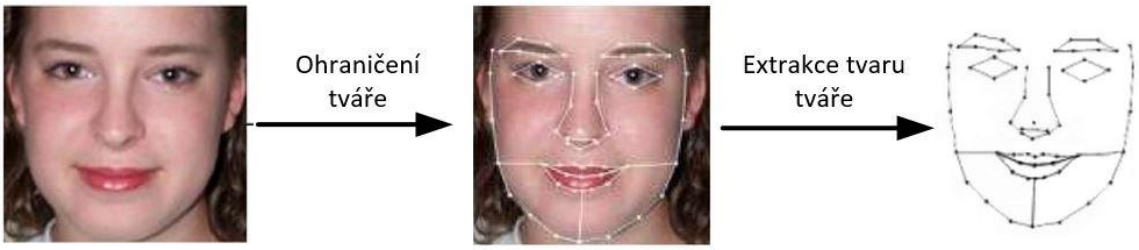
   Vstupní obrázek je normalizován s ohledem na geometrické vlastnosti obličeje (šířka, velikost, tvar, natočení apod.) a také s ohledem na fotometrické vlastnosti (osvětlení, stupnice šedi apod.). (Le, 2011)

   Existují dvě základní metody pro ohraničení tváře, obě metody využívají statistické metody k parametrizaci tvaru obličeje.

   - **ASM** (Active Shape Model, česky Aktivní tvarový model)

   - **AAM** (Active Appearance Model, česky Aktivní vzhledový model) – tento algoritmus zajišťuje nalezení významných bodů tváře. (Cootes a Taylor, 2004)

Metody využívají statistický model PCA (Principal Component Analysis) pro parametrizaci tvaru obličeje (dle Le, 2011) a také model PDM (Point Distribution Model, Model distribuce bodů) pro ohraničení prvků tváře souborem bodů.



Obrázek 10: Extrakce rysů tváře
Zdroj: převzato dle Le, 2011 a upraveno pro potřeby práce

3. **Extrakce rysů** (Feature Extraction) – probíhá na základě extrakce rysů fotometrické a geometrické normalizace.

   Přístupy k extrakci rysů mohou být ale různé:

   a. Metoda založená na geometrických vlastnostech

   b. Skupina metod založená na šablonách

   c. Skupina metoda založená na segmentaci barev

   d. Skupina metod založená na vzhledu

   (Le, 2011)

4. **Porovnání rysů tváře** (Feature Matching) – porovnání extrahovaných rysů tváře analyzované osoby s rysy tváří osob v databázi.

Pro správnou detekci rysů tváře je důležitý úhel, ze kterého je osoba snímána. Pokud analýza rysů tváře zpracovává snímek, kde je osoba vyfocena zepředu a hlava je orientována rovně, je úspěšnost detekce výrazně vyšší než v případě, kdy je snímána z jiného úhlu, či má jinak orientovanou hlavu. (Gruber, 2015)

Detekce emocí z výrazu lidské tváře

Základní kroky detekce emocí:

1. **Detekce tváře** viz výše.

2. **3D modelování tváře s využitím AAM (Active Appereance Model)**

3. **Klasifikace výrazu tváře –** využití neuronových sítí, které jsou naučené na velkém množství snímků (které byly anotované manuálně).

3.3.3 Analýza videa

Audiovizuální záznam se skládá z audio a video stopy.

Audio stopa je složena z jednotlivých snímků, které na sebe v určitém sledu navazují tak, aby vytvářely reálný dojem pohybu objektů videa. Na video analýzu lze tedy použít stejné druhy analýz jako na analýzu fotek, obrázků apod.

Za audio je považován takový zvuk, který je pro člověka slyšitelný (tedy přibližně zvuk mezi 20 až 20 000 hertzy).

Tento druh analýzy představuje efektivní vizualizaci obsahu videa.

Princip analýzy videa je založen na analýze obrázků/snímku (viz předchozí kapitola 3.3.2 Analýza digitálního snímku). Každé video je poskládáno z určitého počtu snímku dle parametru FPS (frames per second, resp. počet snímků za vteřinu), takže nástroje určené pro tuto analýzu zpracovávají snímky videa stejný způsobem jako jakékoliv jiné obrázky/snímky.

Kromě záznamu obrazu videa je možné zpracovat také zvukový záznam videa. Zde záleží na tom, jaký je požadovaný výstup, viz kapitola 3.3.1 Analýza zvuku.

Analýza audiovizuálního záznamu bývá v současné době využívána především na veřejných prostranstvích, kde je vyšší riziko nebezpečí, např. letiště, nádraží, soudy apod. Nástroje pro bezpečnostní analýzu videa dokáží detekovat zbraň, podezřelé chování atd.

3.4 Analýza dat v reálném čase

V reálném čase jsou nejčastěji zpracovávány logové záznamy. Této problematice se věnuje následující kapitola 3.4.1 Zpracování logových záznamů.

3.4.1 Zpracování logových záznamů

Zpracování tzv. proudů dat (jinak streamovaných dat, datových toků) v téměř reálném čase dovoluje reagovat na současnou situaci (využívané např. pro zajištění bezpečnosti, sledování transakcí, sledování zdravotního stavu uživatele apod.). Vstupní proudy dat jsou generovány senzory, měřiči, sociálními sítěmi, výrobními zařízeními atd. Této problematice se věnuje také Kudyba (2014).

Analýza proudů dat se využívá např. v energetice, zdravotnictví, bezpečnosti, ve výrobě, telekomunikaci, e-commerce a v dalších odvětvích. Tento typ zpracování nestrukturovaných dat se využívá např. při detekci zločinů, viz kapitola 2.5.1.1 Fraud Detection (detekce podvodů), jak blíže popisuje Kudyba (2014), ale také při tzv. Condition monitoringu výrobních zařízení.

Výsledkem analýzy proudů dat může být varování, výzva ke konkrétní akci, upozornění apod.

EDA (Event Driven Architecture)

Jedná se o styl návrhu a konstrukce IT systémů. Tento systém je založen na konstrukci systému z nezávislých komponent. Tyto komponenty používají komunikační schéma publish-subscribe. Jednotlivé komponenty mezi sebou sdílí pouze formát událostí.

Tento architektonický styl návrhu IT systémů rozlišuje 3 úrovně zpracování událostí:

- Zpracování jednoduchých událostí (Simple Event Processing)
- Zpracování proudů událostí (Event Stream Processing neboli ESP)
- Zpracování komplexních událostí (Complex Event Processíng neboli CEP)

CEP – Complex Event Processing

Pro komplexní zpracování dat (událostí) v reálném čase se v současné době využívá technologie CEP.

Cílem dané technologie je zpracovat komplexní události z několika zdrojů v reálném čase a za co nejkratší dobu tak, aby bylo možné na probíhající události reagovat s co nejkratší časovou prodlevou (téměř v reálném čase). Tato technologie poskytuje kontinuální vhled na probíhající události, což umožňuje detekci možných hrozeb či příležitostí, jak popisují ve svém článku Filip Nguyen a Tomáš Pitner (2012).

Jak ve své práci uvádí Crha (2012), události dělíme na události vyššího a nižšího řádu. Událostí nižšího řádu v tomto případě rozumíme jako jeden záznam (řádek) z logových záznamů – záznamů aktivity dané aplikace či daného zařízení (např. pokus o přihlášení do systému). Událostí vyššího řádu rozumíme například zablokování uživatele v systému. Jednotlivé události mají mezi sebou určité vztahy, např. vztahy v rámci času (událost A se stala před událostí B), kauzální vztahy (událost A způsobila událost B).

Vzájemné a komplexní propojení událostí z jednotlivých systémů na jednotlivých úrovních umožňuje dávat tyto události do souvislostí (využitím operací jako je korelace, filtrování, detekce vzorů a závislostí).

V systémech využívajících CEP, např. v SIEM (Security Information and Event Management), musí být nastavena korelační pravidla (určitý sled událostí různých řádů), resp. správné detekční scénáře, při kterých má být uživatel upozorněn. na možné nebezpečí.

CEP komponenty - EPN (Event Processing Network)

EPN popisuje, jakými komponentami jsou data přijímána, zpracována a odesílána dalším systémům/aplikacím. Tento model vysvětluje architekturu systémů pro zpracování proudů dat např. CEP. (Sharon a Etzion, 2007)

Tento model se skládá z několika komponent:

- Tvůrce události (Event Producer)
- Příjemce události (Event Consumer)
- Zpracovatel událostí (Event Processing Agent)
- Spojovací agent (Event Channel)

ESP – Event Stream Processing

Jak uvádí Kudyba (2014), jedná se o další techniku pro zpracování proudů Big dat neboli technika pro analýzu dat v pohybu. ESP je používána i pro historickou analýzu.

Tato technika je řazena pod technologii CEP. ESP se řadí k technikám, které využívají model kontinuálních dotazů k analyzování proudů dat v pohybu. (Kreps, 2015)

ESP systémy jsou obvykle navrženy tak, aby byly schopny zpracovat velké objemy proudů dat (událostí) s velmi krátkou časovou latencí. Vzhledem k tomuto faktu jsou tyto systémy využívány např. ke zpracování událostí z kapitálových trhů, kde je potřeba se rychle rozhodovat, ale také i systémech, které se specializují na zajištění kybernetické bezpečnosti či na prediktivní analýzu. (Kudyba, 2014)

3.4.1.1 Condition monitoring

Jedním z druhů analýz, která využívá logové záznamy, je v současné době velmi populární tzv. Condition monitoring (anglicky CM).

Tato analýza zpracovává data v reálném čase a monitoruje, zda data odpovídají stanoveným parametrům (stanoveným podmínkám). Jedná se například o monitorování výrobního stroje, ze kterého lze sbírat data o vývoji teploty stroje, míry opotřebení, hlučnosti, vytížení stroje apod. Poté prostřednictvím správně nastavených pravidel (podmínek) lze určit, zda je stroj správně seřízen, přiměřeně vytížen, zda je nutné provést servis či výměnu nebo lze očekávat poruchu stroje. (Reeves, 1998)

Nejčastěji jsou sledovány parametry stroje, které mohou predikovat nějakou poruchu či chybu stroje – zde je využito prediktivní analýzy.

Tato analýza je využívána především ve výrobě a v logistice, její zásluhou šetří společnost nemalé náklady předcházením rizik – na základě sesbíraných dat je možné plánovat údržbu či výměnu strojů (např. právě výrobních zařízení, nákladních automobilů apod.), jak popisuje Bartoš (2018).

4 Přínosy zpracování nestrukturovaných dat

Pro dosažení kýžených výsledků je nutné definovat zdroje dat, ze kterých budou data dolována, jak a jestli data budou někde ukládána, jak budou dál řízena, jaká data budou analyzována a jak využít získané poznatky.

Před analýzou dat je nutné, položit si otázky na/ohledně:

- **Zdroje dat**

  Při analýze trhu není potřeba dolovat data z výrobních systémů. Zdroje jsou vybírány v závislosti na cílech dané analýzy.

- **Uložení a řízení dat**

  Rozhodnutí zda data ukládat nebo zpracovávat v reálném čase a pokud budou ukládány, jak dlouho je ukládat a jaké technologie k tomu využívat.

- **Data k analýze**

  Určení relevantních dat – zda nějaká data vyloučit, či zpracovávat všechna data (nedoporučuje se). Např. pro analýzu sentimentu je třeba zpracovat volné texty a již není nutné znát podrobnosti o každém z autorů daného textu.

- **Poznatky**

  Čím více relevantních poznatků je získáno, tím mají společnosti větší jistotu při rozhodování.

4.1 Podpora rozhodování

Jak již bylo zmíněno v prvních kapitolách této práce, data jsou velmi cennou komoditou. Data ale sama o sobě neposkytují žádné informace, nejdříve je nutné jim přiřadit potřebný význam.

Hlavním důvodem zpracování Big dat, vlastně jakýchkoliv dat, je získání potřebných odpovědí, nových informací a poznatků. Data jsou zpracována a vizualizována za účelem zprostředkování informací ve srozumitelné, přehledné a interaktivní formě. Lidský mozek je více vnímavý vůči barvám, tvarům, zvukům, vůním a dalším vjemům, nežli vůči číselným údajům. Právě z vizualizace je lidská psychika schopná vytěžit nejvíce informací, jak uvádí Černý (2016). Na základě kvalitní analýzy a vizualizace dat se mohou firmy správně rozhodovat.

Existuje velké množství druhů analýz Big dat. Důležité je určit, jaké informace chceme získat. Pokud chce společnost zjistit, jak se nový produkt líbí lidem, je vhodné využít analýzu sentimentu z dat ze sociálních sítí. Naopak, pokud se firma potřebuje strategicky rozhodnout, zda se dál rozvíjet, či nikoliv, je vhodné použít prediktivní analýzu.

Pokud společnosti zpracovávají Big data, roste jejich produktivita – společnosti dělají správná rozhodnutí na základě kvalitně zpracovaných dat z různých systémů.

V minulosti nebylo možné zpracovávat taková množství dat z tolika různých zdrojů, což způsobovalo, že společnosti nebyly dostatečně informované: neměly potřebné podklady k učinění správného rozhodnutí.

V současné době mají manažeři velmi usnadněnou práci, svá rozhodnutí mohou konat na základě výsledků z nejrůznějších druhů analýz.

Analýza nestrukturovaných dat pomáhá firmám, ale naopak škodí původcům dat, jako jsou uživatelé sociálních sítí. Ti jsou soustavně sledováni, jsou o nich sbírána data. Společnosti poté snáze ovlivňují uživatele, pokud o nich mají potřebné informace.

4.2 Zajištění bezpečnosti

Dalším přínosem pro společnosti při využití správných nástrojů a postupů je zajištění vysoké bezpečnosti napříč celou organizací.

Současné moderní technologie umožňují riziku předcházet, ale i efektivně řešit bezpečnostní incidenty. V této práci jsou popsány dvě nejznámější techniky, které zpracovávají velké množství nestrukturovaných dat a hledají v nich určitě vzory či podezřelá data, která by mohla naznačovat jakékoliv bezpečnostní riziko.

V práci je popsána technologie SIEM a také metoda pro detekci podvodného jednání – Fraud Detection.

4.3 Minimalizace rizika

Při správném nastavení prahových hodnot chytrých zařízení (např. výrobních strojů) je možné předcházet neočekávaným výpadkům strojů či jejich poruše, a tak minimalizovat riziko ztrát. Touto problematikou se zabývá tzv. Condition monitoring, o tomto druhu analýzy blíže pojednává kapitola 3.4.1.1 Condition monitoring výše.

4.4 Optimalizace

Dalším důležitým přínosem je možnost optimalizace výroby, skladových zásob, vytíženosti zaměstnanců, produktového portfolia atd. Např. pokud jsou sbírána data z výrobních či jiných zařízení, je možné porovnat vytíženost jednotlivých strojů a rovnoměrně rozložit zátěž mezi všechny stroje a zabránit tak extrémnímu vytížení jednotlivých strojů.

Pokud je využito pokročilých nástrojů, které získávají data z různých zdrojů, je možné upravit zátěž jednotlivých objektů výroby i dle stavu zařízení či dle efektivity jednotlivých zařízení, zaměstnanců apod. např. s využitím Condition monitoringu, který je popsán výše.

Optimalizace by měla probíhat pravidelně (např. kvartálně), aby bylo možné reagovat na vzniklé změny a byl zefektivněn chod společnosti.

5 Nástroje pro zpracování nestrukturovaných dat

Nástrojů pro zpracování nestrukturovaných dat je v současné době nepřeberné množství a není možné v této práci všechny obsáhnout.

Z tohoto důvodu je tato kapitola zaměřena na nástroje, které jsou užity v praktické části práce – tedy nástroj NTeX, Přepisovatel.cz, Geneea, FaceReader a Clarifai.

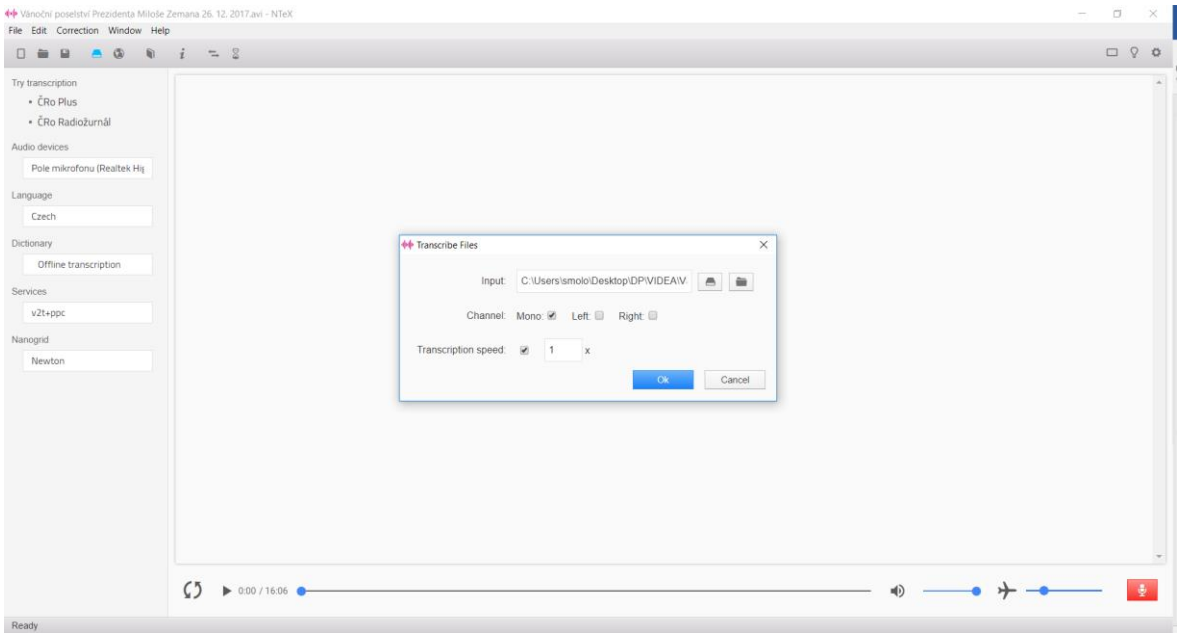
5.1 Nástroj NTeX

Jedná se o demo verzi aplikace NEWTON SpeechGrid společnosti NEWTON technologies.

Podporované jazyky pro přepis: bosenština, bulharština, černohorština, čeština, chorvatština, maďarština, makedonština, polština, ruština, slovenština, slovinština, srbština, švédština, ukrajinština.

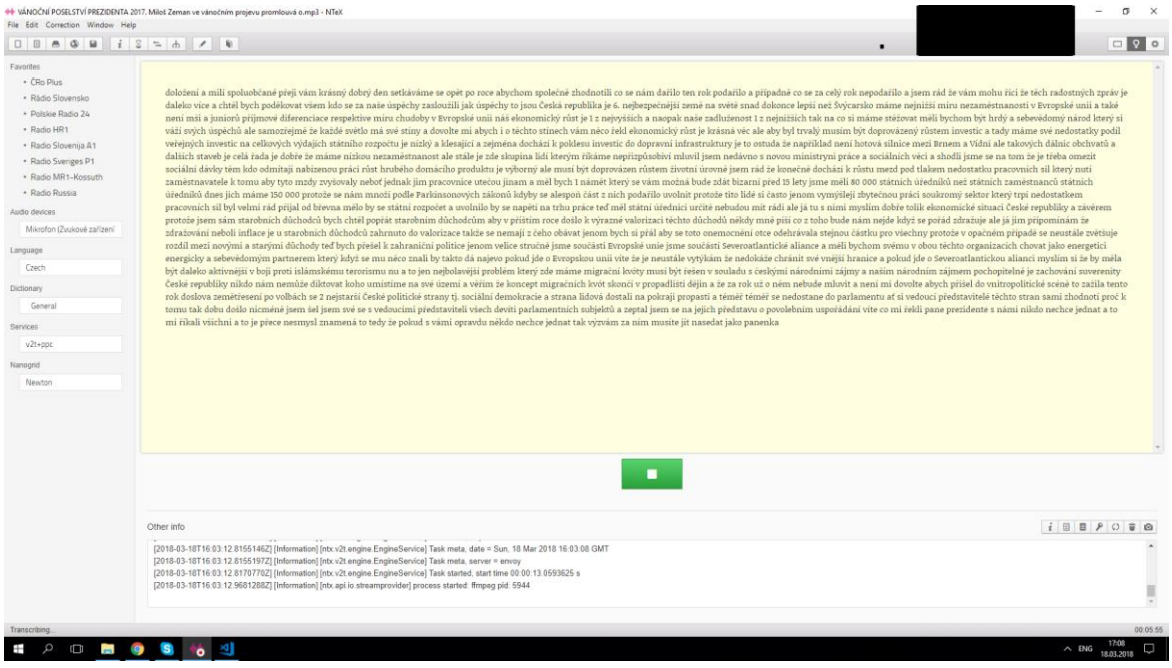
Tento nástroj zprostředkovává přepis videa na text v reálném čase. Lze ho využít i pro přímé diktování.

Postup přepisu



Obrázek 11: NTeX – volba záznamu

Zdroj: služba na stránce firmy NTeX - http://www.ntex.cz/help/en/index.html



Obrázek 12: NTeX – přepis audiovizuálního záznamu na text v reálném čase

Zdroj: služba na stránce firmy NTeX - http://www.ntex.cz/help/en/index.html

Nástroj NTeX podporuje přepis audiovizuálních souborů z těchto formátů: .aac, .avi, .flac, .flv, .ismv, .m4a, .mkv, .mov, .mp2, .mp3, .mp4, .oga, .ogg, .opus, .wav, .webm, .wma, .wmv.

Vygenerovaný přepis je možné uložit ve formátech: .nta, .ntx, .txt, .xml, .trsx, .srt, .tovek, .oga.

Dále nabízí možnost přepisu internetového streamu v reálném čase – po zadání URL adresy záznamu. Aplikace dovoluje i přepis přímo upravovat (dělat např. korekci přepisu), přidávat slova do slovníku apod. Defaultní slovník této služby je vytvořen a spravován tvůrci služby, uživatelé mohou doplnit nová slova.

Nástroj je volně ke stažení, po instalaci ale požaduje přihlášení ke službě NEWTON SpeechGrid.

5.2 Přepisovatel.cz

Tento nástroj nepodporuje přepis audiovizuálního záznamu v reálném čase. Je nutné vybraný soubor přes webové rozhraní nahrát a poté počkat, až bude přepis hotov. Po dokončení přepisu je uživatel upozorněn e-mailem.

Tento nástroj umožní bezplatně zpracovat pouze 60 minut audio záznamu. Pro zpracování záznamu je nutné mít vytvořený uživatelský účet.

Mezi technologie nástroje Přepisovatel.cz patří i detekce řeči, rozlišení mluvčích a zarovnání textu s audiem. V této práci tyto možnosti není třeba využít, jelikož je analyzován monolog – proslov prezidenta České republiky.

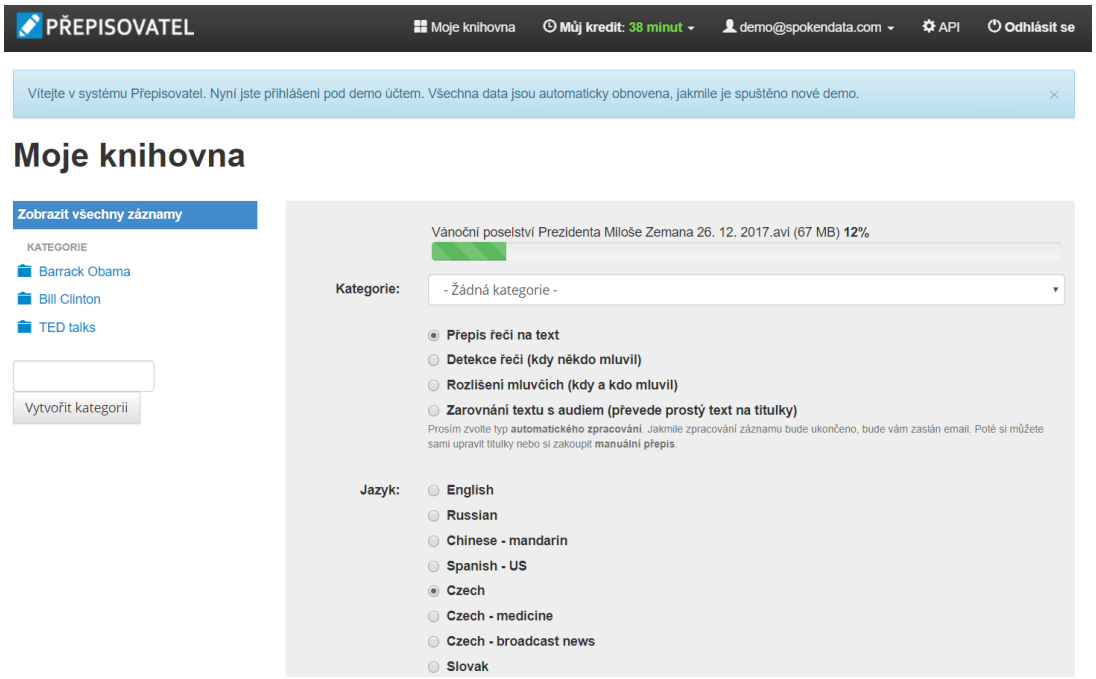
Po vytvoření přepisu je možné textový přepis upravit či objednat manuální přepis.

Provozovatelem je společnost ReplayWell.

Obrázek 13: Překladatel.cz – jak funguje?
Zdroj: https://www.prepisovatel.cz/

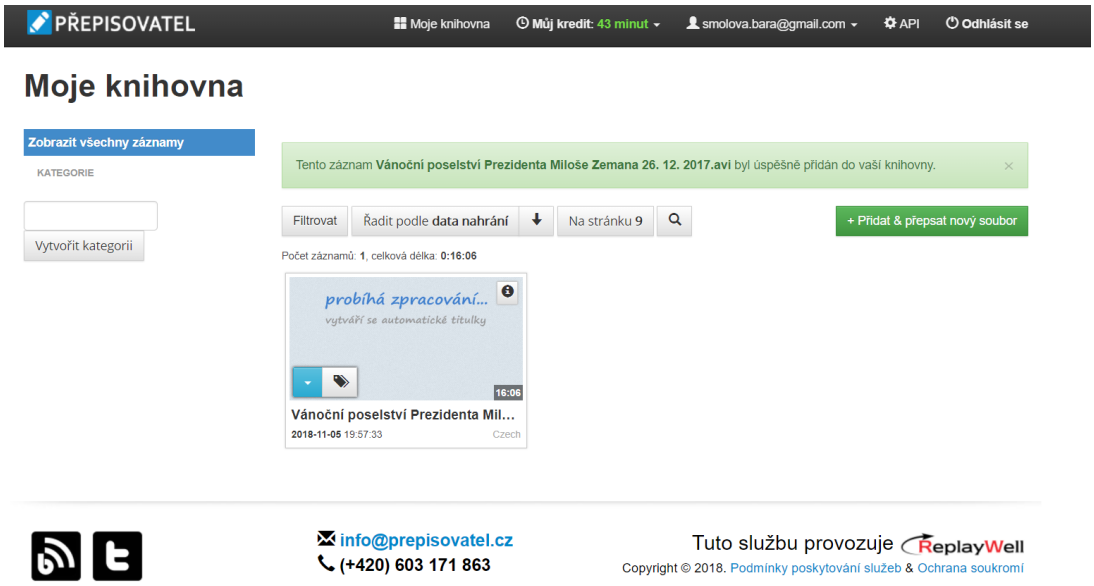
Podporované jazyky pro přepis: angličtina, ruština, čínština, španělština, čeština, slovenština.

Služba Přepisovatel nabízí stažení přepisu v různých formátech – .txt, .html, .srt, .trs, WebVTT, .xml.



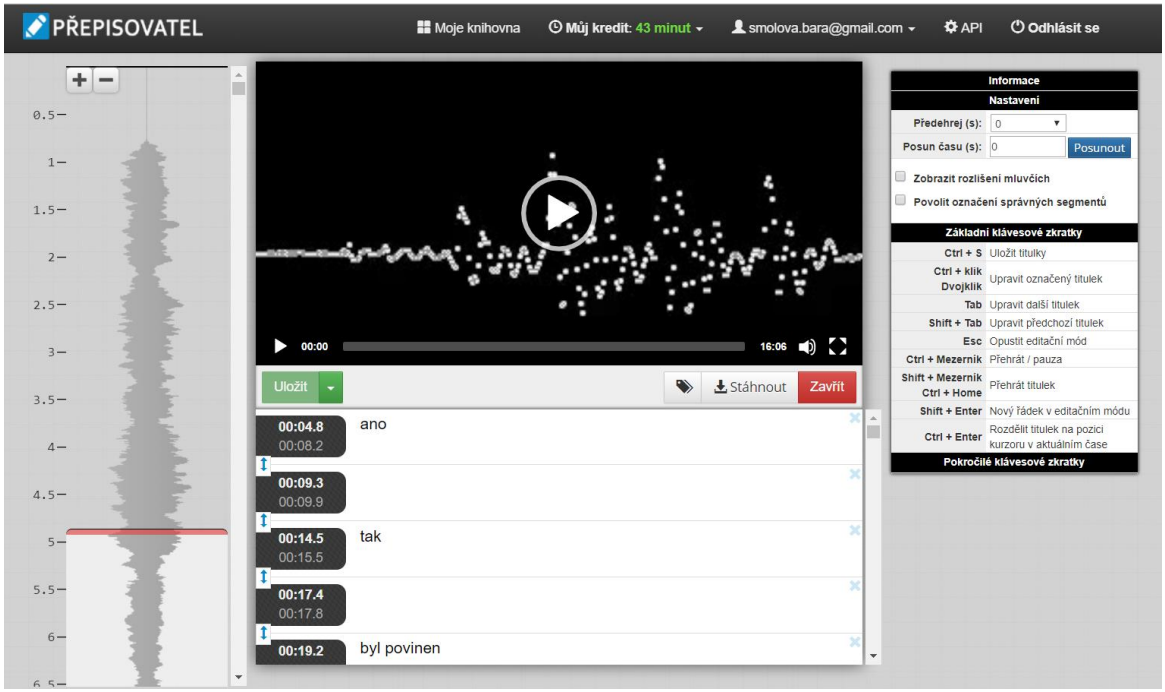
Obrázek 14: Přepisovatel – nastavení parametrů pro přepis projevu na text
Zdroj: Zpracován vybraný audiozáznam Vánočního poselství prezidenta Miloše Zemana ze dne 26. 12. 2017 službou, která je volně dostupná na stránce https://www.prepisovatel.cz.

Po zpracování souborů se záznam objeví v Knihovně uživatele – zde se nachází všechny zpracované soubory přihlášeným uživatelským účtem.



Obrázek 15: Přepisovatel – na pozadí aplikace probíhá zpracování zvoleného záznamu
Zdroj: Zpracování záznamu vánočního poselství prezidenta České republiku roku 2018 službou dostupnou na stránce https://www.prepisovatel.cz.

Po provedení přepisu nám nástroj zobrazí akustické spektrum audio části záznamu a také vygenerovaný přepis, rozdělený do jednotlivých časových intervalů, viz Obrázek níže.



Obrázek 16: Přepisovatel – výsledek přepisu vybraného záznamu
Zdroj: Zpracování záznamu vánočního poselství prezidenta České republiku roku 2018 službou dostupnou na stránce https://www.prepisovatel.cz.

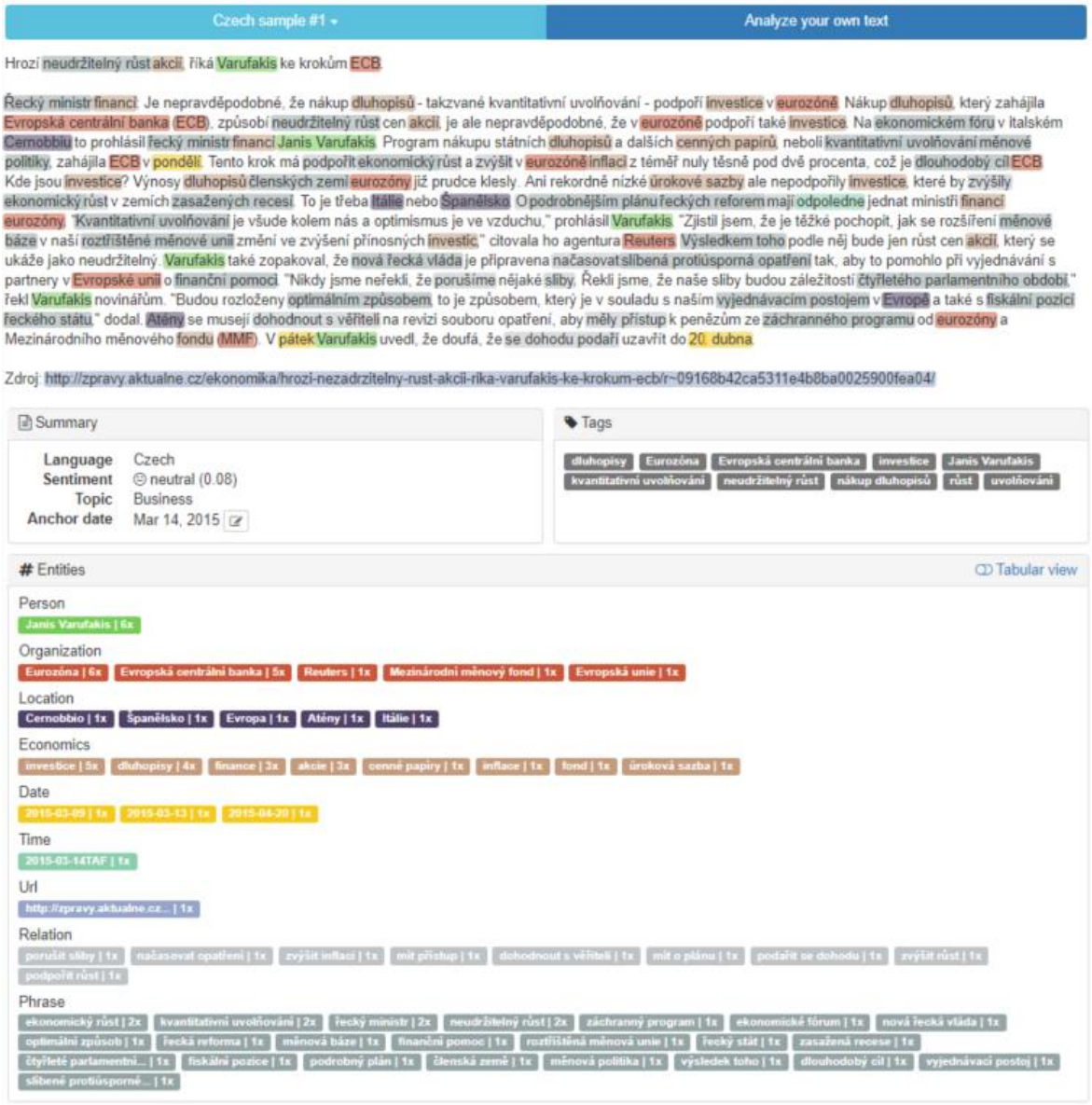
Přepis záznamu je po přihlášení k dispozici na stránce služby Přepisovatel.cz (https://www.prepisovatel.cz/transcription/42372).

5.3 Geneea

Nástroj Geneea nabízí stejně pojmenovaná převážně česká společnost Geneea, která se zabývá analýzou velkého množství textových dat se speciálním zaměřením na analýzu sentimentu.

Společnost dále nabízí i nástroj Frida, který je zaměřen také na textovou analytiku.

Nástroj Geneea kromě detekce sentimentu nabízí i detekci jazyka (český/anglický atd.), identifikaci tématu, automatické štítkování textu (tzv. tagování), extrakce entit textu a opravu diakritiky.



Obrázek 17: Příklad zpracování textových dat nástrojem Geneea

Zdroj: demoverze služby Geneea dostupné z https://demo.geneea.com/

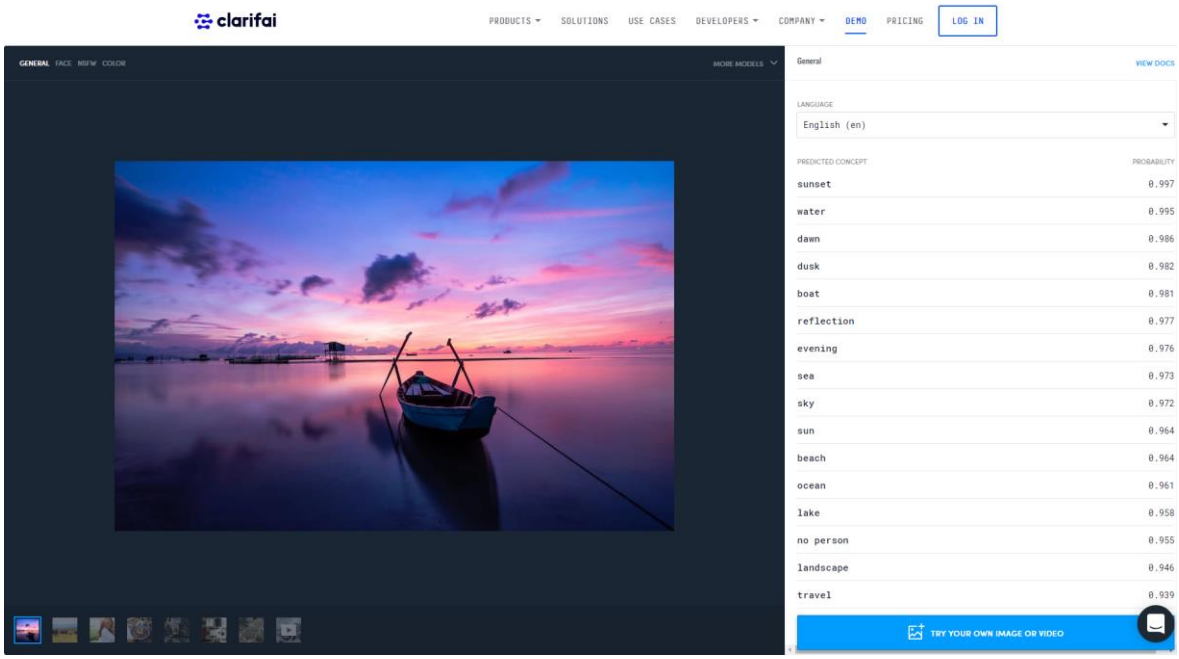
5.4 Clarifai

Clarifai je společnost zabývající se především umělou inteligencí. Ve svých technologiích využívá znalost neuronových sítí a strojového učení k rozpoznání objektů na obrázcích či v audiovizuálních záznamech.

Aplikace nabízená touto firmou je pojmenována stejně jako název firmy – Clarifai.

Aplikace nepodporuje český jazyk, je k dispozici pouze v angličtině.

Aplikace Clarifai nabízí velké množství druhů analýz videa či fotografie/obrázku (např. analýza výrazu tváře, osobnostních rysů člověka, demografických údajů barev apod.).



Obrázek 18: Aplikace Clarifai
Zdroj: https://clarifai.com/demo

5.5 FaceReader

Jedná se o profesionální nástroj pro automatické rozpoznání a analýzu výrazů tváře.

FaceReader dokáže rozpoznat výrazy, které lidská tvář má, když je člověk šťastný, smutný, naštvaný, překvapený, vyděšený, znechucený nebo neutrální. Tyto emoce jsou považovány za základní nebo univerzální lidské emoce. Dané emoce popsal Paul Ekman ve svém článku – Universal Facial Expressions of Emotion. (Ekman, 1970)

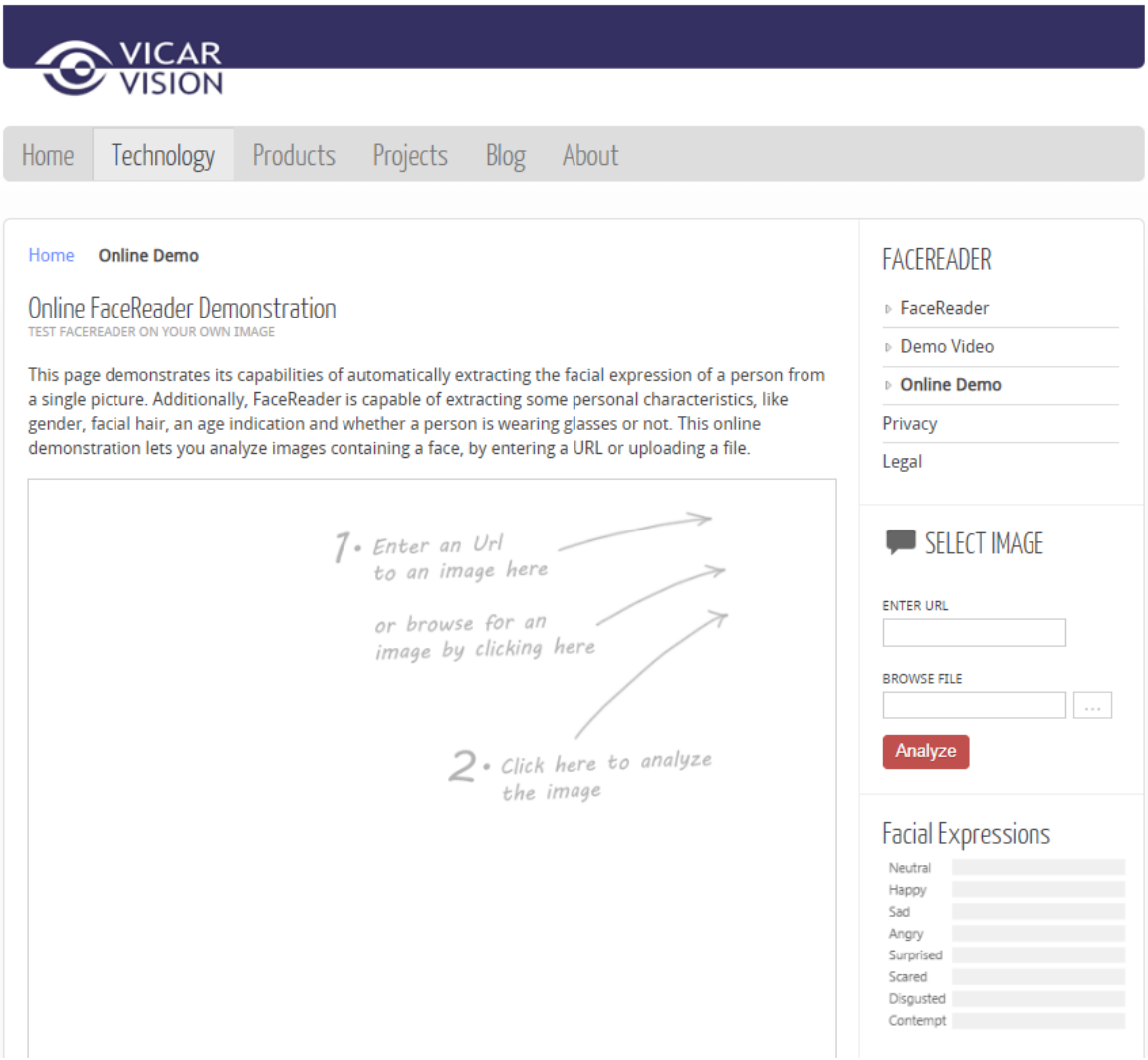
Tento nástroj umožňuje zpracovávat data z videa a fotek v reálném čase – např. prostřednictvím webkamery. V demo verzi tohoto nástroje je k dispozici pouze možnost analýzy fotografií.

Nástroj je vyvinut tak, aby byl schopný zpracovat až 15 snímků za vteřinu při analýze v reálném čase. Umožňuje videa i nahrávat a předzpracovat je.

Do aplikace FaceReader je možné přidávat další pocity, které je potřeba analyzovat. Není ale k dispozici možnost přidat si tento výraz uživatelsky, ale pouze na objednávku.

Praktický příklad využití tohoto nástroje je k dispozici v kapitole 6.3 Rozpoznávání emocí a osobnostních charakteristik z výrazu tváře.

FaceReader je nástrojem společnosti VicarVision.



Obrázek 19: Aplikace FaceReader

Zdroj: http://www.vicarvision.nl/products/online-facereader-demo/

6 Zpracování nestrukturovaných dat vybranými prostředky

Praktická část této diplomové práce je zaměřena na zpracování nestrukturovaných dat, konkrétně audiovizuálního záznamu vybranými nástroji.

V této kapitole jsou použity demo verze nástrojů, freeware či nástroje volně dostupné prostřednictvím webového prohlížeče.

Jako data byl zvolen vánoční proslov pana prezidenta Miloše Zemana, který přednesl 26.12.2017.

Záznam proslovu byl stažen z webu Youtube.com (záznam dostupný zde: https://www.youtube.com/watch?v=sV5PY8-v1I4) prostřednictvím webové služby QDowloander dostupné z https://qdownloader.net/youtube-video-downloader.

Parametry záznamu

Doba trvání:	00:16:03
Formát souboru:	.avi
Velikost:	67,1 MB

Video

Rozlišení:	1280 x 720
Rychlost dat:	439 kb/s
Celková přenosová rychlost (video + zvuk):	564 kb/s
Snímkovací frekvence:	50 fps (snímků za sekundu)

Zvuk

Přenosová rychlost:	125 kb/s
Kanály:	2
Vzorkovací frekvence zvuku:	44 100 kHz

Pro účely této práce bude zpracována nejen zvuková stopa, ale i vybrané snímky video záznamu. Autorce práce se nepodařilo získat přístup k pokročilým nástrojům pro analýzu videa, ale pouze pro analýzu jednotlivých snímků – např. aplikace pro rozpoznání emocí z lidské tváře, věku pohlaví apod.

Záznam je k dispozici na CD, které je přiloženo k diplomové práci.

Daný záznam byl autorkou vybrán cíleně, jedná se o záznam v českém jazyce a v poměrně vysoké kvalitě.

Tento proslov bude prostřednictvím vybraných nástrojů přepsán do textové podoby. Bude zhodnocena kvalita přepisu – nástroje budou vzájemně porovnány.

Kvalitnější přepis mluveného slova v českém jazyce bude podroben analýze sentimentu.

Následně budou jednotlivé snímky video záznamu podrobeny analýze sentimentu a analýze osobnostních charakteristik.

Další část práce je věnována analýze proslovu jako celku i dle jednotlivých témat, ke kterým se pan prezident při svém proslovu vyjadřuje.

6.1 Přepis audiovizuálního záznamu na text

Záznam, který je v rámci této práce zpracován, je v českém jazyce, z tohoto důvodu je nutné zvolit vhodný nástroj pro přepis, který umožňuje přepis českého jazyka na text.

6.1.1 NTeX

Tento nástroj vytváří přepis audiovizuálního záznamu v reálném čase.

Ukázka přepisu vánočního poselství pana prezidenta – realizace nástrojem NTeX:

Přepis audiovizuálního záznamu službou Přepisovatel.cz neobsahuje časové intervaly – nástroj generuje přepis do volného textu, viz Ukázka 1.

„*doložení a milí spoluobčané přeji vám krásný dobrý den setkáváme se opět po roce abychom společně zhodnotili co se nám dařilo ten rok podařilo a případně co se za celý rok nepodařilo a jsem rád že vám mohu říci že těch radostných zpráv je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jak úspěchy to jsou Česká republika je 6. nejbezpečnější země na světě snad dokonce lepší než Švýcarsko máme nejnižší míru nezaměstnanosti v Evropské unii a také není mši a juniorů příjmové diferenciace respektive míru chudoby v Evropské unii náš ekonomický růst je 1 z nejvyšších a naopak naše zadluženost 1 z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomý národ který si váží svých úspěchů ale samozřejmě že každé světlo má své stíny a dovolte* mi *abych i o těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvalý musím být doprovázený růstem investic a tady máme své nedostatky podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury*“ Strojový přepis projevu Vánočního poselství pana prezidenta Miloše Zemana ze dne 26. 12. 2018.

Ukázka 1 – přepis NTeX

Celý přepis tohoto záznamu prostřednictvím nástroje NTeX je dispozici v Příloze A diplomové práce a také na přiloženém CD.

6.1.2 Přepisovatel.cz

Záznam projevu je nahrán do webové služby cca 19:57:33 dne 5. 11. 2018, přepis byl dokončen ve 21:03 dne 5. 11. 2018. Přepis záznamu tímto nástrojem trval 1 hodinu, 2 minuty a 30 sekund.

Ukázka přepisu vánočního poselství pana prezidenta – realizace nástrojem Přepisovatel.cz:

Přepis audiovizuálního záznamu službou Přepisovatel.cz obsahuje časové intervaly, viz Ukázka 2.

Nástroj generuje přepis do struktury. Tato struktura je dělena časovými údaji, pod kterými je volný text rozdělen do jednotlivých řádků v závislosti na délce odmlk mezi slovy pana prezidenta.

Řádky jsou rozděleny do časových úseků dlouhých cca 2–3 sekundy v závislosti na délce odmlk. Jak je nastavena minimální délka odmlky, po které tato služba udělá konec řádku, nelze dohledat.

00:04.8 - 00:08.2

ano

00:09.3 - 00:09.9

00:14.5 - 00:15.5

tak

00:17.4 - 00:17.8

00:19.2 - 00:21.6

byl povinen

00:22.3 - 00:24.3

00:27.0 - 00:30.6

vložení mylný spoluobčané přeji vám

00:31.1 - 00:32.8

krásný dobrý den

00:34.2 - 00:36.8

setkáváme se opět port se

00:37.5 - 00:39.4

abychom společně

00:39.6 - 00:40.8

a zhodnotili

00:41.4 - 00:42.5

co se nám

00:42.8 - 00:47.7

a za ten rok podařilo a případně co se za ten rok nepodařilo

00:48.8 - 00:55.6

a jsem rámec že vám mohu říci takže těch radostných sprav je daleko více

Ukázka 2 – Přepis Přepisovatel.cz

Po odstranění těchto časových údajů dostaneme čistý text:

„ano tak byl povinen vložení mylný spoluobčané přeji vám krásný dobrý den setkáváme se opět port se abychom společně a zhodnotili co se nám a za ten rok podařilo a případně co se za ten rok nepodařilo a jsem rámec že vám mohu říci takže těch radostných sprav je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jaké úspěchy to jsou česká republika je šest a nejbezpečnější země na světě sme dokonce lepší než švýcarsko máme nejnižší mírou nezaměstnanosti evropské uni a pak je nejnižší míru přímo ve diferenciace respektive mýho chudoby evropské unii náš ekonomický růst je jeden z nejvyšších a naopak naše zadluženost jedna z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomí na který si váží svých úspěchu ale samozřejmě že každé světlo má své stíny a dovolte mi abych bych po těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvali musí být doprovázen růstem investic a tady máme sebe nedostatky podíl veřejných investic svá celokovových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury“ Strojový přepis projevu vánočního poselství pana prezidenta Miloše Zemana ze dne 26. 12. 2018.

6.1.3 Hodnocení kvality přepisu

Jednotlivé přepisy, které jsou vytvořeny v předchozích kapitolách 6.1.1 a 6.1.2, porovnáme s přepisem audiovizuálního záznamu vytvořeného člověkem, viz Ukázka 3. Zdrojem tohoto přepisu je proslov pana prezidenta uskutečněný dne 26.12.2017, který byl stažen ze služby Youtube.com (záznam dostupný zde: https://www.youtube.com/watch?v=sV5PY8-v1I4). Výstup byl tvořen ručně autorkou práce.

Námi hodnocené přepisy jsou bez struktury – jedná se o volné texty.

Přepis vánočního poselství Miloše Zemana z 26. prosince 2017 pořízený člověkem
„Vážení a milí spoluobčané, přeji vám krásný dobrý den,

setkáváme se opět po roce, abychom společně zhodnotili, co se nám za ten rok podařilo a případně, co se za ten rok nepodařilo. A jsem rád, že vám mohou říci, že těch radostných zpráv je daleko více. A chtěl bych poděkovat všem, kdo se za naše úspěchy zasloužili. Jaké úspěchy to jsou? Česká republika je šestá nejbezpečnější země na světě. Jsme dokonce lepší než Švýcarsko. Máme nejnižší míru nezaměstnanosti v Evropské unii, a také nejnižší míru příjmové diferenciace, resp. míru chudoby v Evropské Unii. Náš ekonomický růst je jeden z nejvyšších a naopak naše zadluženost jedna z nejnižších. Tak na co si máme stěžovat? Měli bychom být hrdý a sebevědomý národ, který si váží svých úspěchů.
Ale samozřejmě, že každé světlo má své stíny a dovolte mi, abych i o těchto stínech vám něco řekl.
Ekonomický růst je krásná věc, ale aby byl trvalý, musí být doprovázen růstem investic. A tady máme své nedostatky. Podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající. A zejména dochází k poklesu investic do dopravní infrastruktury. Je ostuda, že například není hotová silnice mezi Brnem a Vídní, ale takových dálnic, obchvatů a dalších staveb je celá řada. …… „

Ukázka 3 – Lidský přepis

Celý text přepisu je obsažen v Příloze E diplomové práce a také na přiloženém CD.

Kvalita jednotlivých přepisů je určena na základě porovnání lidského přepisu záznamu se strojovými přepisy vytvořenými prostřednictvím nástrojů NTeX a Přespisovatel.cz.

Stanovená pravidla pro hodnocení přepisů

Chyby byly rozděleny do dvou kategorií. Každá chyba v přepisu je označena červeným podbarvením textu. Každá chyba je v textu opravena a správná podoba slov je podbarvena v závislosti na závažnosti těchto chyb:

- hrubá chyba přepisu (opravný text je podbarven modře)
    - chybějící či přebývající slovo/písmeno (podbarveno červeně – slovo přebývá, podbarveno modře – slovo chybí)

- slovo obsahující více než dvě špatná písmena či dvě špatná diakritická znaménka (nebo kombinaci špatného diakritického znaménka a špatného písmene). Toto slovo neodpovídá proslovu a je podbarveno červeně a hned za ním je doplněno slovo, které odpovídá proslovu – podbarveno modře a je napsáno velkými písmeny.

- drobná chyba přepisu (opravný text je podbarven zeleně)

  - slovo obsahující jedno špatné písmeno na začátku či na konci slova a zároveň tím není změněn význam tohoto slova.

  - Slova, která jsou chybně oddělena mezerou.

- není hodnoceno jako chyba přepisu

  - interpunkční znaménka (neboli členící) – (např. tečka, čárka, vykřičník, otazník, pomlčka apod.).

  - rozdělení do odstavců (struktura textu – přepisy jsou v současné době vždy ve formě volného textu).

  - chyba v dalších znacích (např. %, vícenásobné mezery, teplotní stupně apod.).

Problematiku interpunkčních znamének se dosud nepodařilo vyřešit uspokojivým způsobem. V současnosti se daří úlohu řešit stále lépe s využitím hlubokých neuronových sítí, viz například výzkum společnosti IBM (Pahuja, Laha a spol, 2017). Tato znaménka vyjadřují strukturu a organizaci textu – jsou závislá na řečníkovi a především na jeho intonaci a na přestávkách v řeči. V předchozích řešeních se pauzy a intonace používaly k určení interpunkce, zatímco výše zmíněný článek problém řeší na textových přepisech řeči, tedy bez využití zvukové informace.

Následující ukázka chybovosti jednotlivých přepisů je dlouhá cca 250 slov – cca první dva odstavce proslovu.

Ukázka přepisu textu prostřednictvím služby – NTeX se znázorněnými chybami vybraného vzorku.

„doloženíVÁŽENÍ a milí spoluobčané přeji vám krásný dobrý den setkáváme se opět po roce abychom společně zhodnotili co se nám dařiloZA ten rok podařilo a případně co se za celýTEN rok nepodařilo a jsem rád že vám mohu říci že těch radostných zpráv je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jakÉ úspěchy to jsou Česká republika je 6. nejbezpečnější země na světě snadJSME dokonce lepší než Švýcarsko máme nejnižší míru nezaměstnanosti v Evropské unii a také není mši a juniorůNEJNIŽŠÍ MÍRU příjmové diferenciace respektive míru chudoby v Evropské unii náš ekonomický růst je 1 z nejvyšších a naopak naše zadluženost 1 z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomý národ který si váží svých úspěchů ale samozřejmě že každé světlo má své stíny a dovolte mi abych i o těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvalý musímMUSÍ být doprovázenýDOPROVÁZEN růstem investic a tady máme své nedostatky podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například není hotová silnice mezi Brnem a Vídni ale takových dálnic obchvatů a dalších staveb je celá řada je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidí kterým říkáme nepřizpůsobiví mluvil jsem nedávno s novou ministryni práce a sociálních věci a shodli jsme se na tom že je třeba omezit sociální dávky těm kdo odmítají nabízenou práci“

Celý text s opravenými chybami v přepisu je k dispozici v Příloze B diplomové práce.

Ukázka přepisu textu prostřednictvím služby – Přepisovatel.cz se znázorněnými chybami vybraného vzorku.

„ano tak byl povinen vloženíVÁŽENÍ A mylnýMILÍ spoluobčané přeji vám krásný dobrý den setkáváme se opět port sePO ROCE abychom společně a zhodnotili co se nám a za ten rok podařilo a případně co se za ten rok nepodařilo a jsem rámecRÁD že vám mohu říci takže

těch radostných spravZPRÁV je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jaké úspěchy to jsou českáČESKÁ republika je šest aŠESTÁ nejbezpečnější země na světě smeJSME dokonce lepší než švýcarskoŠVÝCARSKO máme nejnižší mírouMÍRU nezaměstnanosti V evropskéEVROPSKÉ uniUNII a pak jeTAKÉ nejnižší míru přímo veCENOVÉ diferenciace respektive mýhoMÍRU chudoby V evropskéEVROPSKÉ unii náš ekonomický růst je jeden z nejvyšších a naopak naše zadluženost jedna z nejnižších tak na co si máme stěžovat měli bychom být hrdýHRDÍ a sebevědomíSEBEVĚDOMÝ naNÁROD který si váží svých úspěchuÚSPĚCHŮ ale samozřejmě že každé světlo má své stíny a dovolte mi abych bychI poO těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvaliTRVALÝ musí být doprovázen růstem investic a tady máme sebeSVÉ nedostatky podíl veřejných investic sváNA celokovovýchCELKOVÝCH výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například naNENÍ něho tohoHOTOVÁ silnice mezi brnemBRNEM a vídníVÍDNÍ aALE na takových dál nicDÁLNIC A obchvatů a dalších staveb je celá řada je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidiLIDÍ kterým říkáme nepřízpůsobivíNEPŘIZPŮSOBIVÍ mluvil jsem í. nedavnaNEDÁVNO seS mnouNOVOU ministryni práce a sociálníCH věciÍ a shodli jsme se na tom takžeŽE JE třeba omezit sociální dávky těm kdo odmítají nabízenNABÍZENOU úprav siPRÁCI“

Celý přepis s opravenými chybami je k dispozici v Příloze D diplomové práce.

Přepis vánočního poselství pana prezidenta Miloše Zemana službou NTeX je na vybraném vzorku dat výrazně kvalitnější.

Přepis službou NTeX

- Nástroj začal přepisovat tuto nahrávku od začátku proslovu pana prezidenta – z přepisu byla vynechána instrumentální hudba.

- Psaní velkých a malých písmen (např. u názvů měst, státu, společnosti apod.) v tomto přepisu je v našem vzorku bezchybné.

- Volný text je bez časových údajů.

- Před spuštěním přepisu byl uživatelem nastaven jazyk (čeština), přepis se držel stanovených parametrů, tedy používal česká slova.

- Žádné spojky nejsou vynechány, pouze pár jich přebývá.

Přepis službou Přepisovatel.cz

- Přepis byl zahájen ještě před proslovem pana prezidenta – snaha o přepis instrumentální hudby – špatně rozpoznaná instrumentální hudba versus řeč.

- Tento přepis nerozlišuje velká a malá písmena, přepis je celý vytvořen užitím malých písmen.

- Volný text je s časovými údaji.

- Přepis byl uživatelsky nastaven na český jazyk – přesto jsou v textu špatně přepsána slova do anglického jazyka (např. right/rád apod.).

- Je zde velmi časté vynechávání spojek a předložek v textu (např. a, v, z, k, o apod.), malé množství spojek také přebývá.

Celkové zhodnocení kvality přepisu užitých nástrojů

Tabulka 1: Celkové zhodnocení kvality přepisu užitých nástrojů

	Počet slov	Chybných slov v přepisu – hrubá chyba	Chybných slov v přepisu – drobná chyba	Chyb celkem	chybovost v %
Lidský přepis	1260	0	0	0	0
Přepis nástrojem NTeX	1281	70	20	90	7,1%
Přepis nástrojem Přepisovatel.cz	1271	201	91	292	23,2%
Celkem	X	272	110	382	X

Zdroj: vlastní zpracování

Ačkoliv by se na první pohled mohlo zdát, že nástroj Přepisovatel.cz je na pokročilejší úrovni nežli nástroj NTeX, protože generuje svá data do určité struktury a reaguje na délky odmlk, není tomu tak.

Nástroj od společnosti Newton Technologies NTeX má nižší chybovost nežli nástroj Přepisovatel.cz (7,1 % < 23,2 %). Dle výsledku porovnání lze pro přepis textu v českém jazyce doporučit spíše nástroj NTeX.

Přestože je chybovost stále poměrně vysoká (s využitím nástroje NTeX cca 7,1 %), lze očekávat, že se s vývojem nových technologií, postupů a algoritmů bude snižovat.

6.2 Analýza sentimentu přepisu

V této kapitole se zaměříme na provedení analýzy sentimentu přepisu od společnosti NTeX, který jsme v předchozí kapitole označili za kvalitnější. Pro tyto účely je využita demo verze webového nástroje společnosti Geneea, dostupného z https://demo.geneea.com/, ve kterém je možné analyzovat najednou 3000 znaků, včetně mezer, pokud uživatel není registrovaný u této společnosti.

6.2.1 Celý text

V této kapitole bude analyzován přepis jako celek.

Vzhledem k pravidlům nastaveným společnosti Geneea, jejichž nástroj je v práci pro analýzu sentimentu použit, je nutné zkoumaný přepis řeči rozdělit na 3 části a poté vytvořit průměr výsledků sentimentu. Každá z analyzovaných částí má přibližně 2 536 znaků, včetně mezer.

Výsledek analýzy sentimentu je v rozmezí <-1,1>.

Tabulka 2: Interpretace výsledku sentimentu, dle parametrů nástroje Geneea

Výsledek	Význam
< -1; -0,1 >	Negativní sentiment
< -0,1; +0,1 >	Neutrální sentiment
< +0,1; 1 >	Pozitivní sentiment

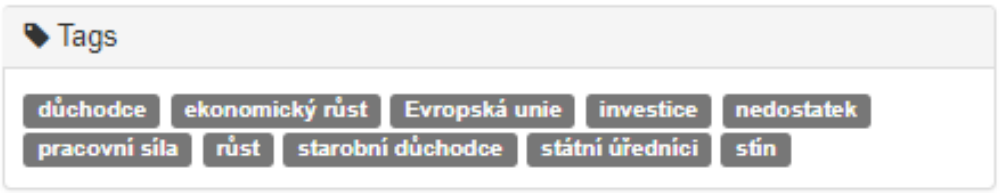
Zdroj: vlastní zpracování

1. **Část**

Jak je patrné z Obrázek 20 níže, nástroj společnosti Geneea v textu znázorní vybrané objekty (které jsou nositelem sentimentu textu) a také analyzovaný text oštítkuje. Štítky 1. části našeho přepisu jsou vyobrazeny, viz *Obrázek **20***, který se nachází níže.

Obrázek 20: Analýza sentimentu přepisu záznamu nástrojem NTeX, 1 část

Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/

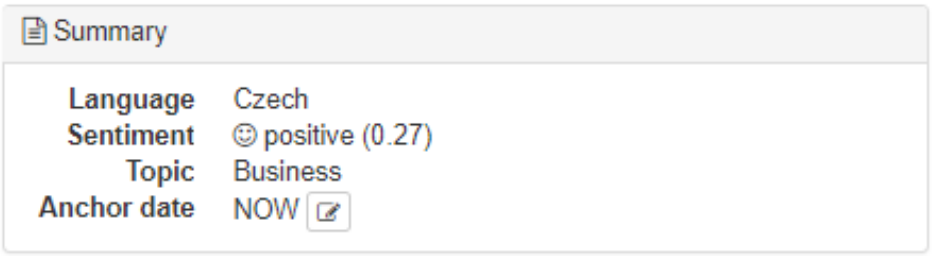


Obrázek 21: Výsledek štítkování přepisu textu službou Geneea, 1. část

Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/

Obrázek 22: Identifikované objekty při analýze přepisu záznamu, 1. část

Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/



Obrázek 23: Výsledek analýzy sentimentu službou Geneea, 1. část přepisu záznamu
Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/

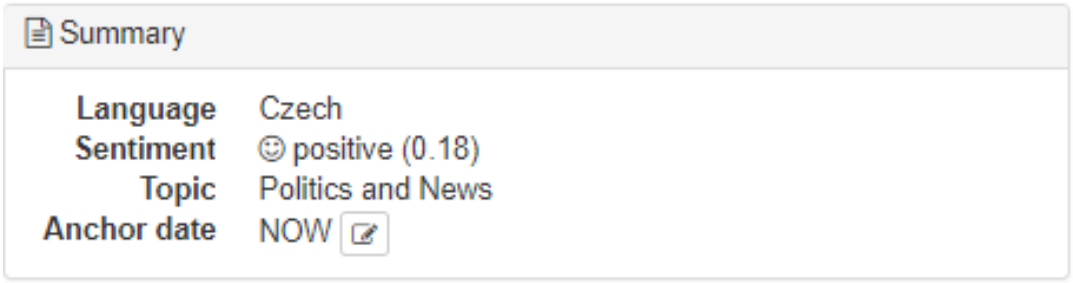
1. část záznamu má výsledek sentimentu 0,27 – tato část textu se tedy řadí mezi pozitivní a téma je obchodní.

2. Část

Druhá část proslovu pana prezidenta se věnuje tématu kvót, inflaci, důchodům, Evropské unii, organizaci NATO a národním zájmům. V této části textu zmiňuje známé osoby české politiky, pana Babiše a pana Sobotku.

Obrázek 24: Identifikované štítky a objekty při analýze přepisu záznamu, 2. část

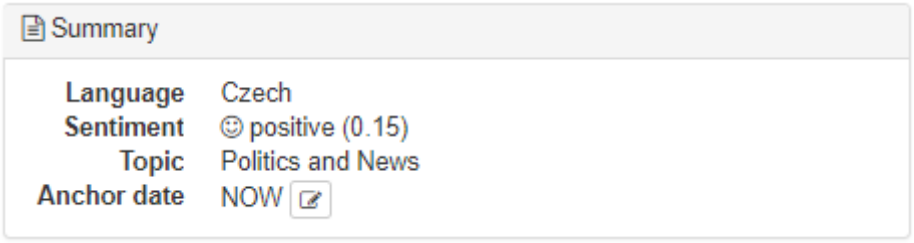
Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/



Obrázek 25: Výsledek analýzy sentimentu službou Geneea, 2. část přepisu záznamu
Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/

2. část analyzovaného přepisu záznamu má výsledek analýzy sentimentu 0,18, tedy pozitivní. Tato část textu je věnována politice a novinkám.

3. **Část**



Obrázek 26: Výsledek analýzy sentimentu službou Geneea, 3. část přepisu záznamu
Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/

Poslední, tj. 3. část přepisu má výsledek 0,15, jedná se tedy také o pozitivní sentiment. Pan prezident při svém proslovu pokračuje v tématu – politika a novinky.

Výsledek a zhodnocení

Výsledek analýzy sentimentu přepisu záznamu společností NTeX bude vytvořen zprůměrováním výsledných hodnot jednotlivých částí.

(0,27 + 0,18 + 0,15)/3 = 0,2

Z analýzy sentimentu celého textu vyplynulo, že proslov pana prezidenta Miloše Zemana mělo pozitivní sentiment (+ 0,20).

Tento výsledek lze interpretovat tak, že pan prezident vnímá vývoj v roce 2017 jako spíše pozitivní.

6.2.2 Jednotlivá témata proslovu

V této části práce rozdělíme vánoční poselství pana prezidenta na jednotlivá odvětví, která ve svém proslovu zmiňuje, např., EU, zahraniční a vnitrostátní politika, zahraničí, starobní důchod, migrační kvóty atd.

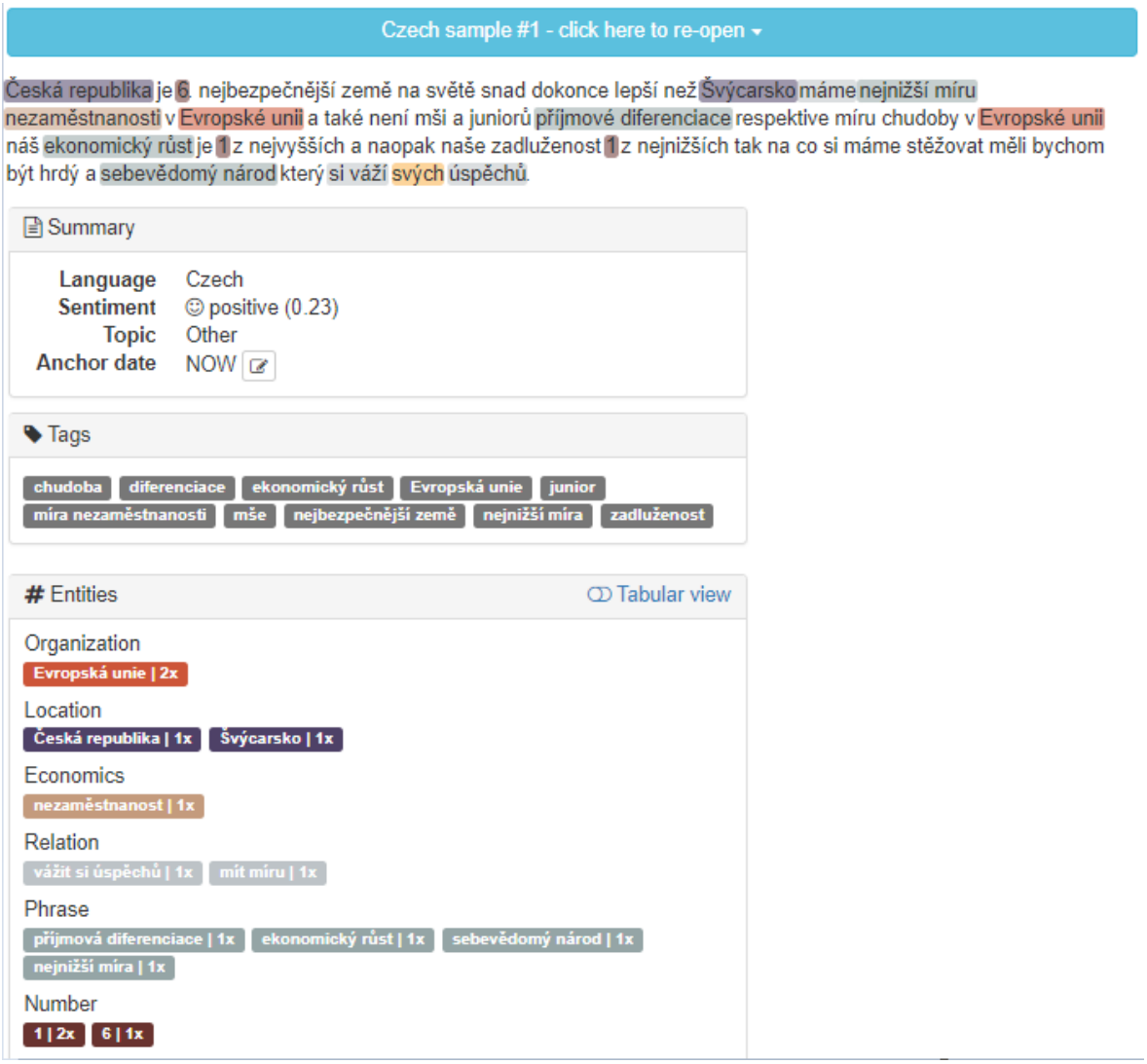
Jak již bylo několikrát zmíněno v této kapitole, je analyzován přepis vánočního poselství pana prezidenta Miloše Zemana ze dne 26.12.2017 – přepis byl realizován nástrojem NTeX. Tento text se vztahuje převážně k roku 2017.

Text přepisu byl rozdělen za účelem této analýzy do základních 4 skupin. Dané skupiny se dále dělí na podskupiny dle tématu analyzovaného záznamu. Jednotlivé podskupiny jsou podrobeny analýze sentimentu – webovým nástrojem společnosti Geneea. Skupiny byly zvoleny autorkou na základě vlastního uvážení.

Text, který spadá pod jednotlivé skupiny, je v textu zvýrazněn barevným podbarvením.

- **Česká republika jako celek**
    - Zhodnocení České republiky jako celku

      Výsledek analýzy sentimentu části textu, zaměřující se na vývoj České republiky v roce 2017, je roven + 0,23, jedná se tedy o pozitivní sentiment, viz Obrázek 27 – Pozitivní sentiment. Výsledná hodnota převyšuje i výsledek zhodnocení celého textu, který je proveden v kapitole výše. Pan prezident tedy hodnotí vývoj České republiky jako celek za rok 2017 velmi kladně.



Obrázek 27: Analýza jednotlivých témat textu – zhodnocení České republiky jako celku

Zdroj: vlastní zpracování s využitím demoverze služby Geneea dostupné z https://demo.geneea.com/

- **Vývoj ekonomické situace v České republice**

  - Ekonomický vývoj České republiky

    Vývoj ekonomické situace je nástrojem společnosti Geneea ohodnocen na + 0,21 – pozitivní sentiment. Tento výsledek i mírně převyšuje výslednou hodnotu sentimentu celého přepisu textu.

  - Nezaměstnanost - nepřizpůsobiví

Téma nepřizpůsobivých pan prezident dle výsledku analýzy sentimentu hodnotí velmi negativně (- 0,29).

- Životní úroveň/Hrubý domácí produkt České republiky (HDP, anglicky GDP)

  Tato část přepisu proslovu je analýzou sentimentu společnosti Geneea zhodnocena jako nejvíce pozitivní ze všech ostatních témat (+ 0,32).

- Státní úředníci

  Vývoj státních úředníků je ohodnocen kladně (+ 0,16).

- Vývoj starobních důchodců

  Nástroj společnosti Geneea tuto část ohodnotil pozitivním sentimentem (+ 0,21). Na základě tohoto zjištění lze usuzovat, že pan prezident se domnívá, že situace pro starobní důchodce se bude vyvíjet pozitivně.

- **Zahraniční politika**

  - Obecné

    Toto téma je ohodnoceno na sentiment + 0,26 – tedy velmi pozitivně. Lze tedy usuzovat, že pan prezident hodnotí stav zahraniční politiky kladně.

  - Migrační kvóty

    Analýza této části textu označila výsledek sentimentu hodnotou 0.00 – zcela neutrální.

- **Vývoj vnitropolitické scény**

  - Zhodnocení vývoje v roce 2017

    Ohodnoceno pozitivním sentimentem + 0,22. Dle výsledku lze tedy usuzovat, že byl pozitivní vývoj vnitropolitické scény v roce 2017.

    - Manipulace s výsledky voleb

      Situaci ohledně údajné manipulace s výsledky voleb pan prezident dle analýzy sentimentu hodnotí neutrálně, ale s negativní hodnotou - 0,07.

  - Předpověď 2018

    Sentiment v této části vychází velmi kladně (+ 0,31). Z toho lze usuzovat, že pan prezident věří v kladný vývoj vnitropolitické scény v roce 2018.

  - Předčasné volby

    V této části textu je sentiment ohodnocen neutrálně (0,00).

doložení a milí spoluobčané přeji vám krásný dobrý den setkáváme se opět po roce abychom společně zhodnotili co se nám dařilo ten rok podařilo a případně co se za celý rok nepodařilo a jsem rád že vám mohu říci že těch radostných zpráv je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jak úspěchy to jsou Česká republika je 6. nejbezpečnější země na světě snad dokonce lepší než Švýcarsko máme nejnižší míru nezaměstnanosti v Evropské unii a také není mši a juniorů příjmové diferenciace respektive míru chudoby v Evropské unii náš ekonomický růst je 1 z nejvyšších

a naopak naše zadluženost 1 z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomý národ který si váží svých úspěchů.

ale samozřejmě že každé světlo má své stíny a dovolte mi abych i o těchto stínech vám něco řekl

ekonomický růst je krásná věc ale aby byl trvalý musím být doprovázený růstem investic a tady máme své nedostatky podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například není hotová silnice mezi Brnem a Vídni ale takových dálnic obchvatů a dalších staveb je celá řada

je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidí kterým říkáme nepřizpůsobiví mluvil jsem nedávno s novou ministryni práce a sociálních věci a shodli jsme se na tom že je třeba omezit sociální dávky těm kdo odmítají nabízenou práci

růst hrubého domácího produktu je výborný ale musí být doprovázen růstem životní úrovně jsem rád že konečně dochází k růstu mezd pod tlakem nedostatku pracovních sil který nutí zaměstnavatele k tomu aby tyto mzdy zvyšovaly neboť jednak jim pracovnice utečou jinam a měl bych 1 námět který se vám možná bude zdát bizarní před 15 lety jsme měli 80 000 státních úředníků než státních zaměstnanců státních úředníků dnes jich máme 150 000 protože se nám množí podle Parkinsonových zákonů kdyby se alespoň část z nich podařilo uvolnit protože tito lidé si často jenom vymýšlejí zbytečnou práci soukromý sektor který trpí nedostatkem pracovních sil byl velmi rád přijal od břevna mělo by se státní rozpočet a uvolnilo by se napětí na trhu práce teď měl státní úřednici určitě nebudou mít rádi ale já tu s nimi myslím dobře tolik ekonomické situaci České republiky

a závěrem protože jsem sám starobních důchodců bych chtěl popřát starobním důchodcům aby v příštím roce došlo k výrazné valorizaci těchto důchodů někdy mně píší co z toho bude nám nejde když se pořád zdražuje ale já jim připomínám že zdražování neboli inflace je u starobních důchodců zahrnuto do valorizace takže se nemají z čeho obávat jenom bych si

přál aby se toto onemocnění otce odehrávala stejnou částku pro všechny protože v opačném případě se neustále zvětšuje rozdíl mezi novými a starými důchody

teď bych přešel k zahraniční politice jenom velice stručně jsme součásti Evropské unie jsme součástí Severoatlantické aliance a měli bychom svému v obou těchto organizacích chovat jako energetici energicky a sebevědomým partnerem který když se mu něco znali by takto dá najevo pokud jde o Evropskou unii víte že je neustále vytýkám že nedokáže chránit své vnější hranice a pokud jde o Severoatlantickou alianci myslím si že by měla být daleko aktivnější v boji proti islámskému terorismu

nu a to jen nejbolavější problém který zde máme migrační kvóty musí být řešen v souladu s českými národními zájmy a naším národním zájmem pochopitelně je zachování suverenity České republiky nikdo nám nemůže diktovat koho umístíme na své území a věřím že koncept migračních kvót skončí v propadlišti dějin a že za rok už o něm nebude mluvit

a není mi dovolte abych přišel do vnitropolitické scéně to zažila tento rok doslova zemětřesení po volbách se 2 nejstarší České politické strany tj. sociální demokracie a strana lidová dostali na pokraji propasti a téměř téměř se nedostane do parlamentu ať si vedoucí představitelé těchto stran sami zhodnotí proč k tomu tak dobu došlo nicméně jsem šel jsem své se s vedoucími představiteli všech devíti parlamentních subjektů a zeptal jsem se na jejich představu o povolebním uspořádání víte co mi řekli pane prezidente s námi nikdo nechce jednat a to mi říkali všichni a to je přece nesmysl znamená to tedy že pokud s vámi opravdu někdo nechce jednat tak výzvám za ním musíte jít nasedat jako panenka v koutě a jednání může zúročit i s vaší iniciativy a mělo by to být jednáni o které se bude týkat jak programových tak personál o nich kompromisu jak zpěv politice zvykem myslím si že návrh programového prohlášení Babišovy vlády je prostě takové diskuse docela dobrým výchozím materiálem když mluvím o Babišově vládě chtěl bych poznamenat že jsem dostával moudré rady které mi říkali nejmenuji Babišovu vládu dřív než Babiš získat důvěru v Poslanecké sněmovně ale to by reálně znamenalo že po týdny možná dokonce po měsíce tady povládne Sobotkova vláda v demisi a to si snad proboha nikdo nepřežil bych si vás strašit Belgií kde jednání o vládě trvalo 2 roky myslím si že během několika málo měsíců se podaří mít shodil vládu s důvěrou Poslanecké sněmovny a nebude-li úspěch hned v 1. kole vytvořím časový

prostor pro důkladná jednání politických partnerů tak aby snad někdy v únoru se uskutečnil 2. pokus který již může být úspěšný

nezapomeňte že v politice jsou 2 extrémy 1. pokus obklíčit vítěze voleb vytvořit proti němu jakousi pseudo koalici těch méně úspěšných a izolovat ho 2. extrém vítěz zašlape do země všechny poražené a vládne islám a myslím si že z hlediska demokratických procedur je dobré usilovat o to aby žádný z těchto extrémů se neuskutečnilo a aby politické strany spolu jednali a CSc. dohodnou vytvořily patrně menšinovou tolerovanou vládu závěrem této části bych chtěl potěšit občany a nepotěšit politiky

občas slyším hlasy že bych měl vypsat předčasné volby což mi ústava v některých situaci umožňuje chci naprosto jasně říci že to nikdy neudělám protože předčasné volby několik měsíců po řádných volbách by byly výsměchem občanů kteří šli k řádným volbám a volili tak jak volili občané rozdali karty to je a politici musí umět s těmito kartami hrát občan vy nemůžete vyměnit olej můžete vyměnit politiky proč ne někteří z nich možná skončí v opozici a jak napsal Rudolf Bechyně starý sociální demokrat za 1. republiky sucha je z piva opozice a já dodávám klatovský 2 nikdy není a ani posolené a

takže nepřeji někomu aby skončilo v opozici ale na 2. straně demokracie a samozřejmě určitou opozici v parlamentu vyžaduje až si každý sám rozhodne zda takovou opozici chce být věřícím že nedojde k tomu že se budeme obviňovat z toho že výsledky minulých parlamentních voleb byly zmanipulovány zahraničními rozvědkami je to ubohé je to trapné a je to urážející volby byly svobodnému a jsem rád že i Bezpečnostní informační služba! A prohlášení kde konstatuje že k žádnému takovému ovlivňování nedošlo jsme svobodný názor na ně se svobodnými občany a dnes už si můžeme říci že jsme i úspěšná země před dvěma lety jsem vám řekl že skončila blbá nálada dnes vám říkám že se nemáme co stydět a že je mnoho věci na které můžeme býti hrdi takže z líhně na hlavu a buď dnes byli vědomi lidmi kteří spoléhají na svůj vlastní rozum a nedají slovy manipulovat ať už údajnými zahraničními rozvědkami nebo zejména českým tiskem Českou televizi a dalším mi vždy spoléhají více

na vlastní rozum a já mám pro to své heslo které jsem vám už jednou říkal věřím že zdravý rozum zvítězí nad závistivou hloupostí

a není mi dovolte abych připil na vítězství zdravého rozumu na úspěch České republiky a na úspěch nás všech

Šťastný a Veselý nový rok

6.2.3 Výsledek a zhodnocení analýzy sentimentu

Výsledek analýzy sentimentu celého textu je 0,20 – jak již bylo zmíněno v kapitole 6.2 Analýza sentimentu přepisu výše, je tedy možné považovat celé vánoční poselství pana prezidenta Miloše Zemana ze dne 26.12.2018 za pozitivní.

V předchozí kapitole byla provedena analýza sentimentu jednotlivých témat proslovu. Většina témat vykazovala kladný sentiment, který se dal očekávat vzhledem ke kladnému sentimentu celého proslovu.

Negativně či neutrálně byly ohodnoceny tato témata:

- nezaměstnanost – nepřizpůsobiví (- 0,29)
- migrační kvóty (0,00)
- manipulace s výsledky voleb (- 0,07)
- předčasné volby (0,00)

Sentiment je vyhodnocen na základě pravidel a metod, které má nástroj společnosti Geneea nastaven, je tedy pravděpodobné, že se v případě využití jiného nástroje pro analýzu sentimentu budou výsledky sentimentu mírně odlišovat.

Analýza sentimentu není v současné době na takové úrovni, aby bylo možné její výsledky považovat za vždy správné. Například téma v předchozí kapitole – Státní úředníci – pan prezident ve svém projevu podotýká, že máme velké množství státních úředníků, kteří nejsou potřeba a kteří si sami vymýšlejí svou práci, a dále, že by bylo vhodné tyto úředníky přesunout do soukromého sektoru, kde je nedostatek pracovních sil. Analýza sentimentu vyhodnotila tuto část s pozitivním sentimentem (+ 0,16), což neodpovídá realitě. Toto téma by dle autorky práce mělo mít výsledný sentiment spíše negativní – pan prezident negativně hodnotí současnou situaci, negativně hodnotí nárůst počtu státních pracovníků za posledních patnáct let.

Problematika nesprávného určení sentimentu je velmi úzce spojena se způsoben užití jazyka a slov, např. pokud člověk používá ironii, či větu poskládá tak, že působí pozitivně.

Přesnost analýzy sentimentu s využitím současných nástrojů stále není 100%, ale platí pravidlo, že čím více textu podrobíme analýze sentimentu, tím přesnější bude náš výsledek. Resp. pokud analyzujeme text o 100 znacích, je přesnost výsledku analýzy sentimentu nižší, než při analýze 500 znaků textu.

Za poslední roky je možné pozorovat určité zlepšení ve vyhodnocení sentimentu textu a lze dále předpokládat, že s dalším vývojem budou výsledky analýzy sentimentu čím dál tím více odpovídat skutečnosti.

6.3 Rozpoznávání emocí a osobnostních charakteristik z výrazu tváře

V této části práce jsou analyzovány vybrané snímky audiovizuálního záznamu vánočního proslovu naše prezidenta.

Audiovizuální záznam byl stažen z webové služby Youtube.com. Snímky záznamu byly sebrány s využitím multimediálního přehrávače VLC ve verzi 2.2.6 Umbrella, kde bylo možné nastavit interval, ve kterém se mají snímky videa ukládat. Vytvořené snímky jsou všechny k dispozici na přiloženém CD. Dále jsou podrobeny analýze pouze vybrané vzorové snímky.

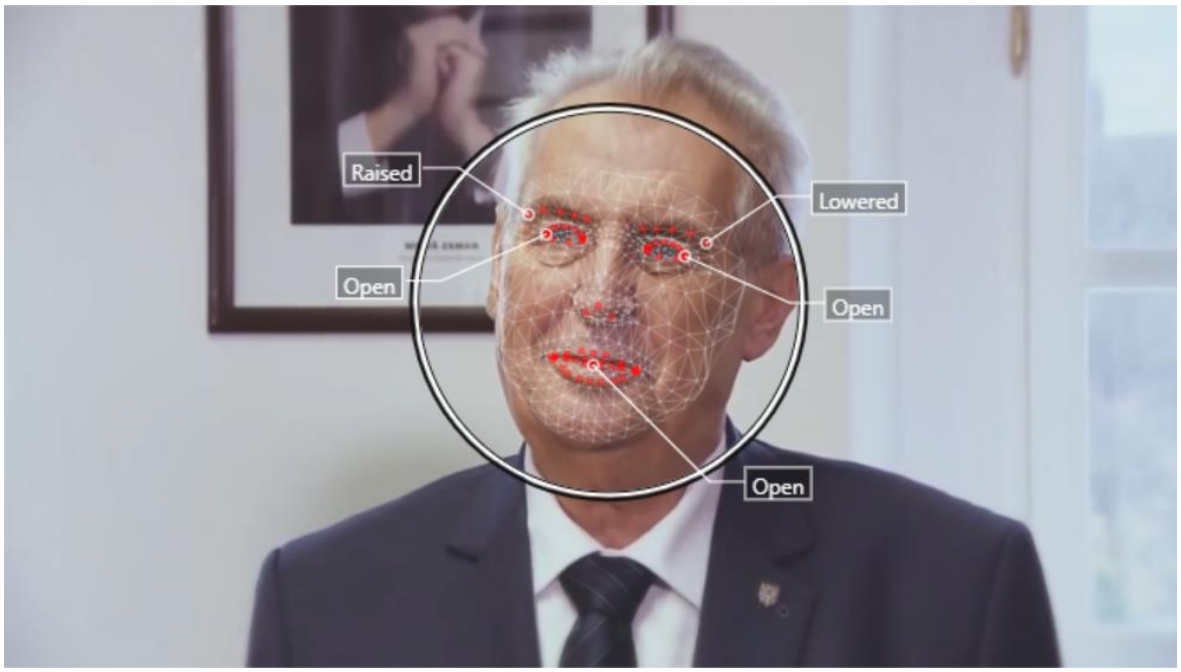
Snímky jsou analyzovány s využitím služby Online demoverze aplikace FaceReader od společnosti VicarVision – http://www.vicarvision.nl/.

Byly zvoleny 3 snímky, snímek ze začátku proslovu, z prostřední části proslovu a z jeho závěru.

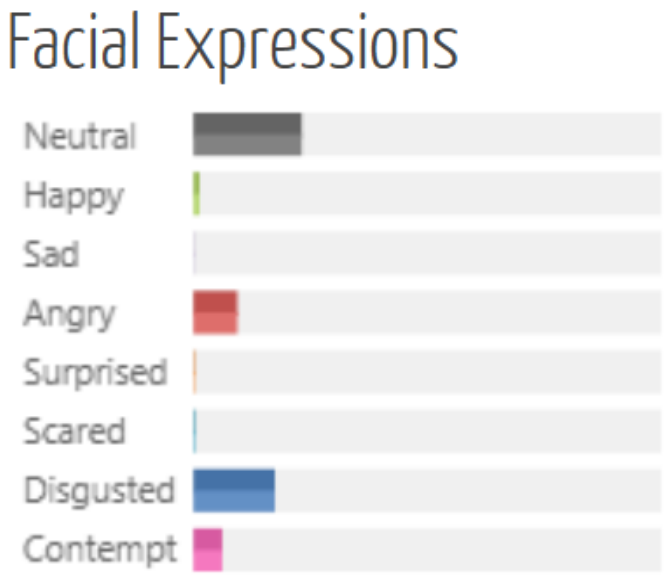
Snímek 1

Obrázek 28: Snímek videa 1

Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player

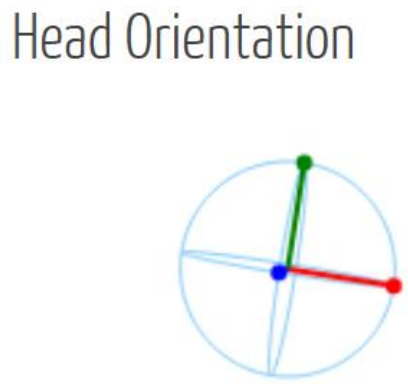


Obrázek 29: Snímek videa 1, podroben analýze sentimentu řečníka
Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player a zpracován online demo verzi aplikace FaceReader - http://www.vicarvision.nl/products/online-facereader-demo/



Obrázek 30: Snímek videa 1 (Obrázek 28), analýza výrazu tváře pana prezidenta
Zdroj:Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/

Snímek z Obrázku 28 je jeden z 50 obrazových snímků v čase 00:16 na vybraném audiovizuálním záznamu. Jak ukazuje výsledek analýzy na Obrázku 30, ze snímku je nejvíce patrný neutrální výraz pana prezidenta, ale také znechucení, zlost a pohrdání.



Obrázek 31: Snímek videa 1 (Obrázek 28), analýza orientace hlavy řečníka
Zdroj: Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/

Na tomto snímku pan prezident natáčí hlavu mírně doprava.

Odhadovaný věk pana prezidenta (45–55 let) neodpovídá skutečnosti, panu prezidentovi v době natočení video záznamu bylo 73 let. Pohlaví, vousy, knír a nošení brýlí nástroj detekoval správně.

Obrázek 32: Snímek videa 1 (Obrázek 28), analýza osobnostních charakteristik
Zdroj: Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/

PREDICTED CONCEPT	PROBABILITY
portrait	0.982
politician	0.971
people	0.962
adult	0.958
leader	0.955
one	0.951
administration	0.945
business	0.944
wear	0.912
man	0.904
election	0.900
festival	0.886
league	0.880

Obrázek 33: Snímek videa 1 (Obrázek 28), analýza scény snímku
Zdroj: vlastní zpracovaní s využitím nástroje Clarifai dostupná z https://clarifai.com/demo

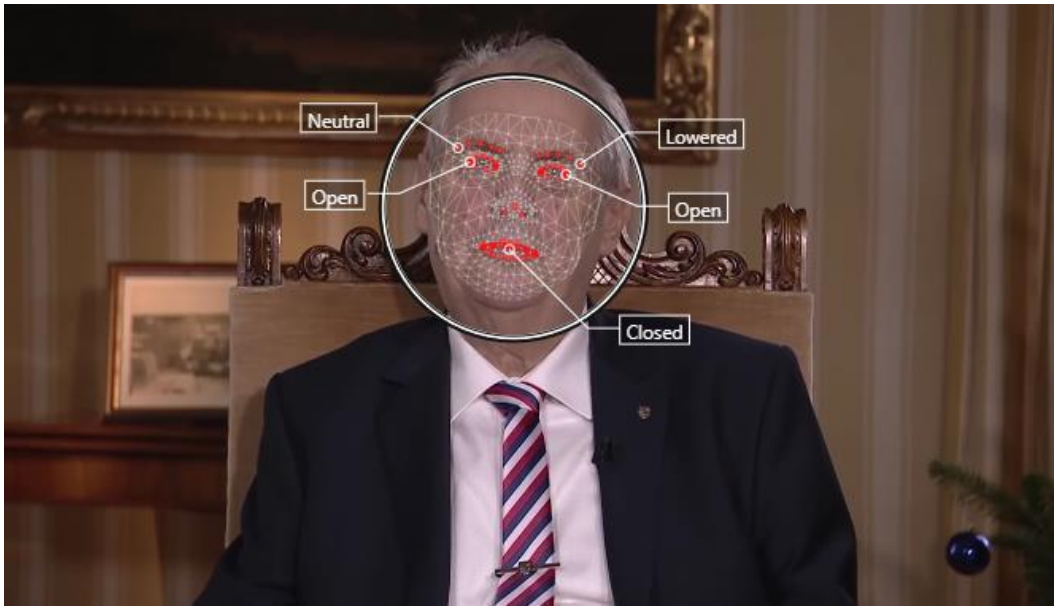
Na snímek 1 byla využita i analýza scény, viz Obrázek 33. Na základě této analýzy byl vytvořen seznam objektů na Obrázek 28 s jejich pravděpodobností správnosti. Mezi objekty tohoto snímku se řadí na první místo s nejvyšší pravděpodobnosti výskytu – portrét s 98,2 %, následuje objekt politik s 97,1 %, člověk s 96,2 % atd. Mezi identifikované objekty se dále řadí – dospělý člověk, vůdčí osobnost, obchod, muž, volby, festival, společnost/liga a další.

Snímek 2

Obrázek 34: Snímek videa 2

Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player

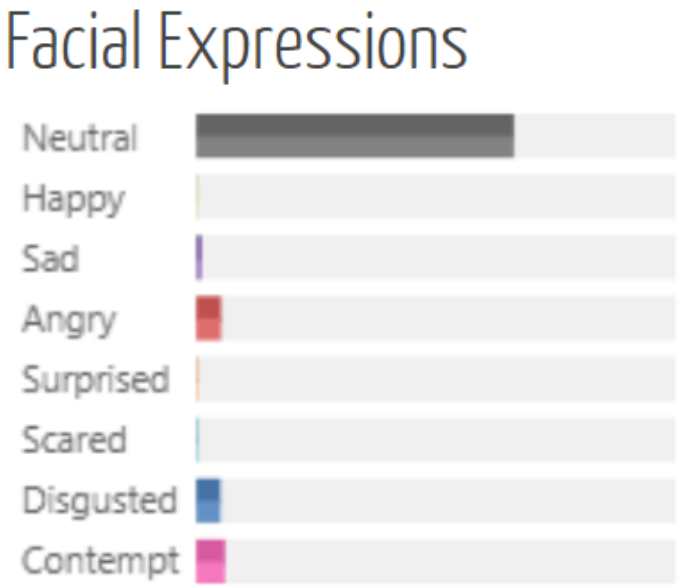
Obrázek 35 byl podroben analýze sentimentu a osobnostních charakteristik. Na Obrázku 35 jsou vytyčené tzv. anotované body.



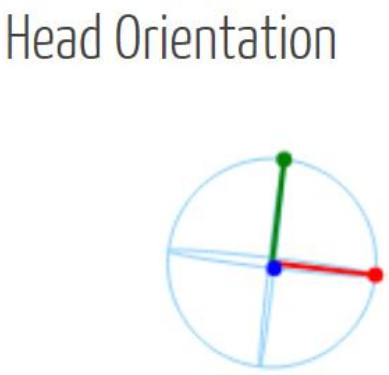
Obrázek 35: Snímek videa 2, podroben analýze výrazu tváře

Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player a zpracován online demo verzi aplikace FaceReader - http://www.vicarvision.nl/products/online-facereader-demo/

Dle výsledku analýzy vybraného snímku na Obrázku 36 lze usoudit, že pan prezident působí především neutrálním dojmem. Lze zde ale i vyčíst náznak zloby, znechucení, pohrdání a opravdu malý náznak smutku.



Obrázek 36: Snímek videa 2 (Obrázek 34), analýza výrazu tváře pana prezidenta
Zdroj: Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/



Obrázek 37: Snímek videa 2 (Obrázek 34), analýza orientace hlavy řečníka
Zdroj: Aplikace FaceReader - http://www.vicarvision.nl/products/online-facereader-demo/

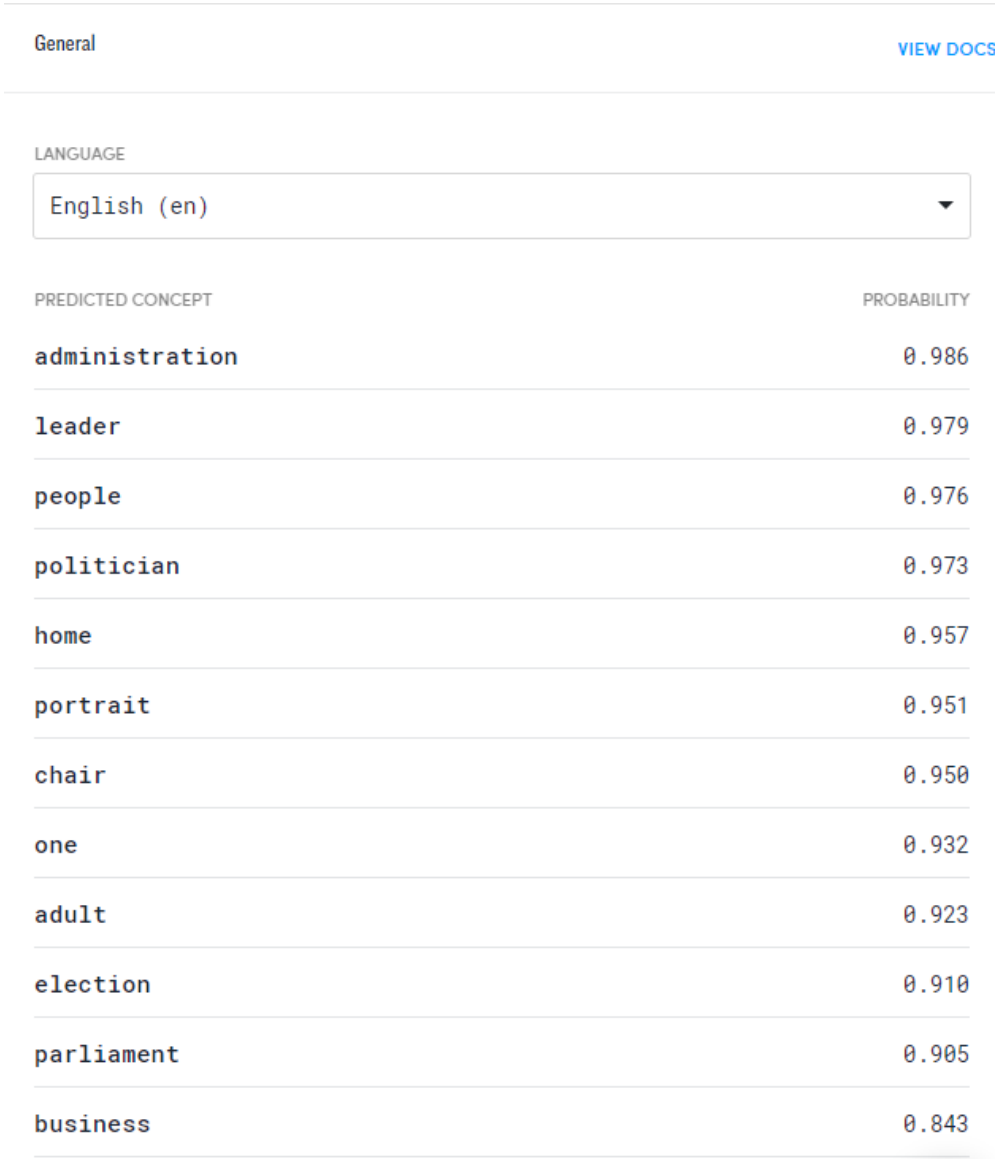
Jak správně určila analýza snímku, pan prezident při proslovu natáčí hlavu mírně doprava, viz Obrázek 37 výše.

Analýza osobnostních charakteristik správně určila pohlaví pana prezidenta.

Odhadovaný věk pana prezidenta se oproti předchozímu snímku mírně snížil, takže neodpovídá věku pana prezidenta ještě více než v analýze na Obrázku 32.

Jak již bylo zmíněno výše, v době natočení analyzovaného audiovizuálního záznamu bylo panu prezidentovi 73 let. Tuto osobnostní charakteristiku analýza neurčila vůbec správně, pokud je brána v potaz vyšší věková hranice, analýza se mýlila o celých 23 let.

Obrázek 38: Snímek videa 2 (Obrázek 34), analýza osobnostních charakteristik
Zdroj: Aplikace FaceReader - http://www.vicarvision.nl/products/online-facereader-demo/



Obrázek 39: Snímek videa 2 (Obrázek 34), analýza scény snímku
Zdroj: vlastní zpracovaní s využitím nástroje Clarifai dostupnou z https://clarifai.com/demo

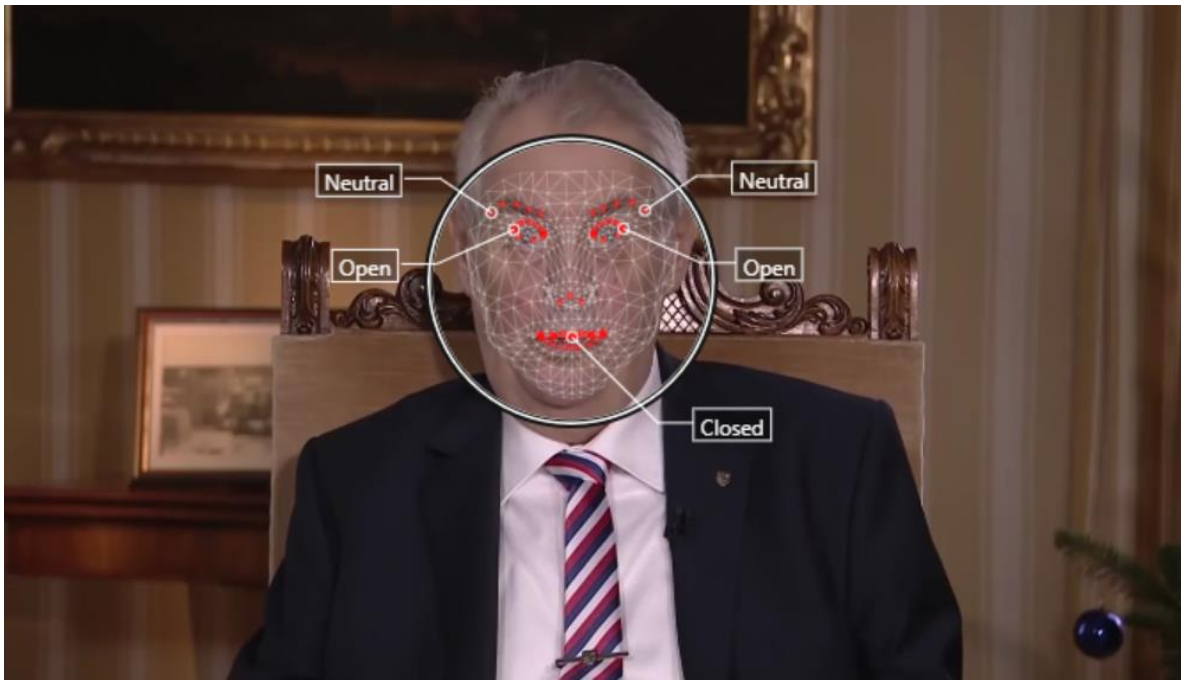
Na Obrázku 39 je analýza scény námi zvoleného snímku, která sestavila seznam předpokládaných objektů, které se na daném snímku nachází. Setřídila je na základě pravděpodobnosti výskytu daných objektů. S nejvyšší pravděpodobností celých 98,6 % označila, že se na snímku nachází vláda, dále určila s pravděpodobností 97,9 %, že se na snímku nachází vůdčí osobnost. S podobně vysokými pravděpodobnostmi detekovala na snímku i člověka, politika, portrét, židli, dospělou osobu a dále například i volby, parlament a obchod.

Analýza scény službou Clarifai tedy odpovídá skutečnosti.

Snímek 3

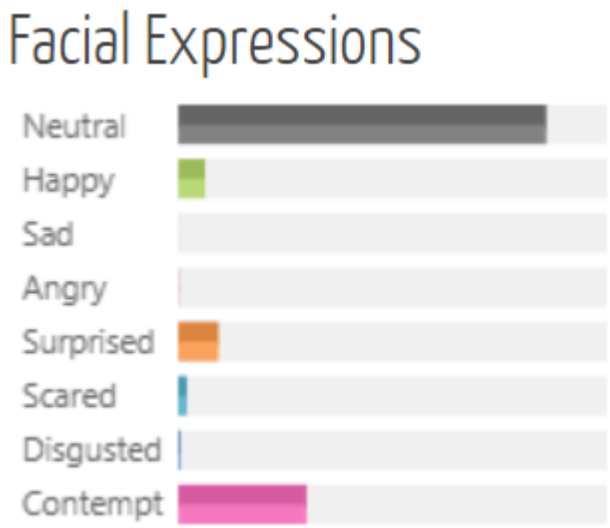
Obrázek 40: Snímek videa 3

Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player



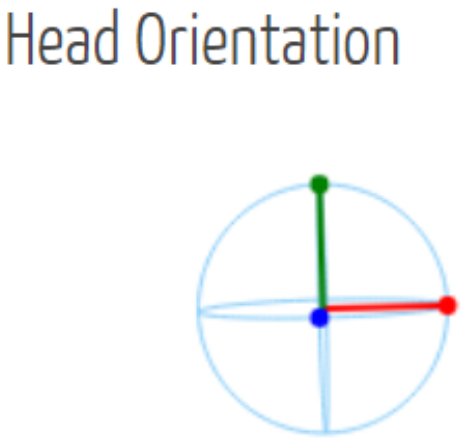
Obrázek 41: Snímek videa 3, podroben analýze výrazu tváře

Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player a zpracován online demo verzi aplikace FaceReader - http://www.vicarvision.nl/products/online-facereader-demo/



Obrázek 42: Snímek videa 3 (Obrázek 40), analýza výrazu tváře pana prezidenta

Zdroj: Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/



Obrázek 43: Snímek videa 3 (Obrázek 40), analýza orientace hlavy řečníka

Zdroj: Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/

Obrázek 44: Snímek videa 3 (Obrázek 40), analýza osobnostních charakteristik

Zdroj: Aplikace FaceReader- http://www.vicarvision.nl/products/online-facereader-demo/

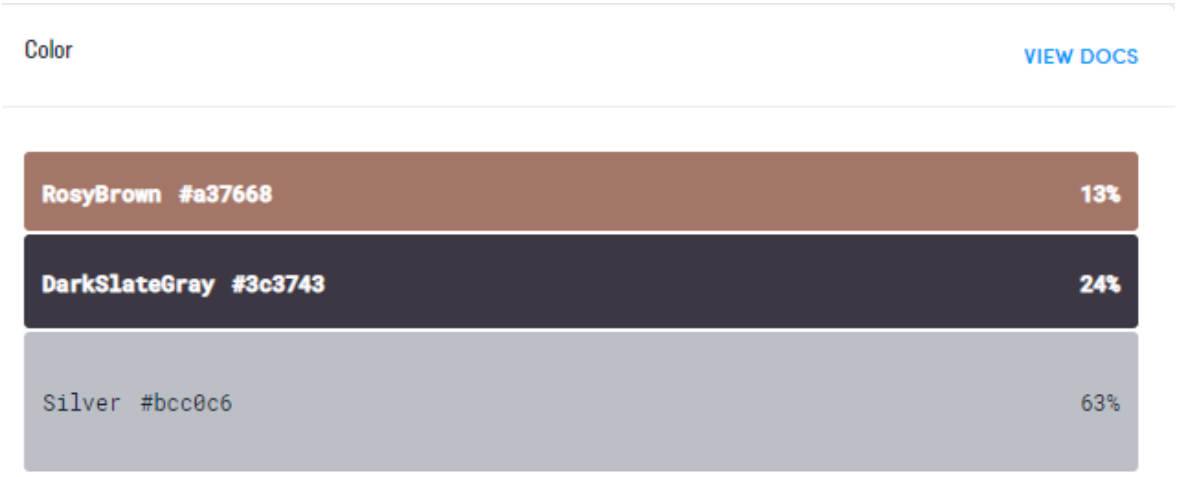
K tomuto snímku 3 (Obrázek 40) nebyla záměrně vytvořena analýza scény, jelikož je totožná se snímkem 2 (Obrázek 34).

6.4 Detekce barev

Prostřednictvím nástroje společnosti Clarifai je možné provést detekci dominantních barev snímků videa.

Snímek 1

Obrázek 45: Snímek 1 záznamu, který bude podroben analýze – detekce barev
Zdroj: Česká televize, snímek byl získán prostřednictvím nástroje VLC Player



Obrázek 46: Výsledek detekce barev ze snímku 1 (Obrázek 45) vybraného záznamu
Zdroj: vlastní zpracování s využitím služby Clarifai dostupné z https://clarifai.com/demo

Analyzovaný snímek obsahuje dle Obrázku 46 z 63 % stříbrnou či šedivou barvu, z 24 % šedou barvu v odstínu DarkSlateGrey a z 13 % hnědou barvu v odstínu RosyBrown.

6.5 Detekce demografických údajů

Nástroj Clarifai vyhodnotil snímek záznamů vánočního proslovu pana prezidenta – snímek 1 – Obrázek 28. Dle výsledku nástroje na Obrázku 47 je pan prezident s pravděpodobností 89,2 % muž, s pravděpodobností 69,1 % je mu 60 let a s pravděpodobností 98,1 % je běloch.

Dle údajů, které jsou o panu prezidentovi dostupné, lze usoudit, že nástroj Clarifai analyzoval daný snímek správně až na odhadovaný věk (panu prezidentovi je na snímku 73 let).

Demographics

1 FACE DETECTED

GENDER APPEARANCE	PROBABILITY
masculine	0.892
feminine	0.108

AGE APPEARANCE	PROBABILITY
60	0.691

MULTICULTURAL APPEARANCE	PROBABILITY
white	0.981
american indian or alaska native	0.012
middle eastern or north african	0.010
hispanic, latino, or spanish origin	0.003

Obrázek 47: Detekce demografických údajů
Zdroj: vlastní zpracovaní s využitím nástroje Clarifai, dostupného z https://clarifai.com/models/demographics-image-recognition-model-c0c0ac362b03416da06ab3fa36fb58e3

6.6 *Zhodnocení*

V této kapitole jsou zhodnoceny výsledky jednotlivých analýz provedených v předchozích podkapitolách kapitoly 6.

Záznam vánočního poselství pana prezidenta Miloše Zemana byl podroben těmto druhům analýz:

- Přepisu (audio)
- Sentimentu textu (audio)
- Sentimentu výrazu tváře (video)
- Osobnostních charakteristik (video)
- Dominantních barev (video)
- Demografických údajů (video)

Tyto analýzy lze rozdělit na analýzu audio záznamu a video záznamu.

Při využití současných technologií pro přepis audio záznamu na text je nutné stále počítat s určitou mírou chybovosti přepisu. Z porovnání, které bylo provedeno, vyplynulo, že nástroj NTeX přepis zpracoval lépe (s nižší chybovostí) nežli nástroj Přepisovatel.cz.

Tento přepis byl dále podroben analýze sentimentu. Z analýzy sentimentu přepisu vyplynulo, že pan prezident hodnotí rok 2017 velmi kladně (ohodnoceno sentimentem +0,2), negativně hodnotí pouze situaci ohledně migračních kvót a také situaci ohledně spekulací nad manipulacemi s výsledky voleb. Velmi negativně hodnotí pouze situaci s nezaměstnaností nepřizpůsobivých (výsledný sentiment - 0,29).

Analýza snímků pana prezidenta při jeho vánočním proslovu indikovala, že pan prezident se tvářil převážně neutrálně, ale občas zde byly i prvky zlosti, znechucení a opovržení.

Přestože výsledky analýzy sentimentu přepisu audia záznamu proslovu pana prezidenta jsou především pozitivní, analýza výrazu tváře pana prezidenta ve snímcích z videa záznamu při proslovu značí především neutrální emoce, či náznaky zlosti, znechucení a opovržení.

Je tedy velmi těžké hodnotit, co způsobilo rozpor ve výsledcích analýzy audio a video záznamu. Lze pouze uvažovat nad třemi variantami:

- Použité nástroje mají vysokou míru chybovosti a nelze tedy jejich výsledky považovat za odpovídající skutečnosti.

- Odlišnosti toho, co pan prezident říká, od toho, co si myslí a co cítí.

- Mohou zde být i kulturní odlišnosti. Nástroj pro detekci může být trénován pro odlišnou kulturu, než je kultura, na které je potom používán. Že kulturní odlišnosti v emocích existují, dokládá článek – Cultural differences in emotion: differences in emotional arousal level between the East and the West. (Lim, 2016)

Na základě získaných informací nelze určit, která varianta je správná.

Závěr

Cíl teoretické části byl splněn, byl popsán současný stav problematiky zpracování Big dat se speciálním zaměřením na data nestrukturovaná, včetně zmapování současných trendů pro zpracování těchto dat.

Hlavní přínos této práce spočívá v uceleném popsání problematiky zpracování Big dat se speciálním zaměřením na data nestrukturovaná. Dalším přínosem práce jsou názorné ukázky zpracování těchto dat vybranými nástroji.

V práci jsou také stanoveny přínosy zpracování nestrukturovaných dat. Dále jsou popsány nástroje, které byly využity pro analýzu vybraného audiovizuálního záznamu v praktické části diplomové práce.

V praktické části práce byl popsán vybraný audiovizuální záznam a byla na něj aplikována dvě řešení pro přepis audio záznamu na text. Tyto nástroje byly porovnány na základě úspěšnosti jejich přepisu. Dle provedeného porovnání byl úspěšnější nástroj NTeX, který měl přepis s nižší chybovostí. Tento přepis byl poté podroben analýze sentimentu textu. Výsledky analýzy sentimentu jsou interpretovány a zhodnoceny. Na snímky vybraného audiovizuálního záznamu bylo aplikováno několik druhů analýz – Analýza sentimentu videa, detekce barev, detekce osobnostních charakteristik a demografických údajů. Výsledky jednotlivých analýz jsou zhodnoceny. V závěru praktické části práce jsou výsledky dílčích analýz seskupeny a je zde zhodnocen vybraný audiovizuální záznam jako celek, včetně interpretace výsledků těchto analýz.

Big data skrývají velký potenciál, který současná společnost teprve začíná objevovat. Analýza těchto dat je velmi prospěšná ve všech odvětvích a je třeba se naučit ji využívat.

Dle autorky této práce má problematika zpracování Big dat – speciálně dat nestrukturovaných – potenciál dál se vyvíjet – především v oblasti analýzy sentimentu a analýzy výrazu lidské tváře. Lze očekávat, že se tyto nástroje budou zdokonalovat a jejich výsledky budou ještě přesnější než doposud.

Seznam použité literatury

Citace

ACL. 2018. Fraud Detection Using Data Analytics in the Banking Industry. Vancouver: *ACL* Services [online]. [cit. 2018-02-15]. Dostupné z: https://www.acl.com/pdfs/DP_Fraud_detection_BANKING.pdf.

AGGARWAL, Charu C. 2018. *Neural Networks and Deep Learning*. Chem: Springer International Publishing. ISBN-13: 978-3319944623.

AGRAWAL, Divyakan, Philip BERNSTEIN, Elisa BERTINO, Susan DAVIDSON a KOL. 2012. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. *Company Research Association* [online]. Washington: CRA, [cit. 2018-11-23]. Dostupné z: https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf.

ANON. 2018. Morfém. In: *Češtinaveslovníku.cz: Kompletní slovník češtinářských pojmů* [online]. cestinaveslovniku.cz, [cit. 2018-04-04]. Dostupné z: https://www.cestinaveslovniku.cz/morfem/.

ANON. 2018. Neuronové sítě: vývoj a testování. *Shrimphood: Vývoj her a speciálních aplikací, programování, technologie, recenze* [online]. Praha: CMS PRO-WEB [cit. 2018-11-22]. Dostupné z: http://www.shrimphood.net/neuronove-site-vyvoj-a-testovani.html.

BARI, Anasse, Mohamed CHAOUCHI a Tommy JUNG. 2014. *Predictive analytics for dummies*. Hoboken, NJ: John Wiley, [2014]. For dummies. ISBN 978-1-118-72896--3.

BARTOŠ, Tomáš. 2018. Přechod z analýzy na predikci ve výrobních podnicích. *SystemOnLine: Ekonomické a informační systémy v praxi* [online]. CCB spol. s r.o. 2018, 6. 6. 2018, [cit. 2018-11-19]. ISSN 1802-615X. Dostupné z: http://www3.systemonline.cz/rizeni-vyroby/prechod-z-analyzy-na-predikci-ve-vyrobnich-podnicich.htm.

BENJAMINS, V. Richard. 2014. Big Data. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) - WIMS '14*. New York, New York, USA: ACM Press, 2014, 1-2. ISBN 9781450325387. DOI: 10.1145/2611040.2611042.

BERNAT, Petr. 2010. Akustika, vznik a šíření zvuku, frekvenční analýza a syntéza, sluchový vjem zvukového signálu. *Anatomie varhan: Osobní stránky Ing. Petra Bernata* [online]. Ing. Petr Bernat, [cit. 2018-11-23]. Dostupné z: https://homen.vsb.cz/~ber30/texty/varhany/anatomie/pistaly_akustika.htm.

BIEHN, Niel. 2013. The Missing V's in Big Data: Viability and Value. *WIRED* [online]. [cit. 2018-02-22]. Dostupné z: https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/.

BREJL, Milan a Vladimír ŠEBESTA. 1999. Analýza zvuku hudebních nástrojů. *Elektrorevue* [online]. Brno: International Science and Engineering Society, o.s., 1999, 13.12.1999, [cit. 2018-11-23]. ISSN 1213-1539. Dostupné z: http://www.elektrorevue.cz/clanky/99011/index.html#algoritmus.

BROWNLEE, Jason. 2016. Supervised and Unsupervised Machine Learning Algorithms. In: *Machine Learning Mastery* [online]. [cit. 2018–11–09]. Dostupné z: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

BRYANT, Randal E., Randy H. KATZ a Edward D. LAZOWSKA. 2008. *Big-Data Computing:* Creating revolutionary breakthroughs in commerce, science, and society. *Computing Research Organization: CRA* [online]. [cit. 2018-02-15]. Dostupné z: https://cra.org/ccc/wp-content/uploadbryans/sites/2/2015/05/Big_Data.pdf.

CETIN, Özgür a Elizabeth SHRIBERG. 2006. Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site. *ICSI - International Computer Science Institute* [online]. International Computer Science Institute, s. 212-224 [cit. 2018-11-24]. Dostupné z: http://www.icsi.berkeley.edu/pubs/speech/CetinShriberg0506.pdf.

COOTES, T. F. a C. J. TAYLOR. 2004. Statistical Models of Appearance for Computer Vision. *FACE RECOGNITION* [online]. Manchester: VCL, Macrh 8 [cit. 2018-11-24]. Dostupné z: http://www.face-rec.org/algorithms/aam/app_models.pdf.

COX, Michael a David ELLSWORTH. 1997. Application-Controlled Demand Paging for Out-of-Core Visualization. *NAS* [online]. Červenec 1997 [cit. 2018-02-15]. Dostupné z: https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf.

CRHA, Aleš. 2012. *Služby zpracování dat ze Smart Grids postavené na technologiích ETL a CEP*. Brno. Diplomová práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Mgr. Filip Procházka, Ph.D.

ČERMÁK, Miroslav. 2014. Co je a není bezpečnostní incident: Kybernetický bezpečnostní incident dle zákona o kybernetické bezpečnosti. *Clever and Smart* [online]. 14. května 2014 [cit. 2018-12-11]. Dostupné z: http://www.cleverandsmart.cz/co-je-a-neni-bezpecnostni-incident/.

ČERNÝ, Michal. 2016. Vizualizace dat: Jak odhalit utajené souvislosti. *VTM.cz: věda, technika, technologie, budoucnost* [online]. Praha: Serafico investment, [cit. 2018-05-02]. Dostupné z: http://vtm.e15.cz/vizualizace-dat-jak-odhalit-utajene-souvislosti.

DAVIES, Jason. 2017. How the Word Cloud Generator Works. *Jason Davies - Freelance Data Visualisation* [online]. Jason Davies, [cit. 2018-02-19]. Dostupné z: https://www.jasondavies.com/wordcloud/about/.

DEL VECCHIO, Pasquale, Alberto DI MININ, Antonio Messeni PETRUZZELLI, Umberto PANNIELLO a Salvatore PIRRI. 2018. Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. *Creativity and Innovation Management*. 2018, **27**(1): 6-22. ISSN 09631690. DOI: 10.1111/caim.12224.

EKMAN, Paul. 1970. Universal facial expressions of emotion. *California mental health research digest*. Sacramento: Dept. of Mental Hygiene, Bureau of Research, 8(4), 151-158. ISSN 0008-1280.

EKOSOFTWARE. 2018. Frekvenční analýza. *Ekosoftware s. r. o.: Software nejen pro zvuk a vibrace* [online]. Liberec: DPOINT.CZ, [cit. 2018-11-23]. Dostupné z: https://www.ekosoftware.cz/frekvencni-analyza.

ETZION, Opher. a Peter. 2011. NIBLETT. *Event processing in action*. Greenwich: Manning. ISBN 978-193-5182-214.

EXPERIAN. 2017. A data powered future. *Experian: Credit Check, Free Credit Score & Comparisons* [online], [cit. 2018-02-23]. Dostupné z: http://www.experian.co.uk/assets/resources/white-papers/data-powered-future-2016.pdf.

FIRICAN, George. 2017. The 10 Vs of Big Data. *TDWI: Transforming Data With Intelligence* [online]. TDWI, 8. února 2017 [cit. 2018-02-15]. Dostupné z: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.

FORSBERG, Markus. 2003. *Why is Speech Recognition Difficult?*. Götebord. White Paper. Chalmers University of Technology, Department of Computing Science.

FRIEDMAN, Uri. 2012. Big Data: A Short History: How we arrived at a term to describe the potential and peril of today's data deluge. *Foreign Policy - the Global Magazine of News and Ideas* [online]. 8. listopadu 2012 [cit. 2018-02-13]. Dostupné z: http://foreignpolicy.com/2012/10/08/big-data-a-short-history/.

GABBATT, Adam. 2011. IBM computer Watson wins Jeopardy clash: Supercomputer outwits US quiz show champions in epic head-to-hard drive battle. *The Guardian: News, sport and opinion from the Guardian´s global edition* [online]. Guardian News and Media Limited, 17. února 2011 [cit. 2018-04-03]. ISSN 0261-3077. Dostupné z: https://www.theguardian.com/technology/2011/feb/17/ibm-computer-watson-wins-jeopardy.

GARETTA, Raúl. 2015. A Gentle Guide to Machine Learning. *MonkeyLearn: Blog*[online], [cit. 2018–11–09]. Dostupné z: https://monkeylearn.com/blog/gentle-guide-to-machine-learning/.

GARTNER. 2005. Inc. *Gartner completes acquisition of META group*. [online] 1. dubna 2005 [cit. 2018-02-19]. Dostupné z: https://www.gartner.com/newsroom/id/492119.

GENEEA. 2018. *Geneea: Inteligent Interpretation* [online]. Praha: Geneea, [cit. 2018-11-13]. Dostupné z: https://www.geneea.com/.

GODSAY, Manasee. 2015. The Process of Sentiment Analysis: A Study. *IJCA - International Journal of Computer Applications*. International Journal of Computer Applications, September 2015, **26**(7): 26-30. ISSN 0975 - 8887. DOI: 10.5120/ijca2015906091.

GRUBER, Ivan. 2015. Detekce významných bodů na lidské tváři pomocí neuronové sítě. *Západočeská univerzita v Plzni: Digitální knihovna* [online]. Plzeň: Západočeská univerzita v Plzni, [cit. 2018-11-23]. Dostupné z: https://dspace5.zcu.cz/bitstream/11025/21295/1/Gruber.pdf.

HELMS, Josh. 2015. Big Data: It's About Complexity, Not Size. *IBM Center for The Business of Government: Connecting research to practice*. [online]. 22. ledna 2015 [cit. 2018-02-13]. Dostupné z: http://www.businessofgovernment.org/blog/business-government/big-data-it%E2%80%99s-about-complexity-not-size.

HOFMANN, Markus a Andrew CHISHOLM. 2015. *Text Mining and Vizualization: Case Studies Using Open-Source Tools*. Boca Raton: Chapman and Hall/CRC. ISBN 978-1482237573.

HOLČÍK, Jiří a Martin KOMENDA. 2015. Koncept umělé neuronové sítě. *Matematická biologie: e-learningová učebnice* [online]. Brno: Masarykova univerzity, [cit. 2018-11-22]. ISBN 978-80-210-8095-9. Dostupné z: http://portal.matematickabiologie.cz.

CHUVAKIN, Anton., Kevin J. SCHMIDT, Chris PHILLIPS a Patricia MOULDER. 2013. *Logging and log management: the authoritative guide to understanding the concepts surrounding logging and log management*. Amsterdam: Elsevier/Syngress. ISBN 1597496359.

IBM. 2015. TAKMI: *Bringing Order to Unstructured Data*. [online]. [cit. 2016-05-02]. Dostupné z: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/takmi/.

IBM. 2014. *The Four V's of Big Data* [online]. USA: IBM press, [cit. 2018-02-13]. Dostupné z: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

ISHIKAWA, Hiroshi. 2015. *Social Big Data Mining*. Boca Raton: CRC Press - Taylor & Francis Group. ISBN 978-1-4987-1094-7.

JAIN, Kunal a Payel Roy CHOUDHURY. 2015. Machine Learning basics for a newbie. *Analytics Vydhia* [online]. [cit. 2018–10–20]. Dostupné z: https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/.

JEYANTHI, Na. 2016. *Internet of Things (IoT) as Interconnection of Threats (IoT).* In: HU, Fei, ed. *Security and Privacy in Internet of Things (IoTs): Models, Algorithms, and Implementations*. Boca Raton: CRC Press - Taylor & Francis Group. ISBN 13: 978-1-4987-2319-0. DOI: 10.1201/b19516-3.

KASÍK, Pavel. 2017. NÁVOD: Sociální sítě vám můžou udělat ze života peklo. Jak se bránit?. *Technet.cz: Technika kolem nás* [online]. Praha: MAFRA, 16. května 2017 [cit. 2018-11-25]. Dostupné z: https://technet.idnes.cz/bezpecne-na-internetu-socialni-site-sdileni-soukromi-kybersikana-phishing-podvody-facebook-twitter-i-i7t-/sw_internet.aspx?c=A170511_120950_sw_internet_pka.

KOONG, Chorng-Shiuh, Tzu-I YANG a Chien-Chao TSENG. 2014. A User Authentication Scheme Using Physiological and Behavioral Biometrics for Multitouch Devices. *The Scientific World Journal*, s. 1-12. ISSN 2356-6140. DOI: 10.1155/2014/781234.

KREPS, Jay. 2015. *I Heart Logs: Event Data, Stream Processing, and Data Integration*. 2. Sebastopol: O'Reilly Media. ISBN 978-1491909386.

KUDYBA, Stephan. 2014. *Big data, mining, and analytics: components of strategic decision making*. Boca Raton. ISBN 14-665-6870-4.

KUMAR, Ashish. 2016. *Learning Predictive Analytics with Python: Gain practical insights into predictive modelling by implementing Predictive Analytics algorithms on public datasets with Python*. Birmingham: Packt Publishing. ISBN 978-1-78398-326-1.

LANEY, Douglas. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Gartner Blog Network*. [online]. 6. února 2001 [cit. 2018-02-16]. Dostupné z: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

LE, Thai Hoang. 2011. Applying Artificial Neural Networks for Face Recognition. *Advances in Artificial Neural Systems*. Ho Chi Minh City: Department of Computer Science, Ho Chi Minh University of Science, 2011, **2011**. s. 1-16. ISSN 1687-7594. DOI: 10.1155/2011/673016.

LIČEV, Lačezar a Štěpán SOJKA. 2009. Rozpoznávání zájmových bodů a objektů na snímcích. *GISAK: Institut geoinformatiky* [online]. Ostrava: Technická univerzita Ostrava, 2018, 25.-28.1.2009 [cit. 2018-11-23]. Dostupné z: http://gisak.vsb.cz/GIS_Ostrava/GIS_Ova_2009/sbornik/Lists/Papers/002.pdf.

LIM, Nangyeon. 2016, Cultural differences in emotion: differences in emotional arousal level between the East and the West. *Integrative Medicine Research*. **5**(2), 105-109. ISSN 2213-4220. DOI: 10.1016/j.imr.2016.03.004.

M-BRAIN. 2018. Big Data Technology with 8 V´s. *M-Brain Market Research, Competitive Intelligence & Media Insights* [online]. New York: M-Brain, [cit. 2018-02-15]. Dostupné z: https://www.m-brain.com/home/technology/big-data-with-8-vs/.

MARR, Bernard. 2014. Big Data: The 5 Vs Everyone Must Know. *LinkedIn* [online]. 6. března 2014 [cit. 2018-02-15]. Dostupné z: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/.

MARR, Bernard. 2015. The Difference Between Big Data and a Lot of Data. *DataInformed: Big Data and Analytics in The Enterprise* [online]. Dedham: Wellesley Information Services, [cit. 2018-02-13]. Dostupné z: http://data-informed.com/the-difference-between-big-data-and-a-lot-of-data/.

MARR, Bernard. 2016. Big Data: The 6th 'V' Everyone Should Know About. *Forbes* [online]. 20. prosince 2016 [cit. 2018-02-13]. Dostupné z: https://www.forbes.com/sites/bernardmarr/2016/12/20/big-data-the-6th-v-everyone-should-know-about/#1bc6eec52170.

MARR, D. a E. HILDRETH. 1980. Theory of edge detection. Proceedings of the Royal Society of London. Series B. Biological Sciences. 207(1167): 187-217. ISSN 1471-2954. DOI: 10.1098/rspb.1980.0020.

McCUE, Colleen. 2007. *Data mining and predictive analysis: intelligence gathering and crime analysis*. Boston: Butterworth-Heinemann. ISBN 07-506-7796-1.

McNULTY, Eileen. 2014. Understanding Big data: The seven V'S. *Dataconomy* [online]. [cit. 2018-02-15]. Dostupné z: http://dataconomy.com/2014/05/seven-vs-big-data/.

MENDELU, Neuronové sítě. 2018. *Univerzitní informační systém MENDELU* [online]. Brno: Mendelova univerzita v Brně, [cit. 2018-11-22]. Dostupné z: https://is.mendelu.cz/eknihovna/opory/zobraz_cast.pl?cast=21471.

MI, Jun, Kun WANG, Bo LIU, Fei DING, Yanfei SUN a Huawei HUANG. 2017. A Multiobjective Evolution Algorithm Based Rule Certainty Updating Strategy in Big Data Environment. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. Singapore, s. 1-6. ISBN 978-1-5090-5019-2. DOI: 10.1109/GLOCOM.2017.8255002.

MONTESINO, Raydel, Stefan FENZ a Walter BALUJA. 2012. SIEM – based framework for security controls automation. Information Management & Computer Security. **20**(4): 248-263. ISSN 0968-5227. DOI: 10.1108/09685221211267639.

NGUYEN, Filip a Tomáš PITNER. 2012. Information system monitoring and notifications using complex event processing. *Proceedings of the Fifth Balkan Conference in Informatics on - BCI '12*. New York, s. 211-216. ISBN 9781450312400. DOI: 10.1145/2371316.2371358.

NIEWELER, Amanda. 2015. 10 Fraud Detection Methods That Will Make You a Hero!. *WhistleBlower Security* [online]. 23. února 2015 [cit. 2018-02-15]. Dostupné z: https://www.whistleblowersecurity.com/10-fraud-detection-methods-that-will-make-you-a-hero/.

NTeX. 2018. Aplikace pro online přepis zvuku na text určená pro prostředí NEWTON SpeechGrid. *NTeX: NEWTON technologies* [online]. Praha: NEWTON Technologies, 2018 [cit. 2018-11-18]. Dostupné z: https://www.newtontech.net/cs/speechgrid/ntex/.

OWENS, Robyn. 1997. Mathematical Morphology. *The University of Edinburgh, School of informatics* [online]. Edinburgh: The University of Edinburgh, 29.10.1997 [cit. 2018-11-24]. Dostupné z: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/OWENS/LECT3/node3.html.

OXFORD UNIVERSITY PRESS. 2018. Big data. *Oxford English Dictionary* [online]. Oxford University Press, 2018 [cit. 2018-02-15]. Dostupné z: http://www.oed.com/view/Entry/18833#eid301162177.

PAHUJA, Vardaan, Anirban LAHA, Shachar MIRKIN, Vikas RAYKAR, Lili KOTLERMAN a Guy LEV. 2017. Joint Learning of Correlated Sequence Labeling Tasks Using Bidirectional Recurrent Neural Networks*:* 1 IBM Research. *IBM Research* [online]. New York: IBM, 2017, 18 July 2017. DOI: 1703.04650v3.

PANDIT, Ketan. 2016. Big data and analytics trends – 2016 [Infographic]. *Big Data Made Simple: One source. Many perspectives.* [online]. Singapur: Crayon Data, February 18, 2016 [cit. 2018-11-30]. Dostupné z: https://bigdata-madesimple.com/big-data-analytics-trends-2016/.

PANG, Bo a Lilian LEE. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, **2**(1-2): 1-71. ISSN 1554-0677. DOI: 10.1561/1500000011.

PANG, Bo, Lilian LEE a Shivakumar VAITHYANATHAN. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia: Association for Computational Linguistics, s. 79-86. DOI: 10.3115/1118693.1118704.

PRESS, Gil. 2016. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. *Forbes* [online]. 23. března 2016. [cit. 2018-02-15]. Dostupné z: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#7cdcad3d6f63.

Přepisovatel.cz. 2018. Textový přepis pro vaše audio/video data. *Textový přepis* [online]. ReplayWell, [cit. 2018-11-18]. Dostupné z: https://www.prepisovatel.cz/dashboard/?_fid=cyqs.

REESE, Richard M. 2015. *Natural Language Processing with Java: Explore various approaches to organize and extract useful text from unstructured data using Java*. Birmingham: Packt Publishing, 2015. ISBN 978-1-78439-179-9.

REEVES, Charles W. 1998. *The Vibration Monitoring Handbook*. Moreton In Marsh: Coxmoor. ISBN 978-1-901892-00-0.

REINSEL, David, John GANZ a John RYDNING. 2017. Data Age 2025: The Evolution of Data to Life-Critical, Don't Focus on Big Data; Focus on the Data That's Big. *Seagate: Storing the world's digital content* [online]. IDC White Paper, duben 2017 [cit. 2018-02-22]. Dostupné z: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf.

RIJMENAM, Mark van. 2015. 7 Important Big Data Trends for 2016. *Datafloq: Driving Innovation through Data* [online]. Hague: Datafloq, December 1. 2015 [cit. 2018-11-30]. Dostupné z: https://datafloq.com/read/7-big-data-trends-for-2016/1699.

SAS Institute Inc. 2018. *Big Data: What it is and why it matters* [online]. North Carolina: SAS Institute, 2018 [cit. 2018-02-13]. Dostupné z: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

SHAFER, Tom. 2017. The 42 V's of Big Data and Data Science. *Machine Learning, Data Science, Big Data, Analytics, AI* [online]. duben 2017 [cit. 2018-02-22]. Dostupné z: https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html.

SHAFFER, Marc. 2017. The Top 7 Big Data Trends for 2017. *LinkedIn* [online]. květen 2017 [cit. 2018-02-22]. Dostupné z: https://www.linkedin.com/pulse/top-7-big-data-trends-2017-m-shaffer/.

SHAH, Aatash. 2015. 7 Big Data Trends That Will Dominate 2016. *Data Science & Training - Machine Learning, Data Science, Business Analytics, Big Data: Edvancer Eduventures* [online]. Mumbai: Edvancer, 23.12.2015 [cit. 2018-11-30]. Dostupné z: https://www.edvancer.in/7-big-data-trends-2016/.

SHARON, Guy a Opher ETZION. 2007. Event Processing Network - A Conceptual Model. *ResearchGate: Share and discover research* [online]. ResearchGate, 2014, [cit. 2018-03-15]. Dostupné z: https://www.researchgate.net/publication/249690470_Event_Processing_Network_-_A_Conceptual_Model.

SHERPA SOFTWARE. 2018. Structured and Unstructured Data: What is It. *Data Governance Solutions*. [online]. Bridgeville: Sherpa Software, [cit. 2018-02-15]. Dostupné z: https://sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/.

SMOLOVÁ, Bára. 2016. *Vytěžování a vizualizace nestrukturovaných a semistrukturovaných dat*. Liberec. Bakalářská práce. Technická univerzita v Liberci. Vedoucí práce Ing. Vladimíra Zádová.

SOME, Kamalika. 2018. TOP 7 BIG DATA ANALYTICS TRENDS FOR 2019. *Big Data, Analytics and Insight* [online]. Nagpur: Stravium Intelligence LLP, 2018, October 29, [cit. 2018-11-30]. Dostupné z: https://www.analyticsinsight.net/top-7-big-data-analytics-trends-for-2019/.

STIEGLITZ, Stefan, Milad MIRBABAIE, Björn ROSS a Christoph NEUBERGER. 2017. Social media analytics: Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*. **39**: 156-168. ISSN 0268-4012. DOI: 10.1016/j.ijinfomgt.2017.12.002.

SUŠICKÝ, Marek a Petr MIKEŠKA. 2015. Big Data aneb Když běžné databázi dochází dech. *IT Systems*. **2015**(11), 18-19. [cit. 2018-02-22]. ISSN 1802-002X. Dostupné z: https://www.systemonline.cz/clanky/big-data-aneb-kdyz-bezne-databazi-dochazi-dech.htm.

ŠMAHAJ, Jan. 2014. *Kyberšikana jako společenský problém*. Olomouc: Univerzita Palackého v Olomouci, Filozofická fakulta, Katedra psychologie. ISBN 978-80-244-4227-3.

TABOADA, Maite, Julian BROOKE, Milan TOFILOSKI, Kimberly VOLL a Manfred STEDE. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*. **37**(2): 267-307. ISSN 0891-2017. DOI: 10.1162/COLI_a_00049.

VESELOVSKÁ, Kateřina. 2017. Analýza nestrukturovaných dat je klíčovou kompetencí. *IT systems* [online]. Brno: CCB spol., 16.10.2017, **2017**(11) [cit. 2018-04-02]. ISSN 1802-615X. Dostupné z: https://www.systemonline.cz/business-intelligence/analyza-nestrukturovanych-dat-je-klicovou-kompetenci.htm.

VOIGHT, Christian, Barbara KIESLINGER a Teresa SCHAEFER. 2017. User Experiences Around Sentiment Analyses, Facilitating Workplace. *Human-Computer Interaction International Conference*. Vancouver, s. 1-14. ISSN 0302-9743. DOI: 10.1007/978-3-319-58562-8_24.

WATTENBERG, Martin a Fernanda VIÉGAS. 2007. The Word Tree, an Interactive Visual Concordance. Word Tree. *IBM Research*, s. 1-8. [cit. 2018-02-15]. DOI: 10.1109/TVCG.2008.172.

Bibliografie

AGGARWAL, Charu C. 2015. *Data Mining: The Textbook 1*. Cham: Springer International Publishing. 2015. ISBN 978-3-319-14141-1.

ANDERSON, Alan a David SEMMELROTH. 2015. *Statistics for Big Data For Dummies*. New Jersey: John Wiley. ISBN 978-1118940013.

BIRD, Steven, Ewan KLEIN a Edward LOPER. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol: O'Reilly Media. ISBN 978-0596516499.

COOPER, William W., Lawrence M. SEIFORD a Kaoru TONE. 2007. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. 2. vydání. New York: Springer. ISBN 978-0387-45281-4.

CVRČEK, Václav. 2015. *Mluvnice současné češtiny. 1, Jak se píše a jak se mluví*. 2. vydání. Praha: Univerzita Karlova v Praze, nakladatelství Karolinum, 416 s. ISBN 978-80-246-2812-7.

ELLIS, Byron. 2014. *Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data*. Indianapolis: Jhn Wiley. ISBN 978-1-118-83791-7.

HÁŠA, Marek. 2016. Big Data 2016: Jak chytře využít data k velkému byznysu. *Markething* [online]. ISSN 1805-4991. Dostupné z: http://www.markething.cz/big-data-2016.

HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. 2015. *Big Data a NoSQL databáze*. Praha: Grada. ISBN 978-80-247-5466-6.

JURAFSKY, Dan a James H. MARTIN. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Upper Saddle River, N. J.: Pearson Prentice Hall. ISBN 978-013-1873-216.

KRATOCHVIL, Marcelle. 2013. *Managing Multimedia and Unstructured Data in the Oracle Database*. Birmingham: Packt Publishing. ISBN 978-1-84968-692-1.

LICHÝ, Alexander. 2016. Objem dat digitálního světa do roku 2020 vzroste desetkrát. *CIO Business World.cz: IT strategie pro manažery* [online]. 17. dubna. 2016 [cit. 2018-02-13]. Dostupné z: http://businessworld.cz/bi-a-data/objem-dat-digitalniho-sveta-do-roku-2020-vzroste-desetkrat-11600.

MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. 2014. *Big Data*. Brno: Computer Press. ISBN 978-80-251-4119-9.

POLÁCH, Eduard. 1999. Textury. *Studentský server: Jihočeská univerzita v Českých Budějovicích - Pedagogická fakulta* [online]. České Budějovice: Jihočeská univerzita v Českých Budějovicích, [cit. 2018-11-24]. Dostupné z: http://home.pf.jcu.cz/~edpo/povray/kap10.html.

PROQUEST. 2017. Databáze článků ProQuest [online]. Ann Arbor, Ml, USA: ProQuest. [cit. 2017-09-28]. Dostupné z: http://knihovna.tul.cz/.

ROUMI, Mahshid. 2009. Implementing Texture Feature Extraction Algorithms on FPGA. *ResearchGate: Share and discover research* [online]. Berlín: ResearchGate, [cit. 2018-11-24]. Dostupné z: https://www.researchgate.net/publication/46140875_Implementing_Texture_Feature_Extraction_Algorithms_on_FPGA.

SAKR, Sherif, Muhammad HAMMOUND, Majd F. SAKR, Anna LIU a další. 2014. ed. *Large Scale: Processing and Management*. New York: CRC Press. ISBN 978-1-4665-8151-7.

Seznam příloh

Příloha A: Přepis službou NTeX

doložení a milí spoluobčané přeji vám krásný dobrý den setkáváme se opět po roce abychom společně zhodnotili co se nám dařilo ten rok podařilo a případně co se za celý rok nepodařilo a jsem rád že vám mohu říci že těch radostných zpráv je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jak úspěchy to jsou Česká republika je 6. nejbezpečnější země na světě snad dokonce lepší než Švýcarsko máme nejnižší míru nezaměstnanosti v Evropské unii a také není mši a juniorů příjmové diferenciace respektive míru chudoby v Evropské unii náš ekonomický růst je 1 z nejvyšších a naopak naše zadluženost 1 z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomý národ který si váží svých úspěchů ale samozřejmě že každé světlo má své stíny a dovolte mi abych i o těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvalý musím být doprovázený růstem investic a tady máme své nedostatky podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například není hotová silnice mezi Brnem a Vídni ale takových dálnic obchvatů a dalších staveb je celá řada je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidí kterým říkáme nepřizpůsobiví mluvil jsem nedávno s novou ministryni práce a sociálních věci a shodli jsme se na tom že je třeba omezit sociální dávky těm kdo odmítají nabízenou práci růst hrubého domácího produktu je výborný ale musí být doprovázen růstem životní úrovně jsem rád že konečně dochází k růstu mezd pod tlakem nedostatku pracovních sil který nutí zaměstnavatele k tomu aby tyto mzdy zvyšovaly neboť jednak jim pracovnice utečou jinam a měl bych 1 námět který se vám možná bude zdát bizarní před 15 lety jsme měli 80 000 státních úředníků než státních zaměstnanců státních úředníků dnes jich máme 150 000 protože se nám množí podle Parkinsonových zákonů kdyby se alespoň část z nich podařilo uvolnit protože tito lidé si často jenom vymýšlejí zbytečnou práci soukromý sektor který trpí nedostatkem pracovních sil byl velmi rád přijal od břevna mělo by se státní rozpočet a uvolnilo by se napětí na trhu práce teď měl státní úřednici určitě nebudou mít rádi ale já tu s nimi myslím dobře tolik ekonomické situaci České republiky a závěrem protože jsem sám starobních důchodců bych chtěl popřát starobním důchodcům aby v příštím roce došlo k výrazné valorizaci těchto důchodů někdy mně píší co z toho bude nám nejde když se pořád zdražuje ale já jim připomínám že zdražování neboli inflace je u starobních důchodců zahrnuto do valorizace takže se nemají z čeho obávat jenom bych si přál aby se toto onemocnění otce odehrávala

stejnou částku pro všechny protože v opačném případě se neustále zvětšuje rozdíl mezi novými a starými důchody teď bych přešel k zahraniční politice jenom velice stručně jsme součásti Evropské unie jsme součástí Severoatlantické aliance a měli bychom svému v obou těchto organizacích chovat jako energetici energicky a sebevědomým partnerem který když se mu něco znali by takto dá najevo pokud jde o Evropskou unii víte že je neustále vytýkám že nedokáže chránit své vnější hranice a pokud jde o Severoatlantickou alianci myslím si že by měla být daleko aktivnější v boji proti islámskému terorismu nu a to jen nejbolavější problém který zde máme migrační kvóty musí být řešen v souladu s českými národními zájmy a naším národním zájmem pochopitelně je zachování suverenity České republiky nikdo nám nemůže diktovat koho umístíme na své území a věřím že koncept migračních kvót skončí v propadlišti dějin a že za rok už o něm nebude mluvit a není mi dovolte abych přišel do vnitropolitické scéně to zažila tento rok doslova zemětřesení po volbách se 2 nejstarší České politické strany tj. sociální demokracie a strana lidová dostali na pokraji propasti a téměř téměř se nedostane do parlamentu ať si vedoucí představitelé těchto stran sami zhodnotí proč k tomu tak dobu došlo nicméně jsem šel jsem své se s vedoucími představiteli všech devíti parlamentních subjektů a zeptal jsem se na jejich představu o povolebním uspořádání víte co mi řekli pane prezidente s námi nikdo nechce jednat a to mi říkali všichni a to je přece nesmysl znamená to tedy že pokud s vámi opravdu někdo nechce jednat tak výzvám za ním musíte jít nasedat jako panenka v koutě a jednání může zúročit i s vaší iniciativy a mělo by to být jednáni o které se bude týkat jak programových tak personál o nich kompromisu jak zpěv politice zvykem myslím si že návrh programového prohlášení Babišovy vlády je prostě takové diskuse docela dobrým výchozím materiálem když mluvím o Babišově vládě chtěl bych poznamenat že jsem dostával moudré rady které mi říkali nejmenuji Babišovu vládu dřív než Babiš získat důvěru v Poslanecké sněmovně ale to by reálně znamenalo že po týdny možná dokonce po měsíce tady povládne Sobotkova vláda v demisi a to si snad proboha nikdo nepřežil bych si vás strašit Belgií kde jednání o vládě trvalo 2 roky myslím si že během několika málo měsíců se podaří mít shodil vládu s důvěrou Poslanecké sněmovny a nebude-li úspěch hned v 1. kole vytvořím časový prostor pro důkladná jednání politických partnerů tak aby snad někdy v únoru se uskutečnil 2. pokus který již může být úspěšný nezapomeňte že v politice jsou 2 extrémy 1. pokus obklíčit vítěze voleb vytvořit proti němu jakousi pseudo koalici těch méně úspěšných a izolovat ho 2. extrém vítěz zašlape do země všechny poražené a vládne islám a myslím si že z hlediska demokratických procedur je dobré usilovat o to aby žádný z těchto extrémů se neuskutečnilo

a aby politické strany spolu jednali a CSc. dohodnou vytvořily patrně menšinovou tolerovanou vládu závěrem této části bych chtěl potěšit občany a nepotěšit politiky občas slyším hlasy že bych měl vypsat předčasné volby což mi ústava v některých situaci umožňuje chci naprosto jasně říci že to nikdy neudělám protože předčasné volby několik měsíců po řádných volbách by byly výsměchem občanů kteří šli k řádným volbám a volili tak jak volili občané rozdali karty to je a politici musí umět s těmito kartami hrát občan vy nemůžete vyměnit olej můžete vyměnit politiky proč ne někteří z nich možná skončí v opozici a jak napsal Rudolf Bechyně starý sociální demokrat za 1. republiky sucha je z piva opozice a já dodávám klatovský 2 nikdy není a ani posolené a takže nepřeji někomu aby skončilo v opozici ale na 2. straně demokracie a samozřejmě určitou opozici v parlamentu vyžaduje až si každý sám rozhodne zda takovou opozici chce být věřícím že nedojde k tomu že se budeme obviňovat z toho že výsledky minulých parlamentních voleb byly zmanipulovány zahraničními rozvědkami je to ubohé je to trapné a je to urážející volby byly svobodnému a jsem rád že i Bezpečnostní informační služba! A prohlášení kde konstatuje že k žádnému takovému ovlivňování nedošlo jsme svobodný názor na ně se svobodnými občany a dnes už si můžeme říci že jsme i úspěšná země před dvěma lety jsem vám řekl že skončila blbá nálada dnes vám říkám že se nemáme co stydět a že je mnoho věci na které můžeme býti hrdi takže z líhně na hlavu a buď dnes byli vědomi lidmi kteří spoléhají na svůj vlastní rozum a nedají slovy manipulovat ať už údajnými zahraničními rozvědkami nebo zejména českým tiskem Českou televizi a dalším mi vždy spoléhají více na vlastní rozum a já mám pro to své heslo které jsem vám už jednou říkal věřím že zdravý rozum zvítězí nad závistivou hloupostí a není mi dovolte abych připil na vítězství zdravého rozumu na úspěch České republiky a na úspěch nás všech Šťastný a Veselý nový rok

Příloha B: Přepis službou NTeX – chyby

doloženíVÁŽENÍ a milí spoluobčané přeji vám krásný dobrý den setkáváme se opět po roce abychom společně zhodnotili co se nám dařiloZA ten rok podařilo a případně co se za celýTEN rok nepodařilo a jsem rád že vám mohu říci že těch radostných zpráv je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jakJAKÉ úspěchy to jsou Česká republika je 6. nejbezpečnější země na světě snadJSME dokonce lepší než Švýcarsko máme nejnižší míru nezaměstnanosti v Evropské unii a také není mši a juniorůNEJNIŽŠÍ MÍRU příjmové diferenciace respektive míru chudoby v Evropské unii náš ekonomický růst je 1 z nejvyšších a naopak naše zadluženost 1 z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomý národ který si váží svých úspěchů ale samozřejmě že každé světlo má své stíny a dovolte mi abych i o těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvalý musímMUSÍ být doprovázenýDOPROVÁZEN růstem investic a tady máme své nedostatky podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například není hotová silnice mezi Brnem a Vídní ale takových dálnic obchvatů a dalších staveb je celá řada je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidí kterým říkáme nepřizpůsobiví mluvil jsem nedávno s novou ministryni práce a sociálních věci a shodli jsme se na tom že je třeba omezit sociální dávky těm kdo odmítají nabízenou práci růst hrubého domácího produktu je výborný ale musí být doprovázen růstem životní úrovně jsem rád že konečně dochází k růstu mezd pod tlakem nedostatku pracovních sil který nutí zaměstnavatele k tomu aby tyto mzdy zvyšovaly neboť jednakJINAK jim pracovnicePRACOVNÍCI utečou jinam a měl bych 1 námět který se vám možná bude zdát bizarní před 15 lety jsme měli 80 000 státních úředníků nežNE státních zaměstnanců státních úředníků dnes jich máme 150 000 protože se nám množí podle Parkinsonových zákonů kdyby se alespoň část z nich podařilo uvolnit protože tito lidé si často jenom vymýšlejí zbytečnou práci soukromý sektor který trpí nedostatkem pracovních sil bylBY JE velmi rád přijal od břevna mélooODBŘEMENIL by se státní rozpočet a uvolnilo by se napětí na trhu práce teď mělMĚ státní úřednociÚŘEDNÍCI určitě nebudou mít rádi ale já tuTO s nimi myslím dobře tolik K ekonomické situaci České republiky a závěrem protože jsem sám starobníchSTAROBNÍ důchodcůDŮCHODCE bych chtěl popřát starobním důchodcům aby v příštím roce došlo k výrazné valorizaci těchto důchodů někdy mně píší co z toho bude nám nejdeBUDEME MÍT když se pořád zdražuje ale já jim připomínám že

zdražování neboli inflace je u starobních důchodců zahrnuto do valorizace takže se nemají z čeho obávat A jenom bych si přál aby se totoTATO onemocnění otceVALORIZACE odehrávala stejnou částkuČÁSTKOU pro všechny protože v opačném případě se neustále zvětšuje rozdíl mezi novými a starými důchody teď bych přešel k zahraniční politice jenom velice stručně jsme součástíSOUČÁSTÍ Evropské unie jsme součástí Severoatlantické aliance a měli bychom svémuSE v obou těchto organizacích chovat jako energeti energickyENERGICKÝ a sebevědomýmSEBEVĚDOMÝ partneremPARTNER který když se mu něco znali byNELÍBÍ takto dá najevo pokud jde o Evropskou unii víte že jeJÍ neustále vytýkám že nedokáže chránit své vnější hranice a pokud jde o Severoatlantickou alianci myslím si že by měla být daleko aktivnější v boji proti islámskému terorismu nu a to jenTEN nejbolavější problém který zde máme migrační kvóty musí být řešen v souladu s českými národními zájmy a naším národním zájmem pochopitelně je zachování suverenity České republiky nikdo nám nemůže diktovat koho umístíme na své území a věřím že koncept migračních kvót skončí v propadlišti dějin a že za rok už SE o něm nebude mluvit a neníNYNÍ mi dovolte abych přišel doPŘEŠEL K vnitropolitické scéně toTA zažila tento rok doslova zemětřesení po volbách se 2 nejstarší České politické strany tj. sociální demokracie a strana lidová dostaliDOSTALY na pokrajiOKRAJ propasti a téměř téměř se nedostaneNEDOSTALY do parlamentu ať si vedoucí představitelé těchto stran sami zhodnotí proč k tomu tak dobu došlo nicméně jsem šelSEŠEL jsem své se s vedoucími představiteli všech devíti parlamentních subjektů a zeptal jsem se na jejich představu o povolebním uspořádání víte co mi řekli pane prezidente s námi nikdo nechce jednat a to mi říkali všichni a to je přece nesmysl znamená to tedy že pokud s vámi opravdu někdo nechce jednat tak výzvámVY ZA za ním musíte jít nasedatNESEDAT jako panenka v koutě a jednání může zúročitZAČÍT i sZ vaší iniciativy a mělo by to být jednáníJEDNÁNÍ které se bude týkat jak programových tak personál o nichPERSONÁLNÍCH kompromisuKOMPROMISŮ jak zpěvJE V politice zvykem myslím si že návrh programového prohlášení Babišovy vlády je prostěPRO takové diskuse docela dobrým výchozím materiálem když mluvím o Babišově vládě chtěl bych poznamenat že jsem dostával moudré rady které mi říkali nejmenujiNEJMENUJTE Babišovu vládu dřív než Babiš získatZÍSKÁ důvěru v Poslanecké sněmovně ale to by reálně znamenalo že po týdny možná dokonce po měsíce tady povládne Sobotkova vláda v demisi a to si snad proboha nikdo nepřežilNEPŘEJE bych siNECHCI vás strašit Belgií kde jednání o vládě trvalo 2 roky myslím si že během několika málo měsíců se podaří mít shodilZDE vládu s důvěrou

Poslanecké sněmovny a nebude-li úspěch hned v 1. kole vytvořím časový prostor pro důkladná jednání politických partnerů tak aby snad někdy v únoru se uskutečnil 2. pokus který již může být úspěšný nezapomeňte že v politice jsou 2 extrémy 1. pokus obklíčit vítěze voleb vytvořit proti němu jakousi pseudo koaliciPSEUDOKOALICI těch méně úspěšných a izolovat ho 2. extrém vítěz zašlape do země všechny poražené a vládne islám aSÁM myslím si že z hlediska demokratických procedur je dobré usilovat o to aby žádný z těchto extrémů se neuskutečniloUSKUTEČNIL a aby politické strany spolu jednali a CSc. A TY KTERÉ SE dohodnou vytvořily patrně menšinovou tolerovanou vládu závěrem této části bych chtěl potěšit občany a nepotěšit politiky občas slyším hlasy že bych měl vypsat předčasné volby což mi ústava v některých situaciSITUACÍCH umožňuje chci naprosto jasně říci že to nikdy neudělám protože předčasné volby několik měsíců po řádných volbách by byly výsměchem občanůOBČANŮM kteří šli k řádným volbám a volili tak jak volili občané rozdali karty to je a politici musí umět s těmito kartami hrát občan vyOBČANY nemůžete vyměnit olejALE můžete vyměnit politiky proč ne někteří z nich možná skončí v opozici a jak napsal Rudolf Bechyně starý sociální demokrat za 1. republiky suchaSUCHÁ je z pivaSKÝVA opozice a já dodávám klatovský 2TATO SKÝVA nikdy není a ani posolené aPOSOLENÁ takže nepřeji někomuNIKOMU aby skončiloSKONČIL v opozici ale na 2. straně demokracie a samozřejmě určitou opozici v parlamentu vyžaduje až si každý sám rozhodne zda takovou opozici chce být věřícímVĚŘÍM že nedojde k tomu že se budeme obviňovat z toho že výsledky minulých parlamentních voleb byly zmanipulovány zahraničními rozvědkami je to ubohé je to trapné a je to urážející volby byly svobodnémuSVOBODNÉ a jsem rád že i Bezpečnostní informační služba ! AVYDALA prohlášení kde konstatuje že k žádnému takovému ovlivňování nedošlo jsme svobodný názor na něSVOBODNÁ ZEMĚ se svobodnými občany a dnes už si můžeme říci že jsme i úspěšná země před dvěma lety jsem vám řekl že skončila blbá nálada dnes vám říkám že se nemáme ZA co stydět a že je mnoho věciVĚCÍ na které můžeme býti hrdi takže z líhně naZDVIHNĚME hlavu a buď dnes byli vědomiBUĎME SEBEVĚDOMÝMI lidmi kteří spoléhají na svůj vlastní rozum a nedají slovySE manipulovat ať už údajnými zahraničními rozvědkami nebo zejména českým tiskem Českou televiziTELEVIZÍ a dalším miDALŠÍMI MÉDII vždy spoléhajíSPOLÉHEJTE více na vlastní rozum a já mám pro to své heslo které jsem vám už jednou říkal věřím že zdravý rozum zvítězí nad závistivou hloupostí a

neníNYNÍ mi dovolte abych připil na vítězství zdravého rozumu na úspěch České republiky a na úspěch nás všech Šťastný a Veselý nový rok

Příloha C: Přepis službou Přepisovatel.cz

ano tak byl povinen vložení mylný spoluobčané přeji vám krásný dobrý den setkáváme se opět port se abychom společně a zhodnotili co se nám a za ten rok podařilo a případně co se za ten rok nepodařilo a jsem rámec že vám mohu říci takže těch radostných sprav je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jaké úspěchy to jsou česká republika je šest a nejbezpečnější země na světě sme dokonce lepší než švýcarsko máme nejnižší mírou nezaměstnanosti evropské uni a pak je nejnižší míru přímo ve diferenciace respektive mýho chudoby evropské unii náš ekonomický růst je jeden z nejvyšších a naopak naše adluženost jedna z nejnižších tak na co si máme stěžovat měli bychom být hrdý a sebevědomí na který si váží svých úspěchu ale samozřejmě že každé světlo má své stíny a dovolte mi abych bych po těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvali musí být doprovázen růstem investic a tady máme sebe nedostatky podíl veřejných investic svá celokovových výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například na něho toho silnice mezi brnem a vídní a na takových dál nic obchvatů a dalších staveb je celá řada je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidi kterým říkáme nepřízpůsobiví mluvil jsem í. nedavna se mnou ministryni práce a sociální věci a shodli jsme se na tom takže třeba omezit sociální dávky těm kdo odmítají nabízen úprav si růst hrubého domácího produktu je výborný ale musí být doprovázen růstem životní úrovně sem hrát takže konečně dochází k růstu mars podtlakem nedostatku pracovních sil který nutí zaměstnavatelé k tomu aby tyto mzdy zvyšovali neboť jinak jim pracovnici utečou jinam a měl bych je daná méně který se vám možná bude zdát bizarní před patnácti lety sme měli osumdesát tisíc státních úředníků nestátních zaměstnanců

státních úředníků dnes bych máme sto padesát tisíc proto zase namnoží podle paniky mezony zákonů kdyby se alespoň část z nich podařilo uvolnit protože tito lidé si často jenom vymýšlejí zbytečnou práci soukromý sektor který trpí nedostatkem pracovních sil byl velmi right přijel dobře méně orbis ostatně rozpočet a uvolňujou by se na pěti na trhu práce teď mně statný úřednici určitě nebudou mít rády ale já to sněm nemyslim dobře tolik ekonomické situaci české republiky a vzali jeden protože jsem sám starobní důchod se bych chtěl popřát starobním důchodcům aby příštím roce došlo k výrazné valorizaci těchto důchodů někdy mě píší co s toho budeme mít když se pořád zdražuje ale já jim připomínám že zdražování neboli

inflace je u stavebních důchodců zahrnuto do valorizace takže se nemají s čeho obávat a jenom bych si přál aby se tato konverzace odehrávala stejnou částkou pro všechny protože f. opačném případě se neustále zvětšuje rozdělil mezi novými a starými důchody teď bych přestřelky zahraniční politice jenom velice stručně sme součásti evropské unie jsme součásti severoatlantické aliance a měli bychom se obou těchto organizacích chovat jako energetici energicky a sebevědomí partner který když se mu něco malý by takto dá najevo pokud jde o evropskou unii víte že ji neustále vytýkám že nedokáže chránit své vnější hranice a pokud jde o severoatlantickou alianci myslím si že by měla být daleko aktivnější v boji proti islámské mu terorismu no a na ten nejbolavější problém který zde máme migrační coding musí být řešen souladu s českými národními zahájeny a vašim národy němu zájmem pochopitelně je zachování suverenity české republiky nikdo nám nemůže diktovat koho umístíme na své území a věřim že koncept legračních koho skončí propadlišti dějin takže za rok ušlého něm nebude mluvit tak na něm jednoho to abych přišel do vnitropolitické scéně ta zažila tento doslova zemětřesení po volbách se dvě nejstarší české politické strany to jest sociální demokracie a stranami doleva dostali na okraji propasti a téměř téměř se nedostane do parlamentu ať si vedoucí představitelé těchto stran sem nezhodnotit proč k tomu tak do došlo nicméně sešel jsem se se vedoucími představiteli vše devíti parlamentních subjektu a zeptal sem se na jejich představu volebním uspořádání víte co mi řekli pane prezidente se nám mi nikdo nechce jedna atomy říkali všichni a to je přece nesmysl znamená to tedy takže pokud s vámi opravdu v někdo nechce jedna takže vy sám s. ani nemusíte nic nese dat jako panenka vkladů tě a jednání může značit i z vaší iniciativy a mejlově to bity jednání které se bude týkat jak programových tak personální jich kompromisu je kyjev politice zvykem myslím si že návrh programového prohlášení babi šavi vlády je pro takové diskuze docela dobrým výchozím materiál když mluvilo obdoby filmy a de chtěl bych poznamenat takže jsem dostal moudré rádi které neřikali nejmenujete aby ševo vládu dřív než babičce získá důvěru poslanecké sněmovně ale to by reálně znamenalo že po týdny možná dokonce půlměsíce tady pokládané sobotková vláda v demisi a to si snad proboha nikdo nepřežil nechci vás strašně dbali v. nejede jedna něho vládě trvalo dva roky myslím si že během několika málo měsíců se podaří mít zde volala do z důvěrou poslanecké sněmovny a nebo byly úspěchem hned prvním kole vytvoříme časový prostor pro důkladná jednání politických partnerů tak aby snad někdy v únoru se uskutečnila druhý pokus který když může být úspěšný nezapomeňte takže politice jsou dva extrémně první pokus obklíčit vítěze vole vytvořit proti němu jakousi pseudo koalici těch méně úspěšných a k izolovat dva druhý

extrém vítěz zašel a pedro zeleně všechny poražené a vládne sám myslím si že z hlediska demokratických procedur je dobré usilovat o to abych žádný z těchto extrému se neuskutečnil a aby politické strany spolu jednali a ty které se dohodnou vytvořili patrně menšinovou tolerováno landů závěrem této části bych chtěl potěšit občany a potěšit politiky občas slyším hlasy že bych méně vypsat předčasné volby což mi ústava v některých situacích umožňuje chci naprosto jasně říci že to nikdy neudělal protože předčasné volby několik měsíců pořádnýho nebyly výsměchem občanů kteří žili v žádným volbám a volili ta je volili občané rozdali kate a politici musí umět s těmito kartami hrát opřeli nemůžete vyměnit ale můžete vyměnit politiky proč ne někteří z nich může na skončí von pozici a jak napsala rudolf bechyně starý sociální demokrat za první republiky suchara irskýma opozice a já dodávám tato stevena někdy není ani posune takže nepři někomu aby skončilo v opozici ale na druhé straně demokracie samozřejmě určitou opozici for elementu vyžaduje ať si každý sám rozhodne zda takovou opozicí chce být věříme jižně nedojde k tomu že se budeme obviňován toho že výsledky minulých parlamentního l. byly zmanipulované zahraničními rozvědka ne je to libovolné je to trapné a je to urážející volby byly svobodné a jsem rand takže vy bezpečnostní informační služba vydala prohlášení kde konstatuje řekli žádnému takovému ovlivňovány nedošlo sme svobodná zejména se svobodnými občané a dnes u si můžeme říci že jsme i úspěšná země před dvěma lety jsem vám řekl že skončila byl byla nám dnes vám říkám že se nemáme za co stydět takže je mnoho věci na které můžeme b. ti hadi takže zdvihy něm na hlavu a buďme sebevědomí jinými lidmi kteří spoléhali na svůj vlastní rozum a dají se manipulovat a čemuž údajnými zahraničními rozvědka my nebo zejména česky útiskem českou televizi a dalšími médii spoléhají té na vlastní rozum a já mám proto sled heslo ptali seznamu štěpnou říkal věřím takže zdravý rozum z vítězi krát závistivou hlouposti a není mi dovolte abych připila na vítězství zdravého no na úspěch české republiky a na úspěch nás všech šťastný a veselý nový jo dobrý nějak

Příloha D: Přepis službou Přepisovatel.cz – chyby

ano tak byl povinen vloženíVÁŽENÍ A mylnýMILÍ spoluobčané přeji vám krásný dobrý den setkáváme se opět port sePO ROCE abychom společně a zhodnotili co se nám a za ten rok podařilo a případně co se za ten rok nepodařilo a jsem rámecRÁD že vám mohu říci takže těch radostných spravZPRÁV je daleko více a chtěl bych poděkovat všem kdo se za naše úspěchy zasloužili jaké úspěchy to jsou českáČESKÁ republika je šest aŠESTÁ nejbezpečnější země na světě smeJSME dokonce lepší než švýcarskoŠVÝCARSKO máme nejnižší mírouMÍRU nezaměstnanosti V evropskéEVROPSKÉ uniUNII a pak jeTAKÉ nejnižší míru přímo veCENOVÉ diferenciace respektive mýhoMÍRU chudoby V evropskéEVROPSKÉ unii náš ekonomický růst je jeden z nejvyšších a naopak naše zadluženost jedna z nejnižších tak na co si máme stěžovat měli bychom být hrdýHRDÍ a sebevědomíSEBEVĚDOMÝ naNÁROD který si váží svých úspěchuÚSPĚCHŮ ale samozřejmě že každé světlo má své stíny a dovolte mi abych bych I poO těchto stínech vám něco řekl ekonomický růst je krásná věc ale aby byl trvaliTRVALÝ musí být doprovázen růstem investic a tady máme sebeSVÉ nedostatky podíl veřejných investic sváNA celokovovýchCELKOVÝCH výdajích státního rozpočtu je nízký a klesající a zejména dochází k poklesu investic do dopravní infrastruktury je to ostuda že například naNENÍ něho tohoHOTOVÁ silnice mezi brnemBRNEM a vídníVÍDNÍ aALE na takových dál nicDÁLNIC A obchvatů a dalších staveb je celá řada je dobře že máme nízkou nezaměstnanost ale stále je zde skupina lidiLIDÍ kterým říkáme nepřízpůsobivíNEPŘIZPŮSOBIVÍ mluvil jsem í. nedavnaNEDÁVNO seS mnouNOVOU ministryni práce a sociálníSOCIÁLNÍCH věciVĚCÍ a shodli jsme se na tom takžeŽE JE třeba omezit sociální dávky těm kdo odmítají nabízenNABÍZENOU úprav siPRÁCI růst hrubého domácího produktu je výborný ale musí být doprovázen růstem životní úrovně semJSEM hrátRÁD takžeŽE konečně dochází k růstu marsMEZD podtlakemPOD TLAKEM nedostatku pracovních sil který nutí zaměstnavateléZMĚSTNAVATELE k tomu aby tyto mzdy zvyšovaliZVYŠOVALY neboť jinak jim pracovniciPRACOVNÍCI utečou jinam a měl bych je danáJEDEN méněNÁMĚT který se vám možná bude zdát bizarní před patnácti lety smeJSME měli osumdesátOSMDESÁT tisíc státních úředníků nestátníchNE STÁTNÍCH zaměstnanců státních úředníků dnes bychJICH máme sto padesát tisíc protoPROTOŽE zaseSE namnožíNÁM MNOŽÍ podle paniky mezonyPARKINSONOVÝCH zákonů kdyby se alespoň část z nich podařilo uvolnit protože

tito lidé si často jenom vymýšlejí zbytečnou práci soukromý sektor který trpí nedostatkem pracovních sil bylBY JE velmi rightRÁD přijelPŘIJAL dobře méně orbisODBŘEMENIL BY SE ostatněSTÁTNÍ rozpočet a uvolňujouUVOLNILO by se na pětiNAPĚTÍ na trhu práce teď mně statnýSTÁTNÍ úředniciÚŘEDNÍCI určitě nebudou mít rádyRÁDI ale já to sněmS NIMI nemyslimMYSLÍM dobře tolik K ekonomické situaci českéČESKÉ republiky a vzali jedenZÁVĚREM protože jsem sám starobní důchod seDŮCHODCE bych chtěl popřát starobním důchodcům aby V příštím roce došlo k výrazné valorizaci těchto důchodů někdy měMNĚ píší co sZ toho budeme mít když se pořád zdražuje ale já jim připomínám že zdražování neboli inflace je u stavebníchSTAROBNÍCH důchodců zahrnuto do valorizace takže se nemají sZ čeho obávat a jenom bych si přál aby se tato konverzaceVALORIZACE odehrávala stejnou částkou pro všechny protože f.V opačném případě se neustále zvětšuje rozdělilROZDÍL mezi novými a starými důchody teď bych přestřelkyPŘEŠEL K zahraniční politice jenom velice stručně smeJSME součásti evropské unie jsme součástiSOUČÁSTÍ severoatlantickéSEVEROATLANTICKÉ aliance a měli bychom se V obou těchto organizacích chovat jako energetici energickyENERGICKÝ a sebevědomíSEBEVĚDOMÝ partner který když se mu něco malý byNELÍBÍ taktoTAK TO dá najevo pokud jde o evropskouEVROPSKOU unii víte že ji neustále vytýkám že nedokáže chránit své vnější hranice a pokud jde o severoatlantickouSEVEROATLANTICKOU alianci myslím si že by měla být daleko aktivnější v boji proti islámské muISLÁMSKÉMU terorismu noNU a na ten nejbolavější problém který zde máme migrační codingKVÓTY musí být řešen V souladu s českými národními zahájenyZÁJMY a vašimNAŠÍM národy němuNÁRODNÍM zájmem pochopitelně je zachování suverenity českéČESKÉ republiky nikdo nám nemůže diktovat koho umístíme na své území a věřimVĚŘÍM že koncept legračníchMIGRAČNÍCH kohoKVÓT skončí V propadlišti dějin takžeA ŽE za rok ušlýhoUŽ SE O něm nebude mluvit tak na němA NYNÍ MI jednoho toDOVOLTE abych přišelPŘEŠEL doK vnitropolitické scéně ta zažila tento ROK doslova zemětřesení po volbách se dvě nejstarší české politické strany to jest sociální demokracie a stranamiSTRANA levaLIDOVÁ dostaliDOSTALY na okraji propasti a téměř téměř se nedostaneNEDOSTALY do parlamentuPARLAMENTU ať si vedoucí představitelé těchto stran semSAMI nezhodnotitZHODNOTÍ proč k tomu tak do došlo nicméně sešel jsem se seS vedoucími představiteli všeVŠECH devíti parlamentních subjektuSUBJEKTŮ a zeptal semJSEM se na jejich představu O volebním uspořádání víte co mi řekli pane prezidente s nám miS NÁMI nikdo nechce jednaJEDNAT atomyA TO MI říkali všichni a to je přece nesmysl znamená to tedy takžeŽE pokud s vámi opravdu v

někdoNIKDO nechce jednaJEDNAT takžeTAK vy sám s.ZA aniNÍM nemusíteMUSÍTE nicJÍT nese datNESEDAT jako panenka vkladů těV KOUTĚ a jednání může značitZAČÍT i z vaší iniciativy a mejlověMĚLO BY to bityBÝT jednání které se bude týkat jak programových tak I personální jichPERSONÁLNÍCH kompromisuKOMPROMISŮ je kyjevJAK JE V politice zvykem myslím si že návrh programového prohlášení babi šaviBABIŠOVY vlády je pro takové diskuze docela dobrým výchozím materiálMATERIÁLEM když mluviloMLUVÍM O obdoby filmyBABIŠOVĚ a deVLÁDĚ chtěl bych poznamenat takžeŽE jsem dostal moudré rádiRADY které neřikaliMI ŘÍKALY nejmenujeteNEJMENUJTE aby ševoBABIŠOVU vládu dřív než babičceBABIŠ získá důvěru V poslanecké sněmovně ale to by reálně znamenalo že po týdny možná dokonce půlměsícePO MĚSÍCE tady pokládanéPOVLÁDNE sobotkováSOBOTKOVA vláda v demisi a to si snad proboha nikdo nepřežilNEPŘEJE nechci vás strašněSTRAŠIT dbali v.BELGIÍ nejedeKDE jednaJEDNÁNÍ něhoO vládě trvalo dva roky myslím si že během několika málo měsíců se podaří mít zde volala doVLÁDU zS důvěrou V poslanecké sněmovny a nebo bylyNEBUDE-LI úspěchemÚSPĚCH hned prvním kole vytvoříméVYTVOŘÍM časový prostor pro důkladná jednání politických partnerů tak aby snad někdy v únoru se uskutečnilaUSKUTEČNIL druhý pokus který kdyžJIŽ může být úspěšný nezapomeňte takžeŽE V politice jsou dva extrémněEXTRÉMY první JE pokus obklíčit vítěze voleVOLEB vytvořit proti němu jakousi pseudo koalici těch méně úspěšných a k izolovat dvaHO druhý extrém vítěz zašel aZAŠLAPE pedroDO zeleněZEMĚ všechny poražené a vládne sám myslím si že z hlediska demokratických procedur je dobré usilovat o to abychABY žádný z těchto extrémuEXTRÉMŮ se neuskutečnil a aby politické strany spolu jednaliJEDNALY a ty které se dohodnou vytvořiliVYTVOŘILY patrně menšinovou tolerovánoTOLEROVANOU landůVLÁDU závěrem této části bych chtěl potěšit občany a potěšitNEPOTĚŠIT politiky občas slyším hlasy že bych méněMĚL vypsat předčasné volby což mi ústava v některých situacích umožňuje chci naprosto jasně říci že to nikdy neudělalNEUDĚLÁM protože předčasné volby několik měsíců pořádnýhoPO ŘÁDNÝCH nebylyBY BYLY výsměchem občanůOBČANŮM kteří žiliŠLI vK žádnýmŘÁDNÝM volbám a volili taTAK jeJAK volili občané rozdali kateKARTY a politici musí umět s těmito kartami hrát opřeliOBČANY nemůžete vyměnit ale můžete vyměnit politiky proč ne někteří z nich může naMOŽNÁ skončí vonV poziciOPOZICI a jak napsalaNAPSAL rudolfRUDOLF bechyněBECHYNĚ starý sociální demokrat za první republiky sucharaSUCHÁ irskýmaJE SKÝVA opozice a já dodávám tato stevenaSKÝVA někdy není

ani posunePOSOLENÁ takže nepřiNEPŘEJI někomuNIKOMU aby skončiloSKONČIL v opozici ale na druhé straně demokracie samozřejmě určitou opozici forV elementuPARLAMENTU vyžaduje ať si každý sám rozhodne zda takovou opozicí chce být věřímeVĚŘÍM jižněŽE nedojde k tomu že se budeme obviňovánOBVIŇOVAT Z toho že výsledky minulých parlamentníhoPARLAMENTNÍCH l.VOLEB byly zmanipulované zahraničními rozvědkaROZVĚDKAMI ne je to libovolnéUBOHÉ je to trapné a je to urážející volby byly svobodné a jsem randRÁD takžeŽE vyl bezpečnostníBEZPEČNOSTNÍ informační služba vydala prohlášení kde konstatuje řekliŽE K žádnému takovému ovlivňoványOVLIVŇOVÁNÍ nedošlo smeJSME svobodná zejméněZEMĚ se svobodnými občanéOBČANY a dnes uUŽ si můžeme říci že jsme i úspěšná země před dvěma lety jsem vám řekl že skončila byl bylaBLBÁ námNÁLADA dnes vám říkám že se nemáme za co stydět takžeA ŽE je mnoho věci na které můžeme b. tiBÝTI hadiHRDI takže zdvihy něm naZDVIHNĚME hlavuHLAVY a buďme sebevědomí jinýmiSEBEVĚDOMÝMI lidmi kteří spoléhaliSPOLÉHAJÍ na svůj vlastníZDRAVÝ rozum a dajíNEDDAJÍ se manipulovat aAŤ čemužUŽ údajnými zahraničními rozvědka myROZVĚDKAMI nebo zejména českyČESKÝM útiskemTISKEM českouČESKOU televiziTELEVIZÍ a dalšími médii spoléhají téSPOLÉHEJTE na vlastní rozum a já mám proto sledSVÉ heslo ptaliKTERÉ seznamuJSEM štěpnouVÁM UŽ říkal věřím takžeŽE zdravý rozum z vítěziZVÍTĚZÍ krátNAD závistivou hlouposti HLOUPOSTÍ a neníNYNÍ mi dovolte abych připilaPŘIPIL na vítězství zdravého noROZUMU na úspěch českéČESKÉ republiky a na úspěch nás všech šťastný a veselý nový jo dobrý nějakROK

Příloha E: Lidský přepis

Vážení a milí spoluobčané, přeji vám krásný dobrý den,

setkáváme se opět po roce, abychom společně zhodnotili, co se nám za ten rok podařilo a případně, co se za ten rok nepodařilo. A jsem rád, že vám mohou říci, že těch radostných zpráv je daleko více. A chtěl bych poděkovat všem, kdo se za naše úspěchy zasloužili. Jaké úspěchy to jsou? Česká republika je šestá nejbezpečnější země na světě. Jsme dokonce lepší než Švýcarsko. Máme nejnižší míru nezaměstnanosti v Evropské unii, a také nejnižší míru příjmové diferenciace, resp. míru chudoby v Evropské Unii. Náš ekonomický růst je jeden z nejvyšších a naopak naše zadluženost jedna z nejnižších. Tak na co si máme stěžovat? Měli bychom být hrdý a sebevědomý národ, který si váží svých úspěchů.

Ale samozřejmě, že každé světlo má své stíny a dovolte mi, abych i o těchto stínech vám něco řekl.

Ekonomický růst je krásná věc, ale aby byl trvalý, musí být doprovázen růstem investic. A tady máme své nedostatky. Podíl veřejných investic na celkových výdajích státního rozpočtu je nízký a klesající. A zejména dochází k poklesu investic do dopravní infrastruktury. Je ostuda, že například není hotová silnice mezi Brnem a Vídní, ale takových dálnic, obchvatů a dalších staveb je celá řada.

Je dobře, že máme nízkou nezaměstnanost. Ale stále je zde skupina lidí, kterým říkáme nepřizpůsobiví. Mluvil jsem nedávno s novou ministryní práce a sociálních věcí a shodli jsme se na tom, že je třeba omezit sociální dávky těm, kdo odmítají nabízenou práci.

Růst hrubého domácího produktu je výborný, ale musí být doprovázen růstem životní úrovně. Jsem rád, že konečně dochází k růstu mezd. Pod tlakem nedostatku pracovních sil, který nutí zaměstnavatele k tomu, aby tyto mzdy zvyšovaly, neboť jinak jim pracovníci utečou jinam. A měl bych jeden námět, který se vám možná bude zdát bizarní. Před patnácti lety jsme měli osmdesát tisíc státních úředníků. Ne státních zaměstnanců, státních úředníků. Dnes jich máme sto padesát tisíc, protože se nám množí podle Parkinsonových zákonů.

Kdyby se alespoň část z nich podařilo uvolnit, protože tito lidé si často jenom vymýšlejí zbytečnou práci, soukromý sektor, který trpí nedostatkem pracovních sil, by je velmi rád přijal, odbřemenil by se stání rozpočet a uvolnilo by se napětí na trhu práce. Teď mě státní úředníci určitě nebudou mít rádi, ale já to s nimi myslím dobře. Tolik k ekonomické situaci České republiky.

A závěrem, protože jsem sám starobní důchodce, bych chtěl popřát starobním důchodcům, aby v příštím roce došlo k výrazné valorizaci těchto důchodů. Někdy mně píší, co z toho budeme mít, když se pořád zdražuje, ale já jim připomínám, že zdražování neboli inflace je u starobních důchodců zahrnuto do valorizace, takže se nemají Z čeho obávat. A jenom bych si přál, aby se tato valorizace odehrávala stejnou částkou pro všechny, protože v opačném případě se neustále zvětšuje rozdíl mezi novými a starými důchody.

Teď bych přešel k zahraniční politice. Jenom velice stručně. Jsme součástí Evropské unie, jsme součástí Severoatlantické aliance, a měli bychom se v obou těchto organizacích chovat jako energeti energický a sebevědomý partner, který, když se mu něco nelíbí, tak to dá najevo. Pokud jde o Evropskou unii, víte, že jí neustále vytýkám, že nedokáže chránit své vnější hranice. A pokud jde o Severoatlantickou alianci, myslím si, že by měla být daleko aktivnější v boji proti islámskému terorismu.

Nu a ten nejbolavější problém, který zde máme - migrační kvóty, musí být řešen v souladu s českými národními zájmy. A naším národním zájmem pochopitelně je zachování suverenity České republiky. Nikdo nám nemůže diktovat, koho umístíme na své území. A věřím, že koncept migračních kvót skončí v propadlišti dějin a, že za rok už se o něm nebude mluvit.

A nyní mi dovolte, abych přešel k vnitropolitické scéně. Ta zažila tento rok doslova zemětřesení. Po volbách se dvě nejstarší české politické strany - to jest sociální demokracie a strana lidová - dostaly na okraj propasti a téměř téměř se nedostaly do Parlamentu. Ať si vedoucí představitelé těchto stran sami zhodnotí, proč k tomu tak do-došlo. Nicméně sešel jsem se s vedoucími představiteli všech devíti parlamentních subjektů, a zeptal jsem se na jejich představu o povolebním uspořádání. Víte, co mi řekli? Pane prezidente, s námi nikdo nechce jednat. A to mi říkali všichni. A to je přece nesmysl. Znamená to tedy, že pokud s

vámi opravdu někdo nechce jednat, tak vy za za ním musíte jít, nesedat jako panenka v koutě, a jednání může začít i z vaší iniciativy. A mělo by to být jednání, které se bude týkat jak programových, tak personálních kompromisů, jak je v politice zvykem. Myslím si, že návrh programového prohlášení Babišovy vlády je pro takové diskuze docela dobrým výchozím materiálem. Když mluvím o Babišově vládě, chtěl bych poznamenat, že jsem dostával moudré rady, které mi říkaly, nejmenujte Babišovu vládu dřív, než Babiš získá důvěru v Poslanecké sněmovně. Ale to by reálně znamenalo, že po týdny, možná dokonce po měsíce tady povládne Sobotkova vláda v demisi, a to si snad proboha nikdo nepřeje. Nechci vás strašit Belgií, kde jednání o vládě trvalo dva roky. Myslím si, že během několika málo měsíců se podaří mít zde vládu s důvěrou Poslanecké sněmovny, a nebude-li úspěch hned v prvním kole, vytvořím časový prostor pro důkladná jednání politických partnerů tak, aby snad někdy v únoru se uskutečnil druhý pokus, který již může být úspěšný.

Nezapomeňte, že v politice jsou dva extrémy. První je pokus obklíčit vítěze voleb, vytvořit proti němu jakousi pseudokoalici těch méně úspěšných a izolovat ho. Druhý extrém, vítěz zašlape do země všechny poražené a vládne sám. Myslím si, že z hlediska demokratických procedur je dobré usilovat o to, aby žádný z těchto extrémů se neuskutečnil a aby politické strany spolu jednaly a ty, které se dohodnou, vytvořily patrně menšinovou tolerovanou vládu. Závěrem této části bych chtěl potěšit občany a nepotěšit politiky.

Občas slyším hlasy, že bych měl vypsat předčasné volby, což mi ústava v některých situacích umožňuje. Chci naprosto jasně říci, že to nikdy neudělám, protože předčasné volby několik měsíců po řádných volbách by byly výsměchem občanům, kteří šli k řádným volbám a volili tak, jak volili. Občané rozdali karty, a politici musí umět s těmito kartami hrát. Občany nemůžete vyměnit, ale můžete vyměnit politiky, proč ne. Někteří z nich možná skončí v opozici, a jak napsal Rudolf Bechyně, starý sociální demokrat za první republiky - suchá je skýva opozice. A já dodávám, tato skýva někdy není ani posolená.

Takže nepřeji nikomu, aby skončil v opozici, ale na druhé straně demokracie samozřejmě určitou opozici v Parlamentu vyžaduje, a ať si každý sám rozhodne, zda takovou opozicí chce být. Věřím, že nedojde k tomu, že se budeme obviňovat z toho, že výsledky minulých parlamentních voleb byly zmanipulovány zahraničními rozvědkami. Je to ubohé, je to trapné a je to urážející. Volby byly svobodné a jsem rád, že i Bezpečnostní informační služba vydala

prohlášení, kde konstatuje, že k žádnému takovému ovlivňování nedošlo. Jsme svobodná země se svobodnými občany. A dnes už si můžeme říci, že jsme i úspěšná země. Před dvěma lety jsem vám řekl, že skončila blbá nálada. Dnes vám říkám, že se nemáme za co stydět a že je mnoho věcí, na které můžeme býti hrdi. Takže zdvihněme hlavy a buďme sebevědomými lidmi, kteří spoléhají na svůj zdravý rozum a nedají se manipulovat, ať už údajnými zahraničními rozvědkami, nebo zejména českým tiskem a Českou televizí a dalšími médii. Spoléhejte na vlastní rozum. A já mám pro to své heslo, které jsem vám už jednou říkal. Věřím, že zdravý rozum zvítězí nad závistivou hloupostí.

A nyní mi dovolte, abych připil na vítězství zdravého rozumu, na úspěch České republiky a na úspěch nás všech.

Šťastný a veselý nový rok.

Příloha F: CD