Univerzita Hradec Králové

Fakulta informatiky a managementu

Katedra informatiky a kvantitativních metod

Text mining v českém prostředí

Bakalářská práce

Autor: Filip Bidlo
Studijní obor: informační management

Vedoucí práce: Mgr. Jiří Haviger, Ph.D.

Hradec Králové duben 2015

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 23.4.2015 Filip Bidlo

Poděkování:

Děkuji vedoucímu bakalářské práce Mgr. Jiřímu Havigerovi, Ph.D. za metodické vedení práce a odbornou konzultaci při její tvorbě.

Anotace

Bakalářská práce se zaměřuje na text mining v českém prostředí. Jejím cílem je zjištění úrovně dostupných prostředků text miningu v českém jazyce. Práce vytváří přehled těch institucí a nástrojů, které se text miningem v českém jazyce zabývají a uvádí i jejich podrobnější popis. Tento přehled slouží i k porovnání jednotlivých akademických pracovišť, firem, služeb a nástrojů. Zabývá se i jejich možnostmi a využitelností. Nalezené aplikace a nástroje byly také otestovány a zhodnocena jejich úroveň. Teoretickými východisky práce jsou informace o text miningu, průběhu procesu zpracování textu, problémy, které řeší a možnosti, které následná analýza zpracovávaného textu přináší. Bakalářská práce má formu přehledové studie dostupných prostředků.

Annotation

Title: Text Mining in Czech Language

Bachelor thesis aims on text mining in Czech language. The main goal is get to know level of available resources of text mining in Czech language. The thesis creates listing of institutions and tools that are concerned about text mining and reviews them. This overview is created to compare universities, companies, services and tools and evaluate its potential. The applications and tools that were found were also tested and evaluated. Theoretical resource of the thesis is information about text mining, data processing, methods for analyzing and benefits which the analysis brings. The bachelor thesis has form of an overview study of all available resources.

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Se stále rostoucí popularitou elektronizace dokumentů a širokým pokrytím internetem také vzniká velké množství dat. Bridgwater (2010) zmiňuje odhady firmy IBM, že 80% vznikajících dat s potenciálním využitím v obchodní sféře je nestrukturovaných a nemusí se jednat pouze o textové zdroje informací – řadí se sem i neméně důležité obrazové materiály, záznamy videa nebo zvuku. Tato data mohou mít velký potenciál, avšak je nutné je nejprve zpracovat a analyzovat. Touto oblastí se zabývá právě text mining. Jedná se v podstatě o *„soubor procedur pro klasifikaci, sumarizaci, shlukování a filtrování dokumentů, extrakci informací z textových dat.“* (Sedláček 2004)

Takto zpracovávány mohou být veškeré elektronické novinové i internetové články, obchodní smlouvy a dokumenty, ale i (textově přepsané) rozhovory se zákazníky, e-mailová komunikace. Pro běžného uživatele je tedy velmi složité orientovat se v takovém množství dat a ztrácí velké množství času nad nepodstatnými informacemi. Nástroje text miningu umožňují získat důležité informace z nestrukturovaných textů a umožnit tak jejich efektivní využití.

Dle Wittena (2004) se konaly první workshopy na téma text miningu v polovině roku 1999. V počátcích se jednalo zejména o záležitost akademické sféry, v průběhu dalších let se však začalo zapojovat i obchodní odvětví.

Obrázek 1: Text mining spojuje několik odvětví;
Zdroj: upraveno autorem dle Al-Ayyouba (2006)

2 Cíl práce

Cílem práce je zjistit úroveň dostupných prostředků a nástrojů text miningu v českém jazyce. Součástí toho je ověření použitelnosti, rozsahu a účelu použití těchto nástrojů, služeb a aplikací. Práce má být přehledovou studií současného stavu a podoby existujících institucí.

3 Metodika zpracování

Cílem této práce je zjistit úroveň dostupných nástrojů, podpůrných aplikací a jiných služeb text miningu v českém prostředí. Fakt, že pojem text mining se začal objevovat a postupně rozšiřovat již na přelomu tisíciletí nahrává hypotéze, že i pro český jazyk bude existovat dostatečná podpora ať už ze strany akademických pracovišť, tak výrobců software. Jelikož toto odvětví má své počátky v anglicky hovořícím prostředí, lze předpokládat, že úroveň nástrojů nebude na tak vysoké úrovni jako v případě angličtiny. Čeština je navíc obecně považována za složitý jazyk, a proto je otázkou, do jaké míry bude potřeba asistence uživatele těchto nástrojů při zpracování textů a zdrojů v českém jazyce a na jaké úrovni již bude jejich plná autonomie.

Pátrání po nástrojích a prostředcích bude zaměřeno na několik míst. Je téměř jisté, že akademická sféra se výzkumem a vývojem týkajícím se text miningu bude zabývat, ať už půjde o lingvistická pracoviště nebo ta spíše technicky zaměřená. Druhým zdrojem informací budou výrobci software, který umožňuje strojové zpracování textu. Dalším zdrojem jsou pak distributoři software a firmy, které se zabývají školeními v této oblasti. Je otázkou, zda se tímto oborem zabývají i jiné další instituce či jednotlivci.

Vzhledem k charakteru text miningu a nelze příliš informací očekávat v knižní podobě, a proto je pátrání v této oblasti částečně potlačeno.

4 Teoretická část – problémy řešené text miningem

Následující část práce se zabývá text mimingem obecně, procesem zpracováním text, problémy, které řeší a výsledky, které text mining. přináší.

4.1 Text mining a data mining

Text mining je podoborem data miningu, který se také zabývá dolováním informací, ovšem jeho zdrojem jsou často strukturovaná data. Pavel (2006) píše: *„důležitou vlastností data miningu je, že se jedná o analýzy odvození z obsahu dat, nikoli předem specifikované uživatelem nebo implementátorem. Jedná se především o odvozování prediktivních informací, nikoliv pouze deskriptivních.“* Hlavním rozdílem těchto dvou odvětví je tedy skutečnost, že výsledkem data miningu jsou často naprosto nová zjištění, predikce nebo modely. Naopak u text miningu jde o vytažení, zvýraznění či zobrazení informací, které jsou, ať už explicitně nebo implicitně, v textu výslovně zahrnuty.

Vzhledem k faktu, že pojem data miningu je obecně více rozšířený než text mining, dochází často ke zjednodušování a nezainteresované osoby mohou úlohy text miningu zařazovat do data miningu.

4.2 Proces zpracování textu

Proces dolování informací je rozdělen do dvou hlavních částí – první je předzpracování textu, během kterého dochází k okleštění od zbytečných slov, výrazů a prvků, které jsou ke zpracování nepotřebné. Ve druhé části přichází na řadu samotná analýza a získávání znalostí ze zpracovávaného textu.

4.2.1 Předzpracování

Během předzpracování textu je prováděno několik operací, jejichž cílem je vytvořit surový text vhodný k následnému zpracování. Široká škála dokumentů, které mohou být zpracovávány, také nabízí mnoho různých formátů. Pro následnou práci s nimi je důležité je převést na formát podporovaný text miningovým softwarem a zajistit požadované kódování. Celá příprava textů může zabrat i více času, než samotné získávání informací a jejich analýza.

Je vhodné dodat, že ne vždy je nutné (a zejména vhodné) provádět veškeré předzpracovací operace, které jsou níže uvedeny. Množství těchto operací se liší dle typu vstupního textu a výsledku, který od následného zpracování očekáváme. Je tedy nutná asistence člověka, který o těchto úkonech rozhodne.

Stop slova

Stop slova jsou výrazy v jazyce, které nenesou žádný význam, což je typické například pro předložky a spojky. Slova, která si takto určíme, jsou ze zpracovávaného textu vymazána a nebudeme s nimi nadále pracovat. Na webových stránkách http://www.ranks.nl/stopwords/ jsou k dispozici slovníky stop slov pro několik světových jazyků. Ten pro češtinu obsahuje 140 výrazů, které je možno dále upravit dle naší potřeby a typu textu, na který je hodláme aplikovat.

Stemming

Stemming je proces, který díky seznamu koncovek odstraní ze slova jeho předpony a přípony. Vzniká tedy základní kořen slova.

Této operace je využívání v internetových vyhledávačích, které dokáží uživateli zobrazit i vyhledané výsledky pro jiná tvar slova, než který konkrétně zadal. Ne vždy totiž uživatel zadá hledaný výraz v přesném tvaru, který se v nějakém textu vyskytuje, ale pokud je tzv. stemmer, který má tuto činnost na

starost, špatně nastaven a implementován, může docházet k nechtěné záměně slov a tím i překroucení celého významu.

Lemmatizátor

Lemmatizátor je podobnou operací jako stemming s tím rozdílem, že lemmatizátor vrátí základní tvar slova. Neodstraní tedy žádnou jeho část, ale případné vyskloňované či vyčasované slovo převede do jeho základního tvaru.

Vzhledem k charakteru českého jazyka obě tyto funkce nejsou stoprocentně spolehlivé. Mnoho slov tedy může být chybně vyhodnoceno a následně narušen výsledek celého dolování informací.

Morfologická analýza

„V rámci procesu automatického zpracovávání korpusu se morfologickou analýzou rozumí ta část, při níž se každému slovnímu tvaru v (korpusovém) textu přiřadí všechna jeho lemmata a všechny morfologické údaje včetně slovního druhu v podobě značky (tagu).“ (Cvrček, pojmy:morfologicka_analyza - Příručka ČNK, 2013)

Morfologická značka poskytuje informace o konkrétním slově v daném kontextu. Díky morfologickým znalostem slov můžeme snáze filtrovat výsledky či dále analyzovat texty.

K dokončení morfologického značení je nutné provést proces takzvané desambiguace (zjednoznačnění), při kterém je vybrán správný slovní tvar u zjišťovaného slova. Tento proces probíhá ručně, poloautomaticky nebo automaticky.

„Např. ve větě ‚Větry vanou od západu.‘ se při morfologické interpretaci věty nejprve přiřadí morfologickou analýzou tvaru vanou dvě lemmata a dvě morfologické interpretace:

- *lemma = vana, subst. fem. sg. instr.*
- *lemma = vát, 3. os. pl. préz,*

a poté se při desambiguaci vybere náležitá 2. interpretace.“ (Cvrček, pojmy:desambiguace - Příručka ČNK, 2014)

Shlukování synonym

V textu se může nacházet mnoho různých slov stejného významu – synonym. Pomocí databáze synonym je možno tato slova sjednotit do jednoho a ucelit výsledné počty klíčových slov.

Přiřazování váhy

Zbylá slova po provedení všech předchozích operacích se nazývají termy. Těmto termům je přiřazena váha *„podle jejich počtu výskytů a podle toho, jak moc jejich přítomnost mění obsah (význam) dokumentu.“* (Vysloužilová, 2014)

Přímá řeč

Není neobvyklé, že se ve zpracovávaném textu vyskytuje také přímá řeč ohraničená patřičnými uvozovkami. Věty vyskytující se v nich mohou být jinak zabarveny a používat naprosto jiná slova, než zbytek textu. Mohou nastat případy, ve kterých se chceme zabývat právě přímou řečí v textu uvedenou stejně jako možnost, že přímá řeč v textu uvedená je nežádoucí a musíme je extrahovat.

N-gramy

V textech se také vyskytují fráze nebo spojení několika slov, odborně nazývané n-gramy. *„N-gram je sekvence n po sobě jdoucích slov (například „strojové učení“ je 2-gram.“* (Grobelnik, 2007)

Hledání n-gramů v textu probíhá pomocí algoritmu, kterému stanovíme minimální frekvenci (počet výskytů) a maximální délku fráze. Algoritmus následně

vygeneruje n-gramy, se kterými lze dále pracovat jako se slovy. Tato operace je důležitá pro nalezení a zajištění ustálených slovních spojení a sousloví, která se v textu mohou vyskytovat.

Term frequency – inverse document frequency (TF-IDF)

Po provedení předchozích operací jsou z původních vět a slov vytvořeny tzv. termy. Pouze samotný počet termů není vhodný k dalšímu zpracování, nejprve je tedy třeba je podrobit frekvenční analýze.

TF-IDF se skládá ze dvou částí – term frequency je dle Čepka (2013) Definován takto:

$$tf(t) = \frac{počet\ výskytů\ termu}{celkový\ počet\ termů}$$

Loria (2013) k tomu dodává *„Jestliže se slovo v dokumentu vyskytuje často, je důležité. Přiřaď tomu slovu vysoké skóre.“*

Inverse document frequency se pak zaměřuje na pohled nad úroveň samotného dokumentu – ukazuje, jak často se term vyskytuje v ostatních dokumentech. Výpočet podle Čepka je

$$idf(t) = log\ \frac{||D||}{||\ \{d : t \in d\}\ ||}$$

kde: $||\ D\ ||$ - Celkový počet dokumentů

$||\ \{d : t \in d\}\ ||$ - Počet dokumentů, ve kterých se term vyskytuje

TF-IDF nakonec získáme, když tyto dvě míry vynásobíme

$$tf - idf(t, d) = tf(t, d) * idf(t)$$

Loria (2013) zjednodušuje *„Pokud se slovo objevuje v mnoha dokumentech, není to unikátní ukazatel. Přiřaď tomu slovu nízké skóre.“*

Jazykový korpus

Výsledkem procesu předzpracování textu (či textů) je jazykový korpus. Korpus je většinou velkého rozsahu, je tak zajištěna jeho reprezentativnost. Dnešní

korpusy jsou uchovány v elektronické podobě a slouží nejen k jazykovědným účelům, ale právě jako zdroje podkladů pro text mining.

Korpusů existuje mnoho druhů, mohou mít zdroje dat z psaného i mluveného projevu, jsou jednojazyčné i vícejazyčné, synchronní (představují záznamy z jazyka z určitého období) nebo diachronní (zaměřují se na vývoj jazyka v čase) apod.

4.2.2 Analýza textu

Po důkladné přípravě textu přichází na řadu aplikace vybraného typu analýzy a vytvoření strukturovaných dat z původně nestrukturovaného dokumentu. Vybraný typ analýzy již přináší konkrétní výsledky a v případě vytvoření strukturovaných dat je lze využít pro další data miningové procesy jako je tvorba modelů či predikcí.

Kategorizace textů

„Kategorizace textů (nebo také klasifikace textů) je zařazení dokumentů do předdefinovaných kategorií na základě jejich obsahu.“ (Sebastiani 2002)

Je velmi praktické mít dokumenty roztřízeny do kategorií, jelikož při dalším zpracováváním textů z určité skupiny můžeme předzpracovací proces i následnou samotnou analýzu dokonale zacílit na vlastnosti určité kategorie.

Kategorizace probíhá na základě výskytu četnosti slov v dokumentu, ten pak může být zařazen do jedné kategorie, ale určité úlohy jej mohou zařadit do více kategorií najednou. Převzato od Wittena (2004)

Dle Hearstové (1999) by kategorizace neměla být spojována s text miningem, protože to *„nevede k objevení nových informací – autor textu pravděpodobně věděl, o čem daný dokument je.“*

Sama ovšem zmiňuje výjimky, které se za text mining dají považovat, jako například iniciativu *DARPA Topic Detection and Tracking*, konkrétně její úlohu *On-*

line New Event Detection. V této službě jsou vstupními daty veškeré (online) novinové články v chronologickém pořadí a výstupem je zjištění, který článek byl na jednotlivá témata publikován jako první.

Analýza sentimentu

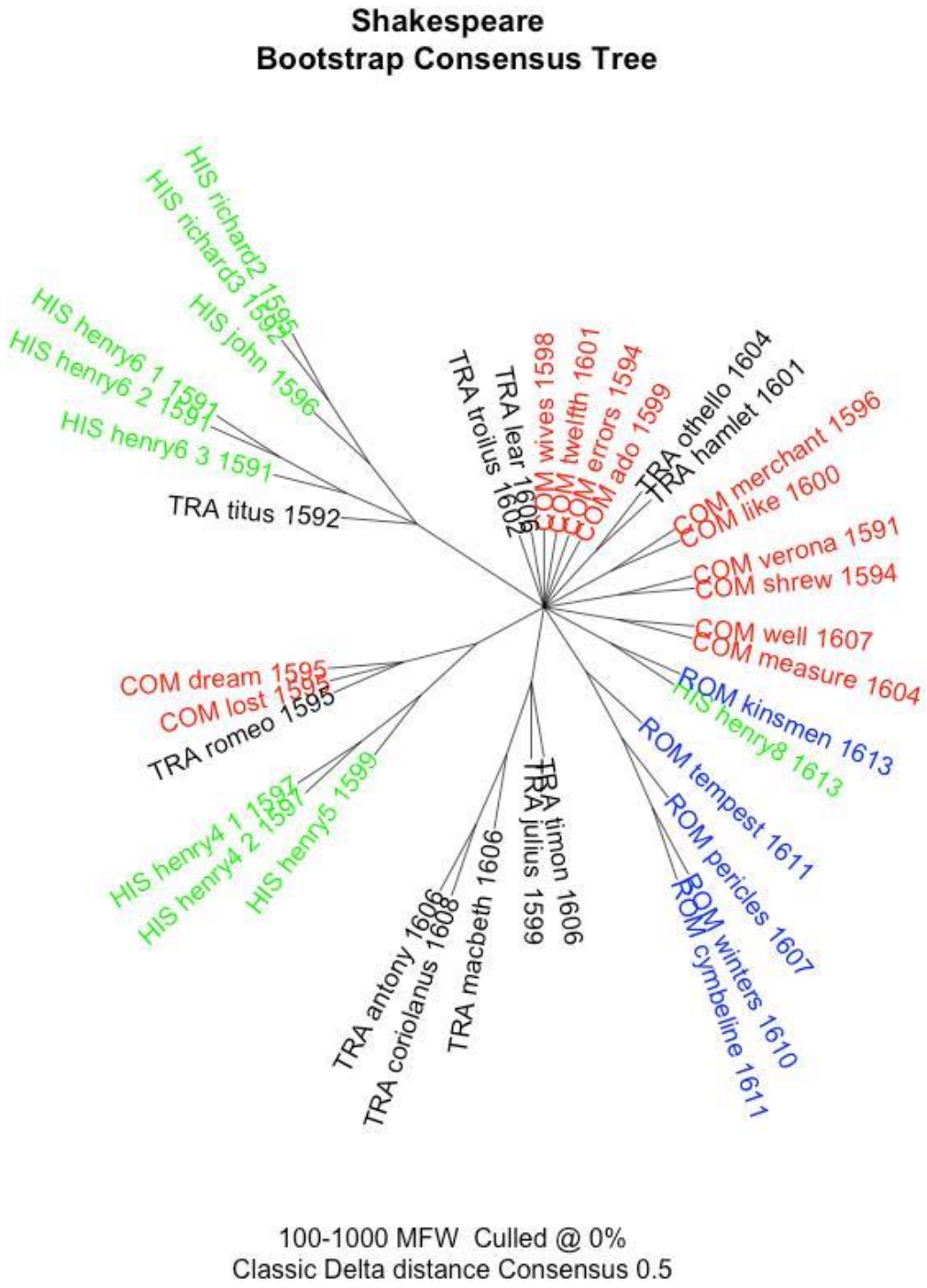
„Analýza sentimentu je technika dolování dat, která systematicky vyhodnocuje textový obsah pomocí technik strojového učení. Analýza sentimentu představuje účinnou a efektivní vyhodnocovací metodu názoru spotřebitelů v reálném času“ (Rambocas, 2013)

Analýza sentimentu má tedy využití především v marketingovém odvětví ať už u komerčních firem nebo například politických stran a jiných uskupení. Z komerčního využití je vhodné sledovat názory uživatelů na sociálních sítích, speciálních hodnotících webech (v Čechách populární heureka.cz) nebo na stránkách koncových prodejců zboží.

Shluková analýza

„Shlukování textů je plně automatický proces, který rozdělí soubor dokumentů do skupin. Dokumenty v každé skupině si jsou v určitém (zvoleném) směru podobné. Pokud je pro rozlišení použit obsah dokumentů, pak různé skupiny korespondují s různými náměty a tématy obsaženými v tomto souboru dokumentů. Na shlukování lze proto nahlížet také jako na způsob, jak zjistit, co daný soubor dokumentů obsahuje.“ (Sedláček 2004)

Zatímco v kategorizaci textů je každému dokumentu přidělen určitý (předem definovaný) štítek, u shlukování jsou sloučeny do jednotlivých skupin (shluků) dle své vzájemné podobnosti, ale bez dalšího popisu.



Obrázek 2: Výsledek shlukové analýzy Shakespearových děl

Zdroj: Whelan, 2013

Na obrázku výše jsou viditelné shluky po zpracování Shakespearových děl. Většina prvků každého shluku má stejný žánr (komedie označena jako COM,

tragédie – TRA, romantické – ROM či historické dílo – HIS) a rok napsání daného díla. Pro každý shluk je specifická vzájemná vnitřní podoba a nepodobnost mezi shluky navzájem.

Existuje množství hierarchických a nehierarchických algoritmů vhodných pro řešení shlukové analýzy.

Autorství textu

Autora dokumentu v celé řadě případů známe ať už díky jeho podpisu v textu nebo informaci uložené v metadatech souboru. S pomocí statistických metod a frekvence slov v textu se dá s určitou mírou pravděpodobnosti autorství zjistit či ověřit.

Určování autora textu je však použitelné pouze tehdy, pokud máme dostatek zpracovaných textů napsaných daným autorem – poté se můžeme pokusit potvrdit či vyvrátit, zda je tento člověk autorem i tohoto nepodepsaného dokumentu.

Za příklad stojí zpracovaná díla Williama Shakespeara, která dohromady čítají 885 000 slov, obsahující 31 500 různých slov, z nichž 14 440 bylo použitou pouze jedenkrát, 4 300 dvakrát atd. U jiných nepodepsaných dokumentů lze tedy na základě analýzy počtu unikátních slov, slov použitých pouze jednou, dvakrát apod. odvodit, zda Shakespeare může být potenciálně autorem i tohoto textu. (převzato od Wittena 2004)

Shrnutí textu

Vzhledem k množství textů a jejich délkám je sumarizace velmi vhodným nástrojem pro rychlé pochopení hlavní myšlenky dokumentu. Tento proces shrnutí definuje Ježek (2010) jako *„vyjmutí nejdůležitější informace ze zdrojového textu, která jej zestručňuje pro účely a úlohy uživatele.“*

Fráze (věty či rovnou celý odstavec) vhodné ke shrnutí dokumentu jsou následně porovnávány a je ověřována jejich vhodnost. Zjišťuje se, zda se vyskytují v názvu práce, nadpisu, abstraktu nebo zda byla použita i v jiných pracích. Tato

forma shrnutí se nazývá extrakt – jedná se o doslovné vytažení vhodných termínů z dokumentu.

Druhou formou výsledku je abstrakt. Abstraktem se rozumí *„souhrn, který nemusí obsahovat a většinou neobsahuje sekvence slov z originálního textu.“* (Ježek 2010)

Vytvoření takového souhrnu je však složité a prakticky se neobejde bez zásahu uživatele. Generování smysluplného abstraktu zachycujícího hlavní myšlenku a zároveň bez gramatických chyb je v dnešní době stále těžce proveditelné.

Filtrování

Filtrování nabízí v text miningu široké možnosti využití a v dnešní době je již hojně využíváno, ať už ke komerčním, soukromým či vědeckým účelům.

Základní rozdělení filtrů jsou tzv. white a black listy. White list specifikuje vybraná slova (věty, fráze), která jsou dovolena užívat a povoluje jejich zobrazení či zpracování. Naproti tomu black list je opačným seznamem, tedy listem zakázaných výrazů, které v konečném výsledku vidět nechceme.

Black listy našly uplatnění u spam filtrů e-mailových schránek, kde po doručení e-mailu je porovnán jeho obsah se spam filtrem. Pokud filtr vyhodnotí doučený e-mail jako spam, je přesunut do zvláštní složky. Moderní e-mailové schránky jsou tzv. učící se – reagují na úkony uživatele při označování jednotlivých e-mailů funkcemi „toto není spam“ nebo naopak „toto je spam“ a aktualizují tak své filtry.

Dnešní e-shopy funkci filtrování využívají také – nabízí konkrétnímu zákazníkovi zboží z kategorie, kterou si již v minulosti prohlížel a mohl by mít o tento produkt potenciálně zájem.

Filtrování využívá i sociální síť Facebook u statusů vytvořených správci fanouškovských stránek. Analyzuje zejména délku statusu, použití velkých/malých písmen a interpunkčních znamének. Status, který vyhodnotí jako nepříjemný (například text napsaný pouze velkými písmeny s mnoha vykřičníky) pak zobrazí menšímu počtu fanoušků než text napsaný korektně.

5 Praktická část – instituce zabývající se text miningem v ČR

5.1 Ústav Českého národního korpusu

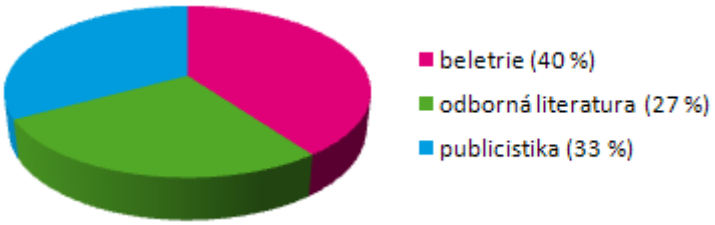
„Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který slouží k jazykovému výzkumu.“ (Čermák, oficiální web ÚČNK)

Český národní korpus byl založen v roce 1994, působí při Filozofické fakultě Univerzity Karlovy v Praze a zabývá se budováním počítačového korpusu psané i mluvené češtiny.

Korpus je sestaven z rozmanitých zdrojů. *„Texty jsou do korpusu zařazovány podle předem určených poměrů tak, aby co nejvěrněji reprezentovaly daný jazyk. Přestože jejich získávání a zařazování bývá často u menšinových typů pracné, do korpusu se vědomě kvůli jeho vyváženosti zařazují.“* (Čermák, oficiální web ÚČNK)

Korpus SYN2010

Korpus SYN2010 je žánrově vyvážený reprezantativní korpus z let 2005 až 2009.



Obrázek 3: Složení korpusu dle žánru
Zdroj: Cvrček, cnk:syn2010, 2015

Dle jeho specifikace se jedná o synchronní korpus, který splňuje následující požadavky na zdrojové texty:

- *„beletristické texty - autor narozen po roce 1880, dílo bylo vydáno po roce 1945 a je neustále čteno a vydáváno*
- *odborné texty - dílo bylo vydáno po roce 1989*
- *publicistika - není starší než 5 let“* (Cvrček, pojmy:synchronni - Příručka ČNK, 2013)

Celý korpus je složen z tzv. tokenů. *„Token je nejmenší jednotka textu, většinou se jedná o grafické slovo (tj. řetězec alfabetických znaků oddělený mezerou v textu), resp. o jednu jeho konkrétní realizaci.“* (Cvrček, pojmy:token - Příručka ČNK, 2014)

Korpus SYN2010 jich bez interpunkce obsahuje celkem 101 219 603. Tokeny jsou dále zpracovávány a rozdělovány na jednotlivé slovní tvary. *„Slovní tvar je jednotka typizovaná, jedná se o typ. Např. slovní tvar* chceme *může mít velmi mnoho různých realizací (tokenů); v korpusu SYN2010 je jich 5627.“* (Cvrček, pojmy:word - Příručka ČNK, 2013)

Dalším zpracováním získáme lemma – základní tvar slova. Například tvary lesům, lesy, lesích má lemma les. Těchto lemmat je v korpusu 785 580, což je méně než jedno procento z celkového množství tokenů.

Korpus ORAL2013

Dalším korpusem je ORAL2013, který obsahuje výrazy z mluvené spontánní češtiny z neformální komunikace a pokrytí celé České republiky.

„Korpus ORAL2013 se skládá z 835 nahrávek z let 2008–2011 a obsahuje 2 785 189 textových slov, tj. celkem 3 285 508 pozic; v sondách vystupuje celkem 2 544 mluvčích, z toho 1 297 unikátních.“ (Cvrček, cnk:oral2013, 2015)

Tento korpus není na rozdíl od typu SYN2010 lemmatizován ani morfologicky označkován.

5.1.1 Nástroje

Na oficiálních stránkách Českého národního korpusu je k dispozici také množství nástrojů, které umožňují praktické použití.

KWords

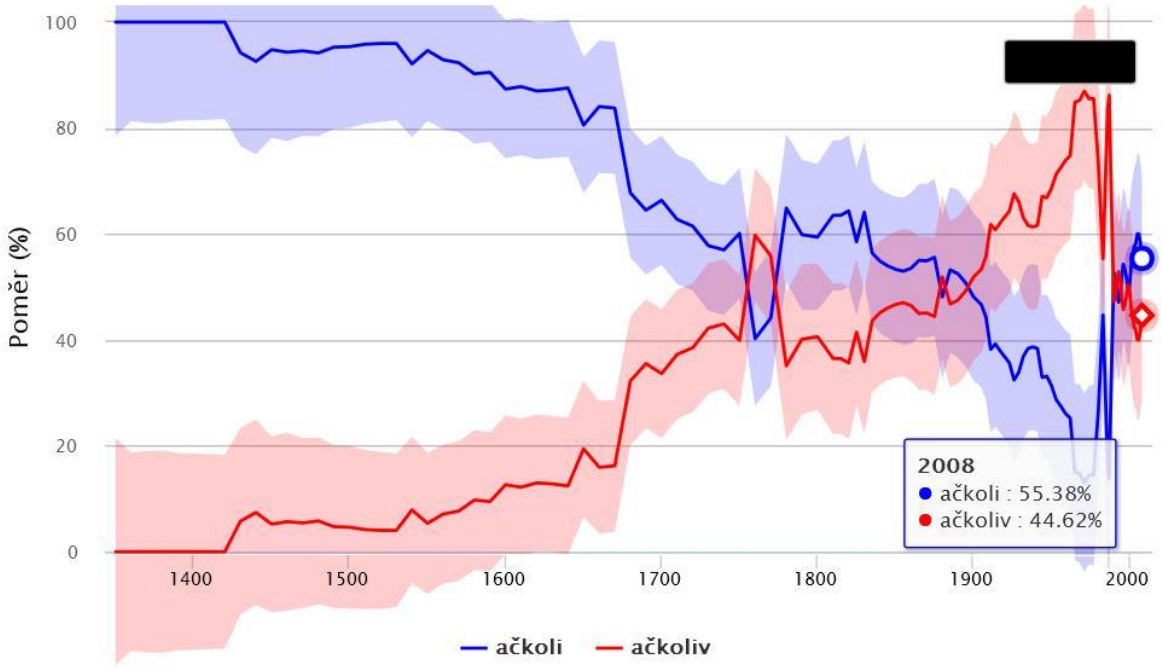
Nástroj KWords analyzuje vložený (uživatelem nahraný) text a vyhledá klíčová slova tohoto textu. Dokáže pracovat se stop-listem – odstraní slova, která nenesou význam (zájmena, předložky, spojky, čísla). Uživatel může vybrat tzv. referenční korpus, se kterým bude vložený text porovnáván a nástroj v zadaném textu označí ta, která byla použita častěji, než bychom je dle referenčního korpusu čekali. K dispozici jsou korpusy z let 2005, 2010, ale i z dob totality (1952 – 1977) nebo mluvený korpus, jejich volba musí pro korektní výsledek odpovídat vloženému textu.

Dalším výstupem z této aplikace jsou tzv. keyword links – vazba mezi klíčovými slovy. Nástroj v tomto případě zkoumá vztahy mezi klíčovými slovy daného textu, konkrétně jejich společný výskyt v textech. Pokud se slova vyskytují ve vzájemném kontextu, indikuje to blízkost témat.

SyD

Tento nástroj vzájemně porovnává četnosti použití dvou (či více) zadaných variant slov. Je vhodný pro porovnání *slov „které si vzájemně konkurují (např. ačkoli × ačkoliv, už × již.“ (Autor neznámý, oficiální web nástroje)* Porovnání lze provádět na současných korpusech psaných i mluvených, nebo z hlediska historického vývoje – synchronní i diachronní perspektiva. Aplikace je veřejně dostupná na https://syd.korpus.cz/.

Obrázek 4 ukazuje grafické znázornění historického užívání slov *ačkoli* a *ačkoliv*. Je patrné, že v době od 15. do 17. století jasně dominovalo slovo *ačkoli*, zatímco v dnešní době jsou tato dvě slova téměř ve vzájemné rovnováze.



Obrázek 4: SyD historie použití slov ačkoli x ačkoliv

Zdroj: Autor, výstup z aplikace SyD

Inverse Text Sort

„Inverse Text Sort čte text z jednoho nebo více textových souborů (prostý text), rozdělí text na slova, odstraní zadané nežádoucí znaky a slova obsahující ony nežádoucí znaky, slova inverzně setřídí (viz níže) a výstup uloží do nového textového souboru (prostý text).“ (Velíšek, 2010)

Inverzní třízení znamená vzájemné porovnávání písmen ve slovech v opačném pořadí, než je to běžné. V tomto případě se tedy jedná o třízení postupně zprava doleva.

Aplikace je k dispozici jako spustitelný .exe soubor pro operační systém Windows. Program obsahuje nastavení zpracování vkládaného textu – vypuštění slov kratších/delších než zadaný počet znaků, převedení všech slov na malá písmena, vložení podporovaných znaků a symbolů či možnosti výstupu.

Mimo to, že program seřadí slova inverzně podle abecedy a vypíše je do souboru, obsahuje také nástroj *Word Endings Picker*, který s tímto výstupním

souborem dále pracuje. Tento nástroj umožňuje slova setřídit do skupin podle jejich koncovek. To lze provést buď dle zadaného počtu shodných (posledních) znaků nebo přímo zadané koncovky.

5.2 Ústav pro jazyk český Akademie věd ČR

Ústav pro jazyk český Akademie věd ČR se zabývá českým jazykem v mnoha ohledech a vytváří i nástroje, které jsou dostupné na webu. Z hlediska text miningu je jejich nejzajímavějším výtvorem Korpus Dialog.

5.2.1 Korpus DIALOG

„Korpus DIALOG je speciální multimediální korpus mluvené češtiny. Shromažďuje veřejné jazykové projevy dialogického typu – nahrávky a přepisy diskusních pořadů českých televizí. Slouží výzkumu mediální komunikace a výzkumu mluvené češtiny v její současné veřejné podobě.“ (Autor neznámý, oficiální web korpusu)

Aktuální verze korpusu je označena jako 1.1 a obsahuje nahrávky z 28 různých televizních diskuzních pořadů, celkem je z nich zaznamenáno 932 373 textových slov.

Korpus byl automaticky morfologicky anotován a lemmatizován pomocí systému z Filozofické fakulty Univerzity Karlovy, odkud pochází také Český národní korpus.

Korpus umožňuje pomocí vyhledávače prohledání veškerých přepsaných záznamů, které jsou zde k dispozici. Vyhledávat lze slovo, lemma, tag, filtrovat je možné mluvčího jako ženu/muže nebo konkrétní osobu, pořad či datum jeho vysílání. Při hledání slova lze omezit hledání zadáním jiného slova, které se musí vyskytovat v levém nebo pravém kontextu hledaného výrazu a to až do maximálního rozsahu 25 slov od hledaného výrazu.

„Korpus DIALOG je budován a spravován na půdě Ústavu pro jazyk český Akademie věd České republiky ve spolupráci s Ústavem formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy.“ (Autor neznámý, oficiální web korpusu)

5.3 Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied

Slovenská akademie věd se primárně zabývá zkoumáním slovenštiny, avšak jejich Slovenský národní korpus se týká i českého jazyka. Na stránkách Slovenského národního korpusu, lze totiž najít Slovensko-český paralelní korpus (http://korpus.juls.savba.sk/skcs.html). Ten umožňuje vyhledávat slovo (případně více slov) v českém nebo slovenském korpusu a výsledky zobrazí v obou jazycích. Současná verze 3.0 obsahuje 119,4 milionů tokenů ve slovenské části a 119,53 milionů tokenů části české.

Tento korpus může být vhodný například k porovnání překladů či různorodosti slov dvou jinak velmi blízkých jazyků.

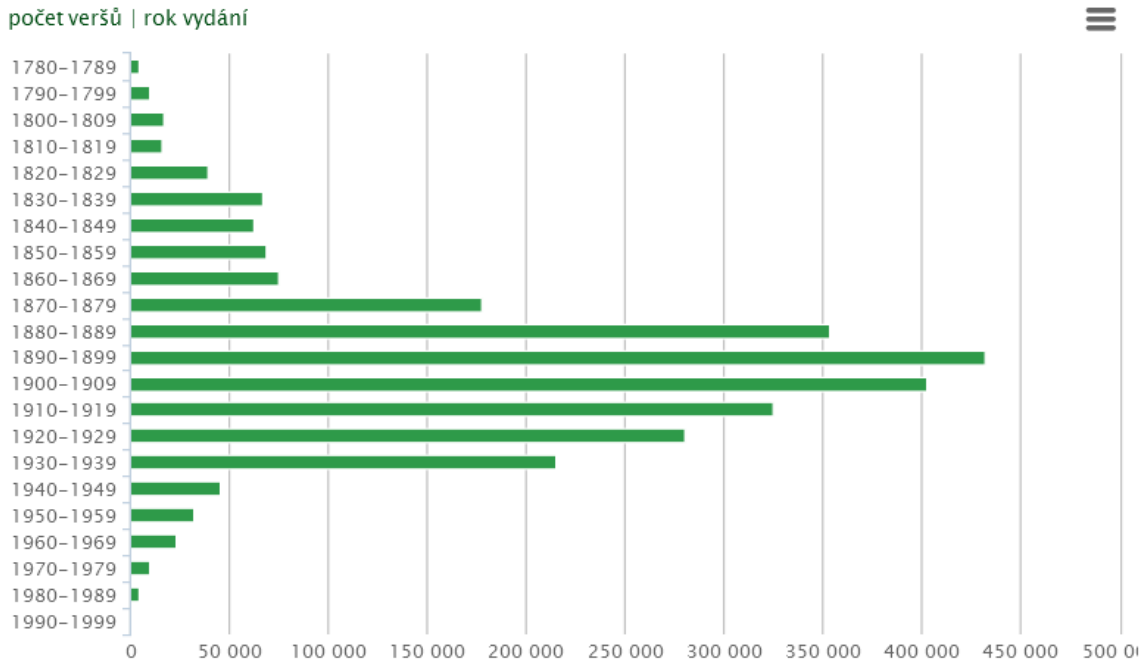
5.4 Ústav pro českou literaturu Akademie věd České republiky

„Jsme vědecký ústav, který zajišťuje základní výzkum české literatury v jazykovém i teritoriálním vymezení od počátků do současné doby.“ (Autor neznámý, oficiální web ústavu) Ústav pochopitelně spadá pod českou Akademii věd a zabývá se literárně vědeckými disciplínami.

Pracují na řadě projektů a jsou zde vyvíjeny i nástroje, které přímo souvisejí s text miningem.

5.4.1 Korpus českého verše

„Korpus českého verše (KČV) je lemmatizovaný, foneticky, morfologicky, metricky a stroficky anotovaný korpus české poezie 19. a počátku 20. století.“ (Plecháč, Korpus českého verše | Versologický tým, 2015)



Obrázek 5: Statistika počtu veršů z jednotlivých let
Zdroj: (Plecháč, Korpus českého verše | Versologický tým, 2015)

„KČV obsahuje celkem

- *1 689 básnických sbírek*
- *78 391 básní*
- *2 664 989 veršů*
- *14 592 037 slov“* (Plecháč, Korpus českého verše | Versologický tým, 2015)

Na stránkách KČV jsou k dispozici nástroje, které pracují s daným korpusem, ale i nástroje, jejž můžeme použít na námi vložený text.

Eufonometr

Eufonometr zjišťuje míru eufonie (libozvučnosti) daného textu. Tento program byl aplikován na veškeré básně v korpusu a u jednotlivých autorů byl eufonický koeficient zprůměrován. Do webového formuláře lze vložit libovolný text a výsledek je možno porovnat s autory básní

Frekvenční slovníky

„Frekvenční slovníky české poezie obsahují údaje o četnosti slov v básnických dílech zahrnutých do Korpusu českého verše.“ (Plecháč, Frekvenční slovníky | Versologický tým, 2015)

Mezi tři nejčastěji užívaná podstatná jména v českých básních tak patří slova *srdce*, *duše* a *láska*.

Frekvenční slovníky jsou zpracovány pro všechny autory jak na bázi lemmat, tak tokenů.

Gunstick – databáze českých rýmů

„V Databázi českých rýmů zpřístupňujeme výsledky automatické analýzy rýmových shod v básnických sbírkách obsažených v Korpusu českého verše a vydaných do roku 1920 (databáze obsahuje přes jeden milion rýmových párů).“ (Plecháč, Gunstick | Versologický tým, 2015)

Po zadání hledaného slova jsou vyhledány rýmy navazující na zadaný výraz. Uživatel se zároveň dozví jednotlivé básně, ve kterých se rým vyskytuje s veršem, ve kterém je obsaženo hledané slovo i veršem následujícím. Nechybí ani grafické znázornění s časovou osou.

5.4.2 Česká elektronická knihovna

„Česká elektronická knihovna (ČEK) zpřístupňuje 1700 básnických knih česky psané poezie 19. a počátku 20. století.“ (Autor neznámý, oficiální web ČEK)

Současná podoba je výsledkem třetího grantového úkolu, projekt ČEK byl tedy dlouhodobě doplňovaný a zpracovávaný.

Z jejich záznamů vychází i Korpus českého verše, který však na svých stránkách uvádí, že ČEK obsahuje i duplicitní záznamy.

Básně zde uvedené lze filtrovat podle autora, místa vydání, nakladatele, autora mota, data vydání nebo podle toho, komu byly věnovány. U všech či pouze vyfiltrovaných básní pak lze zobrazit statistiku (minimální, maximální a průměrná délka verše nebo strofy, jednotlivé počty N-písmenných slov), abecední a frekvenční slovník nebo v nich vyhledávat.

5.4.3 Další projekty ústavu

Bibliografie české literární vědy (od roku 1945)

„Bibliografie české literární vědy zachycuje bohemistickou literárněvědnou produkci od roku 1945 až do současnosti. Excerpční základnu této anotované bibliografie tvoří literárněvědné a literárněkritické publikace (původní i přeložené), odborná literatura dalších společenskovědních oborů s obsahovou vazbou k problematice literatury, předmluvy a doslovy z knih české beletrie, sborníky (vysokoškolské, vlastivědné, z konferencí), specializované literární časopisy, obecné kulturní a společenskovědní časopisy a deníky.“ (Autor neznámý, oficiální web ÚCL, podstránka Oddělení bibliografické a archivní)

Na webových stránkách http://aleph20.lib.cas.cz/ je k dispozici vyhledávání dle množství různých kritérií. Lze používat i logické operátory, vyhledávání části slov *„Například* květ? *vyhledá záznamy, v nichž byla použita slova květina, květinový, květinářský atd.“* (Autor neznámý, oficiální web nástroje, 2011)

Slovník české literatury

„Slovník obsahuje více než tisíc hesel českých spisovatelů a literárních časopisů z období 1945–2000.“ (Autor neznámý, oficiální web slovníku)

Ve slovníku jsou zařazeny informace o spisovatelích, časopisech, dílech a institucích, které mají co dočinění s literaturou zejména mezi lety 1945 a 2000.

Slovník mimo faktografických informací obsahuje i ty bibliografické a jeho základem je především digitalizace již existujících slovníků i další nové aktualizace a doplňování.

5.5 IBM

IBM je jedna z předních světových firem v oboru informačních technologií. Z hlediska text miningu je jejich pole působnosti v oblasti vývoje softwarových nástrojů, které dokáží zpracovávat textová data.

IBM Watson Content Analytics

Nástroj IBM Watson Content Analytics plně podporuje češtinu, jeho hlavním zaměřením nástroje je analýza zpětné vazby zákazníka a sentimentu výpovědí. Dokáže se napojit na webové stránky, diskuzní fóra a sociální sítě, kde analyzuje příspěvky uživatelů a vyhodnocuje jejich obsah.

Neoficiální informace po krátké online konverzaci poskytl Matej Vendrinský (Obchodník IBM pro Českou republiku a Slovensko), ten hovořil o ceně začínající na $80 000 ročně.

„Nejčastější využití IBM Watson Content Analytics

- *Analýza Zpětné vazby zákazníka, analýza sentimentu výpovědí*
- *Detekce podvodů*
- *Predikce odchodu zákazníků*
- *Analýza hovorů v call-centru*

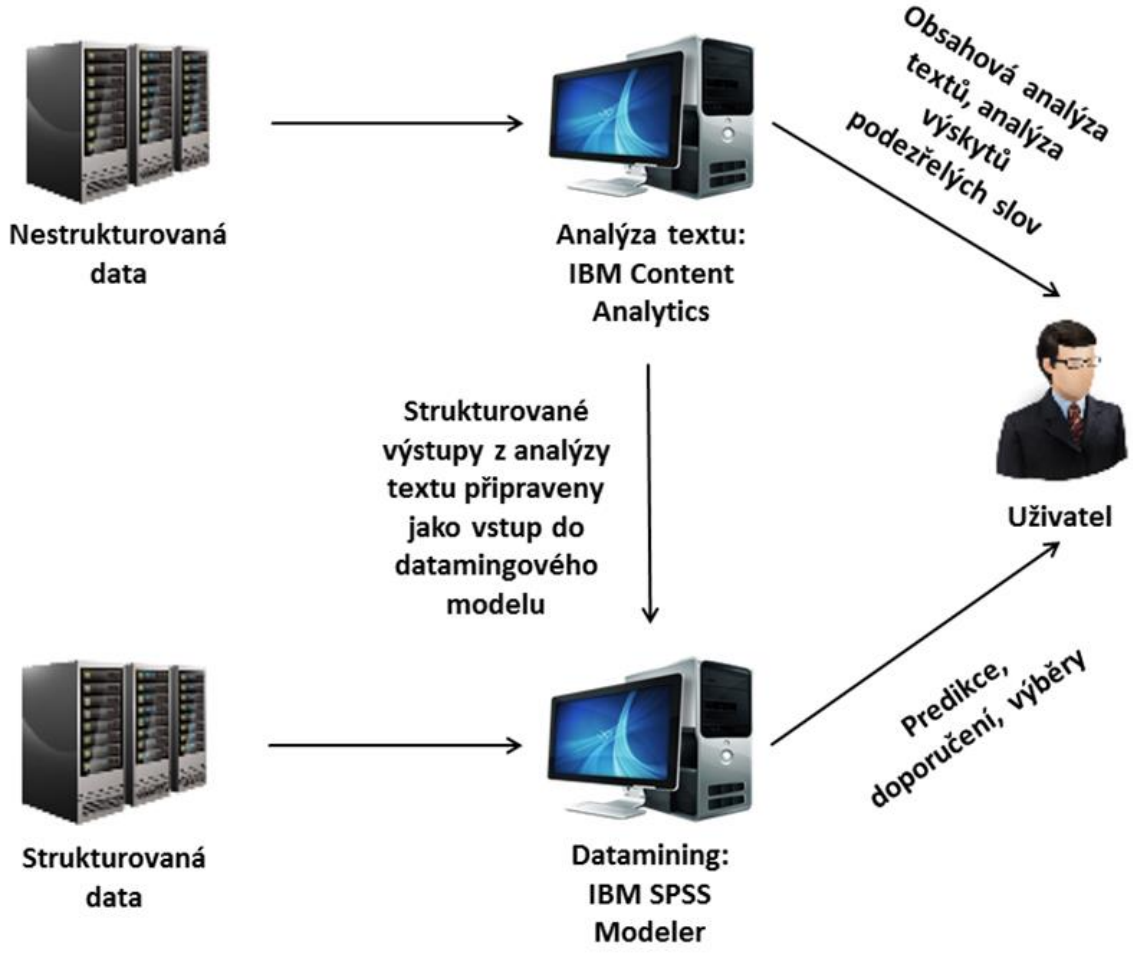
Mezi další úlohy patří:

- *Hledání vhodných kandidátů*
- *Monitorování e-mailové komunikace zaměstnanců*
- *Analýza trhu*
- *Automatická klasifikace dokumentů*
- *Lepší odhad zákaznických požadavků"* (Autor neznámý, oficiální web Acrea)

IBM SPSS Modeler

Program IBM SPSS Modeler spadá do softwarového balíku SPSS Statistics. Informace o cenách těchto programů nejsou veřejné, záleží na konkrétních požadavcích klienta, jako je prostředí, ve kterém software běží, typ licencování nebo případné rozšíření. Neoficiální informace opět od Mateje Vendrinského hovoří řádově o mnohem nižší ceně, než v případě programu Content Analytics. V tomto případě se má cena roční licence pohybovat okolo $2 500 bez DPH.

Ačkoliv se jedná spíše o data miningový nástroj, poslední verze Modeler Premium podporuje také text miningové funkce jako například analýzu sociálních sítí. SPSS Modeler obsahuje vlastní lingvistické knihovny pro angličtinu, holandštinu, němčinu, italštinu, japonštinu, portugalštinu a španělštinu. Čeština mezi podporovanými jazyky tedy chybí.



Obrázek 6: Schéma kombinace analýzy textu a dataminingu
Zdroj: Autor neznámý, oficiální web Acrea

Z uvedeného schématu je patrné, že analýzu nestrukturovaných dat je vhodné provádět pomocí programu IBM Content Analytics, který nám poskytne konkrétní výstupní informace. Za podpory programu SPSS Modeler tyto informace ale můžeme dále zpracovávat a vytvářet různé modely či predikce.

Školení Acrea

Nástroje text miningu jsou určené pro velké množství nestrukturovaných dat různých druhů a typů. Pro účely dokonalého pochopení práce s poměrně složitými aplikacemi jsou vedeny kurzy a školení, jejichž zaměření je zejména pro korporátní klientelu.

Konkrétně kurz firmy Acrea je prováděn s podporou softwaru společnosti IBM, ve kterém se vyučuje práce s výše zmíněné programy Watson Content Analytics a SPSS Modeler.

Firma Acrea je certifikovaným partnerem firmy IBM pro český a slovenský trh. Zabývá se prodejem produktů IBM a poskytováním služeb, jako je i zmíněné školení, k daným programům.

5.6 StatSoft

Tato firma byla založena v USA a nyní je dceřinou společností technologického gigantu Dell. Její česká pobočka (www.statsoft.cz) existuje od roku 1999. Zabývá se především statistikou, správou a analýzou dat či jejich vizualizací.

StatSoft Statistica

Tato aplikace je stěžejní v portfoliu firmy StatSoft, má mnohaletou historii a umožňuje další rozšiřitelnost.

Kromě velkého množství statistických a data miningových nástrojů obsahuje aplikace i text mining a web crawling.

Web crawling, v překladu „lezení po webu“, je nástrojem indexace webových stránek pomocí jejich zdrojových kódů. Uživatel zadá webovou stránku, na které má analýza probíhat, typ dokumentů, který má hledat a hloubku hledání. Hloubka 1 znamená hledání pouze na zadané stránce, hloubka 2 značí hledání na zadané stránce i hledání na podstránkách, na které je z původní stránky odkazováno atd. K zamezení hledání na „cizích“ stránkách je zde funkce k omezení hledání pouze na původně zadané doméně. Výchozí filtr souborů je nastaven na .html a .htm soubory, to lze však změnit na jakýkoliv typ. Například lze tedy vyhledávat pouze dokumenty, které na stránce či podstránkách nacházejí. Díky analýze titulní stránky http://www.uhk.cz zjistíme, že se z ní lze dostat na 105 dalších stránek či podstránek.

Do nástroje text mining lze jako zdroje dat vkládat soubory uložené v počítači nebo jej propojit s web crawlingem a použít tak jako zdroj soubory, které se nacházejí na internetu. Vkládat lze i velké množství souborů, které jsou následně hromadně zpracovány. Text lze očistit o stop slova, sjednotit synonyma a fráze, která mají vystupovat jako jedno slovo. Všechny tyto možnosti se dají volitelně zapnout/vypnout a upravit slovníky, na základě kterých je provedena úprava. Aplikace dále umožňuje další drobné úpravy (délka slov, povolené znaky apod.). Výsledkem je frekvenční analýza s údajem o počtu dokumentů, ve kterých se dané slovo nachází.

Nativní podporu českých textů bohužel aplikace nepřináší – nemá tedy nástroje pro lemmatizování ani stemmer s podporou českého jazyka. České znaky rozeznávat dokáže, ale pouze částečně – znaky s diakritikou zaměňuje za znaky bez ní. (zdroj informací o podpoře češtiny: telefonní kontakt se zástupcem společnosti)

Školení StatSoft

Firma StatSoft nabízí kurz týkající se text miningu. Takto charakterizují svůj kurz na webových stránkách:

„Tento jednodenní kurz je určen těm, kteří se zajímají o možnosti získávání a vyhodnocování informací z textových dokumentů, nejčastěji z webových zdrojů, ale i ze souboru textových dokumentů nebo různých databází. Cílem kurzu je ukázat možnosti jak pracovat s vágními textovými informacemi a jak z nich získat užitečnou informaci pro další zpracování a rozhodování.“ (Autor neznámý, oficiál web firmy StatSoft)

V tomto kurzu, jehož cena činí 6 800 Kč, je vyučována práce se samotným softwarovým nástrojem společnosti – STATISTICA.

5.7 SAS

Firma SAS se zabývá softwarem a službami v business sféře a i pro text mining má vlastní řešení – SAS Text Miner.

SAS Text Miner

SAS Text Miner má podporu celkem 27 jazyků, což je v porovnání s ostatními text miningovými nástroji poměrně velké číslo. Čeština ve výčtu jazyků nechybí.

Vstupní data je možno vkládat ve formátu pdf, HTML, prostý text, formáty používané v Microsoft Office (word, excel apod.) nebo databázové formáty. Zvládá dokonce převod některých proprietárních formátů do vhodnější podoby ke zpracování.

Software umožňuje definovat si vlastní stop slova, která budou z textu odstraněna před jeho dalším zpracováním. Stejně tak lze z textu odstranit (nebo se naopak zaměřit na) speciální textové prvky, kterých je zde předdefinovaných 18. Jedná se o prvky jako adresa, název společnosti, datum, telefonní číslo, čas apod.

Konečné výstupy je možno zobrazit v grafické i v textové (tabulkové) podobě.

Terms

	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
	data	2747	786	☑	0.103	Noun	Alpha
⊟	software	1729	691	☑	0.114	Noun	Alpha
	applications	345	219			Noun	Alpha
	applicaion	1	1			Noun	Alpha
	application	499	264			Noun	Alpha
	softwae	2	2			Noun	Alpha
	software	881	494			Noun	Alpha
	solftware	1	1			Noun	Alpha
⊞	system	1164	565	☑	0.143	Noun	Alpha
⊞	user	645	379	☑	0.19	Noun	Alpha

Obrázek 7: Výsledky filtrování textu

Zdroj: Autor, výstup z programu SAS Text Miner

Na obrázku výše lze vidět výsledky filtrování slov *data*, *software*, *system* a *user* ve velkém množství souborů. Program dokáže sám rozpoznat možné překlepy hledaných slov v dokumentech (*softwae* apod.) a tyto slova také vyfiltruje.

Slovo *application* bylo zařazeno jako synonymum slova *software*, proto se vyskytuje ve výsledcích ve skupině s tímto slovem.

Školení SAS

Na produktu SAS Text Miner probíhá dvoudenní školení, trvající dva dny vždy 8 hodin za cenu 13.870,- Kč resp. 5.605,- Kč pro akademická pracoviště.

„Po absolvování kurzu by měl účastník být schopen:

- *zpracovat a připravit textová data pro analýzu*
- *převádět nestrukturovaná textová data na strukturovaná číselná data*
- *prozkoumávat fráze ve sbírkách dokumentů a třídit dokumenty do tematických skupin podle kritérií podobnosti*
- *hledat dokumenty, které jsou nejblíže asociovány se zadaným slovem nebo frází, a slova/fráze, které jsou nejblíže asociovány se zadaným dokumentem*
- *používat textová data k vylepšení predikčních modelů“* (Autor neznámý, oficiální web firmy SAS)

Obsahem kurzu je tedy zpracování, transformace a konverze dat, analýza sad dokumentů a prediktivní modelování.

5.8 Masarykova univerzita

Na akademické půdě se text miningem do hloubky zabývá Centrum zpracování přirozeného jazyka (CZPj) na Fakultě informatiky Masarykovy univerzity v Brně.

CZPj se zabývá teorií i aplikovanými výsledky zpracování přirozeného jazyka v těchto oblastech:

- Korpusy
- Slovníky
- Morfologie
- Syntaktická analýza
- Sémantika

5.8.1 Vybrané nástroje

Níže jsou uvedeny vybrané nástroje, které jsou online dostupné a byly autorem otestovány.

Detekce témat - TO|P|icks

Nástroj sloužící k vyhledání stěžejních témat (klíčových slov) v textech. Jeho webové demo nabízí možnost vyhledat témata v textu dlouhém až 10 000 znaků. Aplikace je dostupná na https://nlp.fi.muni.cz/projekty/topicks/index.py.

Ohákování – CZ accent

Nástroj ohákování dokáže text zadaný bez diakritiky převést na text s diakritikou. Aplikace je dostupná na http://nlp.fi.muni.cz/cz_accent/. Diakritiku doplňuje na správných místech, ale kvůli četnosti používání diakritických znamének v českém jazyce může vznikat také řada chyb.

Problémy vznikají u slov, které mají různý význam s diakritikou a bez ní, nemusí se přitom jednat o slova v základním tvaru, ale mohou být různě vyskloňovaná či vyčasovaná.

Vstupní text	Výstupní text	Míněný význam
citelny text	citelný text	čitelný text
citelny prohresek	citelný prohřešek	citelný prohřešek
citelne zraneni	citelně zranění	citelné zranění
citelna ztrata	čitelná ztráta	citelná ztráta
piste citelne	pište citelně	pište čitelně
stafetovy kolik	štafetový kolik	štafetový kolík
dostal koliku	dostal kolíku	dostal koliku

Tabulka 1: Testovací data v nástroji Ohákování

Zdroj: Autor

Na uvedených příkladech lze demonstrovat úskalí „ohákování“ textu. Řešením tohoto problému by mohla být změna stylu doplňování diakritiku – nedívat se na každé slovo zvlášť, ale chápat je jako sousloví. Pokud by tedy za zadaným slovem „citelny“ následovalo slovo „text“ tak víme, že se bude jednat o slovo „čitelný“. Avšak ani touto cestou by se nedala zajistit 100% správnost doplnění diakritiky.

Ajka (Majka)

Ajka je morfologický analyzátor češtiny, který vznikl již v roce 1999 jako součást diplomové práce. Je dostupný přes webové rozhraní na stránce http://nlp.fi.muni.cz/projekty/wwwajka/WwwAjkaSkripty/morph.cgi. Jeho funkcí je rozdělit zadaný text na jednotlivá slova a každé morfologicky analyzovat. U každého slova (pokud je tak možno) vypíše seznam možných zařazení do kategorií ve formě morfologického označení. Každé toto označení lze dále rozkliknout
a zobrazit jeho textovou podobu.

Pro správnou interpretaci a orientaci ve výsledcích je nutné znát významy morfologických značek a morfologického názvosloví. Popis značení, dokumentace k aplikaci i kompletní diplomová práce jsou na uvedené adrese k dispozici.

Alternativou k nástroji ajka je majka. Zatímco ajka již není udržována, majku její autor upravuje a vylepšuje. Opravuje chyby předchozího nástroje, je 7x rychlejší a nabízí určitá vylepšení.

K dispozici je zdrojový kód majky s morfologickou databází pro češtinu a 14 dalších cizích jazyků.

5.9 Západočeská univerzita

Na Západočeské univerzitě v Plzni působí pod Fakultou aplikovaných věd výzkumná skupina zabývající se text miningem (Text-Mining Research Group, TMRG).

Jejich dlouhodobým cílem je *„vytvořit robustní systém, který extrahuje znalosti z částečně strukturovaných dat ve vícejazyčném internetovém prostředí, aby takto byly získány nové informace/znalosti, které v původních datech nejsou explicitně obsaženy.“* (Autor neznámý, oficiální web TMRG)

Vydali několik publikací a článků zabývajících se například filtrováním závadného webového obsahu, detekcí plagiátorství, shrnutím dokumentu, jeho klasifikací nebo získávání informací o uživatelích prostřednictvím sociálních sítí.

5.10 Karlova univerzita

Matematicko-fyzikální fakulta

Pod Matematicko-fyzikální fakultou Karlovy univerzity v Praze existuje Institut formální a aplikované lingvistiky, který byl založen v roce 1990.

V institutu pracovali a pracují na mnoha projektech, týkajících se zpracování textu a jazyka. V projektech se často zabývají morfologií, strojovým překladem, vytvořili Český akademický korpus.

Filozofická fakulta - Studia nových médií

Pod Filozofickou fakultou běží magisterský obor Studia nových médií, kde realizovali projekt *Párování pracovních míst s nezaměstnanými skrze sémantická data.*

Na webu http://www.damepraci.cz/ je dostupný vyhledávač pracovních pozic, do kterého lze zadat název pozice, místo výkonu práce nebo co uživatel „umí,

zvládá, baví ho“. Poslední zmíněná kolonka dokáže pracovat se synonymy a prohledá shodu v nabízených pozicích.

Vedoucím oboru je Mgr. Josef Šlerka (ředitel pro výzkum a vývoj Socialbakers), který na svém webu http://databoutique.cz/ pravidelně přidává příspěvky založené na reálných data miningových a text miningových analýzách sociálních sítí. Jeho jméno je spjato i se službou Social Insider (bod 5.12 této práce).

Ústav teoretické a komputační lingvistiky

Ústav teoretické a komputační lingvistiky (ÚTKL) se zabývá počítačovým zpracováním českého jazyka a vytvářením nástrojů pro jeho zpracování.

Jejich hlavním přínosem je podíl na vzniku Českého národního korpusu i dalších nástrojů.

Poziční tagy

Jedním z nástrojů jsou Poziční morfologické tagy, dostupné na http://utkl.ff.cuni.cz/~skoumal/morfo/. Jedná se v podstatě o podpůrnou aplikaci Českého národního korpusu, která interaktivně pomůže uživatelům vyznat se v morfologických značkách, které se v ČNK používají a lépe porozumět slovu, které je analyzované.

Devulgarizátor

Dalším nástrojem, který není tak vědecky zaměřen a má spíše zábavný charakter je Devulgarizátor dostupný na http://utkl.ff.cuni.cz/devulgarizator/. Jeho hlavní (a jedinou) funkcí je nahrazení nevhodných (neslušných a nespisovných) výrazů těmi vhodnými. Jeho použití je však diskutabilní, a i výsledky procesu „devulgarizace“ jsou často neúplné.

Tento nástroj však demonstruje, že použití text miningových nástrojů může být velmi široké, záleží vždy na kreativitě autora. Pochopitelně také na tom, zda má nástroj mít nějaký hlubší smysl či slouží pouze pro zajímavost či pobavení.

5.11 Voyant-tools

Voyant-tools je online nástroj dostupný na webové adrese http://voyant-tools.org/.

Obsahuje sadu funkcí ke zkoumání vloženého nebo nahraného textu.

Obrázek 8: Grafické zobrazení frekvenční četnosti slov
Zdroj: Autor

Voyant-tools v základu nenabízí seznam českých stop slov, ale lze jej jednoduše vytvořit a aplikovat na vložený text. Na obrázku výše je jedna z funkcí nazývající se word cloud. Jedná se o grafické znázornění slov dle jejich frekvenčního výskytu v textu. Jako zdrojový text sloužila seminární práce o systémech pro podporu rozhodování (DSS), což je patrné pokud obrázek začneme číst od slov psaných největším fontem.

Celá tato webová aplikace je vhodná zejména k porovnávání několika vložených textů.

Do nástroje bylo nahráno sedm literárních děl Karla Čapka (Kniha apokryfů, Krakatit, Povídky z jedné kapsy, Povídky z druhé kapsy, První parta, Továrna na absolutno a Válka s mloky) ve formátu .pdf a zanalyzováno.

Jednou ze zjišťovaných a porovnávaných hodnot je úroveň slovní zásoby. Úroveň slovní zásoby měří nástroj Voyant-tools takto (Sinclair, 2015):

$$slovní\ zásoba = \frac{počet\ unikátních\ slov}{celkový\ počet\ slov} * 1000$$

Čím vyšší číslo vyjde, tím je slovní zásoba díla (textu) na vyšší úrovni.

Nejvyšší slovní zásoba z uvedených děl je v románu Továrna na absolutno (330.8), nejnižší pak v díle První parta (189.9). Takto obrovský rozdíl je dán zejména dvěma faktory.

1. V román Továrna na absolutno vystupuje vysokoškolsky vzdělaný vynálezce motorů, který přináší blahobyt pro lidstvo. Pivní parta pojednává o skupině horníků, kteří se snaží zachránit jiné, zasypané horníky. Již z těchto nástinů děje je patrné, že autor musel použít jinou úroveň jazyka a celkovou slovní zásobu.
2. Dílo Továrna na absolutno má 36 500 tokenů, zatímco První parta obsahuje přes 50 000 tokenů. V kratším díle je přitom snazší mít více unikátních slov, jelikož i celková slovní zásoba (českého) jazyka má omezený počet slov. I přes to je v absolutní počtu unikátních slov vítězem Továrna na absolutno (12 000 oproti 9 500), což potvrzuje důležitost literárních vlastností daného díla (prostředí, délka děje apod.).

Voyant-tools zjistí i příznačná slova pro každé dílo. Jedná se o slova, která se v jednom díle vyskytují mnohem častěji, než v ostatních, a tím jsou pro dané dílo jednoznačně určující. V následujícím výčtu je vždy zobrazeno dílo, za kterým následuje pět příznačných slov.

- Kniha apokryfů: bych, děl, hamlet, já, císař
- Krakatit: prokop, mu, carson, princezna, prokopa
- Povídky z jedné kapsy: pan, pane, já, řekl, mně
- Povídky z druhé kapsy: já, mně, paní, té, pane
- První parta: standa, pepek, adam, tam, jo
- Továrna na absolutno: bondy, marek, absolutno, prezident, pan
- Válka s Mloky: mloci, kapitán, mloků, mloky, abe

Nástroj nepodporuje lemmatizaci, proto jsou zobrazena slova vždy v původním znění.

Další funkcí je grafické zobrazení četnosti užití slova (či porovnání více slov) napříč analyzovanými díly nebo zobrazení levého i pravého kontextu ke slovu.

5.12 Social Insider

Social Insider je původně česká společnost, nyní vlastněná firmou Socialbakers. Nástroj Social Insider, dostupný na https://www.socialinsider.cz/, je jedinou aplikací, která zvládá analyzovat sentiment v českém jazyce.

Online přístupná aplikace monitoruje dění a komentáře na sociálních sítích (Facebook, Twitter, Google Plus), českých a slovenských blozích a novinových serverech, ze kterých dále analyzuje data.

V demo verzi dostupné pro veřejnost není nástroj analýza sentimentu dostupný, a proto jej nebylo možnost vyzkoušet.

5.13 Granty

Databáze státních grantů je dostupná na webových stránkách Fondu rozvoje vysokých škol (FRVŠ), který získává finanční prostředky z Ministerstva školství, mládeže a tělovýchovy. Od roku 2006 jsou veškeré schválené i neschválené žádosti o finanční prostředky vystaveny na oficiálních stránkách FRVŠ - http://www.frvs.cz/.

Další udělování grantů probíhá prostřednictvím Grantové agentury České republiky, jejichž výsledky jsou pak dostupné na stránkách http://www.isvav.cz/. Veškeré granty udělené Ústavu pro českou literaturu AV ČR byly uděleny právě prostřednictvím tohoto poskytovatele.

Udělené příspěvky

Zřejmě nejvyšší dotace na jazykovědné účely spojené s počítačovým zpracováním byly uvolněny pro tvorbu, úpravu a udržování Českého národního korpusu. Grantový projekt běží od roku 1999 prozatím do roku 2016 a je rozdělen na tři etapy. V té první, mezi lety 1999 a 2004, bylo ze státního rozpočtu přiděleno 23 212 000 Kč. Druhá etapa trvající o rok déle, tedy do roku 2011 byla o poznání štědřejší – přinesla 88 488 000 Kč. Na tuto navazuje další, trvající do roku 2016 s celkovým rozpočtem ze státu ve výši 110 000 000 Kč. Cílem je udržení bezplatné a volně přístupné internetové databáze, budování jazykových korpusů. Zmiňuje také, že ČNK se řadí k předním korpusovým pracovištím ve světovém měřítku.

V roce 2005 byly uděleny dvě dotace pro Západočeskou univerzitu. Za jejich podpory byla zpracována témata „Extracting Information from Web Content and Structure." Tato práce se zabývala získáváním dat z webových stránek. V první části probíhá detekcí pornografických webových stránek, které chceme z hledání odstranit. Druhá část se zabývá analýzou vztahů webových stránek o počítačových technologiích v akademické sféře.

Druhý projekt „Searching and Summarizing in Multilingual Enviroment" se zabýval vytvářením systému, který by umožňoval multilingvální vyhledávání v textových databázích a automatické shrnutí vyhledaného textu.

Velké množství grantů k výzkumné činnosti v lingvistickém oboru a i k vývoji nástrojů text miningu putuje na Karlovu Univerzitu, jejich kompletní seznam k dispozici na http://ufal.mff.cuni.cz/grants.

Digitalizace a tvorba Korpusu českého verše (bod 5.4 této práce) byla umožněna díky projektu započatému v roce 2011 a trvajícím do konce roku 2015. Tento projekt byl celkem dotován částkou 6 754 000 Kč, byly díky němu zmapovány dějiny českého verše a vytvořeny webové stránky s několika aplikacemi spolupracujícími s těmito daty.

Česká elektronická knihovna (bod 5.5 této práce) byla také podporována státními dotacemi, kdy nejprve obdrželi grant na tři roky (1998–2000) a poté na čtyři roky (2001–2004). Mezi lety 2004 a 2008 byl za podpory Ministerstva školství, mládeže a tělovýchovy realizován projekt „Digitalizace

knižního katalogu Ústavu pro českou literaturu a zpřístupnění databáze osobností české literatury“, který dovedl ČEK do současné podoby. Výše podpory ze státního rozpočtu činila 5 406 000 Kč, celkové uznané náklady pak byly 10 996 000 Kč.

Slovník české literatury (bod 5.6 této práce) vznikl za podpory dvou grantových dotací. Ta první započala v roce 2006 a trvala do roku 2008. Poskytovatelem byla Grantová agentura České republiky a celková výše podpory ze státního rozpočtu činila 1 897 000 Kč. Druhá grantová podpora od stejného poskytovatele probíhala mezi lety 2009 a 2013. Jejím cílem bylo doplnění hesel literárních institucí. Celkové náklady se vyšplhaly na 3 602 000 Kč. (zdrojem výše uvedených informací jsou veřejně přístupná data ze stránek ISVAV, 2015)

6 Shrnutí výsledků

V praktické části práce bylo zjištěno, jaké instituce se text miningem zabývají a do jaké míry. Na dolování dat z textu se v České republice zaměřuje řada akademických pracovišť. Zejména se jím pak zabývá Masarykova univerzita v Brně, kde vznikla řada text miningových nástrojů, Univerzita Karlova v Praze, která zaštiťuje Český národní korpus nebo Západočeská univerzita v Plzni. Dalším akademickým zástupcem je Ústav pro českou literaturu AV ČR. Na všech těchto pracovištích se zabývají dolováním dat na vědecké bázi ale i po stránce praktické, o čemž svědčí množství nástrojů, které vznikají a mají podporu českého jazyka. Řada projektů je financována ze státních zdrojů díky grantovým podporám a pochopitelně podpoře příslušného ústavu.

Velcí hráči na poli vývoje software se text miningem také zabývají. V práci jsou zmíněny firmy IBM, Statsoft a SAS. Jejich produkty češtinu podporují (ať už částečně nebo plně) a poskytují školení pro práci s nimi (v případě firmy IBM poskytuje školení jejich obchodní partner, zbylé dvě poskytují školení přímo).

Za zmínku stojí i nástroj Social Insider, který jako jediný na trhu má podporu českého jazyka v oboru analýza sentimentu zákazníků nebo uživatelů. Dalším zajímavou službou je pak Voyant-tools z rukou kanadských vysokoškolských profesorů. Jejich aplikace nabízí mnoho zajímavých možností analýz a výstupů, které lze použít i pro česky psané texty.

Následující tabulka uvádí přehled všech důležitých nástrojů, resp. center a ústavů zaštiťujících nástroje, zmíněných v této práci a jejich spojení s mateřskými institucemi nebo jinými organizacemi.

Název	Podpora češtiny	Typ instituce	Spojení
Český národní korpus	ano	akademická	Univerzita Karlova
Korpus DIALOG	ano	akademická	Ústav pro jazyk český AV ČR
Slovenský národný korpus	ano	akademická	Jazykovedný ústav Ľ. Štúra SAV
Korpus českého verše	ano	akademická	Ústav pro českou literaturu AV ČR
Česká elektronická knihovna	ano	akademická	Ústav pro českou literaturu AV ČR
IBM Watson Content Analytics	ano	komerční	Acrea (školení)
IBM SPSS Modeler	ne	komerční	Acrea (školení)
StatSoft Statistika	částečně	komerční	StaSoft
SAS Text Miner	ano	komerční	SAS
Centrum zpracování přirozeného jazyka	ano	akademická	Masarykova univerzita
Text-Mining Research Group	ano	akademická	Západočeská univerzita
Ústav teoretické a komputační lingvistiky	ano	akademická	Karlova Univerzita
Voyant-tools	částečně	akademická	University of Alberta; McGill University
Social Insider	ano	komerční	Socialbakers

Tabulka 2: Přehled v práci zmíněných nástrojů a institucí

Zdroj: Autor

Tučně označeny jsou aplikace a instituce, resp. jejich nástroje, se kterými přišel autor do styku, a byly jím osobně vyzkoušeny. Důvod nevyzkoušení všech aplikací bylo omezení z pohledu komerční licence (nedostupnost trial verze) nebo nedostupnost aplikací pro veřejnost mimo dané akademické působiště.

7 Závěry a doporučení

Účelem první části této práce je shrnout teoretické poznatky o text miningu a seznámit s nimi čtenáře. Práce se snaží nastínit postup strojového zpracování textu, který začíná přípravou zpracovávaného dokumentu a pokračuje volbou konkrétního analytického úkonu.

Druhá část práce se snaží poskytnout co nejširší informace o institucích, které se text miningem zabývají i s dostatečně podrobným popisem každé z nich. Zaměřuje se nejen na akademická pracoviště, ale i obchodní společnosti či jiné nástroje týkající se tohoto oboru.

Bylo zjištěno, že úroveň zpracování dat v českém jazyce je na poměrně vysoké úrovni. Existuje totiž množství korpusů, databází a nástrojů pro jejich aplikaci na českojazyčné texty. Možnosti zpracování, následná analýza a její výsledky byly také ukázány na některých praktických příkladech. K tomu byly použity nástroje, které jsou v práci zmíněny.

Práce je odrazem aktuálního stavu mezi českými institucemi a softwarovými nástroji. Některé popsané nástroje jsou stále ve vývoji a je tedy možné, že neduhy, které je dnes trápí, mohou být v dohledné době odstraněny. I zmíněné nástroje firem IBM, StatSoft a SAS jsou aktualizovány a mohou přinášet nové funkcionality, rozšíření nebo opravy stávajícího řešení. Ve starších verzích například češtinu vůbec nepodporovaly, lze tedy předpokládat, že jejich vývoj není zcela jistě ukončen a určitá vylepšení budou postupně přicházet.

Nadstavbou či rozšířením této práce by mohly být informace, které české firmy nebo instituce používají nástroje text miningu. Bylo by zajímavé zjistit, zda se podniky spoléhají na softwarová řešení, v jakých konkrétních situacích je využívají a jaké mají výsledky. Vhodné by bylo i porovnání nákladů na lidské (ruční) procházení dat a počítačové zpracování s částečnou asistencí člověka a srovnání výsledků těchto dvou metod.

Druhou možností rozšíření může být zaměření se na konkrétní analytický úkon text miningu a porovnání výsledků z jednotlivých aplikací a nástrojů. Tím by

mohlo být zjištěno, který nástroj dosahuje v jakém oboru nejlepších výsledků a zkoumáno, z jaké příčiny tomu tak je.

Vzhledem k vysokým cenám zmíněných softwarových aplikací od firem IBM, StatSoft a SAS můžeme očekávat jejich použití zejména ve velkých firmách a podnicích, pro které bude jejich využití rentabilní. Příkladem mohou být nadnárodní řetězce, banky nebo telekomunikační firmy.

8 Seznam použité literatury

Al-Ayyoub, M. (25. Duben 2006). *Text mining.* Načteno z Stony Brook University, New York: http://www.cs.sunysb.edu/~cse634/presentations/TextMining.pdf

Bridgwater, A. (26. Říjen 2010). *IBM: 80 percent of our global data is unstructured (so what do we do?).* Načteno z The Computer Weekly Application Developer Network: http://www.computerweekly.com/blogs/cwdn/2010/10/ibm-80-percent-of-data-is-unstructured-so-what-do-we-do.html

Cvrček, V. (13. září 2013). *pojmy:morfologicka_analyza - Příručka ČNK*. Načteno z Příručka ČNK: http://wiki.korpus.cz/doku.php/pojmy:morfologicka_analyza

Cvrček, V. (13. září 2013). *pojmy:synchronni - Příručka ČNK*. Načteno z Příručka ČNK: http://wiki.korpus.cz/doku.php/pojmy:synchronni

Cvrček, V. (13. září 2013). *pojmy:word - Příručka ČNK*. Načteno z Příručka ČNK: http://wiki.korpus.cz/doku.php/pojmy:word

Cvrček, V. (24. listopad 2014). *pojmy:desambiguace - Příručka ČNK*. Načteno z Příručka ČNK: https://wiki.korpus.cz/doku.php/pojmy:desambiguace

Cvrček, V. (24. listopad 2014). *pojmy:token - Příručka ČNK*. Načteno z Příručka ČNK: http://wiki.korpus.cz/doku.php/pojmy:token

Cvrček, V. (23. leden 2015). *cnk:oral2013*. Načteno z Příručka ČNK: http://wiki.korpus.cz/doku.php/cnk:oral2013

Cvrček, V. (17. únor 2015). *cnk:syn2010*. Načteno z Příručka ČNK: http://wiki.korpus.cz/doku.php/cnk:syn2010

Čepek, M. (5. Listopad 2013). *Vytěžování dat - Texmining.* Načteno z ČVUT v Praze: https://cw.felk.cvut.cz/wiki/_media/courses/a7b36vyd/cviceni/2013/cviceni/cv07.pdf

Čermák, F., & Kocek, J. (nedatováno). *Co je korpus?* Načteno z Ústav Českého národního korpusu: https://ucnk.ff.cuni.cz/co_je_korpus.php

Grobelnik, M., & Mladenic, D. (8. Leden 2007). *Tutorial on Text Mining and.* Načteno z Artificial Intelligence Laboratory: http://ailab.ijs.si/dunja/TextWebJSI/TMLATutorial-IJCAI2007-8Jan2007.pdf

Hearst, M. A. (26. Červen 1999). *Untangling Text Data Mining.* Načteno z University of California, Berkeley: http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

Ježek, K., & Steinberger, J. (1. Říjen 2010). *Sumarizace textů.* Načteno z Západočeská univerzita: http://textmining.zcu.cz/publications/SumarizDATAKON.pdf

Loria, S. (1. Září 2013). *Tutorial: Finding Important Words in Text Using TF-IDF.* Načteno z Steven Loria: http://stevenloria.com/finding-important-words-in-a-document-using-tf-idf/

Neznámý. (2015). *Veřejně přístupná data IS VaVaI*. Načteno z Veřejně přístupná data IS VaVaI: http://www.isvav.cz/

Neznámý, A. (nedatováno). *IBM Watson Content Analytics | ACREA | Analytická kreativita v business analytics*. Načteno z ACREA | Analytická kreativita v business analytics: http://www.acrea.cz/cz/software/ibm-content-analytics

Neznámý, B. (2011). *Souborný katalog - základní vyhledávání*. Načteno z Knihovna Akademie věd ČR: http://aleph20.lib.cas.cz

Neznámý, Č. (nedatováno). *Česká elektronická knihovna*. Načteno z Česká elektronická knihovna: http://www.ceska-poezie.cz

Neznámý, D. (nedatováno). *Úvod | Korpus DIALOG*. Načteno z Korpus DIALOG: http://ujc.dialogy.cz/

Neznámý, S. (nedatováno). *kurz Text Mining - Služby a kurzy - Data mining - Prediktivní modelování - Credit Scoring - Fraud detection - SPC*. Načteno z Statistická analýza dat - Data mining - Prediktivní modelování - Credit Scoring - Fraud detection - SPC: http://www.statsoft.cz/sluzby/1-kurzy-skoleni/5-nabizene-kurzy/kurz-text-mining/

Neznámý, S. (nedatováno). *Slovník české literatury po roce 1945*. Načteno z Slovník české literatury po roce 1945: http://www.slovnikceskeliteratury.cz/

Neznámý, S. (nedatováno). *SyD: Korpusový průzkum variant*. Načteno z Český národní korpus: https://syd.korpus.cz/

Neznámý, S. (nedatováno). *Školení | Přehled školení SAS v ČR*. Načteno z Školení | Přehled školení SAS v ČR: http://www.sas.com/offices/europe/czech/training/dmtm9.html

Neznámý, T. (nedatováno). *Text-mining Research Group - About us*. Načteno z Text-mining Research Group: http://textmining.zcu.cz/?lang=en&section=profil

Neznámý, U. (nedatováno). *Ústav pro českou literaturu AV ČR*. Načteno z Ústav pro českou literaturu AV ČR: http://www.ucl.cas.cz/

Neznámý, U. (nedatováno). *Ústav pro českou literaturu, Oddělení bibliografické a archivní*. Načteno z Ústav pro českou litaraturu: http://www.ucl.cas.cz/cs/oddeleni/stredisko-literarnevednych-informaci/oddeleni-bibliograficke-a-archivni

Pavel, P. (2006). *Data Mining.* Pardubice: Univerzita Pardubice.

Plecháč, P. (2015). *Frekvenční slovníky | Versologický tým*. Načteno z Korpus českého verše: http://www.versologie.cz/slovniky/slovniky.html

Plecháč, P. (2015). *Gunstick | Versologický tým*. Načteno z Korpus českého verše: http://www.versologie.cz/gunstick/

Plecháč, P. (2015). *Korpus českého verše | Versologický tým*. Načteno z Korpus českého verše: http://www.versologie.cz/kcv.html

Rambocas, M., & Gama, J. (duben 2013). *Marketing Research: The Role of Sentiment Analysis.* Načteno z Faculdade de Economia da Universidade do Porto: http://www.fep.up.pt/investigacao/workingpapers/wp489.pdf

Sebastiani, F. (2002). *Machine learning in automated text categorization.* New York, NY, USA: ACM Computing Surveys.

Sedláček, P. (30. Leden 2004). *Text mining a jeho možnosti (aplikace).* Načteno z Fakulta informatiky Masarykovy univerzity: http://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm

Sinclair, S., & Rockwell, G. (březen 2015). *Voyant Tools: Reveal Your Texts*. Načteno z Voyant Tools: Reveal Your Texts.

Velíšek, Z. (5. leden 2010). *Inverse Text Sort (Czech)*. Načteno z Ústav českého národního korpusu: http://ucnk.ff.cuni.cz/inversesort/Cz.htm

Vysloužilová, L. (22. Květen 2014). *Textová data a dobývání znalostí.* Načteno z ČVUT v Praze: https://cw.felk.cvut.cz/wiki/_media/courses/xp33ppd/ppd2014-10-text_mining.pdf

Whelan, K. (20. Prosinec 2013). *A Stylometric Analysis of Shakespeare's Drama.* Načteno z Textua-Lit-y: A Creative Rea-Lit-y: http://kirstyhw.wordpress.com/2013/12/20/en3006-essay-a-stylometric-analysis-of-shakespeares-drama/

Witten, H. I. (29. Listopad 2004). *Text mining.* Načteno z Computer Science Department, University of Waikato: http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

FIM UHK

UNIVERZITA HRADEC KRÁLOVÉ
Fakulta informatiky a managementu
Rokitanského 62, 500 03 Hradec Králové, tel: 493 331 111, fax: 493 332 235

Zadání k závěrečné práci

Jméno a příjmení studenta: **Filip Bidlo**

Obor studia: Informační management (3)
Jméno a příjmení vedoucího práce: **Jiří Haviger**

Název práce:
Textmining v českém prostředí

Název práce v AJ:
Textmining in Czech Language

Podtitul práce:

Podtitul práce v AJ:

Cíl práce: Cílem práce je zjistit úroveň dostupných prostředků a nástrojů textminingu v českém jazyce.

Osnova práce:
1) Úvod

2) Teoretická část - problémy řešené textminingem

3) Praktická část - instituce zabývající se textminingem v ČR

4) Závěr

Projednáno dne: 29.10.2014

Podpis studenta

Podpis vedoucího práce