Univerzita Hradec Králové

Fakulta informatiky a managementu

Katedra informatiky a kvantitativních metod

Data mining a možnosti nekomerčního softwaru

Bakalářská práce

Autor: Andrea Kubcová
Studijní obor: Aplikovaná informatika

Vedoucí práce: prof. RNDr. Hana Skalská, CSc.

Náchod listopad, 2020

Prohlášení:

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

V Náchodě dne 15.11.2020 Andrea Kubcová

Poděkování:

Děkuji vedoucí bakalářské práce prof. RNDr. Haně Skalské, CSc. za metodické vedení práce. Dále děkuji své rodině, která mi poskytla dostatek času a podpory pro dokončení práce.

Anotace

Bakalářská práce se zabývá problematikou využití nekomerčního softwaru pro data mining. V části teoretické shrnuje základní poznatky z oblasti data miningu a věnuje se metodikám, které se využívají v procesech dobývání znalostí z databází. Zaměřuje se na typy úloh, které data mining řeší, popisuje některé statistické metody a metody strojového učení s učitelem i bez učitele. V části výzkumné představuje práce pět volně dostupných nekomerčních softwarů, které jsou vhodné pro řešení data miningových úloh. Pomocí vybraných funkcionalit tyto softwarové nástroje porovnává. Bakalářská práce detailněji představuje možnosti softwaru Orange, pomocí něhož je řešena ukázková úloha, která je zaměřena na oblast text mining a analýzy sentimentu.

Klíčová slova: data mining, nekomerční software, Orange, text mining, analýza sentimentu

Annotation

Title: Data mining and freeware possibilities

The bachelor thesis deals with the use of non-commercial software for data mining. The theoretical part summarizes the basic knowledge in the field of data mining, it deals with methodologies that are used in the processes of knowledge discovery in databases. It focuses on the types of tasks that data mining solves, describes some statistical methods and some supervised and unsupervised machine learning methods. In the research part, the thesis presents five freely available non-commercial software tools suitable for solving data mining tasks. It compares these software tools using selected functionalities. In more detailed way the bachelor thesis presents the possibilities of the Orange software, which is used to solve an example task. Task is focused on text mining and sentiment analysis.

Key words: data mining, non-commercial software, Orange, text mining, sentiment analysis

Obsah

1 Úvod

Tato bakalářská práce se zaobírá problematikou nekomerčního softwaru, který lze využít k data miningu. Zaměřuje se na pět volně dostupných nástrojů, jejich popis a vzájemné porovnání pomocí vybraných charakteristik. Dále se věnuje detailnějšímu popisu jednoho z vybraných softwarů, s jehož pomocí je následně řešena ukázková data mining úloha.

Přehled hlavních informačních zdrojů, z nichž tato práce čerpá, je krátce představen v kapitole 2. Kapitola 3 přibližuje cíle a metodiky práce. V kapitole následující jsou shrnuty základní teoretické poznatky o data miningu, metodikách, základních úlohách a metodách. Pět vybraných softwarových nástrojů, které jsou vhodné pro řešení úloh data miningu, je představeno v kapitole 5. Závěr této části se věnuje porovnání softwarů na základě vybraných charakteristik. Kapitola 6 detailněji představuje nekomerční nástroj Orange. V poslední kapitole je pomocí toho softwaru řešena ukázková data mining úloha z podoblastí text mining a analýzy sentimentu.

2 Literární rešerše

V teoretické části této bakalářské práce jsou nejvíce využívány tři základní informační prameny.

Práce Dobývání znalostí z databází od Berky (1) shrnuje základní teoretické poznatky o dobývání znalostí z databází a objasňuje vybrané statistické metody a metody strojového učení, které se využívají pro analýzu dat v data miningu. Z tohoto zdroje je v teoretické části bakalářské práce čerpáno nejvíce, hlavně z kapitol, které se týkají dobývání znalosti z databází, a kde jsou popisovány konkrétní techniky pro modelování.

Skalská (2) se ve své práci Data mining a klasifikační modely zaměřuje také na vysvětlení základních pojmů z oblasti data miningu, ale jeho hlavní část objasňuje konkrétní metody, které se využívají v klasifikačních úlohách a k hodnocení modelu binární klasifikace pomocí ROC křivek. Tento pramen byl využíván k ověření a objasnění některých teoretických pojmů v oblasti data miningu.

Witten, Frank a Hall (3) se v publikaci Data mining: practical machine learning tools and techniques zaměřují na obsáhlý popis vybraných analytických metod strojového učení, které se využívají v data miningu. Některé z metod jsou demonstrovány na praktických příkladech. V bakalářské práci byl tento zdroj využíván v souvislosti s popisem analytických metod.

Ve výzkumné části bakalářské práce byly pro vypracování přehledu nekomerčního softwaru a jeho následného porovnání používány on-line dokumentace jednotlivých nástrojů (4), (5), (6), (7) a (8). Všechny tyto on-line zdroje byly využívány pro získání základních informací o softwarových nástrojích a také pro ověřování funkcionalit, které byly nalezeny přímým průzkumem grafického uživatelského prostředí jednotlivých programů. Pro výběr a stanovení porovnávaných charakteristik softwarových řešení byl použit příspěvek An overview of free software tools for general data mining (9). Katalog widgetů na webové stránce softwarového nástroje Orange (10) sloužil pro nalezení detailních informací o rozšířených funkcionalitách tohoto softwaru.

3 Cíl práce a metodika

Cílem bakalářské práce je vyhledat, popsat a porovnat nekomerční softwarové nástroje, které lze využít pro data mining. Pro jeden zvolený nekomerční nástroj pak provést podrobnější zhodnocení jeho možností a řešit jeho pomocí ukázkovou úlohu.

Práce se zaměřuje na porovnání pěti nekomerčních softwarových nástrojů, ty byly vybrány na základě třech kritérií. Hlavním kritériem pro výběr byla komplexnost funkcionalit v oblasti data miningu, která by měla umožnit řešení však základních data miningových úloh. Vybrané softwary tedy nejsou úzce specializované na jednu konkrétní oblast typů úloh. Dalšími kritérii pro výběr bylo představit možnosti různých technologických řešení a odlišnosti ve způsobu modelování. Z tohoto důvodu se ve výběru vyskytují nástroje, které jsou postavené na rozdílných softwarových architekturách a využívají různý způsob modelování.

Vybrané softwarové nástroje byly porovnány na základě několika charakteristik, které jsou součástí procesu data miningu. Jedná se zejména o podporu vstupních a výstupních formátů dat a modelů, konektivitu s databázemi, podporu funkcionalit pro výběr, předzpracování a transformaci dat, způsoby vizualizace dat a podporu konkrétních algoritmů pro tvorbu modelu. Průzkum těchto funkcionalit probíhal přímo v grafickém uživatelském prostředí aktuálních verzí softwarových nástrojů a byl ověřován v on-line dokumentacích příslušných softwarů.

Metodika pro výběr, průzkum a porovnání vybraných softwarových nástrojů je detailněji popsána na začátku kapitoly 5, zde jsou rovněž rozpracovány konkrétní a dílčí cíle průzkumu. Při řešení ukázkové úlohy v kapitole 7 je postupováno dle metodiky CRISP-DM, která je podrobněji popsána v podkapitole 4.2 této práce.

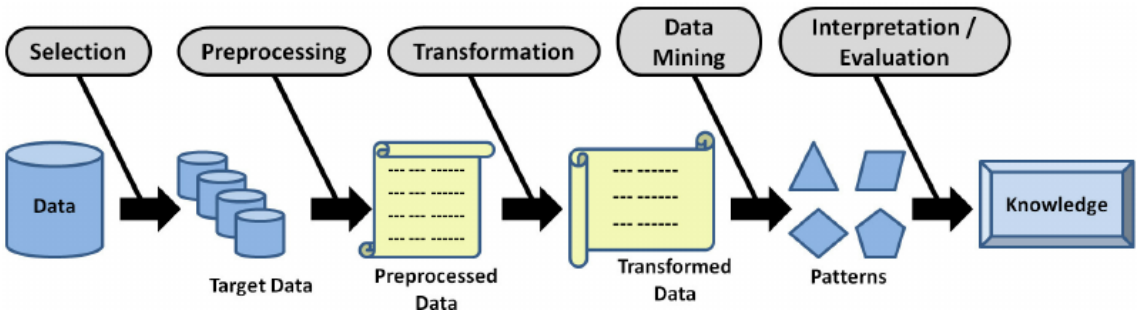
4 Data mining

Data mining (DM) je mezioborovou disciplínou, která využívá poznatků ze statistiky, umělé inteligence a strojového učení.

4.1 Definice

Data mining je charakterizován jako proces, při němž jsou vyhledávány vzory v datech (3). V české odborné literatuře je pro termín „vzory v datech" častěji používáno sousloví „užitečné informace z dat" (1), (2), které bude nadále používáno v tomto textu. DM je tedy vyhledávání užitečných informací z dat, která jsou uložena elektronicky v databázích (3) a jsou obvykle velmi rozsáhlá (2). Proces DM je částečně, nejlépe však zcela, automatizován. DM slouží k řešení konkrétních problémů pomocí analýzy dat. Vyhledané užitečné informace musí být smysluplné a měly by vést k nějaké výhodě například ekonomické. (3).

Na DM je možno nahlížet jako na technologickou součást komplexnější skupiny procesů zvaných dobývání znalostí z databází (Knowledge Discovery in Databases, KDD) (2). KDD lze definovat jako „netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat" (1, s. 15). V KDD je kladen důraz na přípravu dat pro následnou analýzu a na konečnou interpretaci nalezených užitečných informací (1). Součástí KDD jsou procesy: výběr dat, předzpracování dat, transformace dat, data mining a vyhodnocení dat (Obr. 1). DM je pak proces, kdy jsou z transformovaných dat získávány užitečné informace pomocí konkrétních analytických metod.



Obr. 1 Proces dobývání znalostí z databází – technologický pohled

Zdroj: (11)

Na KDD je možno hledět i manažerskou optikou, potom je vstupem do procesu KDD nějaký reálný problém a cílem je najít, co nejvíce užitečných informací, které přispějí k řešení tohoto problému (1). Manažerský řetězec procesu KDD je složen z následujících kroků: sestavení řešitelského týmu, specifikace problému, získání dat, výběr metod, předzpracování dat, data mining a vyhodnocení dat (1).

4.2 Metodiky

S rozvojem KDD začaly vznikat metodiky, jejichž cílem je poskytnout jednotný postup pro řešení různých DM úloh. Některé metodiky byly vyvinuty jako součást komerčních softwarových produktů (SEMMA), jiné jako zcela samostatné (CRISP-DM) (1).

CRISP-DM

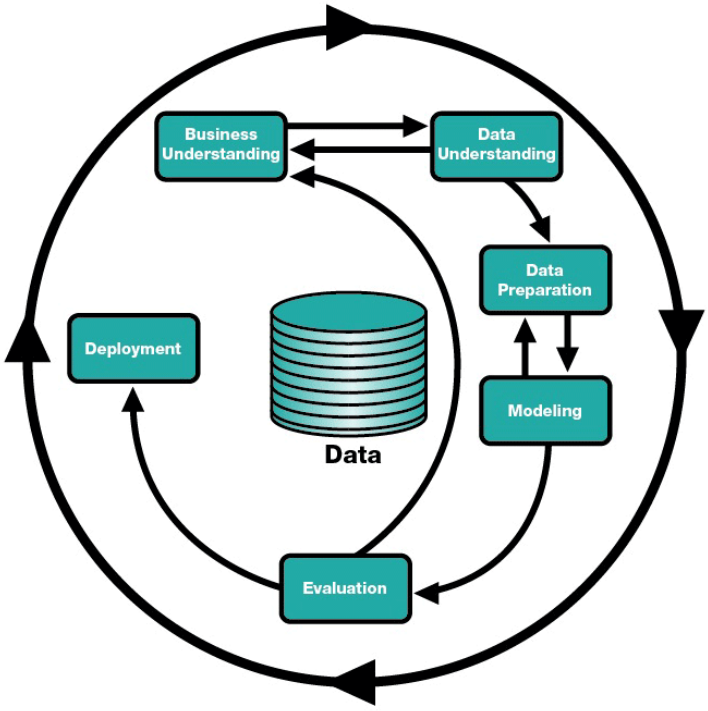
Neproprietární a volně použitelná metodika CRISP-DM (Cross Industry Standard Process for Data Mining) vznikla díky spolupráci 3 velkých firem, které se v 90. letech 20. století zabývaly DM (DaimlerChrysler, ISL později SPSS a NCR). Cílem bylo vytvořit standard, který by byl využitelný napříč různými obory, byl snadno aplikovatelný a volně dostupný (12). Metodika CRISP-DM je v současnosti nejpoužívanější metodikou pro DM (údaj platný k roku 2014) (13). Metodika má šest hlavních fází, jejichž pořadí není pevně dáno, avšak výsledky dosažené v dané fázi ovlivňují výběr té následující. Je vcelku obvyklé se vracet k předchozím částem a opakovat je (1).

- Porozumění problematice (Business understanding) – v první části jde o porozumění problému z obchodní perspektivy (12), nalezení cílů úlohy a jejich následné převedení do zadání DM úlohy (1).
- Porozumění datům (Data understanding) – tato fáze zahrnuje sběr dat k DM úloze, popis, prvotní průzkum a posouzení kvality dat (12).
- Příprava dat (Data preparation) – v této části probíhá výběr, čištění, sestavení, spojování a formátování dat (12). Výsledkem je datový soubor, který bude v následném kroku zpracován konkrétními DM metodami (1).
- Modelování (Modeling) – v této fázi se vybírají nejvhodnější metody pro vlastní DM analýzu, probíhá zde nasazení konkrétních algoritmů na data. Výsledky

jednotlivých metod se srovnávají a kombinují, někdy se aplikace algoritmů opakují s různými parametry. Součástí je i testování na vzorku nezávislých dat a ověřování nalezených informací (1).

- Vyhodnocení (Evaluation) – v této části dochází k posouzení a vyhodnocení nalezených výsledků se zadanými cíli úlohy a rozhodnutí o způsobu jejich využití (1).
- Využití (Deployment) – v posledním fázi dochází k vytvoření návrhu pro využití výsledků DM v praxi (12). Skutečné využití výsledků je pak spíše na straně zadavatele nikoliv analytika (1).

Časově nejnáročnější fází celého procesu je příprava dat, nejdůležitější je fáze porozumění problematice (1). Provázanost jednotlivých fází je možno vidět na schématu metodiky CRISP-DM (Obr. 2).



Obr. 2 Metodika CRISP-DM

Zdroj: (14)

SEMMA

Tato metodika byla vyvinuta firmou SAS pro vlastní DM software Enterprise Miner, jde tedy o metodiku proprietární. Název metodiky je tvořen zkratkou jednotlivých

fází (1). SEMMA má 5 hlavních částí, není třeba je využívat všechny, někdy je nezbytné opakovat kroky předchozí pro dosažení lepších výsledků (15).

- Sample – výběr vhodných dat, vytvoření datového souboru.
- Explore – průzkum dat, vizualizace a shlukování dat.
- Modify – úprava dat, výběr a transformace atributů.
- Model – vlastní analýza dat.
- Assess – vyhodnocení užitečnosti a spolehlivosti výsledků analýzy (15).

Ostatní

Proprietární metodika 5A byla vyvinuta společností SPSS pro vlastní DM nástroj Modeler. Název metodiky odkazuje na 5 jednotlivých fází: Assess – posouzení projektu, Access – sběr a příprava dat, Analyze – provedení analýzy dat, Act – zhodnocení výsledků a Automate – využití výsledků v praxi (1). V roce 2009 byla firma SPSS zakoupena společností IBM (16) a do softwaru IBM SPSS Modeler byla postupně implementována metodika CRISP-DM (17), proto je možno metodiku 5A považovat za historickou a již nevyužívanou.

V posledních letech vznikají mladé metodiky spadající do oblasti DM, které čerpají z agilních metodik a metodiky CRISP-DM. Takové metodiky jsou mimo jiné zaměřeny na efektivnější spolupráci řešitelského týmu. Patří sem TDSP (Team Data Science Process) od firmy Microsoft, která se používá od roku 2016 na platformě Azure, nebo metodika Data Science Lifecycle od firmy Domino, která vznikla v roce 2017. Tyto metodiky nejsou zatím příliš rozšířené (18).

4.3 Zdroje a příprava dat

Data všeobecně je možno rozdělit na strukturovaná a nestrukturovaná. Strukturovaná data jsou obvykle uspořádaná v databázích a navzájem jsou nezávislá. Příkladem strukturovaných dat jsou objekty databáze, dále je možno sem zařadit časová data (vývoj akciových kurzů), prostorová data (geografické informační systémy) a strukturální data (vzorce chemických sloučenin). Typickým příkladem nestrukturovaných dat je potom text (1). Hledáním užitečných informací z nestrukturovaných dat se zabývá například text mining a web mining (2).

Zdrojem dat pro DM jsou nejčastěji databáze, datové sklady, případně datové archivy. Data potřebná k analýze lze získat i přímo a cíleně pomocí dotazníků, hlasování, experimentů a studií (19). Velkým zdrojem dat v nestrukturované formě jsou webové stránky (2).

Příprava dat je považována za velice důležitou a nejobtížnější část celého procesu KDD (1), která je zároveň časově nejnáročnější, může totiž zabírat až 70 % celkového času řešení projektu (2). V procesu KDD je příprava dat rozdělená na tři samostatné kroky: výběr, předzpracování a transformaci dat (11). Cílem přípravy dat je tedy jejich výběr a reprezentace. Výběr dat souvisí s porozuměním problémové domény, někdy je nutná spolupráce specialisty na daný obor, aby došlo k nalezení správných dat (1). Je vcelku běžné, že výběr dat pochází z různých zdrojů, proto je vždy nutné data sjednotit (3). Reprezentace dat se soustředí na uzpůsobení dat pro aplikaci konkrétních DM metod, je tedy nezbytné vědět, jaké metody budou použity pro následnou analýzu. Připravená data mají obvykle podobu datové tabulky s hodnotami atributů jednotlivých objektů (1).

Při předzpracování a transformaci dat je cíleno především na práci s atributy a také na objekty databáze. U atributů je obvyklé setkávat se s problémy, kdy je v datovém souboru obsaženo hodně atributů, a všechny nejsou vhodné pro analýzu. Řešením může být jejich transformace, kdy se z více atributů vytvoří jeden, dále je možno využít výběr nejdůležitějších a nejvhodnějších atributů pro analýzu a odstranění těch nadbytečných. Někdy je třeba řešit chybějící hodnoty atributů, pro tyto účely existuje několik postupů například nahrazení jinou hodnotou (nejčetnější, podílem, libovolnou) (1). U numerických atributů je možno využít matematické transformace, jako jsou normalizace (převod na nový rozsah, minimum, maximum, z-skóre), diskretizace (rozdělení na intervaly) a agregace (vytváření odvozených atributů). Ordinální textové atributy je v některých případech třeba převést na numerické hodnoty (19).

Problémem u objektů databáze může být jejich velká četnost, kdy je datový soubor příliš obsáhlý a výpočetní čas analýzy by se výrazně prodloužil. Řešením může být výběr určitého reprezentativního vzorku dat, případně vytvoření několika

podskupin dat, na nichž jsou zvlášť aplikovány konkrétní DM metody a následně jsou výsledky zkombinovány (1).

Při hledání a získávání dat je možné narazit na některý z těchto problémů: nedostupnost a špatný přístup k datům; nárůst soukromých a licencovaných dat, který vede k nemožnosti jejich využití, nebo k jejich komerčnímu získání; nutnost hledání a využívání účinných a efektivních algoritmů pro extrakci různorodých dat z globálních informačních systémů a velkých databází; nutnost distribuovaného přístupu k velkým datovým setům; problém s velmi rychle se měnícími daty a jejich neustálou aktualizací (20).

V oblasti zpracování velkých dat je možno se setkat s těmito problémy: špatná kvalita dat, kde například chybí hodnoty, existují nepřesné a nesprávné hodnoty, datový soubor má nedostatečnou velikost a data jsou špatně reprezentovaná; velká redundance dat z různých zdrojů a rozdílných formátů, například multimediální obsah jako obrázky, audio a video; všeobecně zpracování velkých komplexních a nestrukturovaných dat do strukturovaného formátu (20).

4.4 Základní metody a úlohy DM

Rozdělení metod

Metody, které se využívají pro analýzu dat v DM úlohách, je možno rozdělit do dvou základních skupin na statistické metody a metody strojového učení (1).

- Statistické metody jsou po teoretické stránce dobře prozkoumané a prověřené dlouholetou praxí. Tyto metody jsou zaměřené na hledání užitečných informací z dat v podobě funkcí, vektorů či podmíněných pravděpodobností. Ke statistickým metodám, které se využívají v DM, patří například kontingenční tabulky, korelační analýza, regresní analýza, logistická regrese, diskriminační analýza, shluková analýza a faktorová analýza (1).
- Metody strojového učení spadají do oblasti umělé inteligence a jsou založené na dvoufázovém přístupu, nejprve se systém učí a vytvoří si obecnou reprezentaci chování (dále tříd) a následně využije znalosti získané učením se pro rozhodování. Učení může probíhat několika způsoby, ke dvěma základním patří

učení se s učitelem a učení se bez učitele. K metodám strojového učení patří rozhodovací stromy, asociační pravidla a rozhodovací pravidla (1) (nazývána též klasifikační pravidla (3)). Tyto tři metody jsou zaměřené hlavně na vyhledávání logických souvislostí a zajímavých vztahů v datech, i díky tomu jsou nalezené užitečné informace uživatelsky srozumitelnější. Dále k metodám strojového učení patří neuronové sítě, Bayesovské sítě a genetické algoritmy (1).

K rozdělení metod DM je nověji přistupováno spíše z pohledu způsobu učení se.

- Učení se s učitelem (Supervised) – u takových metod existuje vždy cílová proměnná, lze sem zařadit například lineární regresi, logistickou regresi, nebo metodu SVM. K typickým úlohám učení se s učitelem patří predikce (21).
- Učení se bez učitele (Unsupervised) – tyto metody naopak nemají cílovou proměnnou, patří sem například shluková analýza (21).

Základní úlohy

Základní typy úloh, které DM řeší lze rozdělit podle schématu strojového učení na klasifikaci, asociaci, shlukování a numerickou predikci (3).

- Klasifikační úlohy se vyznačují tím, že objekty jsou zařazeny do konkrétních tříd, dochází tedy k predikci tříd. Takové typy úloh jsou zvláště vhodné pro nenumerické atributy objektů. Z pohledu strojového učení se jedná o učení se s učitelem, kdy je dopředu specifikovaná třída. K metodám, které se využívají v klasifikačních úlohách patří například rozhodovací stromy a klasifikační pravidla (3).
- Asociační úlohy slouží k nalezení zajímavých struktur v datech, jde v zásadě o predikci atributů. Tyto úlohy jsou rovněž vhodné pro nenumerické atributy. Z pohledu strojového učení se jedná o učení se bez učitele, tedy bez specifikace třídy. K metodám využívaným v asociačních úlohách patří asociační pravidla (3).
- Shlukování slouží pro seskupování podobných dat, jedná se o učení se bez učitele, tedy bez specifikace třídy (3).
- Úlohy numerické predikce jsou druhem klasifikace, jejímž výsledkem je numerická hodnota. Obvykle se využívá metoda regresních stromů (3).

Na typy DM úloh však lze pohlížet i širší optikou, která zahrnuje více analytických metod mimo oblast strojového učení, potom je možno rozdělit úlohy na explorační, deskriptivní, prediktivní, na úlohy soustředící se na hledání vzorů a pravidel a úlohy zahrnující více oblasti (2), (22).

- Explorační úlohy se soustředí na základní popis a vizualizaci dat, což umožňuje získat přehled o struktuře dat (2).
- Deskriptivní úlohy se věnují popisu dat pomocí statistických metod, patří sem například hledání vztahů mezi proměnnými a statistické rozdělení dat (22).
- Prediktivní úlohy se zaměřují na předpověď budoucích hodnot na základě analýzy minulosti. Proto jsou využívány hlavně metody spadající do oblastí klasifikace a regrese (2).
- Úlohy pro hledání vzorů a pravidel v datech využívají hlavně metody pro shlukování a pak také asociační pravidla (22).
- K úlohám, které zahrnují více oblasti potom patří například text a web mining (2), dále analýza obrazu a multimediálního obsahu, tedy audia a videa (22).

Za samostatné typy úloh je možné považovat i verifikační úlohy, jejichž účelem je ověření validity výsledku DM, k tomu je možné využít statistické metody jako analýza rozptylu (používá se zkratka ANOVA anglického termínu Analysis of variance), korelace, lineární regrese, logistická regrese, t-testy a některé další (2). Pro testování správnosti a spolehlivosti modelu jsou dále využívány speciální postupy například křížová validace a bootstrap, nebo Lift a ROC křivky (1).

Statistické metody

Zde jsou stručně popsány některé používané metody v DM, které spadající do oblasti statistiky.

- Kontingenční tabulky slouží k vyhodnocení vztahu mezi dvěma nenumerickými binárními atributy. Pro zjišťování vztahu mezi atributy je možno využít χ^2 test (1).
- Lineární regrese umožňuje sestavit model lineární závislosti numerických atributů (1). V korelační analýze jde o vyjádření síly této závislosti, v regresní

analýze pak o popis této závislosti (23). U regresní analýzy se pro odhad parametrů využívají optimalizační metody, jejichž cílem je odhad parametrů modelu tak, aby rozdíly modelu od skutečného stavu byly minimální podle zvoleného kritéria (1).

- Logistická regrese se zabývá odhadem pravděpodobnosti jevu na základě hodnot nezávislých proměnných, které výskyt jevu mohou ovlivňovat (23). Jde o nelineární regresní metodu, kde sledovaný jev má nenumerické hodnoty (1).
- Diskriminační analýza slouží pro klasifikaci objektů do předem známých tříd. Jde o hledání závislosti nenumerické hodnoty atributu na jiných numerických hodnotách, které musí mít odlišné vlastnosti v každé třídě (1).
- Shluková analýza slouží k rozdělení dat do skupin, které jsou si vzájemně blízké, mají tedy podobné vlastnosti na atributech, které byly použity pro sestavení modelu. Jejím principem je měření vzdálenosti objektů. K metodám shlukové analýzy patří metoda k-means a hierarchické shlukování (1).
- Faktorová analýza zjišťuje závislosti atributů na faktorech, což jsou lineární kombinace jiných atributů (1).
- Různé typy statistických testů umožňují porovnávat dva modely a díky tomu je možné vybrat model vhodnější. Jde o porovnání dvou sad numerických hodnot (například výsledků z křížové validace), kde se zjišťuje, jak významně se liší jejich průměry nebo jiné charakteristiky (1).

Metody strojového učení

V následujících řádcích budou stručně představeny metody strojového učení, které se využívají pro analýzu v DM.

- Rozhodovací stromy aneb metoda „rozděl a panuj“ se využívá k řešení klasifikačních úloh (3). Spočívá v postupném rozdělování dat na menší podmnožiny, až jsou nakonec podmnožiny tvořeny pouze prvky stejné třídy. Pro určení vhodného atributu pro větvení stromu je možné použít výpočet entropie, informační zisk, χ^2 test, nebo Gini index. Využívá se především algoritmus C4.5, případně nejnovější modifikace C5.0, a dále CART (Classification And Regression

Trees), který je vhodný i pro regresní stromy, kde je predikován numerický atribut (1).

- Rozhodovací (klasifikační) pravidla jsou rovněž využívána k řešení klasifikačních úloh (3). Užívané algoritmy rozhodovacích pravidel jsou například pokrývání množin CN4 a ESOD. Klasifikační pravidla lze také získat převodem z rozhodovacího stromu (1).
- Asociační pravidla jsou používána k řešení asociačních úloh (3). Smyslem této metody je nalézt zajímavé souvislosti mezi hodnotami atributů (1). K algoritmů asociačních pravidel patří Apriori, Eclat nebo FP-growth (3).
- Neuronové sítě vycházejí z matematických modelů neuronu. Využívají se pro úlohy klasifikace a predikce, které jsou automatizované. Tvoří alternativu k rozhodovacím stromům a klasifikačním pravidlům, kdy není primárním účelem získat formální matematický popis modelu. Neuronové sítě jsou zvláště vhodné pro analýzu numerických dat (1). Pro tyto metody se používá i souhrnný název rozšířené lineární modely (3).
- Metody založené na instancích patří k metodám založeným na analogii, tedy na podobnosti řešení. Ve fázi učení je předložena databáze s již vyřešenými podobnými problémy, ve fázi usuzování je pak hledán nejpodobnější problém, jehož řešení je použito pro nová data. Metody založené na instancích se používají pro úlohy klasifikace a patří k nim například k-d stromy nebo systém IB3 (1).
- Metody numerické predikce se používají k řešení klasifikačních úloh. Jsou podobné rozhodovacím stromům s tím rozdílem, že je predikována numerická hodnota atributu, takové stromy se nazývají regresní. Využíván je například algoritmus CART (3).
- Bayesovské sítě jsou využívány k řešení klasifikačních úloh, jsou založené na podmíněné pravděpodobnosti. Bayesovskou síť lze zobrazit pomocí acyklicky orientovaného grafu, který pomocí hran detekuje pravděpodobnostní závislosti mezi náhodnými atributy. Znám je především algoritmus EM (1), dále se využívají algoritmy K2 a TAN (3).

K metodám strojového učení, které je možno využít v DM dále patří: genetické algoritmy, případové usuzování, paměťové učení, nebo zcela odlišné induktivní logické programování (1).

4.5 Další metody DM

Testování modelů

Nedílnou součástí procesu KDD je vyhodnocení výsledků data miningu. Pro testování modelů jsou tak využívány nejen tradiční statistické metody jako lineární regrese, ANOVA nebo t-testy, ale také další specializované metody, které byly převzaty z jiných oblastí. Patří k nim ROC křivka, původně využívaná v radiotechnickém oboru a Lift křivka z marketingu. V následujících řádcích je popsáno několik metod, které se často využívají k ověřování validity výsledků DM.

- Křížová validace je způsob testování modelu, kdy se data dopředu rozdělí na určitý počet částí (například 10), kdy jedna část je vyjmuta pro testování a zbývající se využijí pro učení. Celý postup se zopakuje n počtem částí a výsledek testování se zprůměruje (1).
- Lift křivka, někdy též křivka navýšení, je vhodná pro testování modelů klasifikace, kde jsou výstupem numerické hodnoty, lze ji tedy využít třeba pro úlohy řešené pomocí neuronových sítí nebo pomocí Bayesovských metod (1).
- Matice záměn, známá také pod názvem konfusní matice, zachycuje počty správně a nesprávně zařazených příkladů, tedy v kolika případech se klasifikátor shoduje s učitelem, v kolika nikoliv (1).
- ROC křivka je využívána k testování modelů binární klasifikace, umožňuje vyhodnocení klasifikačního pravidla v celé oblasti výsledných hodnot. Vedle vizuálního porovnávání ROC křivek, se též využívá souhrnná číselná charakteristika nazývaná AUC, jedná se o plochu pod křivkou ROC (2).

Kombinování modelů

Možnost zlepšení výsledků DM, které byly dosaženy jednotlivými modely, je jejich vzájemná kombinace. Děje se tak často pomocí některé z variant hlasování (1).

Metody pro kombinaci více modelů spadají do podoblasti strojového učení zvané ensemble learning (sborové učení, případně vzájemné či postupné učení).

- Bagging (bootstrap aggregating) – při tomto učení se z dat vytvoří stejně velké trénovací množiny pomocí náhodného výběru s opakováním (bootstrap). Učení probíhá na každé trénovací množině samostatně, nakonec se nechají výsledné modely hlasovat o přiřazení testovaného příkladu do třídy. U této metody se předpokládá použití stejného algoritmu na všechny trénovací množiny. (1).
- Boosting je metoda, při které se vytvářejí stále nové modely, které vycházejí z vlastností předchozích modelů, konkrétně se zaměřují na příklady, které zatím nebyly správně klasifikovány. V průběhu učení se mění váha modelů, ve fázi klasifikace modely váženě hlasují o přiřazení příkladu do třídy. I u této metody se počítá s použitím jednoho algoritmu. Pro aplikaci této metody se často používá algoritmus AdaBoost (1).
- Náhodný les se obvykle využívá pro velké sady dat s mnoha atributy. Je tvořen desítkami až stovkami rozhodovacích stromů, z nichž každý je sestaven z rozdílného náhodného výběru vzorků trénovacích dat a pomocí stejného klasifikačního algoritmu. Výsledný model vzniká opět hlasováním jednotlivých modelů. Výhoda tohoto algoritmu spočívá mimo jiné v tom, že umožňuje minimalizovat předzpracování dat, dokáže najít nejvíce používané atributy a je odolný vůči odlehlým hodnotám (24).
- Stacking je metoda, která porovnává modely vzniklé na základě různých algoritmů. Provádí učení se z výsledků jednotlivých klasifikátorů, jde o tzv. meta-algoritmus. Cílem metody stacking je rozhodnutí o tom, který porovnávaný klasifikátor je vhodnější pro použití (1).

5 Nekomerční software pro data mining

Počet softwarových nástrojů pro DM je na celosvětovém trhu přibližně 200, z tohoto počtu je zhruba 50 řešení volně dostupných (25). Podle průzkumu serveru KDnuggets (obrazová příloha Obrázek 2), kterého se zúčastnilo 1800 respondentů z řad uživatelů softwaru orientovaného na data mining, se v roce 2019 nejvíce využívaly v této oblasti technologie Python a R. Populárnější se stává hybridní technologie Anaconda, která spojuje přednosti obou těchto programovacích jazyků. Naopak obliba softwarových řešení pro DM v Javě spíše klesá (26).

Jedním z cílů této práce je představit vzorek nekomerčního softwaru, který lze využít pro data mining. Pro účel tohoto výběru je pojem „nekomerční software“ zpřesněn na svobodný a neproprietární software s otevřeným zdrojovým kódem (open source), který je licencován GNU GPLv2 a vyšší, případně jinými kompatibilními a příbuznými licencemi. Takový software je tedy primárně určen pro nekomerční využití.

Kritéria výběru

Při výběru softwaru bylo, vedle svobodného a neproprietárního software, pracováno dále s těmito kritérii:

- Porovnat komplexní DM nástroje, které umožňují řešení minimálně základních klasifikačních, asociačních a shlukových úloh a zároveň nejsou specializované pouze na jednu oblast například klasifikaci pomocí neuronových sítí, deep learning apod.
- Ukázat odlišná technologická řešení softwaru (Delphi / Object Pascal, Java, Python a R).
- Přiblížit rozdílné pohledy na způsob modelování, u vybraných nástrojů se jedná o diagram, skriptování a vizuální programování.

Vybrány byly nástroje: DataMelt, KNIME Analytics Platform, Orange, Rattle a Tanagra. Přehled základních charakteristik zvolených nekomerčních softwarů je v Tab. 1. na straně 17.

Tab. 1 Základní charakteristiky vybraných nekomerčních softwarů

	DataMelt	KNIME	Orange	Rattle	Tanagra
vývojář	Sergei Chekanov	KNIME AG	Univerzita Lublaň	Togaware	Ricco Rakotomalala
aktuální verze	2.5	4.2	3.27.1	5.4.0	1.4.50
programovací jazyk	Java	Java	Python	R	Object Pascal (Delphi)
skriptování	BeanShell, Groovy, Java, JRuby, Jython, Octave	Java	Python	R	ne
platformy	Android, Linux, Mac, Windows	Linux, Mac, Windows	Linux, Mac, Windows	Linux, Mac, Windows	Windows
nekomerční licence	GNU GPLv3	GNU GPLv3	GNU GPLv3	GNU GPLv2	vlastní
komerční licence	ano	ano	ne	Ano	ano
GUI	ano	ano	ano	Ano	ano
CLI	ano	ne	ne	Ano	ne
způsob modelování	jiné	vizuální programování	vizuální programování	Jiné	diagram

Zdroj: Vlastní zpracování

Způsob zkoumání

Zde prezentovaný vzorek softwarů byl testován v posledních stabilních verzích a vždy pouze v základní instalaci vyjma softwaru Orange, který byl rozšířen o funkcionality potřebné pro bližší průzkum v kapitole 6 a řešení ukázkové úlohy v kapitole 7. K ověřování vlastního průzkumu byly využívány aktuální on-line manuály a dokumentace jednotlivých nástrojů, případně další uvedená literatura. U většiny zde popisovaných nástrojů existuje možnost dalšího rozšíření funkcionalit, některé z nich jsou zmíněny v textu, v tabulkách jsou uvedeny v případě, kdy existují oficiální rozšíření od vývojáře.

U každého vybraného nástroje se práce zaměřuje na krátké představení – historii, vývoj a základní určení; technologické řešení včetně možnosti využití skriptovacích jazyků; způsob modelování; stručný popis grafického uživatelského prostředí; zdroje dat, vstupní/výstupní datové formáty a formáty pro načítání/ukládání modelu; možnosti v oblasti výběru, předzpracování a transformace dat; využívané

metody a konkrétní algoritmy DM; testování modelu; vizualizaci dat; možnosti nasazení a využití užitečných informací v praxi; upozornění na specifické funkcionality. Závěrem je u každého softwaru krátké zhodnocení zahrnující slabé a silné stránky, nároky na uživatele a možnosti podpory.

Poslední část této kapitoly se věnuje vzájemnému porovnání těchto zvolených pěti nekomerčních softwarů, vedle slovního zhodnocení je prezentováno tabulkové zpracování prozkoumaných funkcionalit nástrojů. Výběr charakteristik se odvíjel od teoretické části v kapitole 4 této práce a dále na základě příspěvku An overview of free software tools for general data mining (9).

Zkoumané funkcionality

- Externí i vlastní datové formáty, podpora formátů pro načítání a ukládání modelu.
- Funkcionality pro výběr, předzpracování a transformaci dat, možnosti v oblasti nasazení výsledků v praxi.
- Způsoby vizualizace, druhy podporovaných grafů.
- Porovnání algoritmů pro data mining včetně algoritmů pro testování modelu.
- Podpora specifických funkcionalit v oblasti big data, deep learning, image mining, text mining a web mining.

5.1 DataMelt

Software DataMelt (DMelt) je určen pro analýzu a vizualizaci dat, numerické i symbolické výpočty a statistiku (4). Je vyvíjen od roku 2005, původně pod názvem jHepWork, v krátkém období byl přejmenován na SCaVis (2013–2015) a primárně byl určen pro analýzu dat v oblasti částicové fyziky (27). Jedná se o open source software, který je dostupný pod licencí GNU GPLv3 a zároveň pod licencí komerční. Vývoj programu DMelt zaštiťuje sdružení dobrovolných vývojářů jWork.org, jehož supervizorem je Sergei Chekanov, původní autor softwaru (4).

Software DMelt je naprogramován v jazyce Java. Jeho prostředí však umožňuje běh skriptů v dalších integrovaných jazycích jako Groovy, Jython a JRuby, je možné využít také specializované jazyky BeanShell a Octave. DataMelt je přenosný

(nevyžaduje instalaci, je možné jej spouštět z přenosných médií) a multiplatformní software. Dostupný je na Windows, Mac i Linux a na mnoha dalších, které jsou schopny zajistit běh prostředí Java (27). Na operačním systému Android je software v omezené podobě součástí aplikace AWork. Tato mobilní aplikace se zaměřuje na analýzu a vizualizaci dat pomocí skriptovacího jazyku Octave (4).

Prostředí softwaru DMelt je v zásadě vývojové prostředí (IDE), které poskytuje kontrolu a zvýraznění syntaxe, dokončování a analýzu programovacího kódu (27). Modelování je tedy postaveno na tvorbě skriptů, v nichž jsou implementovány konkrétní algoritmy DM metod. Toto prostředí je velice flexibilní, umožňuje načítat, upravovat, ladit a ukládat jednotlivé skripty, které lze aplikovat na různá vstupní data (28). Pro plné využití softwaru je nutná znalost některého ze skriptovacích jazyků.

Grafické uživatelské rozhraní (GUI) DataMeltu se skládá ze třech oken (obrazová příloha Obrázek 2). Levé okno zobrazuje seznam souborů, případně strukturu projektu; v pravém horním okně se nachází hlavní editor skriptů; v pravém spodním okně jsou integrované příkazové řádky pro jednotlivé skriptovací jazyky a systémová konzole pro výstupy z běhu skriptů.

Nástroj DataMelt podporuje standardně nativní Java a Python vstupní/výstupní formáty, umí také pracovat s textovým formátem csv a univerzálním formátem xml. Dokáže získávat data z externích databází pomocí dotazovacího jazyka SQL. V nekomerční licenci neumí DMelt zpracovat formáty Excel a dokáže exportovat dokumenty do formátu pdf. (28).

V oblasti vizualizace dat nabízí software DataMelt možnost tvorby 2D bodových a sloupcových grafů včetně histogramů a 3D plošných grafů (27).

DMelt sám o sobě není specializovaný software pouze pro DM, v nekomerční licenci obsahuje několik předpřipravených skriptů v oblasti klasifikace, regrese, shlukové analýzy a neuronových sítí. Avšak již v základní instalaci integruje volně dostupné knihovny ze tří nástrojů, které jsou určené k DM analýze, jedná se o open source knihovny Weka, Encong a Joone. Všechny tři včetně grafického uživatelského prostředí (28).

Weka

Weka je Java kolekce algoritmů pro strojové učení. Software vznikl a je vyvíjen na Univerzitě Waikato na Novém Zélandu. Poskytuje rozsáhlé funkcionality pro výběr, předzpracování a transformaci dat, pro řešení DM úloh pomocí statistických metod a algoritmů strojového učení v oblasti regrese, klasifikace, shlukové analýzy a asociačních pravidel, nabízí široké možnosti v oblasti validace a vyhodnocení výsledků DM (29).

Encog a Joone

Java frameworky Encog a Joone se specializují na využití neuronových sítí v klasifikačních úlohách a shlukové analýze (27). Encog zahrnuje mimo podpory neuronových sítí také SVM, Bayesovské sítě a genetické algoritmy (30). Joone nabízí v učení se s učitelem algoritmy pro dopředné a rekurzivní neuronové sítě, v oblasti učení se bez učitele Kohonenovy mapy (SOM) (31).

Prostředí DataMelt však umožňuje využívat další externí a volně dostupné knihovny na základě podporovaných skriptovacích jazyků. V jazyce Java je možné dále využít knihovny Datumbox Machine Learning Framework a Smile, obě se specializují na strojové učení a řešení klasifikačních typů úloh (28).

DMelt poskytuje uživateli velkou volnost při modelování, ale klade na něj vyšší nároky v oblasti programování a znalosti jednotlivých algoritmů DM metod. Oproti tomu nabízí možnost využít grafického prostředí programu Weka pro komplexní řešení DM úloh. Hlavní vývojář softwaru Sergei Chekanov společně s iniciativou jWork.org poskytuje k nástroji DMelt obsáhlou on-line nápovědu.

5.2 KNIME Analytics Platform

Software KNIME (Konstanz Information Miner) vznikl na Univerzitě v Kostnici v Německu v roce 2004, první stabilní verze vyšla v roce 2006. Software aktuálně vyvíjí švýcarská společnost KNIME AG (9). Původně byl software určen pro analýzu dat ve farmaceutickém průmyslu, ale velmi rychle se rozšířil mimo tento obor. Název KNIME Analytics Platform (v textu dále zkráceně KNIME) se používá pro open

source verzi programu, která je licencovaná GNU GPLv3, zatímco KNIME Server je komerčně licencovaná verze softwaru (32).

DM nástroj KNIME je technologicky postaven na Java platformě Eclipse (9). Vedle nativní možnosti využití skriptů v jazyce Java, lze doinstalovat integraci skriptovacích jazyků JavaScript, Python a R. KNIME je dostupný na platformách Linux, Mac i Windows, a také v on-line verzi přes cloudová řešení Microsoft Azure a Amazon AWS (32).

Software KMINE stojí na schématu vizuálního programování, kdy jsou do okna umisťovány stavební bloky, zde nazývané uzly (nodes), které svým vzájemným propojením vytvářejí pracovní vlákno (workflow) (9). Vlákno tak přehledně zobrazuje celý proces DM.

Grafické prostředí programu KNIME tvoří hlavní okno a sedm oken vedlejších (obrazová příloha Obrázek 3). V hlavním okně Workflow Editor probíhá modelování pomocí uzlů a jejich vzájemných propojení, případně seskupování uzlů. Vlevo jsou standardně tři okna: KNIME Explorer, Workflow Coach a Node Repository. Vpravo se nachází dvě okna: Description a KNIME Hub Search. Dole pod hlavním oknem jsou Outline a Console. Okna nejsou fixovaná, je možno je dle potřeby minimalizovat, maximalizovat, vypínat a přesouvat, díky tomu je grafické prostředí uživatelsky velice flexibilní.

- Workflow Editor je virtuální plátno, na němž se vytváří pomocí uzlů a jejich propojení pracovní vlákno (5).
- KNIME Explorer zobrazuje strom přístupných pracovních vláken na lokálním disku, ukázková pracovní vlákna na serveru, případně osobní prostor ve službě KNIME Hub (5).
- Workflow Coach zobrazuje seznam doporučených akcí, které jsou vytvořené na základě automaticky sbíraných statistik od komunity uživatelů softwaru. Návrh akcí je aktualizován v závislosti na použitém uzlu a jeho kontextu (5).
- Node Repository je knihovna všech dostupných uzlů a doinstalovaných rozšíření. Uzly jsou pro větší přehlednost rozděleny do kategorií, pro urychlení je možno využít integrované vyhledávací pole (5).

- Description ukazuje popis aktuálního pracovního vlákna, nebo aktuálně vybraného uzlu (5).
- KNIME Hub Search umožňuje vyhledávání ve službě KNIME Hub, po zadání se automaticky přesměruje do internetového vyhledavače, kde jsou zobrazeny výsledky vyhledávání (5).
- Outline zobrazuje zmenšený přehled aktuálního pracovního vlákna a umožňuje v něm rychlý pohyb (5).
- Console zobrazuje zprávy z vykonávání a provádění akcí, nelze aktivně zadávat příkazy, nejedná se tedy o příkazový řádek. V tomto okně je možno na záložkách aktivovat ještě Node Monitor a Error Log (5).

KNIME má užitečné funkcionality v oblasti získávání a přípravy dat. Umí pracovat s mnoha vstupními formáty, standardně s textovými například csv, xls, json a pmml, dále s nestrukturovanými daty, jako jsou obrázky a dokumenty, nebo s časovými řadami. Nástroj KNIME dokáže získávat data z různých zdrojů (datový soubor, databáze, datové sklady, cloudy), má funkcionality pro kombinování dat z odlišných zdrojů do souhrnných tabulek. Software dále umožňuje data čistit a transformovat, a připravit je tak k finální analýze (32).

Vedle standardních grafů, jako jsou bodové, sloupcové, krabicové a histogramy, má KNIME k dispozici pokročilejší grafy například teplotní mapy, vícevrstvé prstencové grafy, grafy rovnoběžných souřadnic a síťové grafy. Software nepodporuje pokročilé 3D vizualizačních techniky například plošné grafy (32).

V oblasti DM poskytuje KNIME základní funkcionality k řešení klasifikačních a asociačních úloh (tyto pouze po instalaci doplňků), také je připraven pro řešení úloh shlukování. Jsou zde bayesovské klasifikátory, některé z algoritmů pro rozhodovací stromy, algoritmy neuronových sítí (MLP a PNN), pro shlukování jsem k dispozici metody k-means, c-means a hierarchického shlukování. Standardně jsou dále k dispozici statistické nástroje pro diskriminační analýzu, lineární a logistickou regresi. K validaci modelů poskytuje software KNIME křížovou validaci, ROC křivky a t-testy. V základní verzi software podporuje sborové učení například metody boosting a náhodný les (32).

Software KNIME obsahuje standardy celého procesu KDD a je proto vybaven funkcionalitami k nasazení výsledků formou exportovaných zpráv s výsledky analýzy v přívětivější formě, která je vhodná i pro zadavatele. Důležitou funkcionalitou je export modelu do pmml s možností připomínkování a zpětného zpracování, aniž by přímo docházelo k zásahu do workflow (32).

KNIME je software, který disponuje širokými možnostmi rozšíření, a dá se přizpůsobit pro různé druhy DM úloh. Nabízí doplňky, které jsou vhodná třeba pro text mining, analýzu sociálních sítí a webů; podporuje integrace různých open source projektů, jako jsou Weka pro strojové učení, Keras pro hluboké učení, Apache Spark pro zpracování velkých dat, H2O pro strojové učení; rovněž lze integrovat prostředí pro skriptování v jazycích Python a R (5).

Software KNIME lze považovat za komplexní nástroj pro data mining včetně předzpracování a nasazení výsledků. To vše v graficky přívětivém a funkčním prostředí. Nástroj má širokou podporu, nejen v softwarové dokumentaci od vývojáře, ale také v uživatelské komunitě, která sdílí konkrétní pracovní vlákna na portále KNIME Hub.

5.3 Orange

Open source software Orange je určen pro strojové učení a vizualizaci dat, je vyvíjen v laboratoři pro bioinformatiku na Univerzitě v Lublani. Na vývoji se podílí i uživatelská komunita. Nástroj je licencován GNU GPLv3, jde výhradně o svobodný a neproprietární software, pro který neexistuje komerční licence (10).

Historie vývoje tohoto softwaru spadá do druhé poloviny devadesátých let 20. století, kdy byl nástroj koncipován jako knihovna jazyka C++, časem bylo do nástroje přidáno rozhraní pro skriptování v jazyce Python. Dnes je Orange programován výhradně pomocí jazyka Python (33). Software je dostupný na platformách Mac, Linux i Windows (10).

Nástroj Orange stojí na principu vizuálního programování, kdy jsou na plátno umisťovány komponenty, zde nazývané widgety, které se vzájemně propojují a vytváří tak pracovní vlákno (workflow) (10).

Grafické uživatelské prostředí programu je velmi jednoduché a intuitivní, skládá se celkem ze třech oken (obrazová příloha Obrázek 4). Vlevo nahoře je okno widgetů, ty jsou roztříděny do kategorií dle použití na data, vizualizaci, modelování, vyhodnocení a učení se bez učitele. V levém dolním okně se zobrazuje stručný popis pro konkrétní widget. Součástí tohoto okna je proužek ikon s šesti dalšími funkcionalitami, které se přímo netýkají modelování, ale jsou určeny pro popis pracovního vlákna. Obě levá okna lze souběžně minimalizovat, poté se zobrazují jako boční pruh s ikonami. Vpravo se nachází okno plátna. V tomto okně probíhá tvorba pracovního vlákna, které tvoří widgety a jejich vzájemná propojení.

Software Orange podporuje textový formát csv, dále formáty tabulkového editoru Excel a Google. Dokáže komunikovat v jazyce SQL s lokálními i on-line databázemi. Má možnost exportu zpráv do formátů html a pdf (6).

V oblasti předzpracování a transformace dat nabízí nástroj Orange standardní funkcionality včetně agregace, diskretizace a normalizace, umí doplnit chybějící hodnoty dle průměrné, nejvíce zastoupené a náhodné hodnoty. Software dokáže komunikovat s externími databázemi a díky tomu je možné importovat do prostředí softwaru jednotlivé tabulky i kompletní sady dat. Dále má zahrnuty funkce pro kombinování a slučování datových tabulek, takže je možné přímo v software přímo vytvořit datovou tabulku vhodnou pro finální analýzu (6).

Orange nabízí vedle běžných vizualizačních technik, jako jsou histogramy, bodové, sloupcové a spojnicové grafy, také speciální grafy například nomogram (zde konkrétně pro vizualizaci Bayesovského klasifikátoru a logistické regrese), graf pro zobrazení stromů i náhodného lesa, graf pro vizualizaci rozhodovacích pravidel, Vennovy diagramy, teplotní mapy a krabicový graf. Software má velmi omezené možnosti v oblasti zobrazování dat pomocí 3D grafů, a to pouze v oblasti bodových třídimenzionálních grafů (6).

Pro data mining má Orange v základní softwarové verzi k dispozici statistické metody lineární regrese, logistické regrese a diskriminační analýzy; v oblasti sborového učení implementuje algoritmy AdaBoost, náhodný les a stacking; z neuronových sítí umí algoritmy MLP a SOM; obsahuje algoritmus CN2 pro klasifikační pravidla. Pro klasifikační úlohy využívá vlastní algoritmus

rozhodovacích stromů Tree, dále k-NN a Naivní Bayes. Pro shlukování slouží metody hierarchického shlukování a algoritmus k-means. Software Orange nemá v základní verzi funkce pro tvorbu asociačních pravidel, existuje však rozšíření od vývojáře. Pro ověřování validity modelů nabízí nástroj křížovou validaci a konfusní matici, Lift a ROC křivky a kalibrační křivku (6).

V nástroji Orange je implementovaná slabší podpora pro případné nasazení a využití výsledků DM, nabízí možnost exportovat výsledky do formátů html a pdf.

Orange disponuje širokými možnostmi rozšíření. V on-line katalogu widgetů jsou připraveny skupiny algoritmů pro text i web mining, analýzu obrazových dat, časových řad a geografických souřadnic, dále také specializované funkce pro bioinformatiku a spektroskopii. Software Orange je navíc nabízen ve dvou oddělených distribucích. Single Cell se specializuje na analýzu dat genového výzkumu, zatímco distribuce Quasar je zaměřena na analýzu spektrálních dat (10).

Software Orange je uživatelsky přívětivou platformou, díky čistému grafickému prostředí, jednoduchosti ovládání a možnostmi rozšíření funkcionalit. Program Orange má vedle tradiční obsáhlé dokumentace i interaktivní formu komunitní nápovědy pomocí kanálů YouTube a Discord.

5.4 Rattle

DM nástroj Rattle (R Analytical Tool To Learn Easily) je volně dostupné grafické uživatelské prostředí statistického programovacího jazyka R. Poskytuje vizuální sumarizaci dat, transformaci dat, tvorbu modelů (učení se s učitelem, učení se bez učitele) a testování modelů. Za vývojem stojí Graham Williams a jeho společnost Togaware sídlící v Austrálii (34). Rattle podléhá licenci GNU GPLv2 (7).

Software Rattle je naprogramovaný v jazyce R a využívá více jak 100 externích specializovaných balíčků pro DM (34). Velmi užitečnou vlastností tohoto softwaru je souběžné zaznamenávání všech interakcí provedených v GUI do skriptu v jazyce R. Tento skript umožňuje další nezávislé využití, například v konzoli běhového prostředí R (7).

Grafické uživatelské prostředí Rattle je vytvořeno pomocí balíčku GTK+ pracovního prostředí GNOME, které se často využívá na unixových systémech (35). GUI tvoří jedno okno (obrazová příloha Obrázek 5), v jehož vrchní části jsou záložky, které přepínají karty s obsahem. Tyto karty jsou rozděleny do devíti podoblastí, které reprezentují proces data miningu od načtení, průzkumu, testování a transformaci dat, přes modelování (to je rozděleno do tří karet na shlukovou analýzu a asociační pravidla, tedy deskriptivní DM, a model, kde jsou soustředěny algoritmy pro predikci), až k vyhodnocení modelu. Na poslední kartě je k dispozici log, kde se průběžně ukládá skript v jazyce R, který zaznamenává celý proces DM.

Modelování v tomto nástroji probíhá na jednotlivých kartách, a to výběrem příslušných metod, konkrétních algoritmů a dodatečných charakteristik. Všechno pomocí jednoduchých akcí pomocí přepínačů, zaškrtávacích polí a rozbalovacích seznamů. Tím je vytvořeno jednoduché pracovní vlákno, které však není zobrazeno celé, ale po částech.

Rattle dokáže načítat data z databází i on-line zdrojů. Přímo podporuje textové formáty csv, txt a tabulkové formáty Excel, dokáže také pracovat s formátem pmml. Všechny tyto formáty umí načítat i ukládat. Do prostředí je možné importovat specializovaný datový formát arff (35).

V oblasti předzpracování a transformace dat je Rattle vybaven možnostmi doplnit chybějící hodnoty, umí normalizaci, diskretizaci a agregaci atributů, dokáže převádět na jiné druhy atributů (35).

Vedle tradičních grafů, jako jsou bodové, sloupcové, krabicové grafy, umí nástroj Rattle zobrazovat také specializované grafy pro data mining například dendrogram, který se využívá pro zobrazení hierarchického shlukování. Pro vizualizaci dále využívá Rattle integrované externí knihovny GGobi a ggraptR, které dokáží zobrazovat různé druhy 2D i 3D grafů. Knihovna GGobi umožňuje data prozkoumávat interaktivně pomocí funkcí přibližování a oddalování, otáčení, sledování kamerou atd. (7).

V oblasti deskriptivního DM má Rattle k dispozici algoritmy pro shlukovou analýzu k-means a hierarchické shlukování; pro asociační pravidla má software připraven

algoritmus Apriori. V oblasti prediktivního DM má nástroj Rattle implementovány algoritmy pro rozhodovací stromy, rozhodovací pravidla (PART), algoritmus SVM; ze statistických metod je k dispozici lineární a logistická regrese; z neuronových sítí je přítomen algoritmus MLP. V oblasti sborového učení nechybí náhodný les a AdaBoost. Rattle poskytuje standardní metody pro vyhodnocení modelů, jako jsou matice záměn, Lift a ROC křivka, t-testy (35).

Nasazení výsledků je v nástroji Rattle podporováno exportem modelu do formátu pmml, který umožňuje další využití mimo jeho prostředí. Z rozšířených funkcionalit v oblasti DM má Rattle k dispozici balíček pro text mining.

Nástroj Rattle má intuitivní uživatelské rozhraní, které snadno dokáže provést kroky data miningu. Účelem softwaru je přiblížit základní možnosti jazyka R využitelné pro DM a nasměrovat uživatele k jeho dalšímu pokročilejšímu využití. Právě proto se tento nástroj hojně využívá v univerzitních prostředích (7). Rattle sám o sobě nedisponuje žádnými rozšiřujícími balíčky funkcí. Podpora a nápověda k nástroji pochází výhradně od jeho vývojáře Grahama Williamse a je k nalezení na webu Togaware (34).

5.5 Tanagra

Specializovaný data mining software Tanagra vzešel z akademického prostředí univerzity v Lyonu v roce 2004, jeho tvůrce je Ricco Rakotomalala. Tanagra je nástupcem programu Sipina, který se specializoval na rozhodovací stromy a pocházel od stejného vývojáře. Jedná se o open source projekt, jehož zdrojový kód je volně k dispozici pro akademické a výzkumné účely (8).

Zdrojový kód softwaru Tanagra je vytvořen v prostředí Delphi v jazyce Object Pascal, je proto primárně určen pro platformu Windows. Prostředí Tanagra je dostupné ve dvou jazycích: angličtině a francouzštině (36). Vývoj softwaru se zastavil v roce 2013 na verzi 1.4.50 (8).

Grafické rozhraní se skládá ze třech oken (obrazová příloha Obrázek 6). Levé horní okno zobrazuje diagram, do něhož jsou během modelování přidávány komponenty, které tvoří posloupnost příkazů, tedy model. Ve spodním okně se nachází komponenty, které se přetažením vkládají do diagramu. Komponenty jsou tvořeny

konkrétními algoritmy DM, statistických a vizualizačních metod a dalšími příkazy pro manipulaci s daty. Komponenty jsou roztříděné do skupin dle možnosti použití (asociace, shlukování, učení se s učitelem atd.). Pravé horní okno slouží k zobrazování výsledků. Tanagra má výlučně grafické rozhraní, absentuje příkazový řádek a editor pro tvorbu a úpravu skriptů.

Software Tanagra umí pracovat se vstupními daty v textovém formátu, v tabulkovém formátu Excel a ve speciálním formátu arff, který je standardem dat v DM softwaru Weka (37).

V oblasti výběru, předzpracování a transformace dat má Tanagra velice omezené možnosti. Komunikace s databázemi není vůbec podporována, práce s atributy například diskretizace, normalizace a agregace zcela absentují.

Tanagra implementuje nejdůležitější algoritmy z oblasti strojového učení a statistiky, které jsou vhodné pro řešení klasifikačních, asociačních a shlukových typů DM úloh. V oblasti strojového učení se s učitelem například algoritmy C4.5, CART, ID3, K-NN, Naivní Bayes, SVM, jeho modifikace BVM a CVM; v učení se bez učitele například Apriori, EM a k-means. Obsahuje rovněž několik metod pro hodnocení kvality modelu například Lift a ROC křivky; v oblasti validity výsledků DM procesu například ANOVA, korelace a t-test.

Výsledky modelovacího procesu lze exportovat do tzv. zprávy, která je automaticky formátována do html (37), toto se dá považovat za omezené možnosti v oblasti nasazení výsledků DM.

Autor nástroje Ricco Rakotomalala průběžně zveřejňuje praktické návody, k nimž dává k dispozici i ukázkové datové tabulky a modelovací diagramy. Do roku 2019 byly tyto návody poskytovány souběžně v angličtině a francouzštině, avšak díky masivnímu rozvoji automatických internetových překladačů je aktuálně nápověda k dispozici jen ve francouzském jazyce s možností strojového překladu (37).

Software Tanagra má jednoduché grafické prostředí, předpřipravené základní algoritmy pro DM analýzu, rozsáhlý systém nápovědy s možností využití ukázkových dat a modelovacích diagramů. Oproti tomu neposkytuje možnost

vkládání a ladění skriptů, nemá příkazový řádek pro ovládání běhu skriptů, úprava algoritmů je možná pouze zásahem do zdrojového kódu programu.

5.6 Porovnání

Vývojář, licence a využití

Všech pět nástrojů původně vzniklo v akademickém prostředí a bylo koncipováno pro studijní a pro specifické vědecké účely (bioinformatika, částicová fyzika a farmaceutický výzkum). U nástrojů Orange, Rattle a Tanagra převažuje i dnes určení pro studijní prostředí případně pro specializované vědecké účely, zatímco softwary DataMelt a KNIME jsou komplexní nástroje, které lze použít pro široké spektrum účelů datové analýzy.

Ve zde představovaném výběru nástrojů je ve čtyřech případech možné využít software zároveň pro účely komerční, existuje pro ně tedy i licence proprietární. Takový software má pak obvykle rozšířené a specializované funkcionality a dostupnější zákaznickou podporu. DMelt například nabízí formu předplatného, díky němuž je přístupná plná on-line podpora a knihovna skriptů. KNIME disponuje proprietární verzí pod názvem Server, je ze všech zde prezentovaných nástrojů nejvíce orientován i do obchodní sféry. Nástroje DMelt, Rattle i Tanagra je možno individuálně zakoupit pro komerční využití 3. stran bez možnosti dalšího šíření kódu.

Technologie

Všechny zde prezentované nástroje vyjma softwaru Tanagra lze považovat za nezávislé na operačním systému, jejich funkčnost je podmíněna multiplatformním běhovým prostředím, jako jsou Java Runtime Environment (DataMelt a KNIME), Miniconda (Orange) a R (Rattle). Nástroj Tanagra je vázaný na vývojové prostředí Delphi, proto je dostupný pouze na operačním systému Windows. Program DMelt má jako jediný svoji verzi pro mobilní platformu Android, aplikace AWork však nenabízí všechny funkčnosti standardní verze DMeltu.

Možnosti vkládání skriptů jsou u vybraných softwarů různé, jejich seznam je k dispozici v Tab. 1 na straně 17. V zásadě platí, že lze standardně vkládat skripty

v programovacím jazyce nástroje, opět to však neplatí pro software Tanagra, který nemá žádnou takovou možnost. DataMelt umožňuje skriptování v dalších jazycích, které jsou implementovány pro Javu (Beanshell, Grovy, Jython a JRuby) a umí pracovat i v jazycích GNU Octave a Matlab.

Způsob modelování

V softwarech Orange a KNIME se modeluje pomocí schématu vizuálního programování. Na ploše je vytvářeno pracovní vlákno, v jednotlivých uzlech se nastavují parametry pro konkrétní funkce, mezi uzly se vytváří propojení. Výsledné pracovní vlákno dokáže přehledně zobrazit celý proces DM. K této koncepci modelování se blíží ještě nástroj Tanagra, který má k dispozici jednoduchý vertikální diagram, jeho možnosti jsou však výrazně omezenější.

Software Rattle má pracovní vlákno reprezentováno kartami, na nichž se nastavují parametry jednotlivých metod, karty jsou řazeny dle logiky procesu modelování.

Nástroj DMelt je sám o sobě vývojovým prostředím, které je vytvořeno pro přímé použití skriptů, avšak díky integrovaným nástrojům Weka, Joone a Encog disponuje i GUI pro data mining. Každý z těchto nástrojů má odlišné způsoby modelování, ale žádný z nich se neblíží paradigmatu vizuálního programování.

Datové formáty, formáty pro model a export zpráv

Možnosti datových formátů a formátů pro načítání ukládání modelu jsou zobrazeny v tabulkové příloze Tabulka 1. Nejvybavenější je v tomto ohledu software KNIME, ten výrazně převyšuje možnosti zbývajících čtyř porovnávaných nástrojů.

Vedle importu tradičních textových a tabulkových formátů pro data, byla prozkoumána i možnost načítání specializovaného datového formátu arff, jenž je původně datovým standardem v softwaru Weka, ale hojně se využívá i mimo něj.

Formát arff je v zásadě upravený textový formát, který se skládá ze dvou částí – hlavičky a datového těla. Hlavička obsahuje název relace a seznam atributů včetně jejich datových typů, druhou část tvoří data, zde jsou uloženy řádkové záznamy, hodnoty jsou odděleny čárkou (38). Import tabulek arff podporují všechny zde

porovnávané nástroje vyjma softwaru Orange, zatímco export podporují jen DataMelt a KNIME.

Z hlediska ukládání a načítání modelu se jeví velmi užitečným formát pmml. Jedná se standard, který zaštiťuje a vyvíjí konsorcium Data Mining Group a využívá se k reprezentaci DM modelů, struktura modelu je zde zaznamenána pomocí xml tagů. V jednom souboru může být i více modelů. Díky tomu je možné přenášet modely mezi různými platformami, softwary a celkově jej díky tomu využívat k nasazení výsledků DM procesu (39). Standard pmml podporují ve vybraném vzorku nekomerčních softwarů pouze KNIME a Rattle.

Formát xml implementuje nástroj DataMelt, jde o možnost kódování dat a struktury modelu do jednoho souboru, podporuje i jednoduché grafy (histogramy) (28). Dá se konstatovat, že v omezené míře nahrazuje specializovanější formát pmml.

Výstupní formáty html a pdf v tabulkové příloze Tabulka 1 reprezentují export zpráv z DM procesu. Tyto funkce lze považovat za omezenou formu nasazení výsledků, jsou však využitelné více pro konečného zákazníka než po analytika. Export do pdf umožňuje software DMelt, KNIME a Orange, do html pak Orange a Tanagra.

Výběr, předzpracování a transformace dat

Důležité funkce pro výběr, předzpracování a transformaci dat absentují v nástroji Tanagra. Zbývající softwary disponují možnostmi importování tabulek případně celých datových setů z databází i on-line databází, mají vestavěné funkcionality pro filtrování, seskupování a kombinování dat. Implementovány jsou funkce pro agregaci, diskretizaci, normalizaci, pro nalezení a nahrazení chybějících hodnot, rovněž je k dispozici analýza hlavních komponent (PCA) ke snížení dimenze dat. Toto srovnání je k náhledu v příloze tabulkové příloze Tabulka 2.

Vizualizace

V oblasti vizualizace dat je nejsilnější nástroj DataMelt, a to hlavně díky podpoře zobrazování pokročilých 3D grafů, k čemuž využívá vestavěnou knihovnu jhplot v součinnosti s OpenGL knihovnou JOGL2. Díky tomu také dokáže zobrazovat data interaktivně a umožňuje jejich detailnější průzkum (28). 3D bodové grafy umí

sestavovat také nástroje KNIME, Orange a Rattle. Podpora třídimenzionálního zobrazování zcela chybí v nástroji Tanagra.

Všechny porovnávané nástroje disponují standardními grafy ve 2D (bodový, histogram, sloupcový, spojnicový), mají též k dispozici dendrogram pro vizualizaci shlukové analýzy, umí některé z pokročilejší grafů (například teplotní mapy nebo rovnoběžné souřadnice). Podporu nomogramu pro zobrazení Bayesovského klasifikátoru případně logistické regrese mají tři nástroje.

Rozdíly v oblasti vizualizace nejsou mezi vybranými softwary moc velké, porovnání jednotlivých podporovaných grafů je k nalezení v tabulkové příloze Tabulka 3.

Data mining algoritmy a metody

Pro porovnání možností nekomerčních softwarových nástrojů bylo vybráno celkem 40 algoritmů a metod používaných pro data mining. Z toho 26 algoritmů z oblasti strojového učení, 6 běžných statistických metod, 4 metody pro testování modelů a 4 algoritmy sborového učení. Vybrané metody a algoritmy pokrývají všechny základní typy DM úloh: klasifikaci, numerickou predikci, asociaci a shlukování, dále zahrnují i oblast pro validaci výsledků. Porovnání algoritmů a metod je k dispozici v tabulkové příloze Tabulka 5.

Softwary byly prozkoumány v základních instalacích bez případných rozšíření. Ve srovnávací tabulce Tabulka 5 jsou zaznamenány i možnosti doplňků, pouze však v případech, kdy tato rozšíření oficiálně poskytuje a podporuje vývojář daného softwaru. U nástroje DataMelt je uveden v závorkách integrovaný software, který podporuje danou funkcionalitu.

V oblasti asociačních úloh jednoznačně vládne nástroj Weka, který je integrován v softwaru DMelt. Podporuje všechny čtyři zkoumané algoritmy pro asociační pravidla Apriori, FP-growth, GSP a Tertius. Slabou podporu asociačních pravidel pak mají nástroje Rattle a Tanagra, oba shodně umí metodu Apriori, zbývající nástroje nemají asociační pravidla integrována v základních verzích, ale je možné doinstalovat rozšíření, která je umí.

U Bayesovské klasifikace bylo porovnáváno zastoupení algoritmů AODE, Naivní Bayes a Bayesovské sítě. Opět se ukázalo, že největší zastoupení pro tuto část DM

má software DataMelt v integrovaném nástroji Weka. Podporu pro Naivní Bayesovský klasifikátor je možno nalézt ve všech zkoumaných nástrojích vyjma softwaru Rattle.

Algoritmy pro kombinaci modelů (bagging, boosting, náhodný les a stacking) jsou podporovány v nástroji DMelt v plném rozsahu (a opět ve Weka), zbývající čtyři softwary mají k dispozici minimálně dvě metody pro sborové učení.

Všech pět porovnávaných nástrojů obsahuje algoritmus MLP pro klasifikační úlohy řešené pomocí neuronových sítí; Kohonenovy mapy pro shlukové úlohy mají k dispozici tři nástroje (DMelt, Orange a Tanagra); ostatní typy neuronových sítí (CNN, PNN a RNN) nejsou moc zastoupeny.

Rozhodovací pravidla nejsou ve vybraných softwarech příliš podporována. Byly hledány algoritmy 1Rule, CN2 (případně CN4), PART a RIPPER. Nejvíce z nich má k dispozici DataMelt v nástroji Weka, zastoupen je dále algoritmus CN2 v softwarech Orange a Tanagra, dále PART v Rattle.

Pro klasifikaci byly prozkoumávány tři algoritmy C4.5 (nebo C5.0), CART a ID3. Byly rovněž hledány vlastní implementace pro rozhodovací stromy, a ukázalo se, že tři nástroje je skutečně mají (KNIME, Orange a Rattle). Nejsilnější zastoupení v tradičních metodách pro klasifikaci mají softwary DMelt a Tanagra, ty zároveň nedisponují vlastním rozhodovacím stromem.

Každý ze zkoumaných softwarů podporuje algoritmus SVM, zajímavostí je, že nástroj Tanagra obsahuje dva další algoritmy BVM a CVM, které jsou odvozené právě z toho algoritmu.

Všech pět nástrojů podporuje v oblasti shlukové analýzy metodu pro hierarchické shlukování a algoritmus k-means. Dále byly vyhledávány algoritmy EM, ten mají k dispozici nástroje DataMelt a Tanagra, a c-means (FCM), který obsahuje pouze software KNIME.

V porovnávání šesti statistických metod (ANOVA, diskriminační analýza, faktorová analýza, lineární a logistická regrese, t-testy) byl zjištěn jediný zásadní nesoulad u faktorové analýzy, kterou podporuje jen software Tanagra, nástroj Rattle jako jediný neumí diskriminační analýzu.

Základní metody pro testování modelů (konfusní matice, křížová validace, Lift a ROC křivky) jsou standardem ve všech porovnávaných nástrojích.

Vyjma nástroje Rattle disponují všechny softwary algoritmem k-nejbližší soused.

Nejvíce algoritmů a metod pro data mining obsahuje nástroj DataMelt, tady je však důležité zdůraznit, že tomu tak je díky integrovaným open source nástrojům Weka, Joone a Encog! Co do počtu zahrnutých metod a algoritmů je zajímavý nástroj Tanagra, který lehce stojí nad zbývajícími třemi nástroji Orange, KNIME a Rattle (řazeno sestupně).

Nasazení výsledků

Nejvíce propracované možnosti pro nasazení výsledků má software KNIME, jednak umí exportovat model do formátu pmml, čímž je zaručena možnost nasadit model do praxe, případně využít ke skóringu, dokáže také generovat zprávy vhodné pro zadavatele do pdf. Do specializovaného formátu pmml dokáže model uložit i nástroj Rattle. Oba výše uvedené nástroje jej také umí zpětně načítat. DataMelt, Tanagra a Orange mají funkce pro export zpráv s výsledky modelování.

Možnosti rozšíření

Do nástroje DataMelt je prakticky možno propojit jakoukoliv knihovnu z podporovaných skriptovacích jazyků, možnosti jsou vskutku široké, ale vždy vyžadují znalosti skriptovacího jazyka.

Vývojáři nástroje KNIME nabízí možnosti tzv. rozšíření (extensions), která se zaměřují například na zlepšení konektivity s databázemi, na zpracování textu a obrázků; a dále tzv. integrace (integrations), které například umí doplnit rozhraní pro skriptování v jazycích JavaScript, Python a R, jsou schopné připojit knihovny volně dostupných nástrojů H2O, Hadoop, Keras a Spark. Zajímavostí je, že pokud by byly do nástroje KNIME zahrnuty všechny oficiální doplňky z rozšíření Weka, směle by tento nástroj předběhl ve srovnání i DataMelt. U obou těchto porovnávaných nástrojů z rodiny Java hraje velkou roli právě software Weka s jeho rozsáhlými funkcionalitami v oblasti DM!

Rozšíření funkcionalit (add-ons) je možné i u softwaru Orange, o těchto možnostech bude více referováno v kapitole 6 této bakalářské práce.

Nástroj Rattle a Tanagra nemají žádné jiné možnosti rozšíření než zásahem do zdrojového kódu programu. U prvního jmenovaného však existuje možnost přenést generovaný skript v jazyce R do konzole jazyka R a dále pracovat s rozšířeními tam.

Speciální funkcionality

Dva zde porovnávané nástroje nabízí možnosti pro text mining. Jde o DataMelt, který má skrze integraci Weka dostupnou sadu funkcionalit určenou pro klasifikaci dokumentů. Druhým nástrojem, který dokáže pracovat s textem je Rattle. Rozšířeními pro případné zpracování textu disponují dále nástroje KNIME a Orange.

Pro zpracování velkých dat a pro hluboké učení jsou připraveny doplňky do softwaru KNIME, pro vytěžování obrazových dat a web mining existují rozšíření pro KNIME a Orange. Tyto speciální funkcionality jsou shrnuty v tabulkové příloze Tabulka 4.

Nápověda a komunita

Pro nástroj DataMelt poskytuje hlavní vývojář softwaru společně s iniciativou jWork.org obsáhlou on-line nápovědu na komunitním webu. Rozsáhlou uživatelskou komunitu má software KNIME, na serveru KNIME Hub sdílí uživatelé konkrétní pracovní vlákna řešených DM úloh, přímo v GUI softwaru je okno s navrhovanými akcemi, které pracuje se statistikami uživatelských akcí. Software Orange má komunitní nápovědu na serveru Discord a propracovanou video nápovědu na YouTube. U nástrojů Rattle a Tanagra se jedná spíše o solitérní projekty autorů softwaru, kteří i sami poskytují nápovědu. Tanagra má blog, kde lze najít plno řešených úloh a vzorových datových tabulek, Rattle má podporu pouze na webu vývojáře.

Pořadí

V závěru této kapitoly bylo kvalitativně porovnáno pět softwarových nástrojů pro data mining s pomocí vybraných charakteristik, které jsou detailně zobrazeny v tabulkových přílohách Tabulka 1**,** Tabulka 2**,** Tabulka 3**,** Tabulka 4 a Tabulka 5**.**

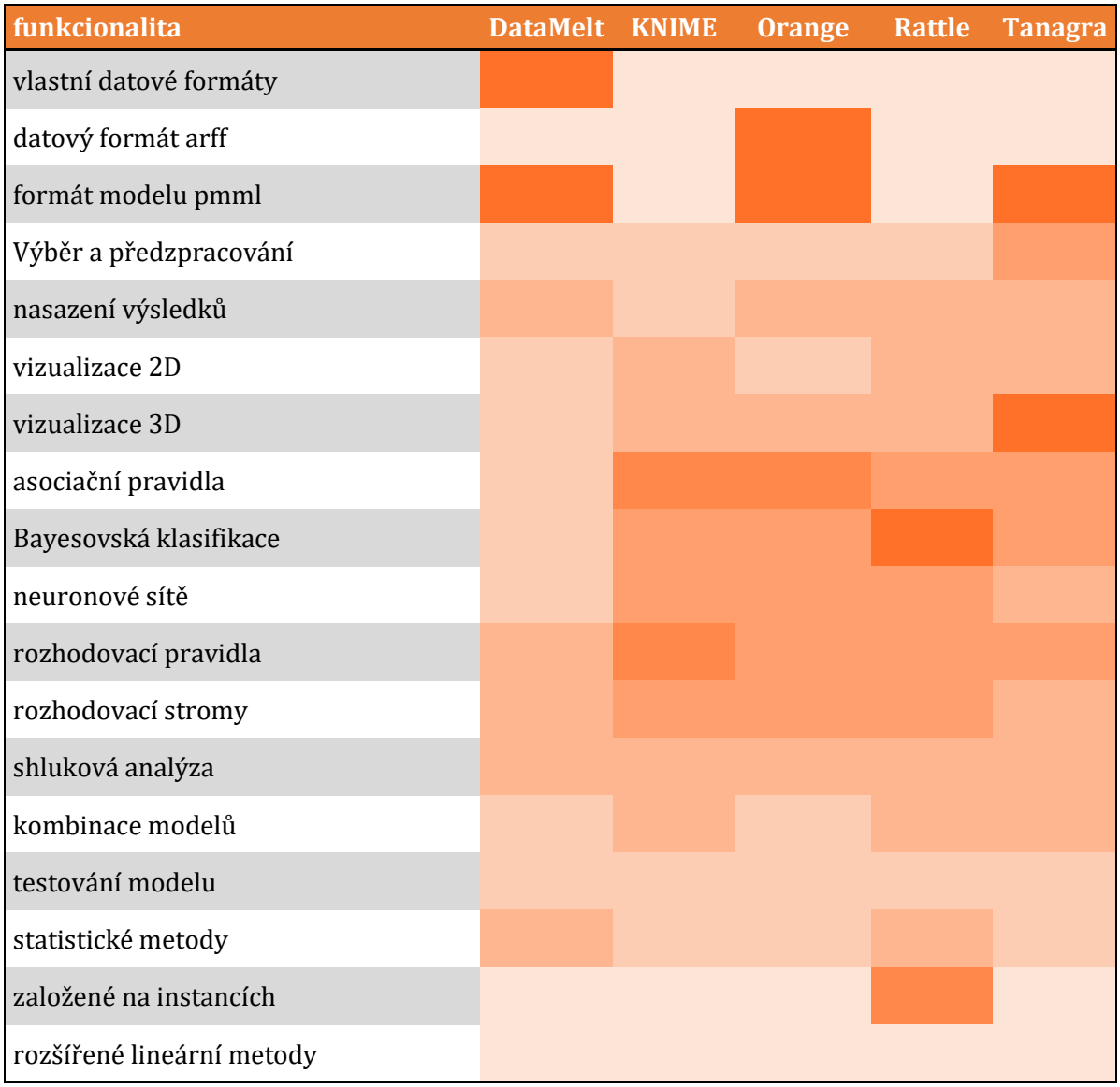
Počet porovnávaných funkcionalit byl 81, z těchto znaků tvořily téměř polovinu algoritmy pro DM, bylo jich 40. Celkové pořadí vybraných softwarových nástrojů při jednoduché sumarizaci všech zkoumaných funkcionalit následuje níže.

1. DataMelt – 76,5 %
2. KNIME Analytics Platform – 74,1 %
3. Orange (49,5) – 61,1 %
4. Rattle a Tanagra – 54,3 %

Při kvantitativní porovnání je vidět, že softwarové nástroje DataMelt a KNIME mají větší odstup od zbývající tří nástrojů.

Každý ze zde prezentovaných softwarů má své výhody i nedostatky, díky bližšímu porovnání je nyní možné vytvořit si představu o funkcionalitách a vybrat si nástroj nejvhodnější pro daný typ DM úloh. K tomu je možné využít souhrnný přehled podpory vybraných funkcionalit zpracovaný pomocí teplotní mapy, která je k dispozici v Tab. 2 na následujíc straně 37.

Tab. 2 Podpora vybraných funkcionalit porovnávaných softwarů



funkcionalita	DataMelt	KNIME	Orange	Rattle	Tanagra
vlastní datové formáty					
datový formát arff					
formát modelu pmml					
Výběr a předzpracování					
nasazení výsledků					
vizualizace 2D					
vizualizace 3D					
asociační pravidla					
Bayesovská klasifikace					
neuronové sítě					
rozhodovací pravidla					
rozhodovací stromy					
shluková analýza					
kombinace modelů					
testování modelu					
statistické metody					
založené na instancích					
rozšířené lineární metody					

Zdroj: Vlastní zpracování

6 Orange

Pro bližší průzkum a následné řešení ukázkové úlohy byl zvolen z porovnaných nekomerčních softwarů nástroj Orange, který patří do rodiny open-source knihoven jazyka Python. V této kapitole jsou popsány základní funkcionality tohoto nástroje, poté následuje zhodnocení možností rozšiřujících balíčků funkcí tzv. add-ons. A z rozšiřujících funkcionalit bude nejvíce představen balíček pro text mining, jehož některé widgety jsou využívány při řešení ukázkové úlohy v kapitole 7.

6.1 Základní funkcionality

Widgety se základními funkcemi jsou v GUI programu roztříděny do pěti kategorií: data, vizualizace, modelování, vyhodnocení a učení se bez učitele, proto bude i následující popis rozdělen takto tematicky. Celá podkapitola o základních funkcionalitách softwaru Orange byla zpracována průzkumem softwaru v GUI a pomocí dokumentace (6).

Data

V záložce Data (obrazová příloha Obrázek 7) jsou základní widgety pro manipulaci s datovými tabulkami, funkcionality pro předzpracování dat a konzole pro možnost vkládání skriptů v jazyce Python.

Orange umí pracovat se základními datovými formáty csv a xls, dále má dva vlastní datové formáty tab a basket (bsk), přičemž druhý je určen speciálně pro použití při text miningu pro tzv. řídká data. V oblasti ukládání/načítání kompletního projektu má nástroj k dispozici vlastní formát ows, dále implementuje formát pickle, což je nativní formát jazyka Python. Orange dokáže načítat data z URL včetně služby Google Sheets, podporuje komunikaci s lokálními i vzdálenými databázemi PostgreSQL a SQL Server.

V oblasti manipulace s datovými tabulkami má Orange standardně k dispozici výběr řádků a sloupců, obarvování dat, prohození sloupců a řádků, spojování datových tabulek, dokáže sestavovat kontingenční tabulky. V předzpracování dat jsou přítomny funkce pro čištění a vzorkování dat, normalizaci, diskretizaci, agregaci,

práci s chybějícími hodnotami. Z dalších funkcí je přítomná možnost znáhodnění dat v datové tabulce.

V balíčku data jsou dále připravené základní metriky pro atributy (rozdělení, rozptyl, minimum a maximum), widget pro skóring klasifikace a regrese (entropie, informační zisk, Gini, ANOVA, χ^2 test), dále funkcionality pro nalezení k-nejbližšího souseda v datech a pro výpočet párového korelačního koeficientu.

Vizualizace

Pod záložkou Vizualizace (obrazová příloha Obrázek 7) jsou k nalezení základní i pokročilé grafy. Z tradičních zobrazovacích metod jsou přítomny histogramy, bodové, krabicové, sloupcové a spojnicové grafy. Orange zároveň nabízí specializovaná zobrazení pro DM, jedná se o grafy pro vizualizaci rozhodovacích stromů, náhodného lesa a rozhodovacích pravidel, graf pro shlukování (silhouette plot), nomogram pro Bayesovský klasifikátor a logistickou regresi, graf lineární projekce pro diskriminační analýzu a PCA. Z ne úplně běžných grafů jsou k dispozici mozaikový a sítový graf, oba pro kontingenční tabulky, Vennovy diagramy pro zobrazení logických vazeb datových tabulek, teplotní mapy k vizualizaci hodnot atributů. Software dále implementuje dva zajímavé interaktivní nástroje pro zobrazování vícerozměrných dat FreeViz a Radviz.

Model

V záložce Model (obrazová příloha Obrázek 7) jsou k dispozici některé z algoritmů strojového učení se s učitelem, dále metody pro kombinování modelů a několik statistických metod.

Pro klasifikační úlohy má Orange k dispozici vlastní algoritmus rozhodovacích stromů Tree, dále využívá algoritmus Naivní Bayes. V softwaru je podporována funkcionalita CN2 Rule Induction pro tvorbu klasifikačních pravidel, konkrétně je implementován algoritmus CN2, jehož výstupem může být rozhodovací list (v pořadí), nebo sada pravidel (bez pořadí).

Ze statistických metod jsou přítomny lineární regrese pro řešení úloh numerické predikce a logistická regrese pro úlohy klasifikační.

Pro kombinování modelů má Orange připraveny tři metody. Nabízí algoritmy sborového učení Náhodný les a AdaBoost, tyto jsou určeny pro klasifikace a regrese, a algoritmus stacking, který je možné použít pro porovnávání různých modelů.

Ke zbývajícím DM metodám, které jsou součástí balíčku Model, patří SVM, neuronové sítí (MLP), k-nejbližší soused (k-NN) a stochastické klesání. Ve widgetu Constant se nachází algoritmus pro nalezení středních hodnot u regresních úloh a nejvíce zastoupených tříd u klasifikačních úloh; widget Calibrated Learner dokáže optimalizovat binární klasifikaci.

V záložce Model je rovněž možné využít funkcionalitu uložení a načítání modelu ve formátu pickle (zde konkrétně pkcls).

Vyhodnocení

Pod záložkou Vyhodnocení (obrazová příloha Obrázek 8) lze nalézt šest widgetů, které jsou zaměřené na validaci výsledků DM. Vedle standardních funkcionalit pro testování modelů, jako jsou matice záměn, ROC křivka, Lift křivka a kalibrační křivka, jsou zde dva widgety, které agregují více funkcionalit, jedná se o widgety Test and Score a Predictions.

Test and Score obsahuje algoritmy pro křížovou validaci, křížovou validaci leave-one-out, náhodný výběr, testování tréninkových dat a testování testovacích dat. Tento widget také zobrazuje metriky například přesnost klasifikace, AUC, přesnost a úplnost. Do widgetu Predictions vždy vstupují data a některý z prediktivní algoritmů, výsledkem je datová tabulka společně s předpovězenými hodnotami.

Učení se bez učitele

V kategorii Učení se bez učitele (obrazová příloha Obrázek 8) jsou zařazeny funkcionality a algoritmy související se shlukováním.

Funkce Distances počítá vzdálenosti mezi řádky, nebo sloupci datové tabulky, výstup pak tvoří matice vzdáleností, pro její vizualizaci je k dispozici mapa vzdáleností. Data z tohoto widgetu mohou být dále použita pro hierarchické shlukování nebo metodu MDS. Spočtené vzdálenosti v matici vzdáleností mohou být dále transformovány (normalizovány a invertovány).

V záložce najdeme několik typů algoritmů pro hierarchické shlukování (Ward, Louvain), také algoritmy pro shlukování k-means a DBSCAN.

Z neuronových sítí určených pro shlukování je k dispozici algoritmus SOM (Self-Organizing Map), který vytváří dvojdimenzionální diskretizovanou reprezentaci dat. K dispozici jsou i další metody mimo oblast neuronových sítí, které slouží ke snížení dimenze dat, jedná se o analýzu hlavních komponent (PCA) a korespondenční analýzu (CA), dále metody t-SNE a MDS.

6.2 Rozšířené funkcionality

Tato část se věnuje rozšířeným funkcionalitám nástroje Orange, které jsou zdarma poskytovány oficiálním vývojářem softwaru. Doplňující tematické balíčky (add-ons) se do programu instalují přes grafické uživatelské rozhraní, nebo přes konzoli jazyka Python. Po nainstalování jsou widgety dostupné pod záložkou, která nese jméno rozšiřujícího balíku. Toto podkapitola věnující se rozšířeným funkcionalitám nástroje Orange byla zpracována průzkumem softwaru v GUI a z dokumentace softwaru (10).

Text mining

Funkcionality v rozšíření Text mining (obrazová příloha Obrázek 8) se dají rozdělit do několika podoblastí. První podoblastí je získávání dokumentů. Orange má díky speciálním widgetům přístup do programových rozhraní pěti databází: The Guardian Open Platform, The New York Times Developer Network, Pubmed, The Twitter Search API a MediaWiki Action API. Textové dokumenty je možné z databází prvních dvou deníků extrahovat na základě data vydání a klíčových slov a je také možné uskutečnit výběr importovaných atributů. U databáze medicínských článků je hledání ještě zpřesněno o autorství a množství importovaných záznamů. Hodně podobně funguje hledání pomocí widgetu pro Twitter. Zatímco u článků na Wikipedii je import dokumentů založen na jednoduchém vyhledávání pomocí klíčových slov.

Další podoblast v rozšíření text mining se týká manipulace s dokumenty a lze sem zařadit funkce pro import dokumentů, náhled dokumentů a předzpracování textu. Připraveny jsou widgety Corpus, ten načítá datové tabulky s textovými korpusy,

Corpus Viewer pro jejich zobrazení, Import Documents, pomocí kterého lze importovat adresáře s dokumenty, kdy adresářová struktura vytváří kategorie dokumentů, Preprocess Text pro předzpracování textu a také Duplicate Detection, který dokáže najít a odstranit duplicity záznamů.

Předzpracování textu probíhá na několika úrovních. Transformace nabízí převedení textů na malá písmena, odstranění diakritiky, odstranění html tagů a URL adres; tokenizace je pro rozdělení textů na malé části (tokeny), obvykle na slova; normalizace se soustředí na převádění tokenů na kořeny (stematizace) nebo základy slov (lemmatizace); filtrování nabízí možnosti vyloučení neužitečných slov (spojky, předložky), výběr množství textu k analýze, výběr počtu nejvíce využívaných tokenů. V předzpracování textu dále najdeme možnost nastavení rozsahu n-gramů, což jsou logická spojení tokenů (například dvojice, trojice atd.).

V podoblasti pro samotnou analýzu textu najdeme v rozšíření pro text mining algoritmy pro modelování témat (Topic Moddeling) a mrak používaných slov (Word Cloud), který spočte výskyt tokenů a zobrazí je graficky. Pro přípravu předzpracovaného textu ke klasifikaci jsou k dispozici dvě metody Bag Of Words a Document Embedding, každá je postavena na odlišném přístupu k převedení tokenů na vektory. Je implementována sada funkcí pro analýzu sentimentu, která obsahuje metody Liu Hu a Vader. Jsou zde widgety hledání kontextu slov (Concordance) a hledání podobností pomocí metody SimHash.

K podpoře vizualizace text miningu je zařazena zajímavá funkce pro zobrazení četnosti dokumentů na mapě světa a možnost převedení korpusu (samotných tokenů nebo celých dokumentů) na síť, které pro zobrazení potřebuje instalaci rozšíření Networks.

Analýza obrazu

V rozšíření pro analýzu obrazu softwaru Orange jsou k dispozici funkce pro manipulaci s obrazovými soubory a jejich následnou klasifikaci. K analýze obrazu využívá software šest modelů pro hluboké učení, z čehož jeden lze využít lokálně a zbývajících pět pouze on-line, kdy se obrázky nahrají na server, kde následně probíhá výpočet a poté jsou výsledky vráceny do GUI softwaru. Základním off-line

modelem je SqueezeNet, který využívá pro trénování web ImageNet. On-line analýzu obrazových dat zajišťují hluboké neuronové sítě Inception v3 od společnosti Google, VGG-16 a VGG-19 vyvíjené na Univerzitě v Oxfordu, tyto tři sítě také využívají pro trénování modelů web ImageNet. Dále jsou zařazeny dvě specializované neuronové sítě Painter, která se učí na sadě 80000 uměleckých děl, a DeepLoc, která se specializuje na buněčné obrázky.

Analýza genetických dat

Jak již bylo zmíněno v kapitole 5.3, Orange je software vyvíjený se specializací na bioinformatiku, díky tomu disponuje možnostmi rozšíření pro zpracování dat z buněčného a genetického výzkumu. Tyto funkcionality jsou k dispozici v rozšiřujících balících Bioinformatics a Single Cell. Součástí těchto balíčků jsou widgety, které slouží k získání dat ze specializovaných databází GEO DataSets, dictyExpress, NCBI Gene, Gene Ontology, KEGG Pathway a PanglaoDB, dále obsahují funkce k vizualizaci těchto dat (sopečný graf), porovnávání genů, specializovanou shlukovou analýzu, skóring genů a buněk. Software Software Orange má k dispozici oddělenou distribuci, která je již v základu vybavená pro druh této specializované analýzy, nese jméno Single Cell, zkráceně scOrange.

Ostatní rozšíření

Software Orange má ještě jednu oddělenou distribuci nazvanou Quasar, která je zaměřená na průzkum dat spektrální analýzy, ta již v základu obsahuje rozšiřující balíček pro spektroskopická data.

Orange má dále připraveny funkcionality pro analýzu časových řad (například finančních nebo předpovědi počasí), využívá k tomu specializované modely ARIMA a VAR, a také vizualizační techniky například pro zobrazení period (kruhová teplotní mapa, periodogram).

Software disponuje dalšími možnostmi rozšíření v oblasti síťové analýzy a tvorby síťových grafů, geokódování a zobrazování dat na mapě světa. Pro data mining má Orange možnost rozšíření v oblast asociačních pravidel, konkrétně lze doinstalovat algoritmus FP-growth a k tomu příslušný zobrazovací widget Frequent Items. Pro studijní účely je připraven balíček Educational, který umožňuje pokročilejší

funkcionality pro propojení s Google Sheets, obsahuje interaktivní zobrazení pro shlukování k-means, klasifikace a regresní analýzy, v oblasti statické vizualizace doplňuje koláčový graf.

Konkrétní funkcionality a možnosti softwaru Orange v oblasti text miningu jsou demonstrovány při řešení ukázkové úlohy, která následuje v kapitole 7.

7 Ukázková úloha

Ukázková úloha je řešena s pomocí metodiky CRISP-DM. Součástí řešení ukázkové úlohy jsou tyto kroky: popis problematické domény, stanovení hlavního cíle a jednotlivých výzkumných hypotéz; popis způsobu výběru a získání dat, popis získaných dat; příprava dat pro analýzu v softwaru Orange; modelování s využitím rozšiřujících funkcionalit pro text mining v softwaru Orange; vyhodnocení získaných výsledků s pomocí nástroje Orange; shrnutí dosažených výsledků a popis možností jejich využití.

7.1 Cíle a výzkumné hypotézy

Cílem ukázkové úlohy je analýza a porovnání informování českých internetových zpravodajských médií (dále média) o českém pracovním trhu v prosinci 2019 a červnu 2020.

Hypotéza

- Média informovala v prosinci 2019 o českém pracovním trhu neutrálně, v červnu 2020 negativně.

Další cíle

- O jakých tématech se nejvíce psalo v prosinci 2019 a červnu 2020?
- Jaké skupiny témat lze v mediálních článcích najít v prosinci 2019 a červnu 2020?

7.2 Získání a základní popis dat

Výběr a získání dat

Data pro tuto ukázkovou úlohu byla získána z celostátních internetových zpravodajských médií. Média byla vybrána dle těchto kritérií:

- zpravodajství,
- deník,
- celostátní,
- informace jsou převážně textové.

Pro účel úlohy bylo vybráno prvních pět nejvíce navštěvovaných médií. Statistika návštěvností zpravodajských webů byla zpracována pomocí projektu NetMonitor (40), data byla získána z kategorie zpravodajství a omezena byla na dvě období 12/2019 a 6/2020.

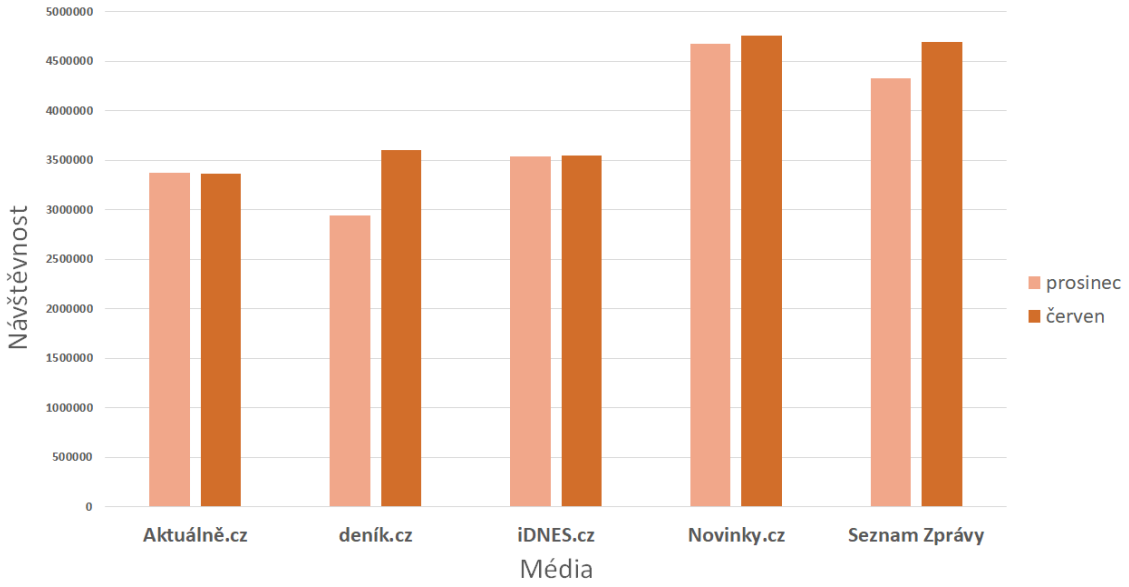
Dle statistik návštěvností bylo zjištěno pořadí pěti nejvíce navštěvovaných zpravodajských webů v měsíci prosinci 2019:

1. Novinky.cz,
2. Seznam Zprávy,
3. iDNES.cz,
4. Aktuálně.cz,
5. deník.cz.

V červnu 2020 bylo pořadí pěti nejvíce navštěvovaných médií následující:

1. Novinky.cz,
2. Seznam Zprávy,
3. deník.cz,
4. iDNES.cz,
5. Aktuálně.cz.

V obou srovnávaných obdobích se tedy na prvních pěti místech nacházejí stejná média jen v obměněném pořadí. Porovnání statistik návštěvnosti je zobrazeno v grafu na Obr. 3 na straně 47.



Obr. 3 Statistiky návštěvnosti médií za měsíce prosinec 2019 a červen 2020

Zdroj: Vlastní zpracování, data návštěvnosti získána z NetMonitor (40)

Na každém zpravodajském webu byl proveden průzkum rubrik. Vybrány byly rubriky zaměřené na zpravodajství pro celou ČR, vyloučeny byly lokální (krajské) rubriky. Na všech pěti webech byly nalezeny podobné rubriky, ty byly určeny pro následnou extrakci článků.

- Aktuálně.cz: domácí, ekonomika a názory
- deník.cz: Česko, ekonomika, komentáře a podnikání
- iDNES.cz: domácí, ekonomika, finance a názory
- Novinky.cz: domácí, ekonomika, finance a komentáře
- Seznam Zprávy: byznys, domácí a názory

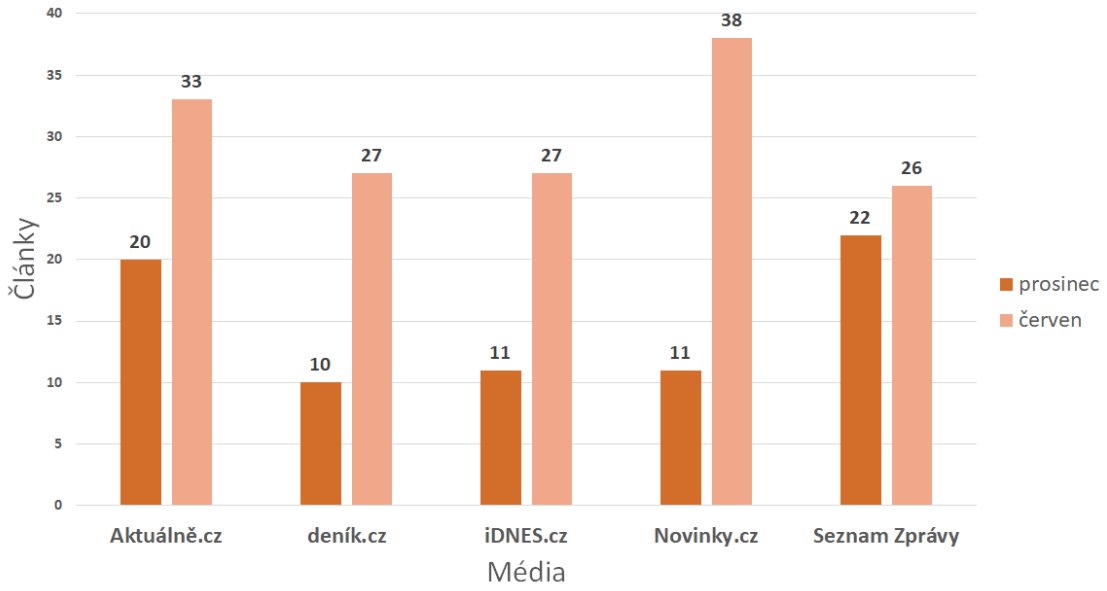
Články byly vybírány dle těchto níže uvedených kritérií.

- Časové omezení: 1. - 31. 12. 2019 a 1. - 30. 6. 2020.
- Klíčová slova: pracovní trh, trh práce, zaměstnanost, nezaměstnanost, pracovní úřad, úřad práce, propouštění, nábor, pracovní místo, pracovní síla (byla zohledněna jednotná i množná čísla, skloňování i pořadí slov).
- Článek musí obsahovat alespoň jedno klíčové slovo.
- Klíčové slovo může být v nadpisu, perexu nebo obsahu článku.

- Klíčové slovo nesmí být v přidaných informacích na html stránce (reklama, odkazy na jiné články, tabulky, grafy, ankety apod.).

Vyhledávání článků proběhlo pomocí integrovaných nástrojů vyhledávání na čtyřech zpravodajských webech (Novinky.cz, deník.cz, iDNES.cz a Aktuálně.cz) (41), (42), (43) a (44); zpravodajský web Seznam Zprávy neumožňuje pokročilé vyhledávací funkce, proto byl použit externí nástroj Google – rozšířené vyhledávání (45) s omezením na doménu seznamzpravy.cz.

Obsah článků byl získán pomocí manuální extrakce z html kódu webových stránek, díky tomu došlo k hrubému očistění dat od html tagů, zachovány zůstaly odkazy URL. Získáno bylo celkem 225 článků, jejich počty dle zdroje a období jsou zobrazeny v grafu na Obr. 4.



Obr. 4 Počty získaných článků z médií za měsíce prosinec 2019 a červen 2020
Zdroj: Vlastní zpracování

Popis dat

Získané články byly uloženy do datové tabulky, která má osm níže uvedených atributů.

- Id obsahuje číslo záznamu, jedná se o datový typ INTEGER.

- Datum nese datum zveřejnění článku, jde o datový typ DATE, který nabývá hodnot 1.12.2019 - 31.12.2019 a 1.6.2020 - 30.6.2020.
- Titulek, perex a obsah jsou textové atributy, jde o datový typ STRING. Tyto tři atributy budou podrobovány textové analýze.
- Zdroj, měsíc a sentiment jsou kategorické atributy. Zdroj tvoří zkratka média, v němž byl článek publikován, může nabývat pěti různých hodnot; měsíc nabývá dvou hodnot (červen a prosinec), je odvozen z atributu datum; sentiment může nabývat třech hodnot, které tvoří zkratky škály sentimentu (pozitivní, neutrální, negativní). Všechny tři tyto atributy jsou datovým typem STRING.

Rozdělení dat

Datová tabulka byla pro řešení úlohy rozdělena na dvě části: trénovací a testovací data. Trénovací data jsou tvořena 2/3 záznamů; testovací data obsahují zbývající 1/3. V trénovací sadě byl s ohledem na nevyvážený počet článků ze dvou různých období dodržen poměr 1 : 2 příspěvků z prosince 2019 a června 2020. Trénovací data budou dále manuálně anotována do kategorií sentimentu (pozitivní, neutrální, negativní), poté budou použita k sestavení a natrénování klasifikačního modelu. Testovací data budou pomocí sestaveného klasifikátoru predikována do jednotlivých tříd sentimentu. Na testovacích datech přiřazených do tříd sentimentu bude vyhodnocena hlavní hypotéza ukázkové úlohy.

7.3 Příprava dat

Předzpracování dat a volba příznaků textových proměnných

V software Orange byly vytvořeny tři textové korpusy:

- trénovací se 148 články,
- testovací se 77 články,
- kompletní s 225 články.

Data do korpusů byla importována z připravených datových tabulek, které jsou uvedeny v kapitole 7.2. Do korpusů bylo vybráno pět atributů, které budou využívány k tvorbě modelu: sentiment, id, měsíc, titulek, perex a obsah. Proměnné titulek, perex a obsah budou podrobeny textové analýze, proto je nutné

předzpracovat a přiřadit jim tzv. příznaky (features). Tímto dojde k vyjádření textu ve formě vektorů, které jsou využívány v algoritmech strojového učení (46). K procesům, které přiřazují textu příznaky patří: transformace, tokenizace, normalizace (stematizace a lemmatizace), filtrování, vytvoření n-gramů a POS tagging (určení slovních druhů) (6). K volbě příznaků byl v softwaru Orange použit widget Preprocess Text.

V oblasti transformace textových dat byla provedena detekce a očištění zbytkových html kódů, které zůstaly v textech po nedokonalé extrakci z webových stránek, byly také odstraněny URL adresy. Velká písmena byla převedena na malá, česká diakritika zůstala zachována.

Následně byly texty rozděleny na slova. Malé textové části se v text miningu nazývají tokeny, proces dělení pak tokenizace. Tokeny byly dále normalizovány na základní tvary slov tzv. lemmata, k tomuto byl využit v softwaru Orange implementovaný lemmatizátor UDPipe s českým slovníkem. Tento nástroj je vyvíjen na Ústavu formální a aplikované lingvistiky (47).

Poté došlo k filtraci pomocí stop slov, což jsou slova, která díky vysoké frekvenci výskytu v textu ztrácejí význam pro analýzu, jedná se obvykle o spojky, předložky a interpunkční znaménka (48). K této filtraci byl využit vlastní slovník stop slov, který byl vytvořen na základě stoplistu z Centra pro zpracování přirozeného jazyka (48) a doplněn o několik dalších výrazů.

Byly vytvořeny n-gramy do maximálního počtu dvou slov.

Software Orange má k dispozici pro klasifikaci textů dva widgety, každý z nich představuje odlišný způsob práce s tokeny.

- Bag of Words – počítá výskyt tokenů (n-gramů) pro každou instanci (dokument, zde článek). Počet může být vyjádřen absolutním výskytem, binárním (je přítomen / není přítomen), nebo logaritmicky (6).
- Document Embedding – vytváří mnoharozměrný vektor pro každou instanci (dokument, zde článek), k tomu používá natrénovaný model fastText (6).

K převzetí příznaků z předzpracovaného textu a následné vektorizaci byl vybrán widget Bag of Words, protože Document Embedding v nástroji Orange nedisponuje českým slovníkem.

Anotace korpusu trénovacích dat

Pro anotaci tříd sentimentu trénovacího korpusu byly použity dva slovníky českých subjektivních výrazů.

- Czech SubLex od Veselovské a Bojara (49), ten obsahuje přes 4500 slov rozdělených na dvě polarity (negativní a pozitivní) (50);
- AFFIN.CZ od Řezníčka (51), který obsahuje téměř 20000 slov rozčleněných na dvě polarity (negativní a pozitivní) (52).

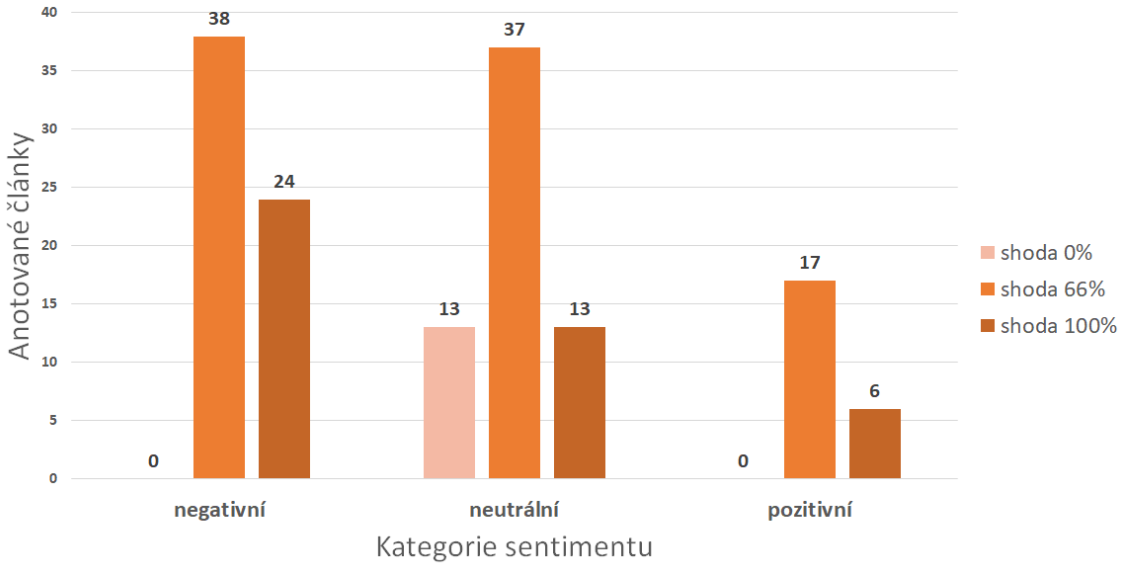
Předzpracovaný korpus kompletních dat byl podroben průzkumu na úrovni četnosti tokenů, k tomu byl využit widget Word Cloud, který spočte výskyt jednotlivých tokenů v celém korpusu a seřadí je od nejfrekventovanějších po nejméně časté. Pro tento účel byl prozkoumán obsah textových proměnných titulek a perex. Z prvních 200 nejfrekventovanějších slov byl exportem vytvořen seznam, v němž byly následně vyhledány všechny subjektivní výrazy, zahrnut byl každý výraz, který byl nalezen alespoň v jednom ze slovníků (49), (51). Z nalezených subjektivních slov byl vytvořen vlastní slovník subjektivních výrazů. Z 200 ověřovaných tokenů zůstalo 36 termínů s polaritou negativní a pozitivní.

Tato citově zabarvená slova byla následně vyhledávána pomocí widgetu Corpus Viewer v titulcích a perexech článků předzpracovaného trénovacího korpusu. Negativnímu výrazu byla přiřazena hodnota -1, pozitivnímu +1. Článek byl vyhodnocen jako pozitivní, pokud součet hodnot nalezených výrazů dosáhl kladné hodnoty, pokud záporné, byl přiřazen do kategorie negativní. Kategorie neutrální byla přiřazena, pokud součet dosáhl hodnoty 0. Tímto způsobem se nepodařilo anotovat 31 případů, kdy nebyl v článku nalezen žádný výraz, u těchto případů byla kategorie nastavena jako neutrální. Z tohoto důvodu byla anotace vzniklá na základě vlastního slovníku dále porovnána s anotacemi dvou hodnotitelů.

Dva nezávislí hodnotitelé zařadili každý článek z trénovacího korpusu do jedné z kategorií: pozitivní, neutrální, negativní. Přiřazení do kategorie sentimentu probíhalo na základě četby nadpisu a perexu článku. Ke kategorizaci nevyužívali hodnotitelé žádný podpůrný slovník, jednalo se tedy o subjektivní hodnocení.

Tři hodnoty anotací (metoda vlastního slovníku a dva hodnotitelé) byly nakonec vzájemně porovnány. Absolutní shoda nastala v 29 % článků, shoda ve dvou případech nastala u 62 % článků, ve zbývajících 9 % nedošlo k žádné shodě. Výsledná kategorie sentimentu byla přiřazena na základě absolutní nebo 2/3 shody, pokud nedošlo k žádné shodě, byla přiřazena kategorie neutrální. Počty článků rozřazených do kategorií sentimentu dle shody hodnotitelů jsou zobrazeny v grafu na Obr. 5.

Kategorie sentimentu byla přiřazena ve všech 148 záznamech trénovacího korpusu do proměnné sentiment. Tímto byl korpus trénovacích dat připraven k sestavení trénovacího modelu.



Obr. 5 Shoda hodnotitelů při anotaci trénovacích dat do kategorií sentimentu

Zdroj: Vlastní zpracování

7.4 Modelování

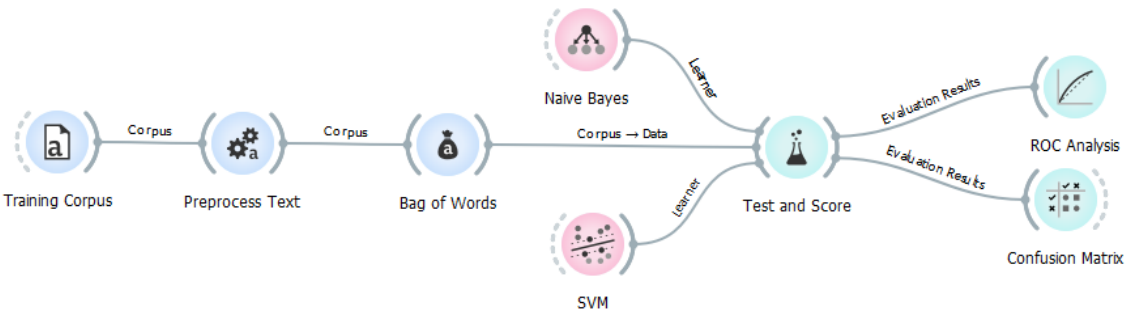
Sestavení klasifikačního modelu pro analýzu sentimentu

Analýza sentimentu v této ukázkové úloze je řešena pomocí klasifikační metody učení se s učitelem. Pro výběr do klasifikačního modelu byly nejprve porovnány na testovacím korpusu tyto dvě metody:

- SVM (Support Vector Machine),
- Naivní Bayes,

které jsou vedle metody maximální entropie často zmiňovány v odborné literatuře, jako algoritmy vhodné pro analýzu sentimentu (50), (53). Tyto dvě metody byly také vybrány s ohledem na jejich implementaci v softwarovém nástroji Orange, který je používán při řešení úlohy.

Oba algoritmy jsou v trénovacím modelu implementovány společně s widgetem Bag Of Words, jak je vidět na Obr. 6. Jejich úspěšnost byla otestována pomocí křížové validace a vzájemně byly porovnány pomocí třech vybraných metrik: AUC – oblast pod křivkou ROC, CA – přesnost klasifikace a F1 – f-skóre, které zohledňuje hodnoty přesnosti (precision) a úplnosti (recall).



Obr. 6 Orange workflow – porovnání klasifikátorů na trénovacím korpusu

Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obě metody byly ověřovány na trénovacím korpusu nejprve za použití všech tří textových atributů (titulek, perex a obsah), následně pouze na dvou (titulek, perex), poté jen na atributu obsah. Oba algoritmy dosáhly nejlepších výsledků při klasifikaci za použití atributů titulek a perex. Ve finálním modelu budou v trénovacím korpusu

využity pouze tyto dva textové atributy. Oba algoritmy byly také testovány pomocí widgetu Bag of Words, ukázalo se, že obě metody dosahují nejlepších výsledků při binárním vyjádření počtu tokenů. Do finálního modelu bylo zvoleno toto nastavení. Souhrn metrik při změnách vstupních parametrů textových atributů trénovacího korpusu a widgetu Bag of Words je k nahlédnutí v tabulkové příloze Tabulka 6.

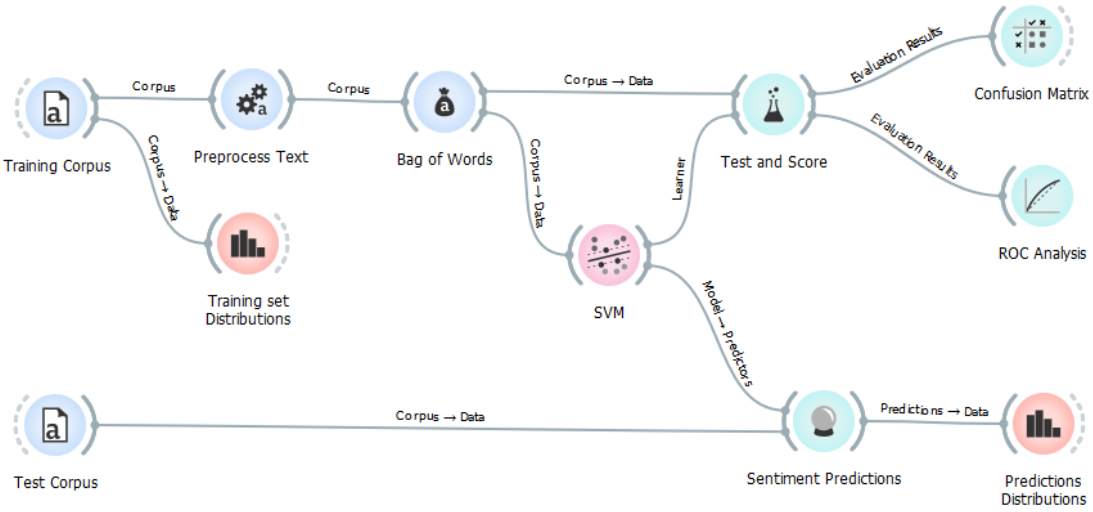
Metoda SVM dosáhla na trénovacím korpusu lepších výsledků než Naivní Bayes. Algoritmus SVM výrazně lépe klasifikoval příklady do tříd sentimentu, Naivní Bayes měl velké množství nesprávně klasifikovaných případů, jak je vidět v konfusních maticích obou metod v obrazové příloze Obrázek 9 a Obrázek 10. Porovnání ROC křivek obou metod na trénovací množině ve všech třech třídách sentimentu je k dispozici v obrazové příloze Obrázek 11, Obrázek 12 a Obrázek 13. Klasifikátor Naivní Bayes dosáhl na trénovacím korpusu neuspokojivých výsledků okolo 30% přesnosti klasifikace a f-skóre, zatímco SVM téměř 60 %. Tyto vybrané metriky jsou zaznamenány v Tab. 3.

Tab. 3 Porovnání metrik klasifikátorů

korpus	metoda	AUC	CA	F1
trénovací	Naivní Bayes	0,727	0,311	0,271
trénovací	SVM	0,761	0,595	0,579
testovací	SVM	0,725	0,649	0,631
kompletní	SVM	0,865	0,747	0,740

Zdroj: Vlastní zpracování výstupů ze softwaru Orange
Poznámka: AUC – oblast pod křivkou ROC, CA – přesnost klasifikace a F1 – f-skóre, které zohledňuje hodnoty přesnosti (precision) a úplnosti (recall).

Do finálního modelu byla vybrána metoda SVM, která byla také použita pro predikci tříd sentimentu testovacího korpusu. Schéma výsledného klasifikačního modelu je zobrazeno na Obr. 7. Výsledný model s algoritmem SVM byl po natrénování na trénovacím korpusu se 148 články využit k predikci tříd sentimentu testovacího korpusu se 77 články. Testovací korpus s predikovanými třídami a kompletní korpus s anotovanými i predikovanými třídami sentimentu byl opět klasifikován metodou SVM. Nejlepších výsledků bylo dosaženo na kompletním korpusu, kde přesnost klasifikace a f-skóre dosáhly 74% úspěšnosti, jak může být vidět v Tab. 3.



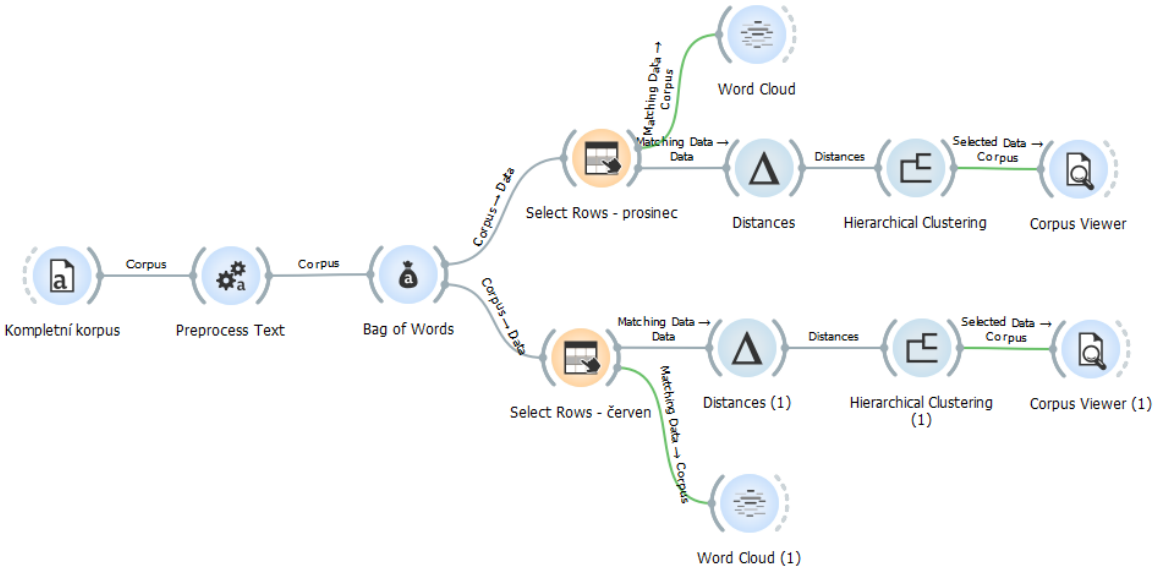
Obr. 7 Orange workflow – výsledný klasifikační model

Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Lepších výsledků by se pravděpodobně dalo docílit rozšířením trénovacího korpusu, kde by se například problematická doména z pěti internetových zpravodajských médií rozšířila například na deset a bylo by extrahováno více článků k textové analýze. Toto naznačuje provedená klasifikace na kompletním korpusu dat, kde jsou u všech záznamů přiřazeny hodnoty sentimentu a dosahuje nejlepších výsledků klasifikace.

Klasifikace do neznámých tříd, hledání okruhů témat a témata

Vedle sestavení klasifikačního modelu pro analýzu sentimentu byly v této ukázkové úloze využívány i metody učení se bez učitele, které umí klasifikovat text do předem neznámých tříd a díky tomu je umožněno najít zajímavé okruhy témat. K tomu byla zvolena metoda pro hierarchické shlukování. Pro získání přehledu o základních tématech, tedy o četnosti tokenů, a jejich vizualizaci, byl použit widget Word Cloud. Schéma klasifikace do neznámých tříd v softwaru Orange je zobrazeno na Obr. 8.



Obr. 8 Orange workflow – model klasifikace do neznámých tříd

Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Do modelu pro klasifikaci do předem neznámých tříd byl vybrán kompletní korpus, který obsahoval všech 225 záznamů, z něj byly k textové analýze využity atributy titulek a perex. Korpus byl předzpracován a poté byl pomocí výběru řádků vyfiltrován na dvě části – články z prosince 2019 (74 záznamů) a června 2020 (151 záznamů). Klasifikace dále probíhala na každé množině dat odděleně.

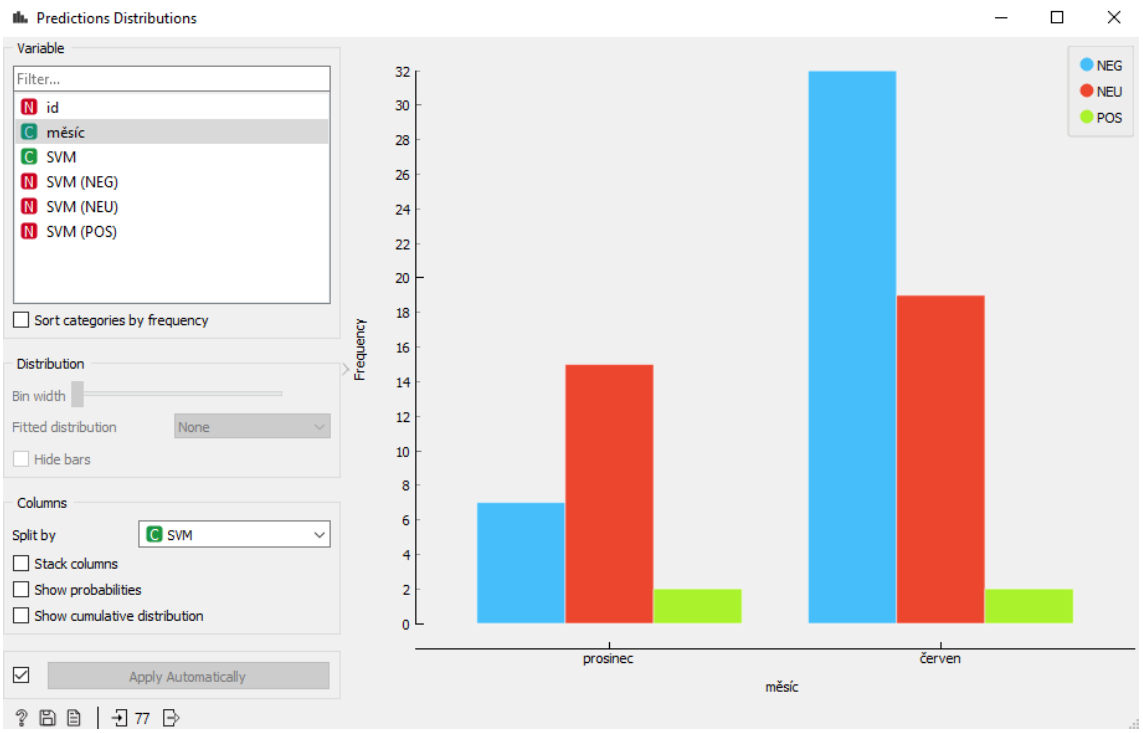
Widget Bag of Words byl použit pro výpočet absolutního výskytu tokenů, poté byla spočtena vzdálenost řádků a pomocí metody hierarchického shlukování (vybrána byla konkrétně metoda Ward) byl vytvořen dendrogram. Bližším průzkumem dendrogramu a jednotlivých shluků v součinnosti s widgetem Corpus Viewer byly nalezeny zajímavé okruhy témat v jednotlivých časových obdobích. Do srovnávací tabulky Tab. 4 na straně 58 bylo vybráno vždy prvních pět shluků z daného období, ty reprezentují hlavní okruhy témat pro články z každého měsíce.

Na dendrogramech v obrazových přílohách Obrázek 16 a Obrázek 17 je vizualizováno vždy prvních pět shluků daného souboru. V každém období se vždy nachází jeden větší shluk (C5 v prosinci 2019, C4 v červnu 200), v němž je koncentrováno výrazně více článků, u takových shluků bylo obtížné najít jednotící okruh témat, i z tohoto důvodu je u některých klastrů uvedeno více témat, jak může být vidět v Tab. 4 na straně 58.

Pomocí widgetu Word Cloud byly spočteny a zobrazeny nejfrekventovanější tokeny a n-gramy, pro tento účel byly zvoleny do maximálního počtu dvou tokenů, jednalo se tedy o bigramy. Díky této jednoduché analýze byla nalezena hlavní témata, o nichž se v médiích informovalo v měsících prosinec 2019 a červen 2020. V Tab. 5 je zobrazeno prvních pět témat z každého období, témata jsou rozdělena na jednoslovná – tokeny a dvouslovná – bigramy. Vizualizace prvních 100 jednoslovných témat z každého období, tzv. mrak slov, je k dispozici v obrazové příloze Obrázek 18 a Obrázek 19.

7.5 Vyhodnocení

Predikce tříd sentimentu testovacího korpusu pomocí sestaveného klasifikačního modelu potvrdila s přesností klasifikace 64 % hypotézu, že média informovala v prosinci 2019 o českém pracovním trhu neutrálně a v červnu 2020 negativně. Výsledné rozdělení predikovaných tříd sentimentu je zobrazeno v grafu na Obr. 9.



Obr. 9 Predikované třídy sentimentu testovacího korpusu

Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Rozdělení tříd sentimentu v trénovacím i kompletním korpusu je podobné, jak může být vidět v obrazové příloze na Obrázek 14 a Obrázek 15.

Textovou analýzou kompletního korpusu článků, který byl pro tento účel rozdělen na dvě části dle období, bylo pomocí metody hierarchického shlukování nalezeno prvních pět okruhů témat z období prosinec 2019 a červen 2020. Okruhy nalezených témat jsou zobrazeny v Tab. 4.

Tab. 4 Vybrané okruhy témat v prosinci 2019 a červnu 2020

shluk	prosinec 2019	články	červen 2020	články
C1	Zvýšení rodičovského příspěvku	4	Nezaměstnanost v květnu	9
C2	Nezaměstnanost v listopadu, vývoj trhu práce	11	Smartwings	18
C3	Průměrná a důstojná mzda, migrace	7	Státní rozpočet a schodek	11
C4	Odebrání dávek	2	Propad ekonomiky, pracovní místa, propouštění	91
C5	Propouštění, prognózy 2020, Česká pošta	50	Sociální odvody, podpora (sociální, pracovní), kurzarbeit, výživné	22

Zdroj: Vlastní zpracování výstupů ze softwaru Orange

Pomocí widgetu Word Cloud byly na kompletním předzpracovaném korpusu článků, rozděleném po měsících, spočteny a vizualizovány nefrekventovanější tokeny a bigramy. Prvních pět tokenů a bigramů z každého období reprezentuje hlavní témata. Nalezená témata z prosince 2019 a června 2020 jsou zobrazena v tabulce Tab. 5.

Tab. 5 Vybraná témata v prosinci 2019 a červnu 2020

typ	prosinec 2019	výskyt	červen 2020	výskyt
token	rok	51	práce	77
	člověk	37	mít	64
	práce	35	člověk	59
	procento	23	procento	51
	koruna	22	firma	49
bigram	úřad práce	9	miliarda koruna	19
	tisíc člověk	7	úřad práce	18
	pracovní místo	5	skupina smartwings	18
	trh práce	5	ministryně práce	14
	nezaměstnanost procento	5	práce sociální	13

Zdroj: Vlastní zpracování výstupů ze softwaru Orange

Některé okruhy témat a témata v jednotlivých obdobích vzájemně korespondují, případně se doplňují. Například v prosinci 2019 byl nalezen okruh „nezaměstnanost v listopadu, vývoj trhu práce“ a v tématech pak bigramy „nezaměstnanost procento“

a „trh práce“, dá se také usuzovat, že o trhu práce a nezaměstnanosti se informuje pomocí nalezeného tokenu „procento“. V červnu 2020 pak můžeme v okruhu témat vidět „státní rozpočet a schodek“ a k tomu odpovídající bigram „miliarda koruna“.

7.6 Využití

Klasifikační model pro analýzu sentimentu vytvořený při řešení této ukázkové úlohy by mohl být využit i na jiné příklady z vybrané problematické domény – tedy internetového zpravodajství, které se zaměřuje na informace o pracovním trhu.

Klasifikační model analýzy sentimentu použitý v této úloze však nedosahuje příliš velké přesnosti, nejspíše i z důvodu nedostatečného rozsahu anotovaných testovacích dat, pro jeho lepší výkon by bylo třeba jej rozšířit. Dalším potenciálně důležitým aspektem v tomto modelu, který může ovlivnit přesnost klasifikace, je samotná anotace článků testovacího korpusu do tříd sentimentu. Pro ni byla snaha nalézt objektivní metodu s tvorbou vlastního slovníku, ale ta nedokázala anotovat celou množinu testovacích článků, proto byla přidána dvě další subjektivní hodnocení, která dokázala přiřadit třídy sentimentu pro všechny články testovacího souboru.

V průběhu tvorby modelu analýzy sentimentu bylo zjištěno, že klasifikace do tříd sentimentu dosahuje lepších výsledků, pokud jsou klasifikovány kratší textové atributy, v tomto případě pouze titulky a perexy článků.

8 Shrnutí a závěr

V teoretické části bakalářské práce bylo popsáno, co je to data mining a k čemu slouží. Bylo objasněno, že DM patří do širší skupiny procesů zvaných dobývání znalostí z databází. Byly přiblíženy dvě základní metodiky CRISP-DM a SEMMA, které se využívají v procesu dobývání znalostí z databází a byl nastíněn vývoj metodik nových. Práce objasnila rozdělení dat na strukturovaná a nestrukturovaná a stručně popsala základní zdroje dat. Dále se práce zaměřila na popis přípravy dat, která předchází aplikaci analytických metod a je velmi důležitá v celém procesu KDD. Následně byly rozebrány typy DM úloh, s nimiž je možno se běžně setkat. Poté byly přiblíženy konkrétní statistické metody a některé z běžných metod strojového učení, které se využívají pro analýzu dat v data miningu. Nakonec byly také představeny základní způsoby testování a kombinování modelů.

Ve výzkumné části bakalářské práce bylo vybráno, popsáno a porovnáno pět nekomerčních softwarů DataMelt, KNIME, Orange, Rattle a Tanagra, které jsou vhodné pro DM. Každý software byl krátce představen. Poté byly softwary porovnány pomocí několika vybraných charakteristik. Jednalo se o datové formáty, funkcionality pro výběr, předzpracování a transformaci dat, nasazení výsledků, způsoby vizualizace, algoritmy pro data mining a podporu specifických funkcionalit (například text a web mining).

Detailněji byl popsán software Orange. Byly rozebrány jeho možnosti v oblasti datových formátů a formátů modelu, získání a předzpracování dat a vizualizace. Byly popsány všechny obsažené funkcionality pro data mining, včetně metod pro kombinování a vyhodnocení modelů. Následně byly představeny rozšířené funkce pro analýzu obrazových a genetických dat a také, s ohledem na řešenou ukázkovou úlohu, funkcionality v oblasti text miningu.

Pomocí softwaru Orange byla v poslední části této práce řešena ukázková úloha z oblasti text mining, konkrétně analýzy sentimentu. Úloha se soustředila na způsob informování českých internetových zpravodajských médií.

9 Použitá literatura

(1) BERKA, Petr. *Dobývání znalostí z databází*. Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.

(2) SKALSKÁ, Hana. *Data mining a klasifikační modely*. Vyd. 1. Hradec Králové: Gaudeamus, 2010, 154 s. Recenzované monografie, 4. ISBN 978-80-7435-088-7.

(3) WITTEN, I., Eibe FRANK a Mark HALL. *Data mining: practical machine learning tools and techniques*. 3rd ed. Burlington: Morgan Kaufmann, 2011, 629 s. Morgan Kaufmann series in data management systems. ISBN 978-0-12-374856-0.

(4) *DataMelt: computation & visualisation* [online]. [Illinois, US]: jWork, c2005-2020 [cit. 2020-07-09]. Dostupné z: https://datamelt.org/

(5) *KNIME documentation* [online]. Zurich: KNIME AG, 2018 [cit. 2020-07-22]. Dostupné z: https://docs.knime.com

(6) *Orange visual programming: [documentation]* [online]. [Lublaň]: Orange Data Mining, 2015 [cit. 2020-08-22]. Dostupné z: https://orange-visual-programming.readthedocs.io/

(7) *Rattle: a graphical user interface for data minining using R* [online]. [Australia]: Togaware, c2006-2020, last modified 2020-08-09 [cit. 2020-08-15]. Dostupné z: https://rattle.togaware.com

(8) *Tanagra: a free data mining software for teaching and research* [online]. Lyon: Ricco Rakotomalala, 2008 [cit. 2020-07-09]. Dostupné z: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

(9) JOVIC, A., K. BRKIC a N. BOGUNOVIC. An overview of free software tools for general data mining. In: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*

[online]. [Opatia, Croatia]: IEEE, 2014, s. 1112-1117 [cit. 2020-07-18]. ISBN 978-953-233-077-9. Dostupné z: doi:10.1109/MIPRO.2014.6859735

(10) *Orange: data mining fruitful and fun* [online]. Ljubljana: University of Ljubljana, 2020 [cit. 2020-07-30]. Dostupné z: https://orange.biolab.si/

(11) The Knowledge Discovery in Databases (KDD) process [scheme]. In: *ReasearchGate* [online]. Berlin: ResearchGate, 2019 [cit. 2019-07-29]. Dostupné z: https://www.researchgate.net/figure/The-Knowledge-Discovery-in-Databases-KDD-process_fig1_274425359

(12) *CRISP-DM: cross industry standard process for data mining* [online]. London: Smart Vision Europe, 2015 [cit. 2019-07-20]. Dostupné z: http://crisp-dm.eu/

(13) PIATETSKY, Gregory. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KD nuggets: Machine Learning, Data Science, Data Mining, Big Data, Analytics, AI* [online]. [Massachusetts, US]: KDnuggets, 2019 [cit. 2019-07-31]. Dostupné z: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

(14) Data mining phases [scheme]. In: *CRISP-DM: cross industry standard process for data mining* [online]. London: Smart Vision Europe, 2015 [cit. 2019-07-20]. Dostupné z: http://crisp-dm.eu/reference-model/

(15) Introduction to SEMMA. *SAS Enterprise Miner 15.1: reference help* [online]. Cary: SAS Institute, 2018 [cit. 2019-07-20]. Dostupné z: https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=15.1&locale=en

(16) IBM to acquire SPSS Inc. to provide clients predictive analytics capabilities. *IBM newsroom* [online]. Armonk: IBM, 2019 [cit. 2019-07-20]. Dostupné z: https://www-03.ibm.com/press/us/en/pressrelease/27936.wss

(17) CRISP-DM in IBM SPSS Modeler. *IBM Knowledge Center: SPSS Modeler 18.2.1 documentation* [online]. Armonk: IBM, 2019 [cit. 2019-07-20]. Dostupné z: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_18.2.1/modeler_crispdm_ddita/clementine/crisp_help/crisp_using_in_clem.html

(18) *Data Science Project Management* [online]. [Phoenix, AZ]: Saltz, Hotz, 2019 [cit. 2019-07-31]. Dostupné z: http://www.datascience-pm.com/

(19) MYATT, Glenn. *Making sense of data: a practical guide to exploratory data analysis and data mining*. Hoboken: Wiley, 2007, 280 s. ISBN 978-0-470-07471-8.

(20) Top 12 common problems in data mining. *Big data made simple* [online]. [Singapore]: Crayon Data, 2020 [cit. 2020-11-14]. Dostupné z: https://bigdata-madesimple.com/12-common-problems-in-data-mining/

(21) JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An introduction to statistical learning: with applications in R* [online]. New York: Springer, 2013 [cit. 2020-11-14]. Springer texts in statistics. ISBN 978-1-4614-7137-0.

(22) PADHY, Neelamadhab. The survey of data mining applications and feature scope. *International Journal of Computer Science, Engineering and Information Technology* [online]. 2012, **2**(3), 43-58 [cit. 2020-11-11]. ISSN 22313605. Dostupné z: doi:10.5121/ijcseit.2012.2303

(23) PETR, Pavel. *Metody data miningu*. Část I. Vyd. 1. Pardubice: Univerzita Pardubice, 2014, , 85 s. ISBN 978-80-7395-872-5.

(24) WILLIAMS, Graham. *Data mining: desktop survival guide* [online]. Camberra (Australia): Togaware, c2004-2010 [cit. 2020-08-25]. Dostupné z: https://www.togaware.com/datamining/survivor/

(25) Software suites/platforms for analytics, data mining, data science, and machine learning. *KDnuggets* [online]. [Illinois, US]: KDnuggets, 2020 [cit.

2020-08-06]. Dostupné z: https://www.kdnuggets.com/software/index.html

(26) PIATETSKY, Gregory. Python leads the 11 top data science, machine learning platforms: trends and analysis. In: *KDnuggets* [online]. [Massachusetts, US]: KDnugetts, 2020 [cit. 2020-08-06]. Dostupné z: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

(27) Software: DataMelt. In: *HandWiki: Encyclopedia of Science and Computing* [online]. [Illinois, US]: jWork, 2019 [cit. 2020-08-05]. Dostupné z: https://handwiki.org/wiki/Software:DataMelt

(28) *DataMelt online manual* [online]. [Illinois, US]: jWork, 2020 [cit. 2020-07-09]. Dostupné z: https://handwiki.org/wiki/DMelt:Start

(29) IAN H., Witten, Frank EIBE, Hall MARK A. a Pall CHRISTOPHER J. *The Weka workbench: online appendix for "Data mining: practical machine learning tools and techniques" Morgan Kaufmann, fourth edition, 2016* [online]. [4th ed.]. [Amsterdam]: [Morgan Kaufmann], 2016 [cit. 2020-08-24]. ISBN 978-0123748560. Dostupné z: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

(30) Encog machine learning framework. *Heaton research* [online]. [Washington]: Heaton Research, Inc., 2020 [cit. 2020-08-24]. Dostupné z: https://www.heatonresearch.com/encog/

(31) *Joone: Java object oriented neural engine* [online]. [s.l.]: Paolo Marrone and the Joone team, 2004, last published 07/23/2017 [cit. 2020-08-24]. Dostupné z: https://www.jooneworld.com/

(32) *KNIME: open for innovation* [online]. Zurich: KNIME AG, 2018 [cit. 2020-07-22]. Dostupné z: https://www.knime.com

(33) DEMŠAR, Janez a Blaž ZUPAN. Orange: data mining fruitful and fun : a historical perspective. *Informatica: an international journal of computing and*

informatics. Ljubljana: Slovensko društvo Informatika, 2013, **37**(1), 55-60. ISSN 0350-5596.

(34) *Togaware* [online]. Canberra (Australia): Togaware, 2017 [cit. 2020-08-15]. Dostupné z: https://togaware.com/

(35) WILLIAMS, Graham J. Rattle: a data mining GUI for R. *The R Journal*. 2009, **1**(2), 46-55. ISSN 2073-4859.

(36) RAKOTOMALALA, Ricco. Tanagra under Linux. In: *Tanagra: data mining and data science tutorials* [online]. Lyon: Ricco Rakotomalala, 2009 [cit. 2020-07-09]. Dostupné z: http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_under_Linux.pdf

(37) *Tanagra: data mining and data science tutorials* [online]. [Lyon]: Ricco Rakotomalala, 2019 [cit. 2020-07-09]. Dostupné z: http://data-mining-tutorials.blogspot.com/

(38) Arff stable. *Weka Wiki* [online]. Hamilton: University of Waikato, 2020 [cit. 2020-08-30]. Dostupné z: https://waikato.github.io/weka-wiki/formats_and_processing/arff_stable/

(39) Pmml version 4.4. *Data mining group* [online]. [Illinois, US]: Center for computational science research, 2020 [cit. 2020-08-30]. Dostupné z: http://dmg.org/pmml/pmml-v4-4.html

(40) Online data (OLA). *Netmonitor* [online]. Praha: SPIR, 2016 [cit. 2020-09-05]. Dostupné z: http://www.netmonitor.cz/online-data-ola

(41) Vyhledávání. *Novinky.cz* [online]. Praha: Borgis, 2003-2020 [cit. 2020-10-14]. Dostupné z: https://www.novinky.cz/hledani

(42) Hledání. *Deník.cz* [online]. Praha: VLTAVA LABE MEDIA, 2020 [cit. 2020-10-14]. Dostupné z: https://www.denik.cz/hledani

(43) Vyhledávání. *IDNES.cz* [online]. Praha: MAFRA, c1999-2020 [cit. 2020-10-14]. Dostupné z: https://hledej.idnes.cz

(44) Vyhledat. *Aktuálně.cz* [online]. Praha: Economia, c1999-2020 [cit. 2020-10-14]. Dostupné z: https://www.aktualne.cz/hledani

(45) Rozšířené vyhledávání. *Google* [online]. [San Francisco]: Google, 2020 [cit. 2020-10-14]. Dostupné z: https://www.google.cz/advanced_search

(46) SYCHRA, Martin. *Analýza sentimentu s využitím dolování dat*. Brno, 2016. Diplomová práce. Vysoké učení technické, Fakulta informačních technologií, Ústav informačních systémů. Vedoucí práce Vladimír Bartík.

(47) *Institute of Formal and Applied Linguistics* [online]. Praha: ÚFAL, 2020 [cit. 2020-09-05]. Dostupné z: http://ufal.mff.cuni.cz/

(48) Český stoplist. *NLP: Centrum zpracování přirozeného jazyka* [online]. Brno: Fakulta informatiky Masarykovy univerzity, c2001-2020 [cit. 2020-09-01]. Dostupné z: https://nlp.fi.muni.cz/cs/StopList

(49) VESELOVSKÁ, Kateřina a Ondřej BOJAR. Czech SubLex 1.0. In: *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University,* [online]. Praha: ÚFAL, 2013 [cit. 2020-09-15]. Dostupné z: http://hdl.handle.net/11858/00-097C-0000-0022-FF60-B

(50) VESELOVSKÁ, Kateřina. *Sentiment analysis in Czech*. [1. vyd.]. [Praha]: Institute of Formal and Applied Linguistics, 2017. Studies in computational and theoretical linguistics. ISBN 978-80-88132-03-5.

(51) ŘEZNÍČEK, Vilém. AFFIN.CZ: AFFIN-like database of czech words for sentiment analysis. *GitHub* [online]. [San Francisco]: GitHub, 2020 [cit. 2020-11-02]. Dostupné z: https://github.com/vilemr/affin.cz

(52) ŘEZNÍČEK, Vilém. Analýza sentimentu: databáze českých slov s polaritou (AFINN.CZ). *Root.cz: [informace nejen ze světa Linuxu]* [online]. Praha: Internet Info, s.r.o., 2020 [cit. 2020-11-02]. Dostupné z:

https://blog.root.cz/hadoop-kdy-uz-ma-cenu-o-nem-uvazovat-a-kdy-jeste-n/analyza-sentimentu-databaze-ceskych-slov-s-polaritou/

(53) HEBERNAL, Ivan, Tomáš PTÁČEK a Josef STEINBERG. Supervised sentiment analysis in Czech social media. *Information Processing & Management: [an international journal]* [online]. [Elsevier], 2014, **50**(5), 693-707 [cit. 2020-09-15]. ISSN 0306-4573. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S0306457314000399

(54) KDnuggets 2019 poll [table]. In: *KDnuggets* [online]. [Illinois, US]: KDnugetts, 2020 [cit. 2020-08-06]. Dostupné z: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html/2

Seznam použitých zkratek

Seznam obrázků

Seznam tabulek

Seznam příloh

Tabulková příloha

Tabulka 1 Vstupní/výstupní formáty

I/O	název	DataMelt	KNIME	Orange	Rattle	Tanagra
vstup	**arff**	ano	ano	ne	ano	ano
	csv nebo txt	ano	ano	ano	ano	ano
	obrázky (svg, png)	ne	ano	ne	ne	ne
	pmml	ne	ano	ne	ano	ne
	vlastní datový formát	ne	ano	ano	ano	ano
	xls nebo xlsx	ne	ano	ano	ano	ano
	xml	ano	ne	ne	ne	ne
výstup	**arff**	ano	ano	ne	ne	ne
	csv nebo txt	ano	ano	ano	ano	ano
	html	ne	ne	ano	ne	ano
	pdf	ano	ano	ano	ne	ne
	pmml	ne	ano	ne	ano	ne
	vlastní datový formát	ne	ano	ano	ano	ano
	xls nebo xlsx	ne	ano	ano	ne	ne
	xml	ano	ne	ne	ne	ne

Zdroj: Vlastní zpracování

Tabulka 2 Výběr, předzpracování a transformace dat, nasazení výsledků

data	název	DataMelt	KNIME	Orange	Rattle	Tanagra
výběr	**databáze**	ano	ano	ano	ano	ne
	on-line databáze	ano	ano	ano	ano	ne
předzpracování a transformace	**filtrování**	ano	ano	ano	ano	ano
	seskupování	ano	ano	ano	ano	ne
	agregace	ano	ano	ano	ano	ne
	diskretizace	ano	ano	ano	ano	ne
	normalizace	ano	ano	ano	ano	ne
	chybějící hodnoty	ano	ano	ano	ano	ne
	PCA	ano	ano	ano	ano	ano
nasazení	**export modelu (pmml)**	ne	ano	ne	ano	ne
	export zpráv	ano	ano	ano	ne	ano

Zdroj: Vlastní zpracování

Tabulka 3 Podpora vizualizačních technik

	název	DataMelt	KNIME	Orange	Rattle	Tanagra
2D	**bodový graf**	ano	ano	ano	ano	ano
	histogram	ano	ano	ano	ano	ano
	krabicový graf	ano	ano	ano	ano	ne
	sloupcový graf	ano	ano	ano	ano	ano
	spojnicový graf	ano	ano	ano	ano	ano
	dendrogram	ano	ano	ano	ano	ano
	nomogram	ano	ne	ano	ne	ano
	pokročilé 2D grafy	ano	ano	ano	ano	ano
3D	**bodové grafy**	ano	ano	ano	ano	ne
	pokročilé 3D grafy	ano	ne	ne	ne	ne

Zdroj: Vlastní zpracování

Tabulka 4 Speciální funkcionality

speciální funkce	DataMelt	KNIME	Orange	Rattle	Tanagra
big data	ne	ne (doplněk)	ne	ne	ne
deep learning	ne	ne (doplněk)	ne	ne	ne
image mining	ne	ne (doplněk)	ne (doplněk)	ne	ne
text mining	ano (Weka)	ne (doplněk)	ne (doplněk)	ano	ne
web mining	ne	ne (doplněk)	ne (doplněk)	ne	ne

Zdroj: Vlastní zpracování

Tabulka 5 Porovnání algoritmů a metod

typ	algoritmus	DataMelt	KNIME	Orange	Rattle	Tanagra
asociační pravidla	**Apriori**	ano (Weka)	ne (doplněk Weka)	ne	ano	ano
	FP-growth	ano (Weka)	ne (doplněk Weka)	ne (doplněk)	ne	ne
	GSP	ano (Weka)	ne (doplněk Weka)	ne	ne	ne
	Tertius	ano (Weka)	ne (doplněk Weka)	ne	ne	ne
Bayesovská klasifikace	**AODE**	ano (Weka)	ne (doplněk Weka)	ne	ne	ne
	Bayesovská síť	ano (Weka)	ne (doplněk Weka)	ne	ne	ne
	Naivní Bayes	ano (Weka)	ano	ano	ne	ano
kombinace modelů	**bagging**	ano (Weka)	ano	ano	ne	ano
	boosting (AdaBoost)	ano (Weka)	ne (doplněk Weka)	ano	ano	ano
	náhodný les	ano (Weka)	ano	ano	ano	ano
	stacking	ano (Weka)	ne (doplněk Weka)	ano	ne	ne
neuronové sítě	**CNN**	ne	ne (doplněk Keras)	ne	ne	ne
	MLP	ano (Weka)	ano	ano	ano	ano
	PNN	ano (Joone)	ano	ne	ne	ano
	RNN	ano (Joone)	ne (doplněk Keras)	ne	ne	ne
	Kohonenovy mapy (SOM)	ano (Joone)	ne (doplněk Weka)	ano	ne	ano
rozhodovací pravidla	**1Rule**	ano (Weka)	ne (doplněk Weka)	ne	ne	ne
	CN2 nebo CN4	ne	ne	ano	ne	ano
	PART	ano (Weka)	ne (doplněk Weka)	ne	ano	ne
	RIPPER	ano (Weka)	ne (doplněk Weka)	ne	ne	ne
rozhodovací stromy	**vlastní implementace**	ne	ano	ano	ano	ne
	C4.5 nebo C5.0	ano (Weka)	ne (doplněk Weka)	ne	ne	ano
	CART	ano (Weka)	ne (doplněk Weka)	ne	ne	ano
	ID3	ano (Weka)	ne (doplněk Weka)	ne	ne	ano
rozšířené lineární metody	**SVM**	ano	ano	ano	ano	ano
shluková analýza	**EM**	ano (Weka)	ne (doplněk Weka)	ne	ne	ano
	hierarchická	ano	ano	ano	ano	ano

shluková analýza	**k-středy**	ano	ano	ano	ano	ano
	fuzzy c-means (FCM)	ne	ano	ne	ne	ne
statistické metody	**analýza rozptylu (ANOVA)**	ne	ano	ano	ano	ano
	diskriminační analýza	ano (Weka)	ano	ano	ne	ano
	faktorová analýza	ne	ne	ne	ne	ano
	lineární regrese	ano (Weka)	ano	ano	ano	ano
	logistická regrese	ano (Weka)	ano	ano	ano	ano
	t-testy	ano (Weka)	ano	ano	ano	ano
testování modelu	**konfusní matice**	ano (Weka)	ano	ano	ano	ano
	křížová validace	ano (Weka)	ano	ano	ano	ano
	Lift křivka	ano (Weka)	ano	ano	ano	ano
	ROC křivka	ano (Weka)	ano	ano	ano	ano
založené na instancích	**k-nejbližší soused**	ano (Weka)	ano	ano	ne	ano

Zdroj: Vlastní zpracování

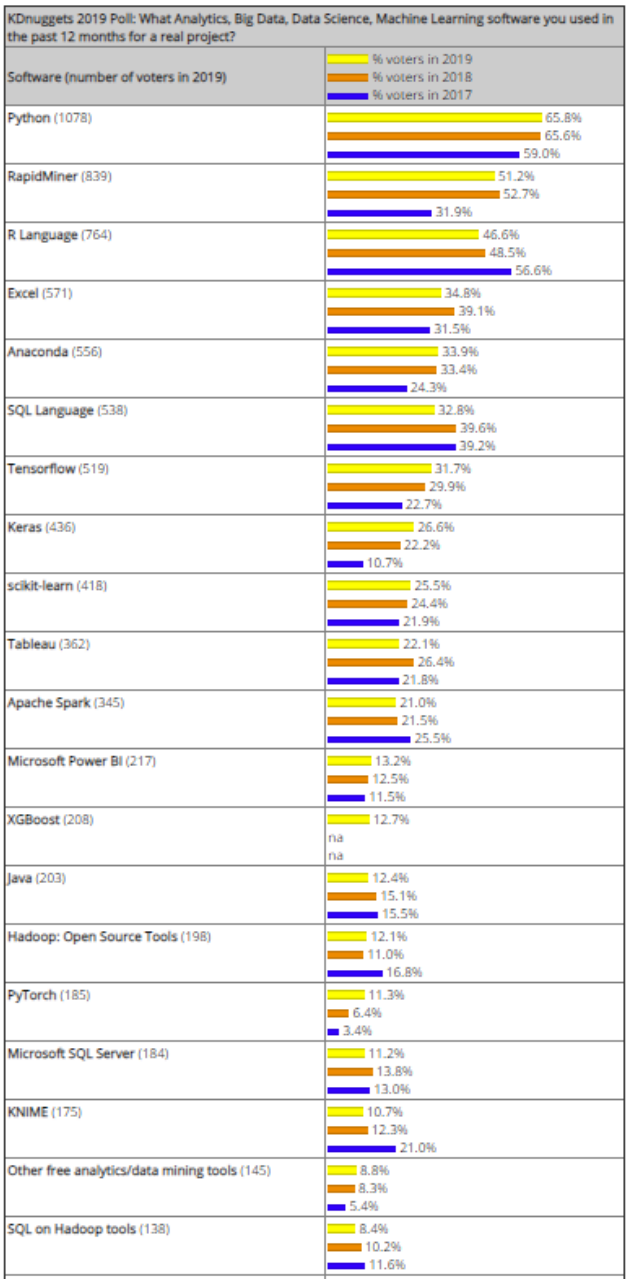
Tabulka 6 Porovnání metrik klasifikátorů při změně parametrů

metoda	textové atributy	Bag of Words	AUC	CA	F1	průměr	průměr CA, F1
Naivní Bayes	obsah	absolutní	0,587	0,155	0,042	0,261	0,099
	obsah	binární	0,605	0,155	0,042	0,267	0,099
	titulek, perex	absolutní	0,707	0,257	0,198	0,387	0,228
	titulek, perex	**binární**	**0,727**	**0,311**	**0,271**	**0,436**	**0,291**
	titulek, perex, obsah	absolutní	0,595	0,155	0,042	0,264	0,099
	titulek, perex, obsah	binární	0,610	0,155	0,042	0,269	0,099
SVM	obsah	absolutní	0,685	0,534	0,525	0,581	0,530
	obsah	binární	0,721	0,568	0,565	0,618	0,567
	titulek, perex	absolutní	0,750	0,601	0,571	0,641	0,586
	titulek, perex	**binární**	**0,761**	**0,595**	**0,579**	**0,645**	**0,587**
	titulek, perex, obsah	absolutní	0,673	0,520	0,514	0,569	0,517
	titulek, perex, obsah	binární	0,725	0,574	0,568	0,622	0,571

Zdroj: Vlastní zpracování výstupů ze softwaru Orange
Poznámka: AUC – oblast pod křivkou ROC, CA – přesnost klasifikace a F1 – f-skóre, které zohledňuje hodnoty přesnosti (precision) a úplnosti (recall). Tučně zvýrazněny metriky dosahující nejlepších výsledků.

Obrazová příloha

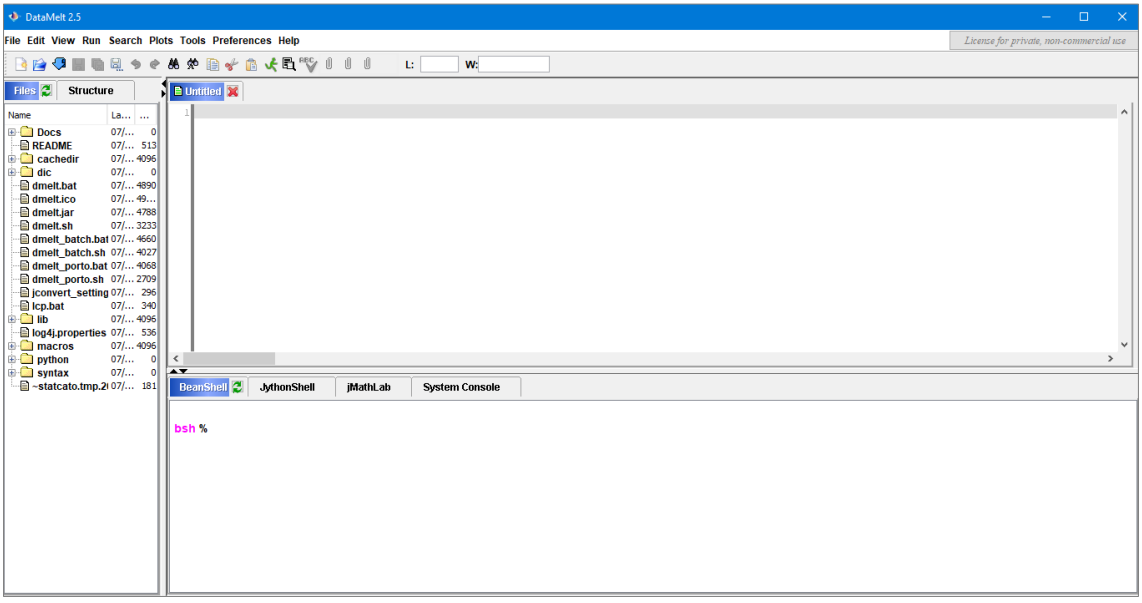
Obrázek 1 Průzkum KDnugetts



Platform	2019 % share	2018 % share	% change
Python	65.8%	65.6%	0.2%
R Language	46.6%	48.5%	-4.0%
SQL Language	32.8%	39.6%	-17.2%
Java	12.4%	15.1%	-17.7%
Unix shell/awk	7.9%	9.2%	-13.4%
C/C++	7.1%	6.8%	3.7%
Javascript	6.8%	na	na
Other programming and data languages	5.7%	6.9%	-17.1%
Scala	3.5%	5.9%	-41.0%
Julia	1.7%	0.7%	150.4%
Perl	1.3%	1.0%	25.2%
Lisp	0.4%	0.3%	46.1%

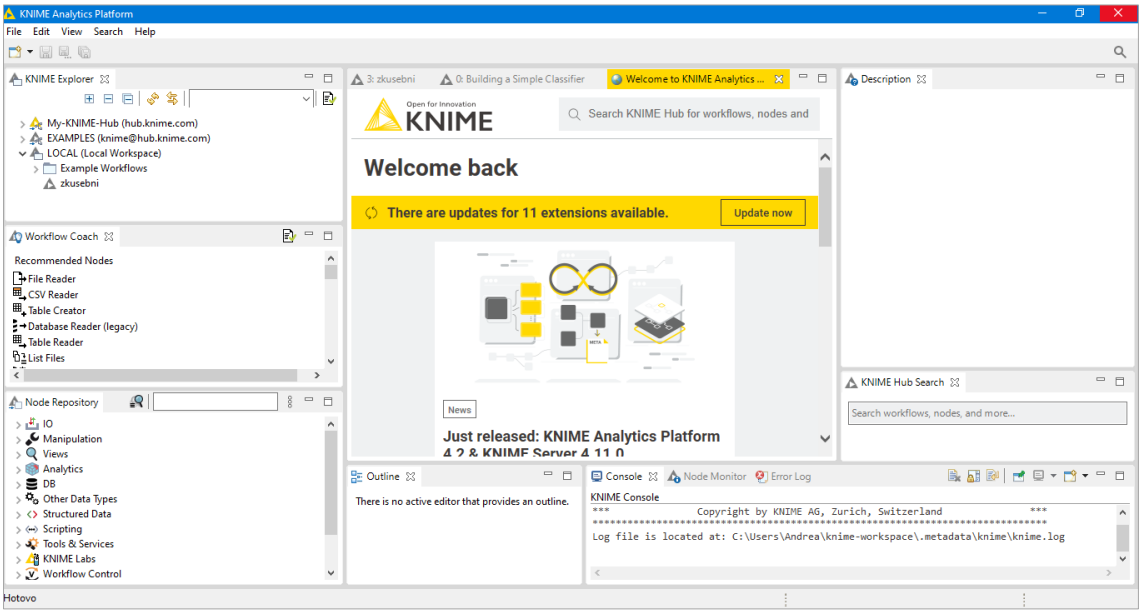
Zdroj: (54)

Obrázek 2 GUI DataMelt



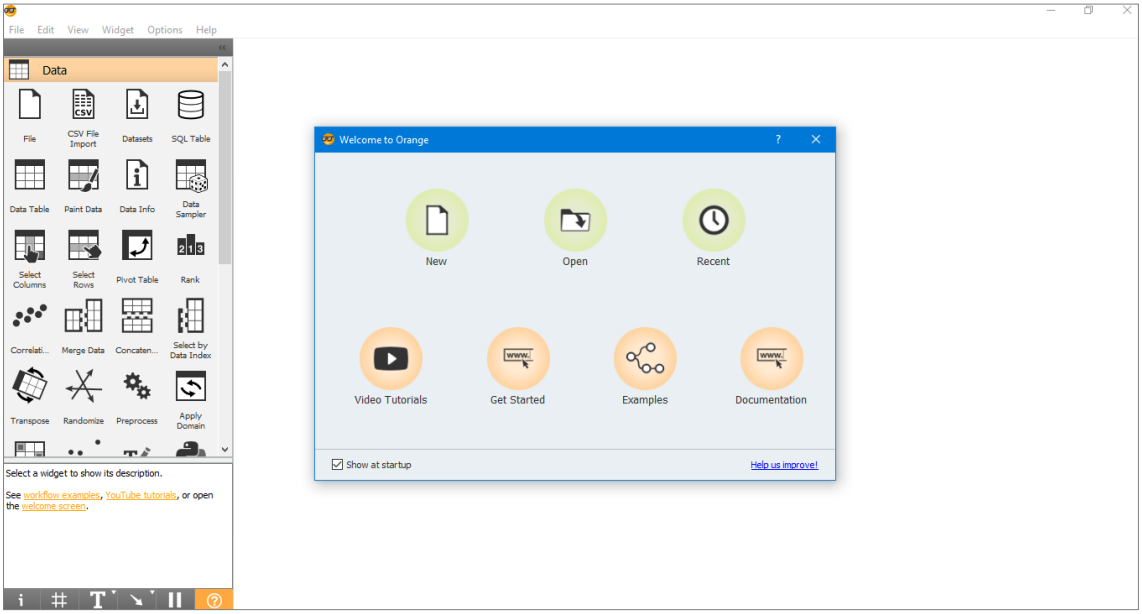
Zdroj: Vlastní zpracování

Obrázek 3 GUI KNIME Analytics Platform



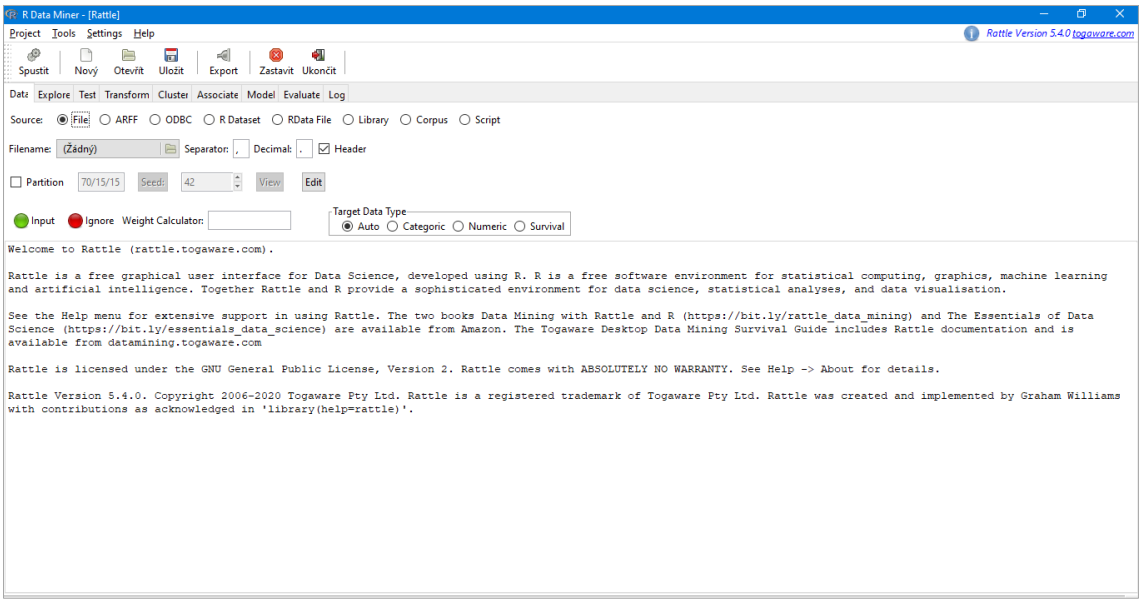
Zdroj: Vlastní zpracování

Obrázek 4 GUI Orange



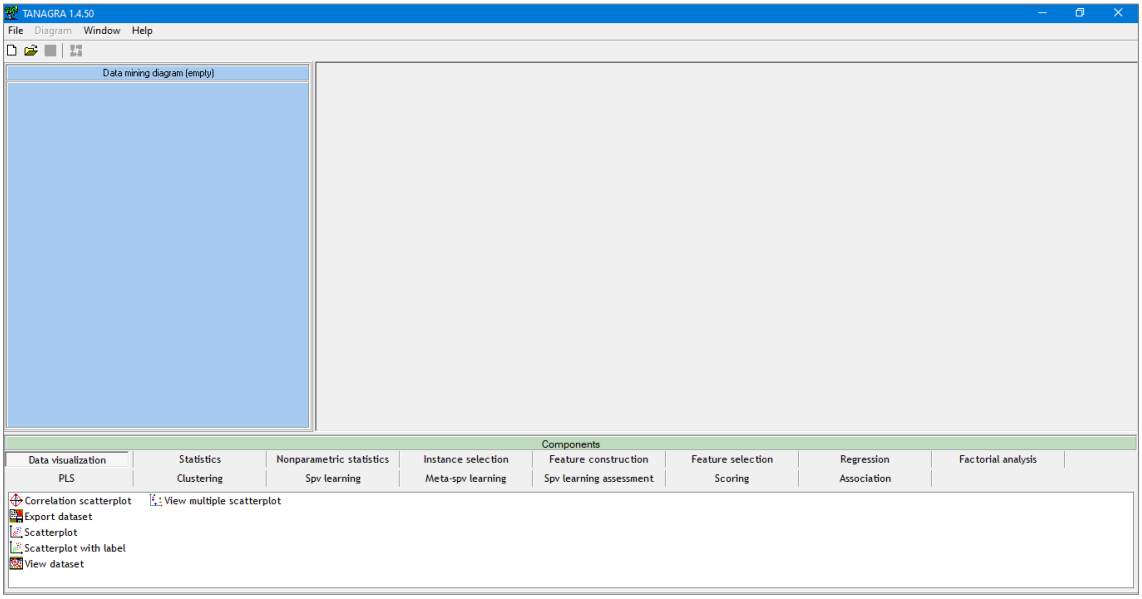
Zdroj: Vlastní zpracování

Obrázek 5 GUI Rattle



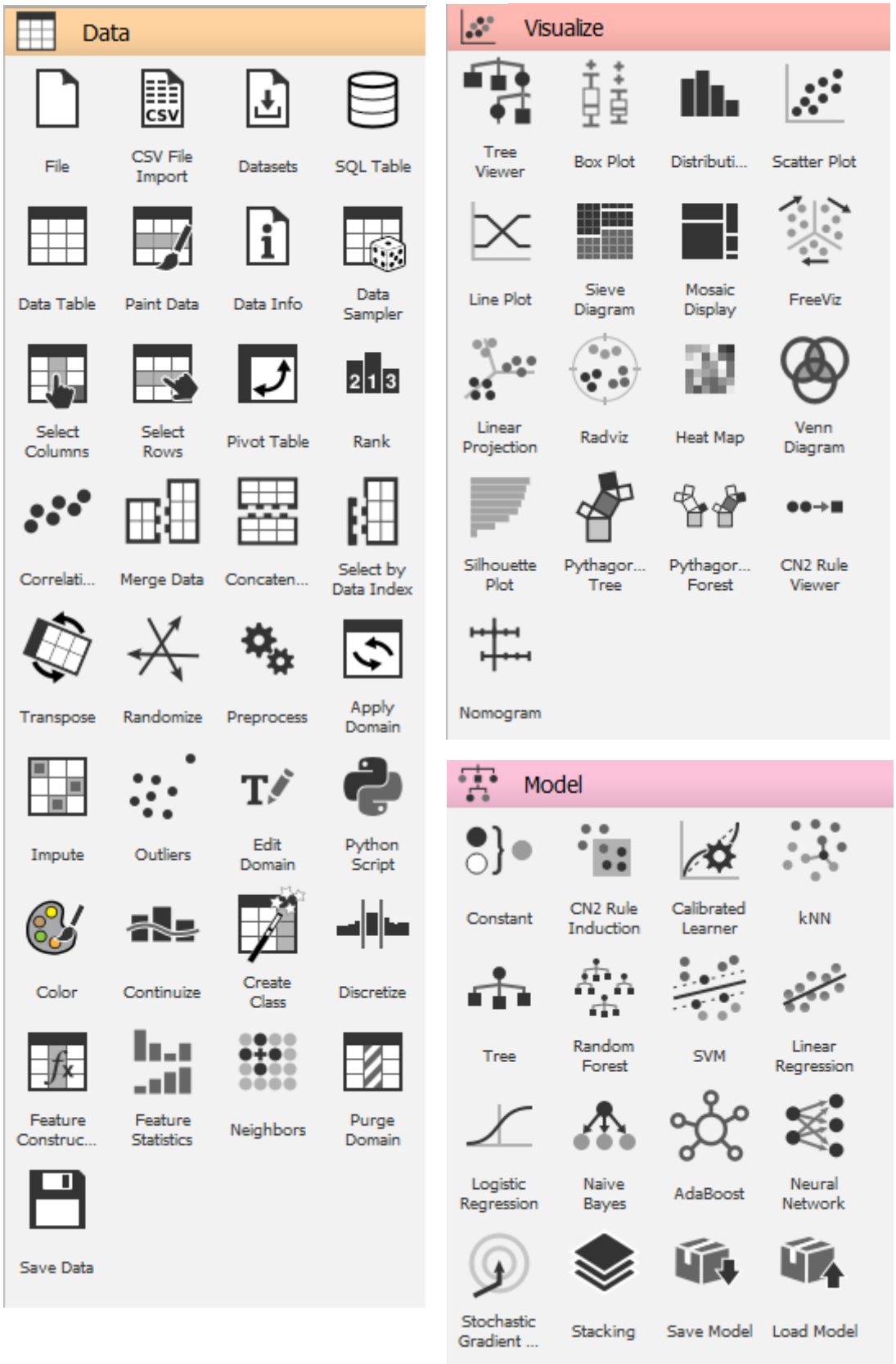
Zdroj: Vlastní zpracování

Obrázek 6 GUI Tanagra



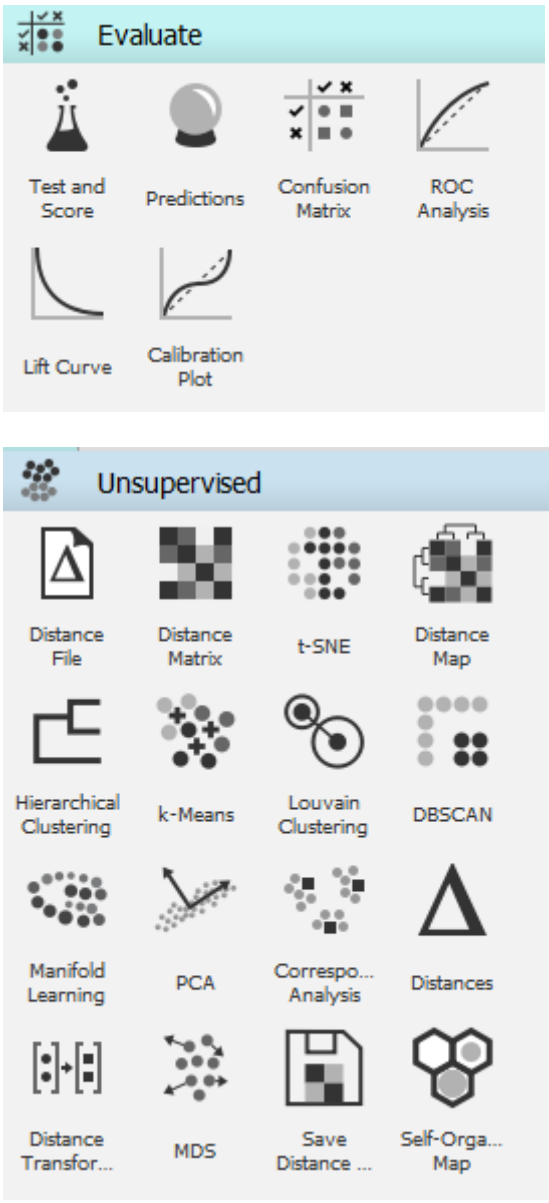
Zdroj: Vlastní zpracování

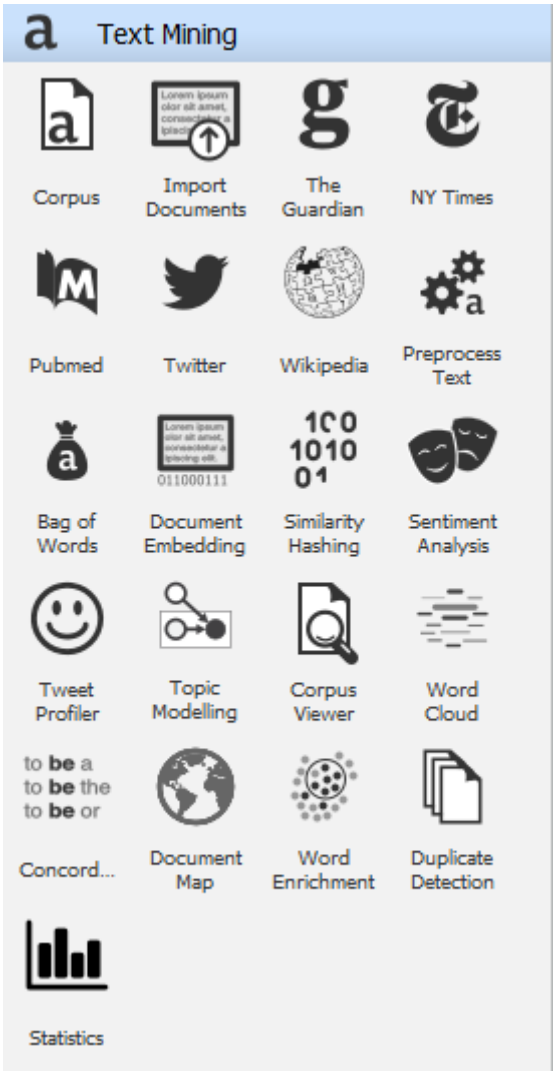
Obrázek 7 Orange - Data, Visualize, Model



Zdroj: Vlastní zpracování

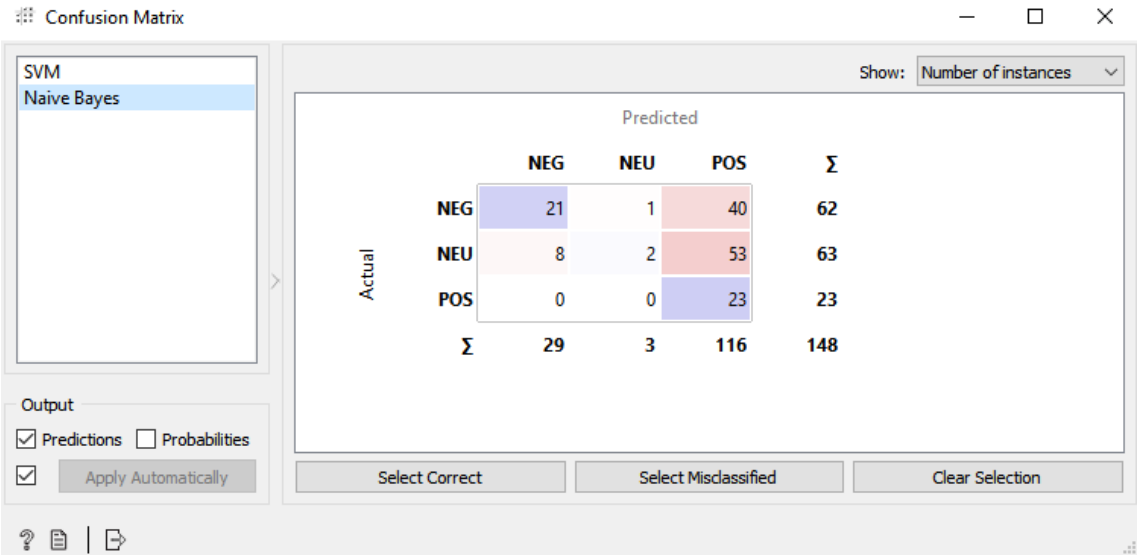
Obrázek 8 Orange – Evaluate, Unsupervised, Text mining





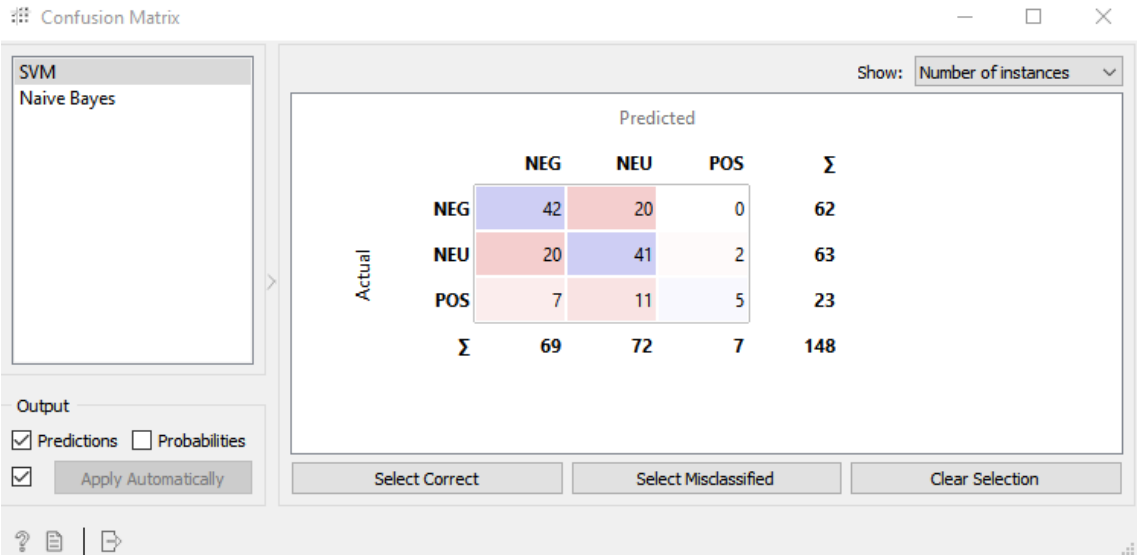
Zdroj: Vlastní zpracování

Obrázek 9 Konfusní matice – klasifikátor Naivní Bayes



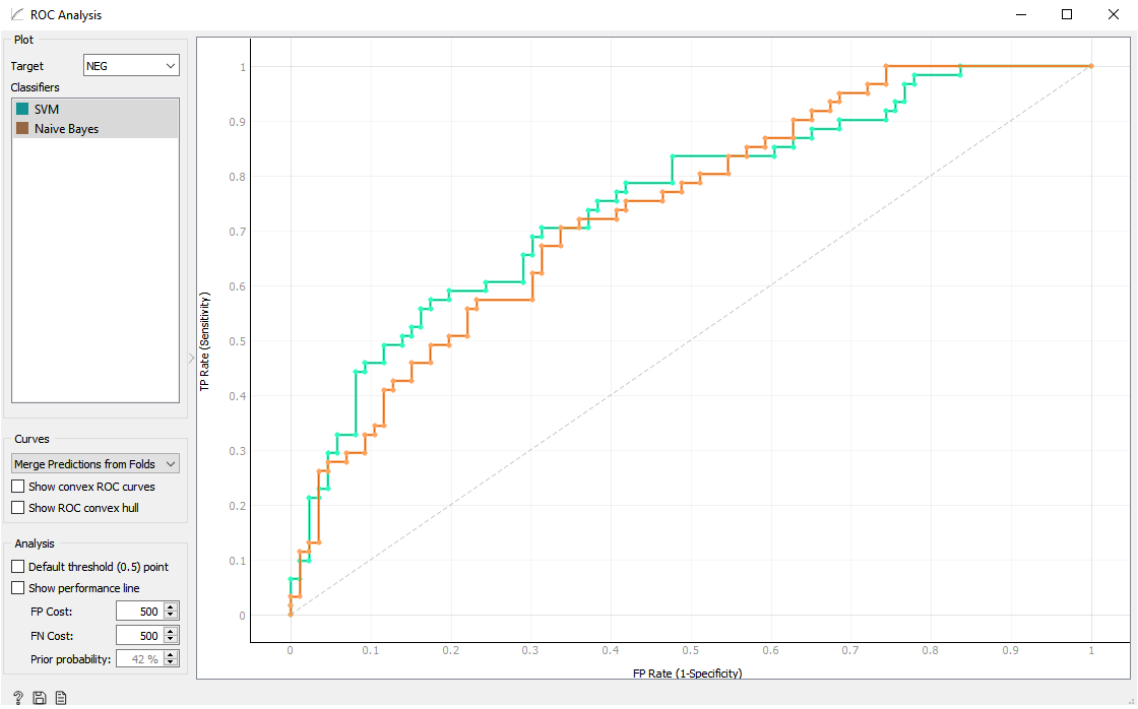
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 10 Konfusní matice – klasifikátor SVM



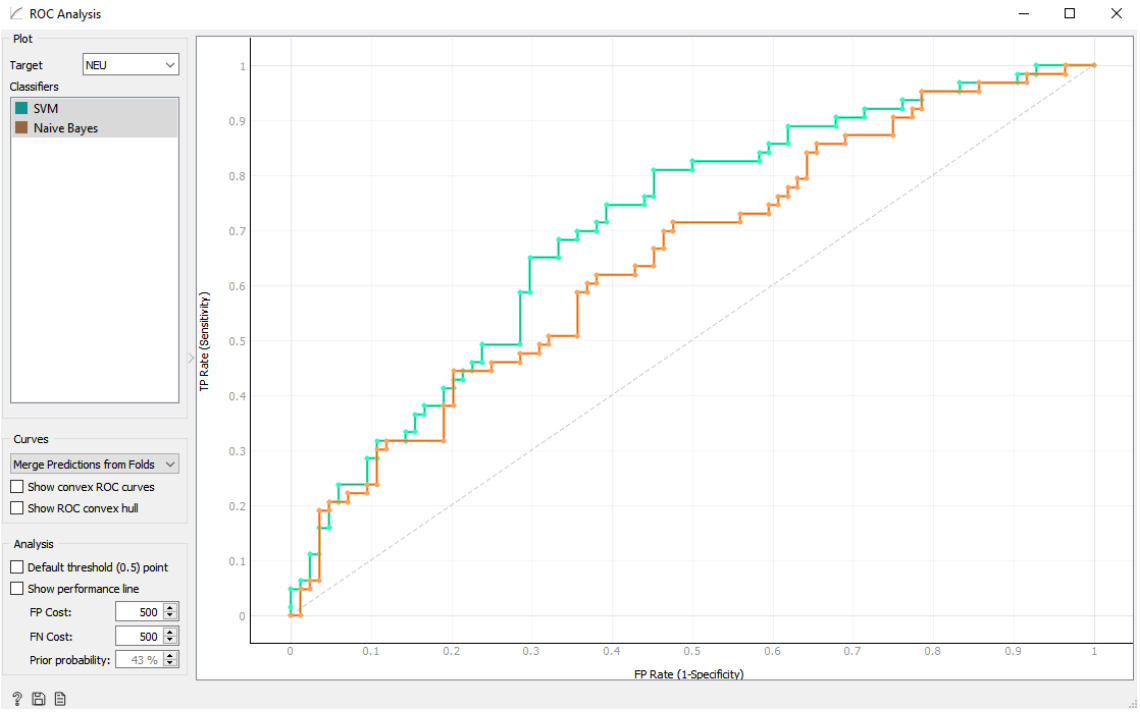
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 11 ROC křivka porovnávaných klasifikátorů – kategorie NEG



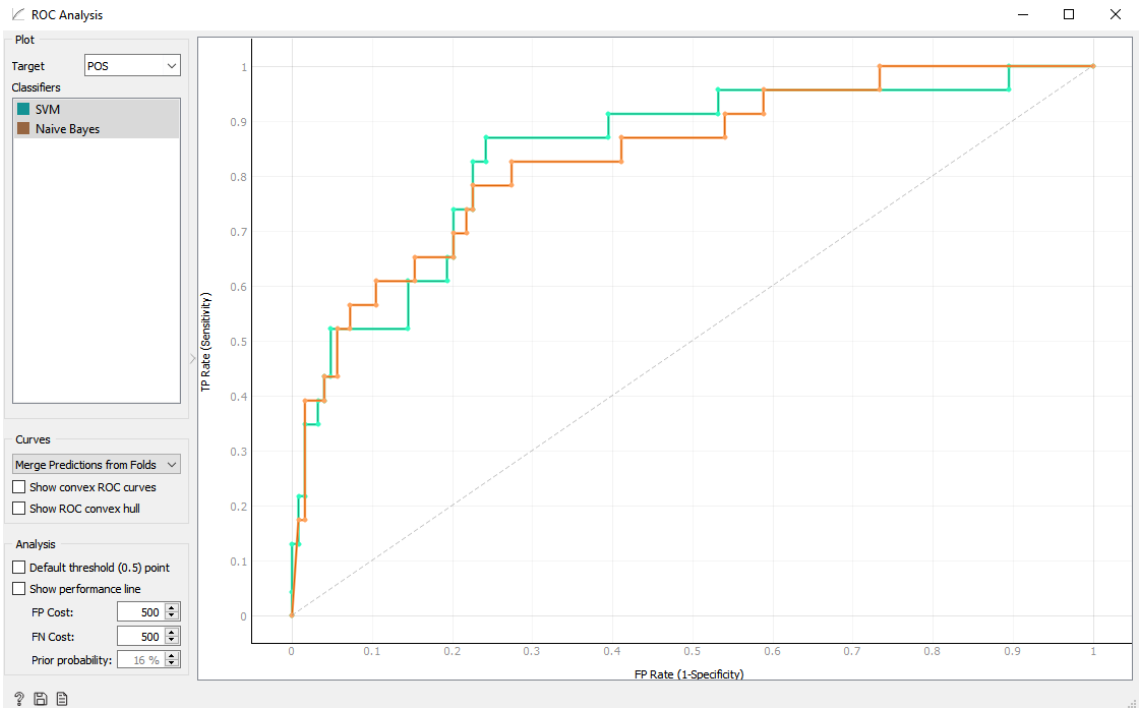
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 12 ROC křivka porovnávaných klasifikátorů – kategorie NEU



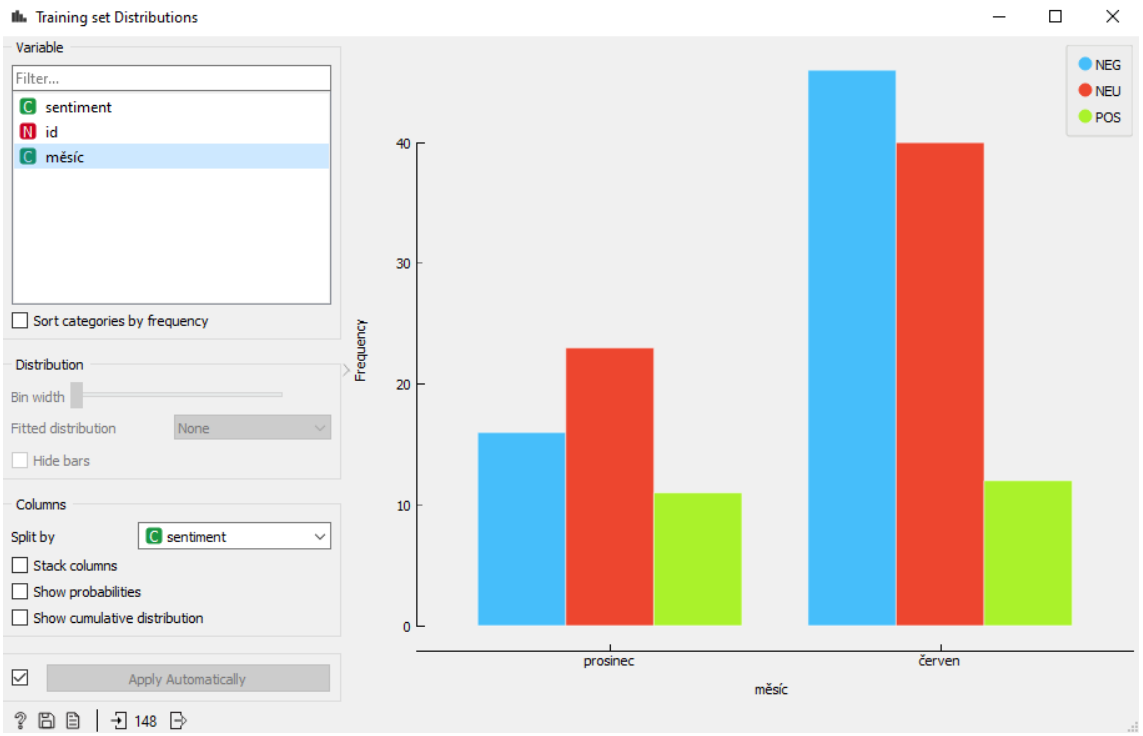
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 13 ROC křivka porovnávaných klasifikátorů – kategorie POS



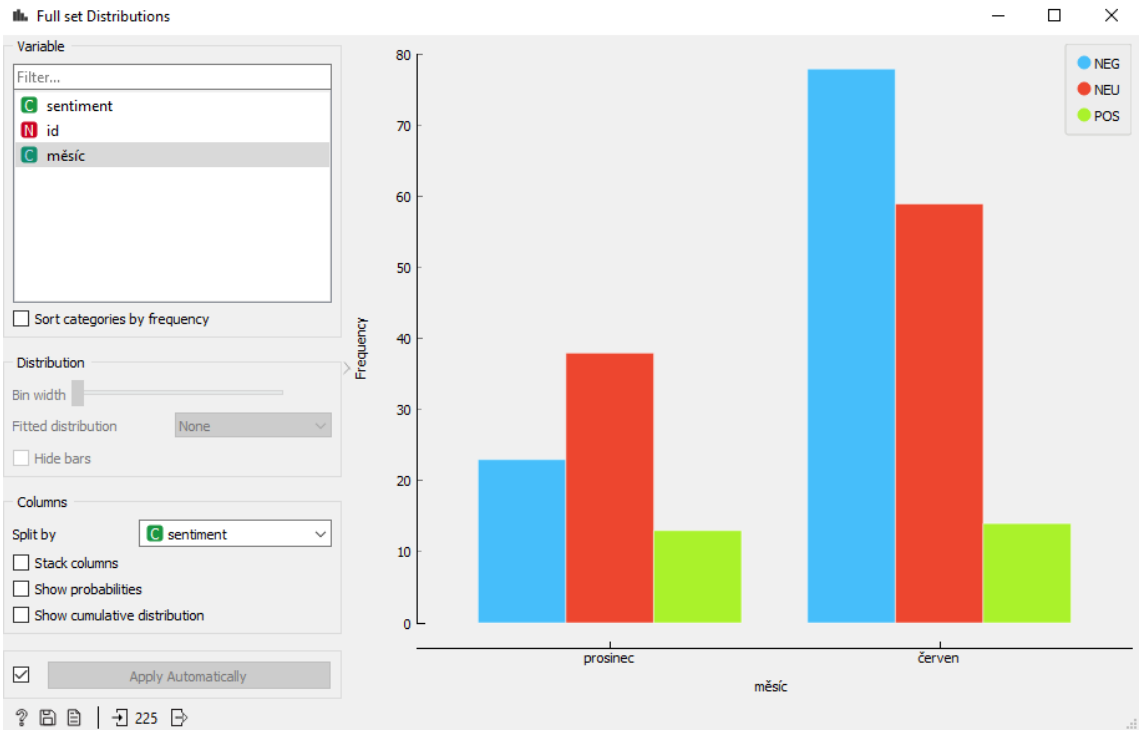
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 14 Anotované třídy sentimentu testovacího korpusu



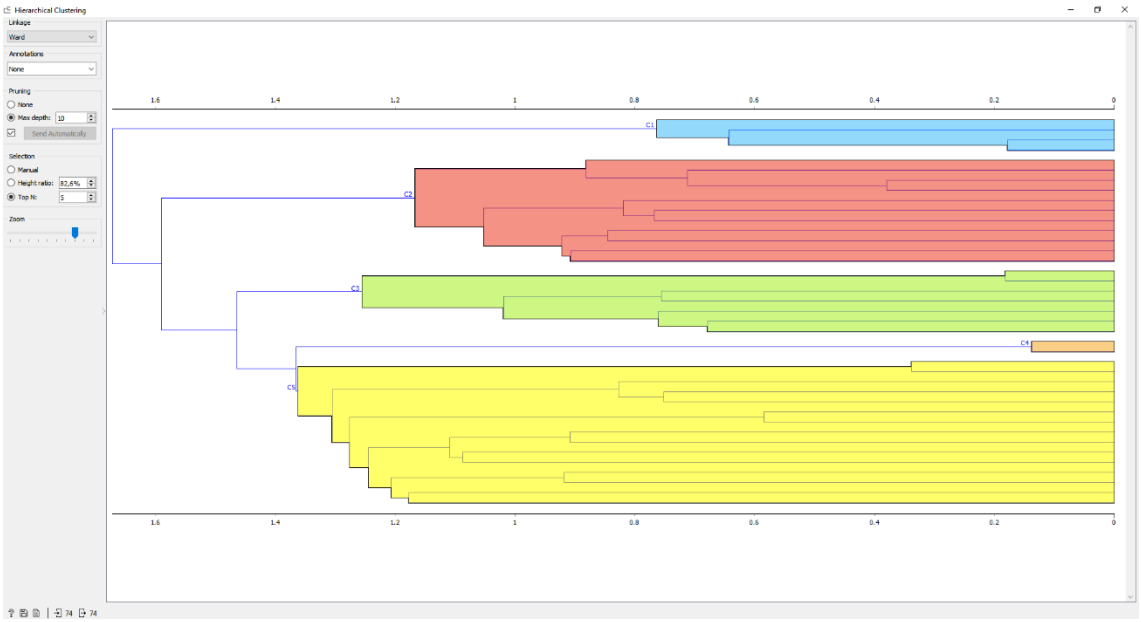
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 15 Anotované a predikované třídy sentimentu kompletního korpusu



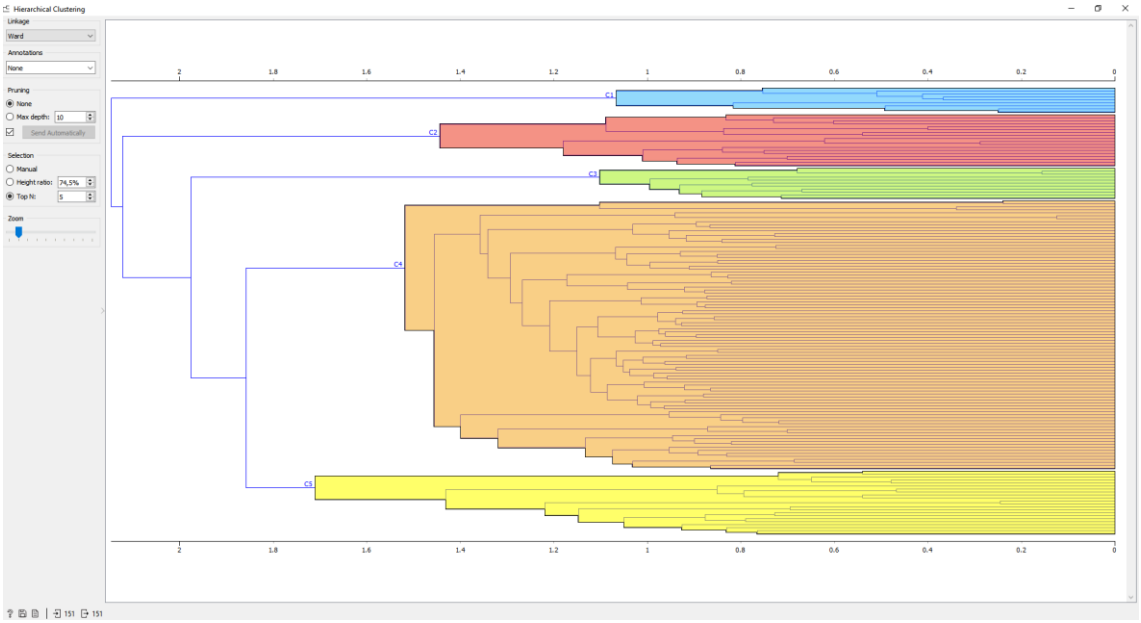
Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 16 Dendrogram – prosinec 2019



Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 17 Dendrogram – červen 2020



Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 18 Mrak slov – prosinec 2019

Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Obrázek 19 Mrak slov – červen 2020

Zdroj: Vlastní zpracování, výstup ze softwaru Orange

Zadání práce

Zadání bakalářské práce

Autor:	**Andrea Kubcová**
Studium:	I1600472
Studijní program:	B1802 Aplikovaná informatika
Studijní obor:	Aplikovaná informatika
Název bakalářské práce:	**Data mining a možnosti nekomerčního softwaru**
Název bakalářské práce AJ:	Data mining and freeware possibilities

Cíl, metody, literatura, předpoklady:

Cíl práce: Cílem práce je vyhledat, popsat a případně porovnat nekomerční typy softwarových nástrojů, které lze využít pro data mining. Pro zvolený nekomerční nástroj provést podrobnější zhodnocení jeho možností a řešit jeho pomocí ukázkovou úlohu. Osnova práce: 1. Úvod 2. Zpracování rešerše literatury 3. Popis metodiky práce Data mining, definice, zdroje dat, metodiky a základní metody 4. Přehled nekomerčních typů software pro data mining 5. Podrobnější studium a popis jednoho typu softwaru 6. Volba dat, typu úloh a ukázková úloha 7. Shrnutí a závěr

BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003. 366 s. ISBN 80-200-1062-9. PETR, Pavel. Metody Data Miningu. Vyd. 1. Pardubice: Univerzita Pardubice, 2014. 1 sv. (85 s.). ISBN 978-80-7395-872-5. PETR, Pavel. Metody Data Miningu. Vyd. 1. Pardubice: Univerzita Pardubice, 2014. 2 sv. (84 s.). ISBN 978-80-7395-873-2. SKALSKÁ, Hana. Data mining a klasifikační modely. Vyd. 1. Hradec Králové: Gaudeamus, 2010. 154 s. Recenzované monografie; 4. ISBN 978-80-7435-088-7. WITTEN, I. H., FRANK, Eibe a HALL, Mark A. Data mining: practical machine learning tools and techniques. 3rd ed. Burlington: Morgan Kaufmann, 2011. xxxiii, 629 s. Morgan Kaufmann series in interactive technologies. ISBN 978-0-12-374856-0. ZHAO, Yangchang a CEN, Yonghua. Data mining applications with R. Amsterdam: Elsevier, 2014. xxi, 470 stran. ISBN 978-0-12-411511-8.

Garantující pracoviště:	Katedra informatiky a kvantitativních metod, Fakulta informatiky a managementu
Vedoucí práce:	prof. RNDr. Hana Skalská, CSc.
Datum zadání závěrečné práce:	14.1.2018