Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky

Extrakce informace z textových dokumentů pro potřeby České obchodní inspekce

Bc. Lukáš Rejmont

Diplomová práce
2018

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2017/2018

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Lukáš Rejmont**
Osobní číslo:	**E160000**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Extrakce informace z textových dokumentů pro potřeby České obchodní inspekce**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Cílem práce je charakterizovat současné přístupy k extrakci informace z textu (zejména ručně definované šablony a strojové učení), charakterizovat zvolenou metodu extrakce, na příkladu textů pro potřeby České obchodní inspekce provést předzpracování dokumentů, navrhnout jmenné entity pro extrakci, provést asociaci jmenných entit do relací a zhodnotit přesnost modelu na reálných datech.

Osnova:
- Extrakce informace z textu
- Charakteristika České obchodní inspekce
- Sběr a zpracování dokumentů
- Extrakce zvolených entit z dokumentů České obchodní inspekce
- Zhodnocení přesnosti extrakce

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 60 stran**

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

BERKA, Petr. Dobývání znalostí z databází. Praha: Academia, 2003. ISBN 80-200-1062-9.
HAN, Jiawei, KAMBER, Micheline. Data mining: concepts and techniques. 2nd ed. San Francisco, CA: Morgan Kaufmann, c2006. ISBN 15-5860-901-6.
AGGARWAL, Charu C., ZHAI, ChengXiang (ed.). Mining text data. New York: Springer, c2012. ISBN 978-1-4614-3222-7.
VOŘÍŠEK, Jiří, PAVELKA, Jan, VÍT, Miroslav. Aplikační služby IS/ICT formou ASP: proč a jak pronajímat informatické služby. Praha: Grada, 2004. Management v informační společnosti. ISBN 80-247-0620-2.

Vedoucí diplomové práce: **doc. Ing. Petr Hájek, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **1. září 2017**

Termín odevzdání diplomové práce: **30. dubna 2018**

doc. Ing. Romana Provazníková, Ph.D.
děkanka

L.S.

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 1. září 2017

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako Školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 9/2012, bude práce zveřejněna v Univerzitní knihovně a prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 11. 12. 2018 Bc. Lukáš Rejmont

PODĚKOVÁNÍ:

Tímto bych rád/a poděkoval/a svému vedoucímu práce doc. Ing. Petru Hájkovi Ph.D. za jeho odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování diplomové práce. Dále bych chtěl poděkovat své rodině, která mě po celou dobu vždy podporovala.

ANOTACE

Cílem práce je charakterizovat současné přístupy k extrakci informací z textu a aplikovat je na textových dokumentech pro potřeby České obchodní inspekce. První část práce je zaměřená na jednotlivé techniky extrakce informací z textových dokumentů. Tyto techniky jsou následně využity v druhé části práce, která se věnuje předzpracování dokumentů, návrhu jmenných entit pro extrakci a asociaci jmenných entit s ohledem na využitelnost Českou obchodní inspekcí. V samotném závěru práce je provedeno zhodnocení získaných výsledků.

KLÍČOVÁ SLOVA

Extrakce informací, umělá inteligence, zpracování přirozeného jazyka, jmenné entity

TITLE

Extraction of information from text documents for the needs of the Czech trade inspection

ANNOTATION

The goal of this thesis is to characterize current approaches of the Information Extraction from text and apply them on text documents for the Czech trade inspection needs. The first part is focused on individual techniques of information extraction from the text documents. These techniques are used in the second part of this thesis, which focused on preprocessing documents, proposal of named entity for extraction and association with regard to usability of Czech trade inspection. At the end of this thesis the obtained results are evaluated.

KEYWORDS

Information extraction, artificial intelligence, natural language processing, named entity

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM GRAFŮ

SEZNAM ZKRATEK

ACE – Automatic Content Extraction (Automatická extrakce obsahu)

CO – Coreference Resolution (Rozlišení koreference)

CoNLL – Conference on Natural Language Learning (konference o zpracování přirozeného jazyka)

CRF – Conditional random fields (Podmíněná náhodná pole)

ČOI – Česká obchodní inspekce

FN – False Negative (Chybně negativní)

FP – False Positive (Chybně pozitivní)

HMM – Hidden Markov Models (Skryté Markovovy modely)

IE – Information Extraction (Extrakce informace)

IR – Infromation Retrieval (Získávání informací)

MUC – Message Understanding Conference

NER – Named Entity Recognition (Rozpoznávání jmenných entit)

NLP – Natural Language Processing (Zpracování přirozeného jazyka)

RBF – Radial Basis Function (Radiální bazické funkce)

ST – Scenario Template production(Šablona scénáře)

SVM – Support Vector Machines (Podpůrné vektorové stroje)

SW – Software

TDIDT – Top Down Induction of Decision Trees (Indukce rozhodovacích stromů shora dolů)

TE – Template Element (Prvek šablony)

TN – True Negative (Správně negativní)

TP – True Positive (Správně pozitivní)

TR – Template Relation (Relace šablony)

URL – Unique Resource Locator (Jednotná adresa zdroje)

VAT – Value Added Tax (Daň z přidané hodnoty)

Úvod

V dnešní době je člověk obklopen množství textů v nejrůznějších podobách, kdy v těchto textech se skrývá velké množství informací. Získání těchto informací však může být velmi zdlouhavý proces s nejistým koncem, který představuje dlouhé hodiny, dny, týdny či dokonce roky strávených nad těmito dokumenty. Dalším úskalím je, že i přes tuto snahu se k daným informacím nemusí člověk dopracovat. Výše uvedené problémy spojené s oblastí extrakce informací z textu se snaží řešit text mining neboli dolování v textech [1].

Extrakce informací představuje automatickou extrakci informací (faktů) z nestrukturovaných či jen částečně strukturovaných dokumentů, kdy takto získanou informaci převádí do strukturované podoby. Extrakcí informací se zabývá celé široké spektrum metod a technik od ručně psaných regulárních výrazů a až po pravděpodobnostní grafické modely [3]. Právě tyto metody a techniky budou v této práci popsány. Jednotlivé metody pak mohou být použity na textech u celého spektra zdrojových materiálů jako jsou novinové články, sociální sítě, blogy, webové stránky apod. A právě webové stránky se staly zdrojem pro extrakci informací této práce.

České obchodní inspekci pak může extrakce informací z nestrukturovaných dokumentů posloužit jako nástroj při odhalování prohřešků proti občanskému zákoníku. Pomocí extrakce informací mohou být například prověřovány internetové obchody. U těchto internetových obchodů může být prověřeno, zda zveřejňují obchodní podmínky a zda tyto obchodní podmínky obsahují všechny povinné náležitosti.

V první části této práce jsou představeny jednotlivé techniky text miningu, jenž nabízí nejrůznější využití jako je kategorizace textů, shlukování textů, analýza sentimentu, získávání informací a extrakce informací. Technika extrakce informací nabízí různé přístupy, jak získat informace, které člověku nejsou na první pohled zcela zřejmé, a právě proto jsou v práci blíže popsány. Dále budou v práci představeny metriky pro vyhodnocení systémů na extrakci informací. Cílem bude představené techniky následně využít při tvorbě modelů pro potřeby České obchodní inspekce. Cílem modelů bude provést získání dokumentů, které budou sloužit jako korpus dokumentů pro zpracování, jejich následné předzpracování a extrakce a asociace jmenných entit. Výše uvedené modely a návrhy jmenných entit budou vytvářeny s ohledem na využitelnost pro Českou obchodní inspekci. V samotném závěru práce bude provedeno vyhodnocení vytvořených modelů na reálných datech a shrnutí celé práce.

1. TEXTOVÁ DATA A JEJICH PŘEDZPRACOVÁNÍ

Mezi hlavní odlišnosti text miningu v rámci data miningu lze považovat povahu analyzovaných dat. U text miningu mají data podobu nestrukturovanou, jedná se o volný text, který je psaný přirozeným jazykem [1]. Tím se odlišují od strukturovaných dat, která jsou ukládána do přesně specifikovaných struktur, kdy tento způsob uložení umožňuje provádět nad těmito daty velké množství operací. Příkladem strukturovaných dat mohou být databáze, formuláře nebo jiné dokumenty s pevně danou strukturou. Na rozdíl od nestrukturovaných dat jsou strukturovaná data bez smysluplného kontextu (postrádají informační souvislosti a jazyková zbarvení). Z výše uvedeného vyplývá, že nestrukturovaná data nejsou ukládána v přesně definovaných jednotných strukturách. Příkladem nestrukturovaných dat jsou textové dokumenty, novinové články, www stránky, prezentace, elektronická pošta, audio a video dokumenty a další. Strukturovaná data tak obsahují informace v souvislostech a velkou mírou redundance. Právě z důvodu vysoké míry redundance, nespecifikované struktury a absence klíčových slov, je vyhledávání relevantních informací v těchto datech složité [1].

Další odlišností je, že u strukturovaných dat jsou znalosti v datech ukryté a nejsou tak pro člověka na první pohled jasné. Oproti tomu u nestrukturovaných dat je znalost velmi často explicitně a jasně uvedena. Problémem je však velká časová náročnost, kterou vyžaduje samotné přečtení celého textu člověkem.

V rámci text miningu se často využívají totožné metody a postupy jako při zpracování strukturovaných dat. Aby však bylo možné těchto metod a postupů využít i v případě nestrukturovaných dat, je nutné provést jejich předzpracování, kdy jednotlivé způsoby předzpracování jsou blíže popsány v kapitole 2 Textová data a jejich předzpracování.

V současné době je odhadována, že až 80 % dat je uloženo v textové podobě, kdy tyto jsou jen velmi málo nebo vůbec strukturované [40]. Analýza textových dat představuje velmi komerčně zajímavou oblast. Analytické úlohy, zabývající se textovými daty, následně nacházejí uplatnění v oblastech marketingu, počítačové bezpečnosti, informačních službách, řízení lidských zdrojů, boje proti terorismu apod. Práce s těmito daty je však obtížná. To je způsobeno samotných charakterem těchto dat. Mezi hlavní charakteristiky těchto dat patří vysoká dimenzionalita (mnoho atributů – slov), řídkost dat a žádná chybějící data [40].

Informace, které jsou obsaženy v množství textových dokumentů, lze obvykle vyjádřit v mnohem stručnější podobě, jelikož ve většině případů je informace obsažená v celém textu podstatně menší (dá se vyjádřit menším počtem slov, než je celý dokument). A právě text mining je nástroj, který takovéto záznamy dokáže automaticky zpracovávat, poskytne stěžejní

informace obsažené v textu a dokumenty setřídí podle jejich obsahu nebo zabarvení, aniž by bylo nutné je zdlouhavě číst.

Jak již bylo řečeno v předcházející kapitole, aby bylo možné nad daty provádět jednotlivé analýzy, je nutné provézt jejich důkladné předzpracování. Jednotlivým krokům předzpracování textových dokumentů je věnována tato kapitola.

1.1. Korpus

Ještě před tím, než budou představeny jednotlivé fáze předzpracování nestrukturovaných dat, je potřeba objasnit pojem korpus. Korpus představuje označení pro počítačový soubor, ve kterém jsou obsaženy uložené texty, které dále slouží k jazykovému výzkumu [29]. Z důvodu snadného vyhledávání slov a slovních spojení mají korpusy jednotný formát a jejich obsah je totožný se všemi jazykovými jevy přirozeného jazyka. Korpus je tak možné využívat pro jazykový výzkum na reálných datech v rozsahu, který dříve nebyl možný. Korpus je charakteristický tím, že se jedná o neměnnou referenční entitu, u které se dá vždy zjistit její přesná velikost, počet různých jazykových jevů apod. [29]. Tyto vlastnosti se pak v oblasti zpracování přirozeného jazyka jeví pro algoritmy jako klíčové z důvodu natrénování vnitřních konstrukcí [29].

Pro snadnější zpracování korpusu se velmi často provádí anotace jeho textové části [29]. Anotace pak přidává korpusu metainformace o textech, jako například kdo je autorem, odkud text pochází apod. Anotace může také přidávat informace jednotlivým jazykovým jevům, kdy mezi anotace jazykových jevů patří tagování a lemmatizace [29], které budou blíže popsány dále.

1.2. Segmentace textu

Člověk se s rozeznáváním jednotlivých větných celků v textu setkává prakticky denně. Z toho důvodu pro něj není tolik složité určit například, zdali tečka v textu určuje konec věty, datum, titul apod. Naproti tomu, pro počítač se text jeví jako pouhá sekvence znaků, a proto je nutné provádět předzpracování dat. Jako první se obvykle provádí segmentace textu na jednotlivé lingvistické jednotky [42]. Hlavní úlohou segmentace je rozdělení přirozeného textu na menší jednotky s co možná největší přesností. Nejčastějším případem segmentace je pak segmentace větná [42]. Aby byla segmentace textu provedena co možná nejpřesněji, je důležité, aby počítač uměl správně rozeznat začátek a konec věty čili správně určil interpunkční znaménko, které ho označuje.

1.3. Tokenizace a tagování

Následujícím krokem při zpracování textu pro extrakci informací je rozdělení vět na jednotlivé větné členy. A právě tento proces obstarává *tokenizace*, jejímž úkolem je tedy rozeznat a označit sekvence znaků, které spolu reprezentují slovo s určitým sémantickým významem – *token* [30]. Standardně jsou od sebe jednotlivé tokeny odděleny pomocí bílých znaků (mezer). Výsledek tokenizace je předveden na následující větě [4]:

Extrakce informací z nestrukturovaného textu.

Výsledek, kdy závorky vždy značí jeden token, pak vypadá následovně:

[Extrakce] [informací] [z] [nestrukturovaného] [textu] [.]

Podoba tokenizace je závislá na analyzovaném jazyku. U anglických textů je efektivní použít k oddělení jednotlivých tokenů bílé znaky a interpunkci. Výhodou této strategie je její snadná implementace, existuje však mnoho případů, kdy tato strategie neodpovídá požadovanému chování (například u akronymů a zkratek) – tomu lze zabránit jejich dopřednou detekcí [30]. Kvalita provedení a výsledek tokenizace je důležitá, neboť tento výsledek je dále použit jako vstup pro další metody z oblasti *Natural language processing* (dále NLP, zpracování přirozeného jazyka). Případná chyba by se tedy dále přenášela a ovlivnila by všechny následující algoritmy [1].

Dále je pak možné ke zpracovávanému textu přidávat metainformace, které blíže specifikují význam jednotlivých tokenů na základě kontextu věty. Tomuto kroku se říká *Part-of-Speech tagging – tagování* [30]. Během tagování (přiřazování tagu) se uvažuje kontext věty, ve které se token vyskytuje a jeho vztah s ostatními tokeny. Jednotlivým tokenům se přidá informace o slovním druhu, jmenný rod, číslo, přivlastňovací rod, osoba a čas. Pro tagování je možné využít algoritmů založených na ručně psaných pravidlech (*Rule-based taggers*) nebo pravděpodobnostních algoritmů (*Probabilistic tagger*) [19]:

- Ručně psaná pravidla – algoritmy, jež závisí na slovnících, kde jsou určeny správné tagy vzorových tokenů a na seznamu pravidel, které určují tagy v závislosti na pozici ve větě.
- Pravděpodobnostní algoritmy – k naučení daného modelu textu používají vzorové korpusy, které jsou správně otagovány. Jsou založeny na principu skrytých Markovových modelů (bude vysvětleno v kapitole 5.1.3 Skryté Markovovy modely).

1.4. Stemming

Za stemming lze označit proces, který má za úkol, na základě jazykových pravidel odstranit z původního slova všechny morfologické části, které nejsou součástí kořene slova a nalezený kořen je pak vrácen jako výstup [21]. Konkrétně se jedná o nalezení kmene (stemu) slova. Kmen slova je velmi často zaměňován za kořen, se kterým se ovšem může a nemusí shodovat [21].

U stemmingu není využíváno morfologické analýzy, jako tomu je v případě lemmatizace (bude vysvětleno v následující kapitole), ale aplikace sady přepisovacích pravidel na data [21]. Z tohoto důvodu je proces stemmingu výrazně časově úspornější než lemmatizace, kdy je nutné zpracovávat velké množství dat tvořící kontext, který umožňuje určit význam termů. Stemmer pracuje pouze s jednotlivými slovy a pokud se toto slovo shoduje s levou stranou přepisovacího pravidla, pak se dané pravidlo aplikuje. Díky tomuto přístupu však může docházet k nesprávným transformacím, které mají za následek vznik nekorektních slov. Velkým problémem je také fakt, že existují slova se stejným kořenem, která mají rozdílný význam [21].

Stemming najde uplatnění především u jazyků z Indo-Evropské skupiny, kdy jednotlivá slova vznikají flexí či obecně gramatickými pravidly, typicky však užitím afixu – předpony (prefixu) a přípony (sufixu). U přípon jsou nejčastěji rozlišovány tři základní typy [42]:

- A-přípony (attached) – jedná se o slovo připojené k jinému slovu. Příkladem může být portugalština, kdy jsou tato slova připojená pomocí pomlčky (jejich odstranění je poměrně snadné). Složitější situace je pak například u italštiny, kde se takto slova připojují bez oddělovače.
- I-přípony (inflectional) – úprava slov na základě pravidel jazyka, za existence výjimek. Příkladem může bát tvorba minulého času v angličtině za využití přípony *ed*.
- D-přípona (derivational) – z jednoho slova je vytvořeno slovo, které pak dokonce může být i jiného slovního druhu (readable, washable).

V následující části budou představeny některé přístupy a algoritmy, které jsou pro stemming a lemmatizaci využívány [42]:

- **Algoritmy hrubé síly (vyhledávací)**

  Algoritmy hrubé síly představují slovníkovou metodu, která využívá tabulku, kde se nachází jednotlivé dvojice vyskloňovaný tvar – kořen slova. Tato tabulka je pak pro

každé slovo prohledávána, z čehož plyne i největší nevýhoda této metody. Jedno slovo může být odvozeno od více základů, které by měly mít v tabulce dva různé záznamy. Hledání stemů touto metodou je tak vhodné především u jazyků, u kterých je nepravidelná flexe (například angličtina: mouse, mice, …).

- **Algoritmy pro odstranění přípony (Suffix-stripping algorithms)**

  Tyto algoritmy využívají seznam pravidel a nepotřebují tak žádný seznam slov (slovník). Optimalizace a vývoj seznamu pravidel představuje jednodušší činnost, než je sestavení tabulky pro algoritmus hrubé síly. Podmínkou pro sestavení kvalitního seznamu pravidel je dobrá znalost daného jazyka. Problém pak představují jednotlivé nepravidelnosti a výjimky, daného jazyka, jež nelze užitím této metody vyřešit. Tyto algoritmy tak jsou aplikovány pouze na jazyky, pro které jsou pravidla jasně daná a obsahují velmi malou množinu výjimek.

- **Stochastické algoritmy**

  Stochastické (statistické) metody jsou postaveny na základu strojového učení. Obvykle se pak jedná o metody učení bez učitele, pomocí nichž jsou odhadnuty parametry stematizačního modelu. Nalezení stemu je provedeno na základě pravděpodobnostního modelu, který je obvykle představován množinou pravidel, jež jsou podobná těm u předchozího modelu. Nejpravděpodobnější stem daného slova je pak nalezen na základě naučených pravidel. Z výše uvedeného jsou patrné hlavního výhody této metody, není nutná spolupráce jazykového experta či tvorba pravidel.

- **Hybridní systémy**

  Výsledku je dosaženo pomocí kombinace výše popsaných přístupů. Příkladem může být kombinace vyhledávacích tabulek a odstraňování přípon (pokud slovo není ve vyhledávací tabulce, je použit algoritmus pro odstranění přípony).

U stemmingu jsou rozlišovány následující dvě chybové metriky [42]:

- **Over-stemming** – chyby způsobené nadměrným stemmingem, jedná se o přehnaně agresivní ořezávání slov, což má za následek přiřazení příliš mnoha stemů jednomu lexému (jedná se o chybu typu *false positive - FP*)
- **Under-stemming** – způsobené nedostatečným stemmingem. Existence slov, která by měla být přiřazena stejnému lexému, ale nejsou (chyba typu *false negative - FN*).

Cílem stemmingu je minimalizovat tyto chyby, což představuje najít jakousi rovnováhu mezi nimi, neboť zmenšení jedné chyby obvykle znamená zvětšení té druhé.

1.5. Lemmatizace

Cíl lemmatizace je velmi podobný jako cíl stemmingu, avšak je dosažen odlišnými prostředky. Jak lemmatizace, tak i stemming mají za úkol zmenšit prostor slov pro danou úlohu (minimalizovat slovník), a z toho důvodu jsou občas v některých případech zaměnitelné [42]. Proces lemmatizace spočívá v převodu slov, která vznikla odvozováním, skloňováním nebo časováním do jejich základního tvaru (lemma). Tento převod je prováděn na základě morfologické analýzy, která tento základní tvar slova určuje [42]. Morfologická analýza není prováděna nad jednotlivými slovy, ale vždy se posuzuje větší část textu, aby bylo možné na základě kontextu správně určit význam slova a jeho základní tvar, který může být shodný s odvozeným či skloňovaným tvarem slova s odlišným významem. Výstupem lemmatizace jsou jazykově korektní slova, která v daném jazyce skutečně existují, což znamená, že v případě potřeby zachování smysluplných slov na výstupu je lemmatizace nezastupitelná [27]. Z výše uvedeného je patrné, že je nemožné provést učení za pomoci metod učení bez učitele. Vždy je nutný ručně anotovaný korpus či vytvořená sada pravidel [42]. Rozdílné výsledky stemmingu a lematizace jsou předvedeny na dvou slovech z anglického jazyka *walk* a *meeting* v Tabulce 1:

Tabulka 1: Výsledek stemmingu a lemmatizace

	Stemming	Lemmatizace
walk	*walk*	*walk*
meeting	*meet*	*meet* nebo *meeting*

Zdroj:[27]

Rozdílné výsledky jsou patrné z výše uvedené Tabulky 1. U slova *walk* je výsledek stemmingu a lemmatizace stejný. Ovšem u slova *meeting*, které může být buď podstatným jménem nebo slovesem, je však výsledek rozdílný. Stemming pro obě varianty vrací kmen *meet*, oproti tomu lemmatizátor sloveso *meet* transformuje na *meet* a pro variantu podstatného jména ponechává tvar *meeting* [27].

1.6. Odstranění Stop-words

Stop-words (stopslova) představují slova, která se v textových dokumentech vyskytují bez závislosti na konkrétním tématu [27]. V případě českého jazyka mohou být stop-words

například spojky, předložky, zájmena apod. Tato slova se tak vzhledem k dolování znalostí z textu považují za irelevantní a jejich odstranění je velmi často používanou metodou předzpracování textových dat, která může zvýšit efektivitu a účinnost aplikovaných metod dolování znalostí [27].

Jednotlivé seznamy stop-words neboli negativní slovníky mohou být buď obecné, nebo doménově orientované. Obecné slovníky bývají veřejně dostupné a zahrnují standardní slova bez významu. Doménově orientované seznamy obsahují doménově specifická stop-words, která nenesou informační hodnotu z pohledu zkoumané domény či kontextu.

Výsledkem aplikace negativního slovníku na textový dokument je mazání irelevantních slov, čímž dojde ke snížení celkového počtu termů ve slovníku [27].

2. OBLASTI VYUŽITÍ TEXT MININGU

Po předzpracování dokumentů je další fází text miningu analýza textu, a právě této fázi byla věnována tato kapitola. Analýza textu představuje automatické analyzování dokumentů, získávání strukturované informace a její následný rozbor. Jinými slovy v této fázi dochází k analýze předzpracovávaného dokumentu a vygenerování termů, které představují základní prvky sloužící k analýze.

Text tak poté může být například roztříděn podle témat, která danou oblast reprezentují, nebo mohou být vyhledána klíčová slova nebo může být vytvořen jeho souhrn. Nástroje, které jsou pro analýzu textu využívány, jsou velmi různorodé, může se jednat například o [13]: kategorizaci textů, shlukování dokumentů, filtrování dokumentů, detekci duplikace, extrakci informací, sumarizaci textů apod. O výše zmíněných nástrojích pojednává tato kapitola.

2.1. Kategorizace textů

Kategorizace textů představuje úkol, který řeší zařazení dokumentů do předem definovaných kategorií, jako jsou politika, sport, ekonomika apod. Do jednotlivých kategorií je dokument zařazován na základě jeho obsahu, tématu, názvu nebo klíčových slov [42]. Každý text může být zařazen do jedné, více nebo žádné kategorie. Toto zařazení se typicky provádí na základě četnosti slov výskytu slov nebo roztřídění pomocí stejného názvu dokumentu [42].

Automatická detekce tématu dokumentu najde uplatnění například při správě rozsáhlých uložišť nebo k eliminaci nevyžádané pošty. Dále lze také využít při analýze webového průzkumu nebo při reklamacích, kdy SW [13] roztřídí jednotlivé typy odpovědí do kategorií a příslušné oddělení se pak zabývá jen těmi odpověďmi, které mají informační hodnotu [13].

Z výše uvedeného je patrné, že kategorizace textů představuje úlohu klasifikace. Na základě extrakce slov z dokumentu vznikne jejich množina, která následně tvoří profil dokumentu. Profily jednotlivých kategorií jsou potom vytvořeny pomocí profilů dokumentů, které do jednotlivých kategorií náleží. Profily kategorií lze vytvořit manuálně, což je ovšem časově velmi náročné, a tudíž i drahé, proto se vytváří automatické klasifikátory, které profily tříd vytváří automaticky. Podle způsobu výběru kategorií se rozlišuje [13], [5], [37]:

- *mono-klasifikace* (každý dokument patří do jedné třídy) – dokumentu je přiřazena ta třída, která dosáhla nejvyšší ohodnocení. Speciálním případem je pak binární klasifikace, kdy existují jen dvě třídy – patří × nepatří.

- *multi-klasifikace* (dokument může patřit do žádné nebo více tříd) – dokumentu jsou přiřazeny třídy, jejichž ohodnocení přesáhne určitou mez.

Pro kategorizaci textů lze využít některou z klasifikačních metod, které mohou být rozděleny do následujících skupin podle toho, jak přistupují k trénování a hledání odpovídajících kategorií [42]:

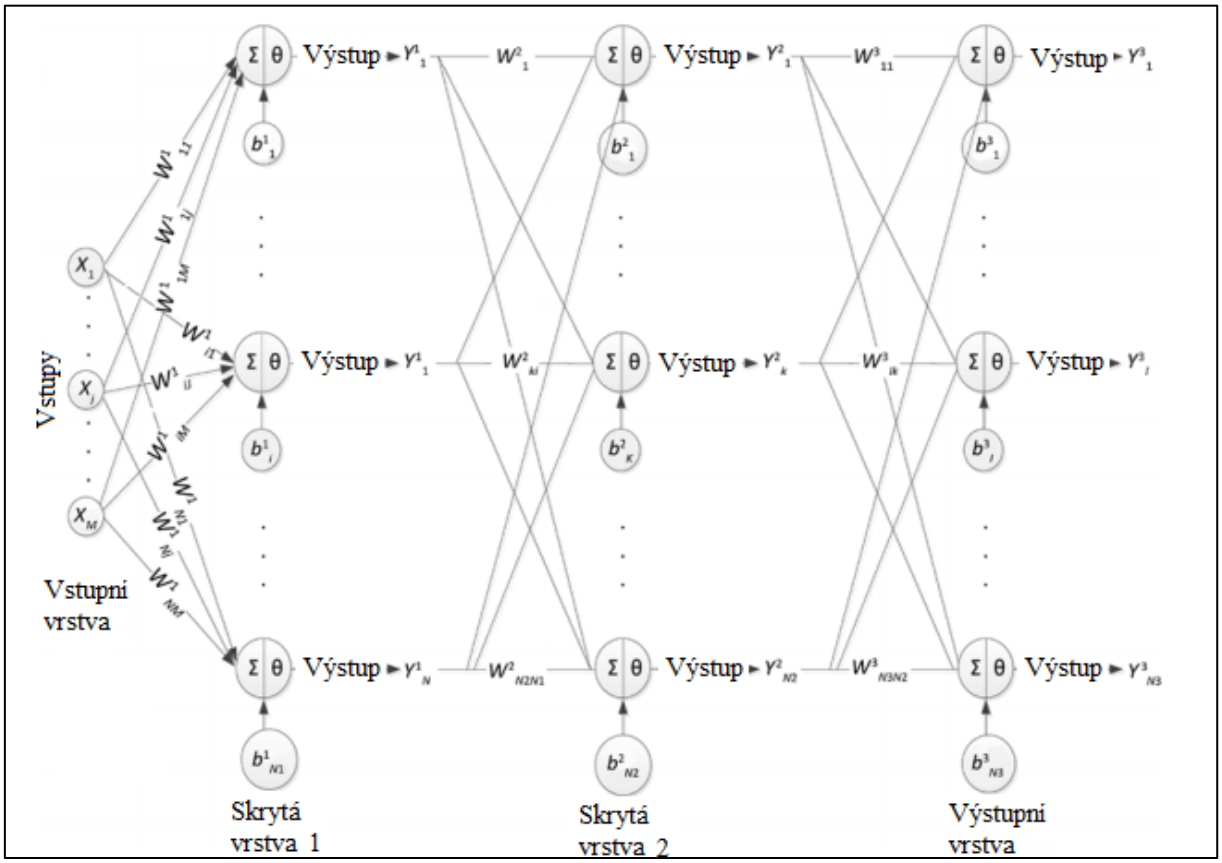
- Metody s klasifikačními pravidly – při trénování hledají pravidla, která danou kategorii popisují. Pravidla pak mohou být využita pro vytvoření rozhodovacího stromu.
- Lineární klasifikátory – dokumenty i kategorie popisují pomocí lineárního vektoru. Při trénování se vektory nastavují tak, aby nejlépe odpovídaly trénovacím dokumentům. Zařazování dokumentů pak probíhá na základě porovnávání vektoru dokumentu s vektory všech kategorií. Následně je vybrána ta kategorie, která má nejvyšší skóre, popřípadě kategorie, u nichž bylo překročeno určité skóre.
- Metody založené na příkladech – nejprve se k dokumentu naleznou dokumenty jemu podobné a na základě jejich zařazení se určí kategorie novému dokumentu.
- Upravené algoritmy strojového učení – například genetické algoritmy, fuzzy množiny apod.

2.1.1. Umělé neuronové sítě

Umělé neuronové sítě (dále jen neuronové sítě) představují jakési napodobení myšlení lidského mozku. Jedná se o modely, jež jsou inspirovány biologickými nervovými sítěmi. Neuronové sítě jsou taktéž velmi často označovány za „černé skříňky“, jelikož detailní popis vnitřní struktury systému není dobře interpretovatelný [1]. Na vnitřní strukturu systému jsou tak kladeny předpoklady, které umožní popsat chování systému funkcí provádějící transformaci vstupu na výstup. Neuronové sítě lze použít všude tam, kde v hlavní roli modelovaného procesu vystupuje náhoda a deterministické závislosti jsou natolik složité, že je nejsme schopni separovat či analyticky identifikovat.

Práci neuronové sítě lze rozčlenit do dvou fází. Ve své první fázi se neuronová síť snaží naučit nastavit své parametry tak, aby co možná nejlépe vyhovovaly požadované topologii sítě. Následuje druhá fáze, kdy neuronová síť produkuje výstupy na základě znalostí, které se naučila v první fázi [11].

Jak již bylo zmíněno výše, důležitou vlastností neuronové sítě je schopnost učení se, což zjednodušeně představuje automatické nastavování vah mezi jednotlivými neurony. Pokud by byla řešena velmi jednoduchá úloha (porovnání 1 a 0), mohl by k jejímu vyřešení stačit pouze jeden neuron. V drtivé většině případů jsou však řešené úlohy mnohem složitější a k vyřešení je potřeba větší množství neuronů ve více vrstvách, což představuje hlubokou neuronovou síť. Tyto sítě mají kromě vstupní a výstupní vrstvy i více skrytých vrstev [2]. Optimální počet vrstev neuronové sítě představuje otázku, na kterou nelze zcela jednoznačně odpovědět, jelikož tento počet závisí na složitosti řešené úlohy. Příklad neuronové sítě se dvěma skrytými vrstvami je na Obrázku 1, kde W^k_{ij} označuje váhu spojení mezi j-tým neuronem vrstvy k-1 a i-tým neuronem vrstvy k. Y_i^k pak představuje výstup i-tého neuronu k-té vrstvy. Výstup neuronové sítě je výstup poslední vrstvy, která je nazývána jako vrstva výstupní [2].



Obrázek 1: Hluboká neuronová síť

Zdroj: Upraveno dle [2]

2.1.2. Rozhodovací a regresní stromy

Cílem rozhodovacích stromů je sekvenčně rozdělit data tak, aby rozdíly v závislé proměnné byly maximální. Rozhodovací strom postupně třídí data do odlišných skupin či větví, které vytvoří nejsilnější separaci hodnot závislé proměnné [1]. Jinými slovy jsou trénovací data

postupně rozdělována na menší a menší podmnožiny tak, aby v těchto podmnožinách převládaly příklady jedné třídy. Tento postup, kdy na počátku je jedna množina trénovacích dat a na konci podmnožiny složené z příkladů téže třídy, je často nazýván *top down induction of decision trees* (dále TDIDT). Jak vyplývá z názvu tohoto postupu, postupuje se metodou specializace v prostoru stromů shora dolů počínaje stromem s kořenem. Jednotlivé kroky algoritmu TDIDT jsou [5]:

1. Zvol jeden atribut jako kořen dílčího stromu,
2. data v tomto uzlu rozděl na podmnožiny na základě hodnot zvoleného atributu a přidej uzel pro každou podmnožinu,
3. pokud existuje uzel, pro který nepatří všechna data do téže třídy, pak pro tento uzel opakuj postup od bodu 1, jinak skonči.

Klíčovou otázkou výše uvedeného algoritmu je, jaký atribut vybrat, aby došlo k nejlepšímu odlišení příkladů různých tříd. K tomu slouží charakteristiky atributů převzaté z teorie informace nebo pravděpodobnosti, kdy se jedná o: *entropie*, *informační zisk*, *poměrný informační zisk* nebo *Gini index* [5].

Entropie je pojem, který se v přírodních vědách používá pro vyjádření míry neuspořádanosti nějakého systému. V teorii informace je pak definována jako funkce dle [5]:

$$H = -\sum_{t=1}^{T}(p_t log_2 p_t), \tag{1}$$

kde p_t je pravděpodobnost výskytu třídy *t* počítaná na určité množině instancí a *T* je počet tříd. Existují-li dvě třídy a *p*=1, což znamená, že všechny instance patří do jedné třídy nebo všechny atributy do třídy nepatří (*p*=0), pak je hodnota entropie nulová. Jsou-li však obě třídy zastoupeny stejným počtem instancí (*p*=0,5), je entropie maximální [5].

Výpočet entropie pro jeden atribut se provede tak, že pro každou hodnotu *v*, kterou může uvažovaný atribut *A* nabýt, se spočítá podle uvedeného vzorce entropie $H\big(A(v)\big)$, vzorec (2) na množině instancí, které jsou pokryty kategorií $A(v)$ [5].

$$H\big(A(v)\big) = -\sum_{t=1}^{T}\frac{n_t\big(A(v)\big)}{n\big(A(v)\big)} log_2 \frac{n_t\big(A(v)\big)}{n\big(A(v)\big)}. \tag{2}$$

Následně se dle vzorce (3) provede výpočet střední entropie *H(A)*, jako vážený součet entropií $H\big(A(v)\big)$, kdy váhy v součtu jsou relativní četnosti kategorií *A(v)* na trénovacích datech [5]:

$$H(A) = -\sum_{v\in Val(A)}\frac{n\big(A(v)\big)}{n} H\big(A(v)\big). \tag{3}$$

Pro větvení stromu se pak vybere ten atribut, který má nejmenší entropii *H(A)*.

Informační zisk i ***poměrný informační zisk*** jsou míry, které jsou odvozené z entropie [5]. Informační zisk podle (4) představuje rozdíl mezi entropií pro celá data (pro cílový atribut) dle (5) a pro uvažovaný atribut. Měří tak redukci entropie, která je způsobená volbou atributu *A*:

$$Zisk(A) = H(C) - H(A), \tag{4}$$

kde

$$H(C) = -\sum_{t=1}^{T} \frac{n_t}{n} log_2 \frac{n_t}{n}\,. \tag{5}$$

V případě informačního zisku se hledá atribut s maximální hodnotou. Výše uvedená kritéria neberou v úvahu počet hodnot zvoleného atributu, a právě z tohoto důvodu se pro volbu atributu používá poměrný informační zisk, který kromě entropie bere v úvahu i počet hodnot atributu, vypočtený dle [5]:

$$Poměrný\ zisk(A) = \frac{Zisk(A)}{Větvení(A)}\,, \tag{6}$$

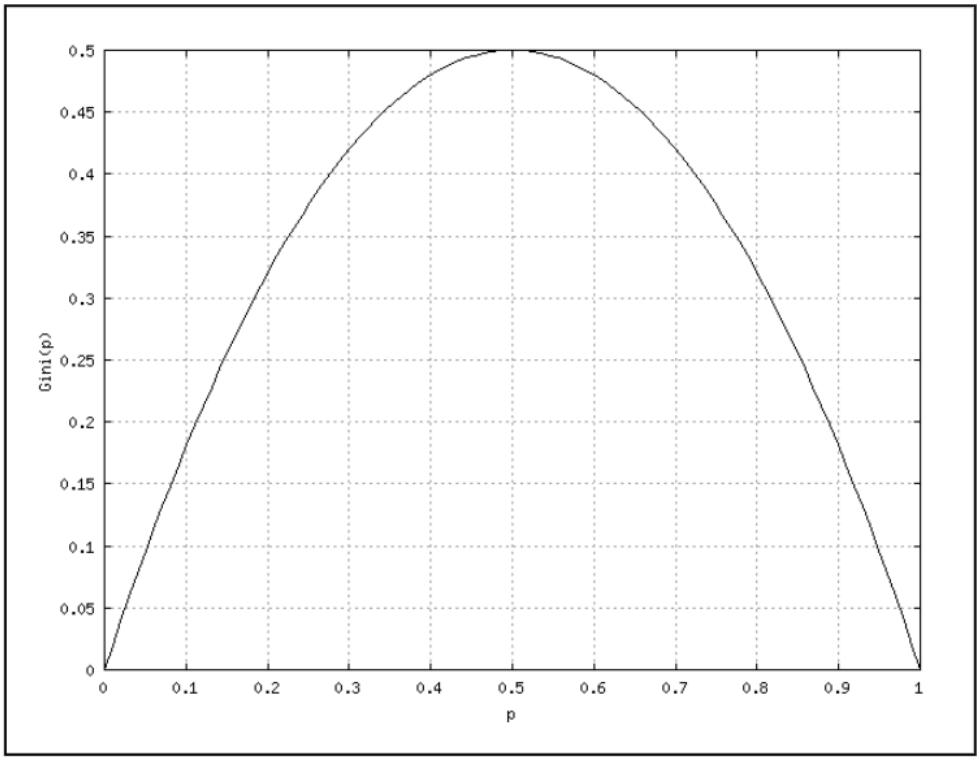
kde

$$Větvení(A) = -\sum_{v \in Val(A)} \left(\frac{n(A(v))}{n} log_2 \frac{n(A(v))}{n} \right). \tag{7}$$

Gini index představuje obdobnou roli, jakou měla v předchozích případech entropie, spočítá se podle vzorce (8) jako [5]:

$$Gini = 1 - \sum_{t=1}^{T} (p_t{}^2), \tag{8}$$

kde p_t je relativní počet instancí *t*-té třídy zjišťovaný na některé jejich podmnožině. Pokud se budou uvažovat dvě třídy, pak je hodnota Gini indexu maximální v tom případě, kdy jsou jednotlivé instance rovnoměrně rozděleny mezi obě třídy. Hodnota indexu je minimální, pokud všechny instance patří do jedné třídy. Závislost Gini indexu na pravděpodobnosti jedné ze dvou tříd je patrná z Grafu 1 [5].



Graf 1: Gini index

Zdroj:[5]

Pro jeden atribut se hodnot Gini indexu vypočítá podle vzorce (9) obdobně jako entropie, a to tak, že pro každý atribut se spočítá vážený součet indexu pro jednotlivé hodnoty atributu, kdy váhy budou relativní četnosti příslušných hodnot [5]:

$$Gini(A) = \sum_{v \in Val(A)} \frac{n(A(v))}{n} Gini(A(v)), Gini(A(v)) = 1 - \sum_{t=1}^{T} \left(\frac{n_t(A(v))}{n(A(v))} \right)^2 . \qquad (9)$$

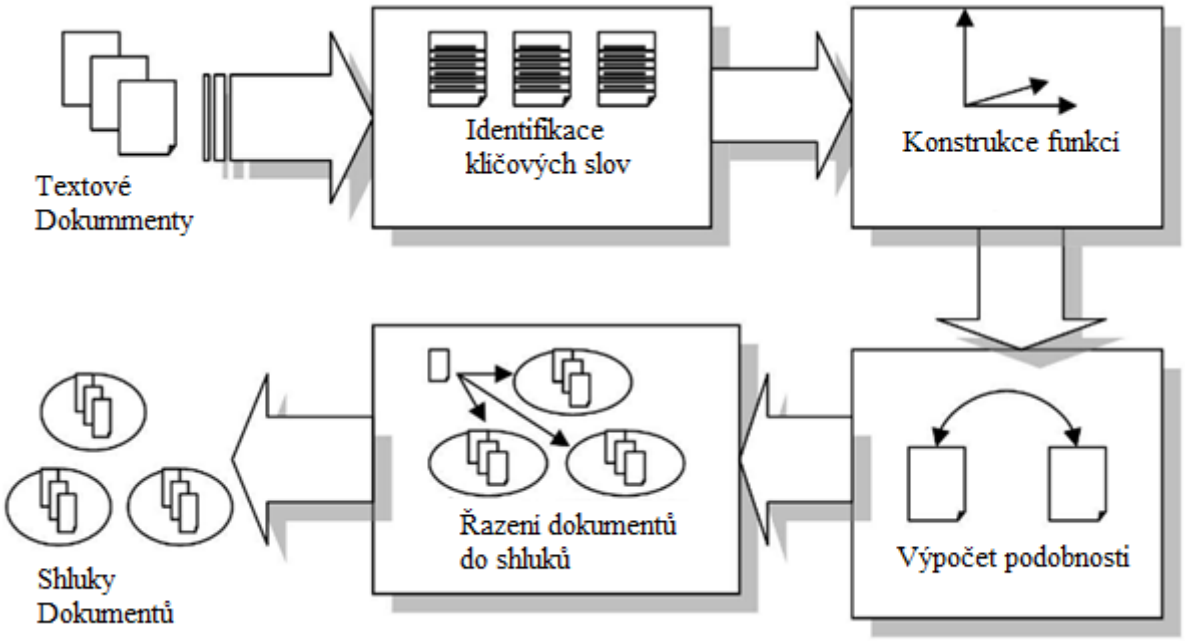
Pro větvení se následně použije ten atribut, který bude mít hodnotu indexu nejmenší.

2.2. Shlukování textů

Dalším způsobem analýzy textů je identifikace textových dokumentů pomocí shlukové analýzy, což představuje rozdělování množiny instancí do podmnožin tak, že v každé podmnožině jsou instance s podobnými vlastnostmi a zároveň rozdíly mezi jednotlivými podmnožinami jsou co možná největší. Z výše uvedeného je patrné, že každý dokument musí být zařazen do jedné skupiny. Pokud je pak například pro rozlišení použit obsah jednotlivých dokumentů, pak různé skupiny korespondují s různými náměty a tématy obsaženými v tomto souboru dokumentů [5]. Celý proces shlukování textů je dobře patrný z Obrázku 2.

Mezi shlukováním textů a kategorizací textů lze najít jistou míru podobnosti, rozdíl spočívá v tom, že kategorizace dokumentů jednotlivé dokumenty třídí do předem stanovaných skupin,

zatímco shlukování textů jednotlivé skupiny vytváří na základě podobnosti analyzovaných dokumentů [1]. Před samotným rozdělováním nejsou známé vlastnosti a velmi často ani počet jednotlivých podmnožin. V případě shlukování textů není potřebná žádná trénovací množina, jedná se tak o učení bez učitele. Dokumenty jsou popsány lineárními vektory, které jsou vytvořeny na základě jejich atributů a výpočet podobnosti je pak prováděn pomocí vzdálenostní funkce [1].



Obrázek 2: Proces shlukování textů

Zdroj: Upraveno dle [44]

2.3. Analýza sentimentu

Sentiment představuje lidský pocit, postoj či názor na určitý podnět, kdy tento je zpravidla buď pozitivní, nebo negativní čili vyjadřuje buď náklonost, nebo odpor. Analýza sentimentu si dává za úkol automaticky extrahovat tyto subjektivní informace z textu a určení tak postoj autora [1].

V rámci analýzy sentimentu existují dva základní typy analýzy. Prvním typem je analýza, jež je zaměřená na rozhodování, zdali je úsek přirozeného jazyka objektivní nebo subjektivní – určuje, zda se jedná o prostý fakt, nebo o názorově zbarvené sdělení [23]. Druhý typ je detekce polarity [23]. Cílem této analýzy je zjistit, zda sentiment obsažený v textu je pozitivní či negativní. V jeho nejjednodušší formě se jedná o pouhou klasifikaci do těchto dvou základních skupin, označujeme ji jako bipolární detekce. Kategorií však může být i více, velmi často jsou uvažovány tři kategorie, kromě výše uvedených je to navíc ještě kategorie neutrální, která značí

objektivní příspěvek bez sentimentu. Další možností je vytvoření většího počtu kategorií, pomocí nichž můžeme vytvořit jakousi stupnici sentimentu (lze hodnotit, do jaké míry je postoj pozitivní či naopak negativní). Analýzy sentimentu lze dále dělit dle objemu dat v přirozeném jazyce, se kterými autor pracuje [23]:

- sentiment na úrovni jednotlivých slov,
- fráze/*n*-gramy – jedná se o dvě či více slov, která k sobě mají sémantický vztah,
- větný sentiment – věty či úseky textu,
- celek – příspěvek, článek, dokument.

2.4. Shrnutí textu

Tato analýza najde uplatnění především tam, kde jsou zpracovávány rozsáhlé textové dokumenty v poměrně krátkém čase. Manuální přečtení každého textu je velmi zdlouhavá činnost, a právě tuto časovou náročnost tohoto problému řeší tato analýza, která vytvoří shrnutí originálního dokumentu nebo jen některé jeho části (kapitoly, odstavce) [23].

Princip této metody je v tom, že software (SW) skenuje text a z dokumentu vybere jen nejdůležitější části. Důležitost je ve většině případů definována uživatele, ale není to podmínkou, uživatel stanoví tzv. koncepty, kterými jsou regulární výrazy nebo gramatická pravidla a podle nich jsou pak dokumenty prohledávány [23]. Tento postup je vhodný v případě, kdy je známo, co je pro nás důležité, abychom zjistili co nejvíce informací o dané oblasti zájmu. SW poté najde požadované informace a výstup bude představovat smysluplná informace vytěžená z rozsáhlého textu [23].

2.5. Extrakce informací

Extrakce informací (dále IE) představuje ve své podstatě proces, který je založený na automatickém získávání strukturovaných dat z nestrukturovaného či částečně strukturovaného textu. IE může být použita jako jedna z metod předzpracování, kdy jsou jednotlivé informace dále analyzovány, ale i jako metoda samotné textové analýzy. Pokud se mluví o IE jako o metodě textové analýzy, jedná se o extrakci předem specifikovaných informací, které jsou následně organizovány, čímž vznikne strukturovaný soubor informací, se kterým lze vykonávat další operace. Systémy IE mohou být následně aplikovány na řešení následujících úloh [28], [9]:

- Rozpoznávání jmenných entit (Named entity recognition, NER) – identifikace a klasifikace jmen, názvů měst, institucí apod.
- Rozlišení koreference (Coreference resolution, CO) – představuje identifikaci vztahů mezi jednotlivými entitami.
- Konstrukce prvků šablony (Template Element, TE) – k NER přidává popisy atributů.
- Konstrukce relací šablony (Template Relation, TR) – identifikuje vztahy mezi TE.
- Vytvoření šablony scénářů (Scenario Template, ST) – vytváří událostní scénáře, do nichž vkládá TE a TR.

Rozpoznávání jmenných entit

NER představuje nejjednodušší a zároveň nejzákladnější úlohu IE. Pomocí NER jsou identifikovány jména osob, názvy měst, institucí, datumy, množstevní jednotky apod. Tato metoda je při svém správném použití schopna dosáhnout vysoké přesnosti okolo 95 % [28].

Rozlišení koreference

CO slouží k identifikaci identických vztahů mezi entitami, které se v textu nacházejí, nejčastěji se vyskytuje jako podstatné jméno a zájmeno, které ho v textu dále zastupuje [28]. Dále je také možné, že některá entita má více jmen či zkratky. Hlavní význam CO spočívá v tom, že vytváří stavební bloky pro procesy konstrukce TE a vytváření ST. CO asociuje popisné informace o entitách, které jsou různě rozmístěny po celém textu. CO musí řešit následující problémy [28]:

- k zájmenu najít správnou entitu,
- různé názvy pro stejnou entitu,
- časová koreference.

Konstrukce prvků šablony

TE vychází z procesů NER a CO, kdy jednotlivým entitám přiřazuje atributy [9]. Jedná se o přiřazení různých aliasů entitě, nad rámec toho, jak to dělá NER. Příkladem může být vláda, která může být také jako státní představitel nebo exekutiva. Dále TE přidává různým entitám jejich typ (osoba, město, úřad apod.).

Konstrukce relací šablony

Hledá vztahy mezi prvky šablony, které byly nalezeny pomocí TE. Tyto vztahy mohou být například: zaměstnanec a zaměstnavatel, nadřízený a podřízený nebo například entita, která geograficky leží v jiné entitě. TR představuje jednu z hlavních úloh systémů IE, přičemž v reálném světě mohou být vztahy téměř nekonečně rozmanité [9].

Vytvoření šablony scénářů

Šablona scénáře představuje nejžádanější výstup ze systémů pro IE. ST spojuje entity (vytvořené pomocí TE) a relace (vytvořené pomocí TR) do popisu událostí. Tato úloha je však také úlohou nejtěžší, což je způsobeno především tím, že se na ní podílejí všechny předcházející úlohy [9].

Na základě výše uvedeného je patrné, že z textu mohou být extrahovány následující čtyři prvky [42]:

- Entity – základní stavební bloky,
- Atributy – znaky extrahovaných entit,
- Fakta – relace mezi entitami,
- Události – aktivity nebo výskyty zájmu, na kterém entity participují.

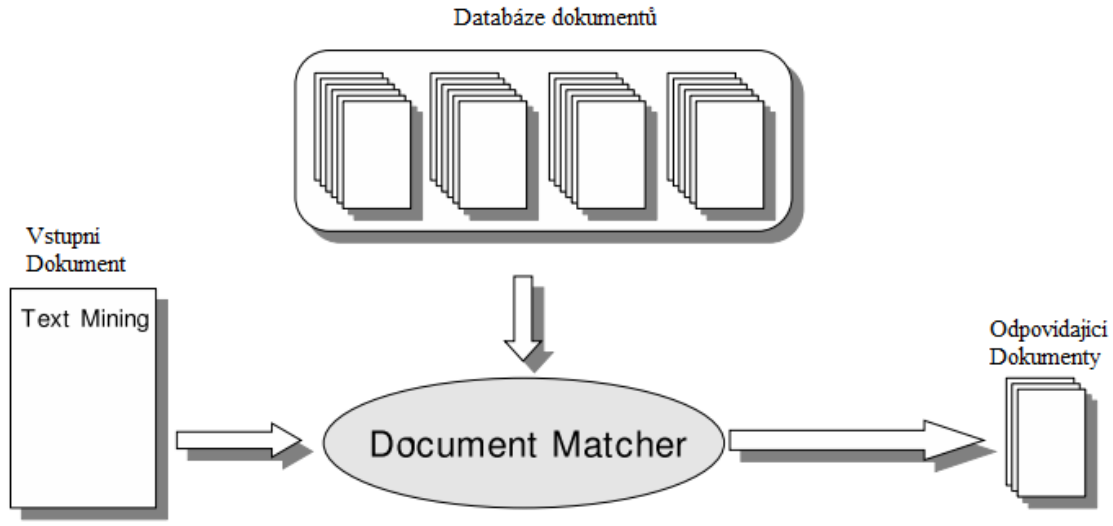
2.6. Získávání informací

Hlavním úkolem získávání informací (Information retrieval, IR) je co možná nejpřesněji a v co nejkratším čase získat dokumenty obsahující užitečné informace pro uživatele [42]. Vyhledané dokumenty jsou následovně řazeny dle určitého kritéria, příkladem takového kritéria může být podobnost mezi vektorem dokumentu a vektorem dotazu.

Získávání informací je tématem, které je nejčastěji spojováno s online dokumenty. Dokumentům jsou přiřazována určitá pojítka, kdy tyto pojítka jsou následně porovnávána s kolekcí dokumentů. Dokumenty, jež mají shodu s těmito pojítky, pak představují odpovědi na dotazy uživatele [42].

Pod výše zmíněnými pojítky si můžeme představit slova, která pomáhají identifikovat význam uložených dat. V případě internetového vyhledávání jsou hledaná slova přiřazena uloženým dokumentům. Nejpřesnější shody jsou pak prezentovány jako výsledky vyhledávání. Výše uvedený proces může být generalizován na dokument, kde je místo pojítek rozuměn celý dokument. Znamená to, že vstupní dokument je potom připojen ke všem prohledávaným

dokumentům a výsledek představují dokumenty s největší shodou [42]. Tento základní princip získávání informací je vyobrazen na Obrázku 3.



Obrázek 3: Získávání shodných dokumentů

Zdroj: Upraveno dle [42]

Z výše uvedeného je patné, že základním konceptem IR je měření a porovnávání podobnosti mezi dvěma dokumenty. Srovnání se provádí mezi dvěma dokumenty, a měří se, jak si jsou podobné. V případě vyhledávání na internetu, je za takový dokument považována i pouhá sada slov zadaných do vyhledávače. Měření podobnosti je možné zařadit mezi metody pro učení a klasifikaci, které jsou nazývány metody nejbližšího souseda [42].

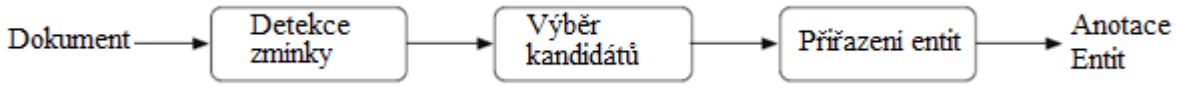
2.7. Asociace entit

Strojové zpracování textu představuje poměrně náročný problém. Schopnost označit jednotlivé entity v textu je klíčem k tomu, aby bylo možno danému dokumentu správně porozumět. Stejně tak jako slova, tak i jednotlivé entity mohou mít leckdy mnohoznačný význam [3]. Lidé jsou schopni při čtení dokumentů využívat jejich předchozích znalostí k rozšifrování této mnohoznačnosti a identifikaci správného významu. Pro strojové zpracování textu může tato mnohoznačnost způsobovat poměrně složité problémy. Klíčová součást strojového zpracování přirozeného jazyka je napojení na rozsáhlé báze znalostí jako je například Wikipedia. Mnohoznačný význam entit je pak řešený asociací entit pomocí unikátního klíče dané databáze [3].

Asociace entit tak dává čtenáři nástroj, jak o entitách zjistit další podrobné informace, které nemusí být obsaženy. Toto sémantické obohacení textových dokumentů má za následek

i zlepšení dalších úloh strojové zpracování přirozeného jazyka jako jsou sumarizace textu, kategorizace textu, detekce a sledování témat apod. [3]. Asociace entit je úlohou rozpoznávání významu entit v textu a jejich spojení s odpovídajícími záznamy v znalostní databázi.

Vstupní text, na kterém chceme provádět extrakci entit a jejich asociaci, budeme dále označovat jako dokument *d*. Úkolem tedy bude vygenerovat anotace entit pro celý dokument označený jako A_d. Každá anotace $a \in A_d$ je dána jako trojice $a = (e, m_i, m_t)$, kde *e* představuje konkrétní entitu (odkaz a záznam do dané znalostní databáze), m_i a m_t označují počáteční a konečný znak zmínky o entitě v dokumentu *d*. Zmínky o entitách se pak v A_d nesmějí překrývat [3]. V průběhu let se k asociaci entit objevil přístup, který je založený na zřetězení třech procesů, které jsou patrné z Obrázku 4.



Obrázek 4: Asociace entit

Zdroj: Upraveno dle [3]

Detekce zmínky je první procesem známým pod pojmem extraktor. Extraktor má za úkol označit část textu (zmínku), který může být spojen s entitou. Zmínka představuje rozpětí textu (souvislou sekvenci pojmů) v dokumentu *d*, která odkazuje na konkrétní entitu, přičemž uvedená entita může či nemusí mít odkaz na znalostní databázi. Běžně je tato detekce založená na rozsáhlém slovníku entit a jejich variací na které se odkazuje. Označování zmínky velmi úzce souvisí s rozpoznáváním jmenných entit a je zde kladen důraz na vysokou úplnost [3]. V praxi tento se tento krok provádí na surovém textu, a to před standardními kroky předzpracování jako jsou například tokenizace, odstranění stopslov apod.

Následujícím krokem je *Výběr kandidátů* označovaný jako Searcher. V tomto kroku dochází k vygenerování souboru entit ke každé zmínce. Vzhledem k tomu, že následující krok *Přiřazení entit* je označován jako nejnákladnější krok ze všech, Searcher by měl vyvážit ideální poměr mezi přesností a úplností. Ve výsledku by tak mělo dojít k zachycení správné entity (ke každé zmínce) z co možná nejmenšího počtu kandidátů [3].

Přiřazení entit představuje krok, kdy je na základě kontextu ke každé zmínce vybrána právě jedna nebo žádná entita. Moderní přístupy k *Přiřazení entit* uvažují tři typy důkazů [3]:

- předchozí význam entit a zmínek,
- podobnost v kontextu mezi textem, který obklopuje zmínku a uvažovanou entitou,

- soudržnost mezi všemi asociacemi entit v dokumentu.

V současnosti době existuje k asociaci entit celá řada systémů. Některé z nich jsou například Alcemy API, AYLIEN Text Analysis API, Google Cloud Natural Language API, Microsoft Entity Linking service, Open Calais a Rosette Entity Linking API [3]. Právě poslední jmenovaný nástroj Rosette Entity Linking API bude využit v této práce, kdy jeho bližší charakteristika bude provedena v kapitole 5.4.2 Nástroj Rosette Text Analytics.

3. PŘÍSTUPY K EXTRAKCI INFORMACE Z TEXTU

Extrakce informací (IE) představuje činnost, která spadá pod tzv. dolování v textech (TM), což představuje disciplínu na pomezí DM, strojového učení a počítačové lingvistiky. Hlavním úkolem extrakce informací je automatické převedení informací z nestrukturovaného textu či částečně strukturovaného textu do strukturované podoby [1].

Případem extrakce informací z textu je rozpoznávání pojmenovaných entit (NER). Jedná se o úlohu, kdy se v textu vyhledávají konkrétní informace a je tak umožněno jejich další zpracování. Pojmenované entity mohou představovat slova a sousloví, která označují jednotlivé objekty reálného světa, jako jsou například jména osob, organizací a míst či čísla a datum [25].

Pro NER existuje celá řada metod, většina těchto metod je závislá na použitém jazyku, existují však i metody, jak vytvořit systém detekující entity bez ohledu a použitý jazyk. Tyto metody však potřebují poměrně velký označkovaný korpus, na kterém se systém může naučit vztahy a závislosti mezi slovy a označenými entitami. Následující část této kapitoly je věnována nejvýznamnějším metodám IE, které zde jsou blíže vysvětleny [24].

3.1. Metody extrakce informací

Zpracování přirozeného jazyka a extrakce informací využívá pro správné vyhodnocování především metody strojového učení s učitelem. Systém se nejprve naučí principy, podle kterých zpracovává text, ze správně zpracovaného textu v korpusu, který poskytuje učitel a následně je může aplikovat při zpracovávání přirozeného jazyka. Z důvodu většího pokrytí jazykových konstrukcí, které přirozený jazyk nabízí, musí být korpusy, ze kterých se systém učí, velmi rozsáhlé [28]. Jelikož je přirozený jazyk velice rozmanitý, není možné postihnout všechny jazykové konstrukce. Systém se proto velmi často musí rozhodnout, jaké je nejlepší zpracování dané části textu, aby byl výsledek co možná nejpřínosnější pro uživatele [28].

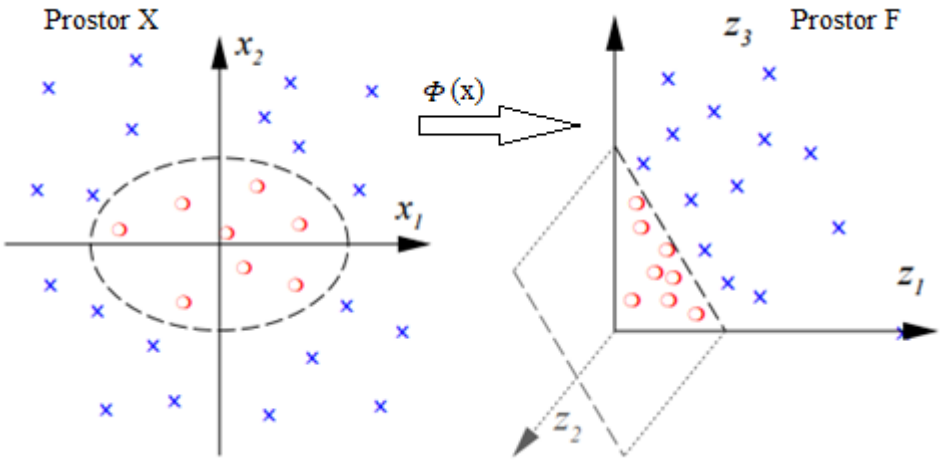
3.1.1. Pravidlový přístup

Pravidlový přístup se využije, pokud se rozlišují entity, jejichž podoba je dopředu známá a v rámci celého textu je charakteristická a jedinečná. Pokud se jedná o entity tohoto typu je nejsnazší cesta, pro jejich rozlišení, ruční sestrojení pravidel, které je nejlépe zachycují. Takovéto pravidlo může představovat regulární výraz – popis délky, kombinace malých a velkých písmen, pozice interpunkce apod. Systém se následně tyto pravidla naučí, aby se na jejich základě mohl rozhodovat při zpracovávání textu [35].

Nejobtížnějším úkolem u pravidlového přístupu je udržování sady pravidel [35]. Jejich velkou nevýhodou představuje ruční redefinice pravidel pro kvalitní výstup při změně charakteru dat. Velmi často je tak potřeba přidat nové informace o kontextu či příliš obecná pravidla dále více konkretizovat. Z výše uvedených důvodů bude následující část věnována metodám strojového učení, které tento nedostatek odstraňují [35].

3.1.2. Podpůrné vektorové stroje

Podpůrné vektorové stroje (SVM) představují metodu strojového učení, která je založená na zobrazení trénovacích dat do vysokorozměrného prostoru a jejich následnému oddělení pomocí nadroviny, která maximalizuje vzdálenosti od bodů reprezentujících zobrazená data [41]. Při klasifikaci se poté testovaná instance promítne do prostoru a třída se určí podle toho, na které straně bod skončil. Tyto metody jsou tak schopné nalézt nejenom triviální lineární hranice, ale i ty velmi složité, nelineární, za využití projekce původního prostoru do prostoru o vyšší dimenzi, ve které jsou třídy lineárně separovatelné [41]. SVM z pravidla provádí klasifikaci vstupů do dvou tříd, pokud je však tříd více, použije se více binárních klasifikátorů. Příkladem může být separace dvou tříd oddělených kružnicí či elipsou, které lze separovat lineárně přidáním dimenze, viz Obrázek 5. Na obrázku jsou původní příklady popsány pomocí dvou atributů x_1 a x_2, kdy tyto instance jsou pomocí elipsy rozděleny do dvou tříd. Pomocí datové transformace je následně úloha převedena do třírozměrného prostoru z_1, z_2, z_3 [5]. Tohoto principu lze obecně využít pro libovolná data. Při projekci dostatečným počtem dimenzí lze vždy najít separující nadrovinu, to znamená, že lze vždy oddělit N různých bodů v prostoru o dimenzi alespoň N-1 [17], [41].



Obrázek 5: Separace tříd s využitím SVM

Zdroj: Upraveno dle [43]

Zásadním problémem je, jak umístit lineární hranici mezi třídami, tak aby byla vedena co nejefektivněji z hlediska kategorizace budoucích dat. Řešení tohoto problému je komplikováno faktem, že v *d*-rozměrném prostoru je lineární oddělovač definován rovnicí, která má *d* parametrů, a tudíž zde hrozí nebezpečí, že dojde ke ztrátě obecnosti klasifikátoru „přeučením“ [41]. Z tohoto důvodu se algoritmy snaží najít takový lineární oddělovač, který má co možná nejširší pásmo mezi ním, pozitivními případy na jedné straně a negativními případy na straně druhé. Tento optimální oddělovač se v algoritmu SVM hledá pomocí metody kvadratického programování.

Nechť existuje *n* k-dimenzionálních vektorů reálných čísel x_i a celá čísla y_i, kde $y_i = \{-1, 1\}$. Hodnota y_i indikuje třídu vektoru i. Problém nalezení optimálního oddělovače pak lze popsat jako hledání hodnot parametrů α_i (váha příslušného atributu), které maximalizují výraz [41]:

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \left(x_i x_j\right), \tag{10}$$

přičemž platí následující omezení:

$$\alpha_i \geq 0; \sum_i \alpha_i y_i = 0. \tag{11}$$

Tento výraz má dvě významné vlastnosti [41]:

- má jedno globální maximum, které lze efektivně najít,
- datové body vstupují do výrazu ve formě skalárního součinu jednotlivých dvojic (toto platí i pro rovnici lineárního oddělovače).

Jsou-li spočteny optimální parametry α_i, je oddělovač definován pomocí rovnice [41]:

$$h(x) = sign\left(\sum_i \alpha_i y_i \, (x \cdot x_i)\right). \tag{12}$$

Vzhledem k omezení, které je vyjádřené dle (11) platí, že lineární oddělovač má nulové váhy α_i pro každý datový bod kromě těch, které jsou nejblíže vlastnímu oddělovači. Právě tyto nejbližší body se pak nazývají podpůrné vektory (support vectors). Ostatní datové body a vektory, kterých je mnohem více, nejsou pro SVM nikterak podstatné a počet parametrů popisujících optimální oddělovač je mnohem menší než *N*.

Běžně se nestává, že bude lineární oddělovač nalezen přímo v originálním vstupním prostoru *X*. Obecně se oddělovač hledá ve vícerozměrném prostoru *F*(*x*) tak, že se nahradí člen $x_i \cdot x_j$ členem $F(x_i) \cdot F(x_j)$. Dále platí, že $F(x_i) \cdot F\left(x_j\right) = K(x_i, x_j)$ není většinou nutné stanovovat přes výpočet pro všechny body. Výraz $K(x_i, x_j)$ se nazývá jádrová funkce (kernel function), kdy tato funkce v souvislosti s SVM může být aplikována na dvojice vstupních dat

k vyhodnocení skalárního součinu v odpovídajícím prostoru. Z toho vyplývá, že lineární oddělovač lze ve vícerozměrném prostoru *F(x)* najít náhradou $x_i \cdot x_j$ jádrovou funkcí $K(x_i, x_j)$. Jádrová funkce musí odpovídat tzv. Mercerovu teorému, kdy tento určuje pro *K* určité podmínky, za kterých existuje skalární součin $\langle \cdot, \cdot \rangle_H$ a mapování $\Phi: L \rightarrow H$ takové, že platí [41]:

$$\langle \Phi(x_1), \Phi(x_2) \rangle_H = K(x_1, x_2) \tag{13}$$

kde $K(x_1, x_2)$ je jádrová funkce a $\Phi(x_i)$ představují mapovací funkce.

Nalezené lineární oddělovače lze zpět mapovat do původního vstupního prostoru, čímž vznikne libovolně zvlněná nelineární hranice mezi jednotlivými třídami [16]. Jakýkoliv algoritmus, který lze převést na uvedený skalární součin, může být náhradou tohoto součinu jádrovou funkcí převeden na jádrovou verzi např. *k*-nejbližšího souseda, perceptron apod [16].

Klasifikace se pak provádí pomocí vztahu:

$$f(x_{new}) = sign\left(\sum_{i=1}^{\#SV} \alpha_i y_i K(x_i, x_{new})\right). \tag{14}$$

Příklady jádrových funkcí:

- lineární, $K(x_1, x_2) = \langle x_1, x_2 \rangle$ (15)
- polynomická, $K(x_1, x_2) = (\gamma \langle x_1, x_2 \rangle + c_0)^d$ (16)
- radiální bázické funkce (RBF), $K(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2)$ (17)

kde $\langle \cdot, \cdot \rangle$ představuje skalární součin, γ šířku u RBF a u polynomické funkce je standardně $\gamma = 1$, d stupeň polynomu, c_0 aditivní konstantu u polynomu (standardně = 0) a $\#SV$ je počet podpůrných vektorů.

SVM je efektivní metoda, která byla již v minulosti mnohokrát využita především na řešení úloh, které se vyznačují velkým množstvím atributů [43]. SVM jsou tyto úlohy schopné provádět výrazně rychleji než například neuronové sítě [41], [43], [26].

3.1.3. Skryté Markovovy modely

Skryté Markovovy modely (Hidden Markov models, HMM) je statistická metoda, kde z modelovaného stavu není známa posloupnost stavů, které vedou k výsledku, ale je známa pravděpodobnostní funkce pro jednotlivé stavy [36]. Mezi největší výhody HMM lze zařadit silné podložení statistikou, které je velmi vhodné pro úlohy zpracování přirozeného jazyka umožňující poměrně robustní a rychlé zpracování dat. Jako nevýhodu HMM lze označit nutnou znalost topologie modelu a stejně jako u ostatních statistických metod potřebu velkého množství trénovacích dat [36].

HMM bývají použity k zařazení předložených příkladů do některé ze tříd, které pro uživatele obsahují užitečné informace. Bude-li se předpokládat množina diskrétních výstupů, HMM se skládají z množiny stavů Q, s určenými počátečními a koncovými stavy q_I a q_F, danou množinou přechodů mezi stavy $(q \rightarrow q')$ a diskrétním slovníkem výstupních symbolů $\sum \sigma_1, \sigma_2, \ldots, \sigma_M$. Model pak generuje řetězec $x = x_1, x_2, \ldots, x_l$. Celý výpočet začíná v počátečním stavu, který následuje výpis výstupního symbolu a přechod do následujícího stavu, kdy tento proces se následně opakuje tak dlouho, dokud není dosažen cílový stav [36]. Parametry modelu jsou reprezentovány pravděpodobnostmi jednotlivých přechodů $P(q \rightarrow q')$ s jakou je stav q' následníkem stavu q a emisní pravděpodobnosti $P(q \uparrow \sigma)$, že daný stav vydává určitý výstupní symbol. Pravděpodobnost řetězce x vydaného HMM je spočítána pomocí vztahu [1], [36]:

$$P(x|M) = \sum_{q_1, q_2, \ldots, q_l \in Q^l} \prod_{k=1}^{l+1} P(q_{k-1} \rightarrow q_k) P(q_k \uparrow x_k), \tag{18}$$

kde q_0 a q_{l+1} musí po řadě náležet q_I a q_F a x_{l+1} představuje konec řetězce tokenů.

Pozorovaným výstupem systému je sekvence symbolů, které vydaly stavy. Základní sekvence je však skryta. Jedním z cílů učení pomocí HMM je znovuzískání sekvence stavů $V(x|M)$, která má největší pravděpodobnost, že produkuje zobrazenou posloupnost [36]:

$$V(x|M) = argmax_{q_1 \ldots q_l \in Q^l} \prod_{k=1}^{l+1} P\,(q_{k-1} \rightarrow q_k) P(q_k \uparrow x_k). \tag{19}$$

HMM mohou být používány pro extrakci informací, kdy struktura základního modelu může být stanovena: každý stav je asociován s třídou, kterou je potřeba extrahovat a vydává slova z množiny pro třídu specifického rozdělení unigramů. Množiny a pravděpodobnosti jednotlivých přechodů získáme z trénovacích dat. Je-li třeba klasifikovat řetězec slov, pak je pro každé slovo provedeno ohodnocení a pomocí Viterbi algoritmu [36] je získána nejpravděpodobnější posloupnost stavů. Stav, který je produkován jednotlivými slovy, představuje třídu pro dané slovo.

Učení struktury modelu

Pro tvorbu modelu, který bude vhodný pro extrakci informací je nejprve nutné učinit rozhodnutí, kolik stavů a jaké hrany bude model obsahovat. Obstojný počáteční model má jeden stav pro jednu třídu s přechody mezi všemi stavy – plně propojený model [36]. Tento model však nemusí být optimální pro všechny případy. Další variantu představuje model, který umožňuje mít pro jednotlivé třídy více než jeden stav a pouze několik přechodů do dalších stavů [36].

Při tvorbě modelu je každému slovu ze zpracovávaného příkladu přiřazen jeho vlastní stav s přechody do stavu odpovídajícímu slovu, které jej následuje. Každý stav je asociován s třídou slov, které obsahuje. Následně je přidán přechod ze stavu počátečního do prvního stavu každé trénovací instance a z posledního stavu každé instance do koncového stavu. Tento model může sloužit jako počáteční model pro různé metody omezení počtu stavů, například lze použít [1], [36]:

- Spojování sousedů – toto kombinuje všechny stavy, které sdílí spoj a mají stejnou třídu. Množina sousedních uzlů se stejnou třídou tak může být spojena do jednoho uzlu se stejným označením. Takto vzniklému novému stavu je přidána hrana do sebe sama, jejíž pravděpodobnost reprezentuje možnost výpočtu setrvat v daném stavu pro danou třídu.
- V-slévání – spojuje jakékoliv dva stavy, které mají stejnou třídu a sdílí přechod z nebo do normálního stavu. Tato metoda redukuje větvení maximálně specifického modelu. Na místo startu v počátečním stavu a vybíráním mezi přechody do stavů tříd, může V-slévání spojit jeho potomky do jednoho stavu tak, že z počátečního stavu potom vede pouze jeden přechod.
- Bayesovské slévání – cílem je pomocí iterativního slévání stavů nalézt model s maximální pravděpodobností pro daná trénovací data $P(M|D)$. Toto se opakuje do té doby, dokud není dosažen optimální poměr mezi pokrytím dat a velikostí modelu. Vztah je pak vyjádřen pomocí Bayesovského pravidla [5]:

  $$P(M|D) = P(D|M)P(M)\ . \tag{20}$$

Použití HMM s sebou přináší i některé problémy, kdy tři nejvýznamnější jsou [2]:

1) Problém vyhodnocení – jak efektivně, pro konkrétní pozorovanou posloupnost, vypočítat pravděpodobnost, že bude vygenerována modelem.
2) Dekódovací problém – jakým způsobem vybrat odpovídající posloupnost stavů k zadanému modelu, která by co nejlépe popisovala pozorovnou posloupnost.
3) Problém učení – tento problém představuje, jak odhadnout parametry modelu λ, k pozorované posloupnosti O tak, aby pravděpodobnost $P(O|\lambda)$ byla maximální – této posloupnosti, která je používána k odhadování parametrů modelu se říká posloupnost trénovací, neboť slouží k učení HMM.

3.1.4. Maximální entropie

Model maximální entropie je založen na principu vytvoření co možná nejméně jiných předpokladů o datech než těch, která jsou daná z trénovacích dat vytvořenými omezeními, na základě kterých tvoří pravděpodobnostní model. Pomocí této metody se vytváří množina příznaků nebo funkcí, které popisují vztah mezi příznaky a výstupem. Pravděpodobnostní rozdělení splňující takovéto podmínky má potom nejvyšší možnou entropii. Takovéto rozdělení pravděpodobnosti je jedinečné a má exponenciální tvar [1], [18]:

$$P(o|h) = \frac{1}{Z(h)} \cdot \prod_{j=1}^{k} \alpha_j^{f_j(h,o)}, \tag{21}$$

kde *o* představuje výstup, *h* kontext, α_i jsou parametry modelu, $f_j(h, o)$ je příznaková funkce a *Z(h)* je normalizační faktor zaručující, že *P(o|h)* bude pravděpodobnostní rozdělení, kdy odpovídá maximální věrohodnosti.

U metody maximální entropie jsou příznaky binární. Příkladem binární příznakové funkce může být [18]:

$$f_j(h, o) = \begin{cases} 1 & slovo = Pardubice \\ 0 & jinak \end{cases} \tag{22}$$

Parametry modelu α_i představují váhu jednotlivých příznaků. K jejich nalezení může být využito různých trénovacích algoritmů. Mezi běžně používané algoritmy patří iterační algoritmy Generalized Iterative Scaling a Improved Iterative Scaling. Jedná se o iterativní postupy, kdy je zlepšován odhad parametrů při každé iteraci [18]. Klasifikátor založený na maximální entropii následně přiřazuje každému slovu nezávisle jednu z následujících tříd [18]:

- začátek pojmenované entity (tag B),
- slovo nacházející se uvnitř pojmenované entity (tag C),
- poslední slovo pojmenované entity (tag L),
- jediné slovo v rámci pojmenované entity (tag U).

Je možné, že systém v průběhu klasifikace může vytvořit nepřípustné třídy [18]. Takovéto třídy lze eliminovat, pokud se uvažuje pravděpodobnost přechodu $P(c_i|c_j)$ mezi třídami c_i a c_j přiřazenými dvěma po sobě jdoucím slovům. Pokud bude takovýto přechod možný, bude jeho pravděpodobnost rovna 1, v opačném případě bude její hodnota rovna 0. Pravděpodobnost tříd $c_1, \ldots, c_n$ přiřazených slovům, která se nacházejí v dokumentu *D* a větách *s* se určí následovně podle [18]:

$$P(c_1, \dots, c_n|s, D) = \prod_{i=1}^{n} P(c_i|D) \cdot P(c_i|c_i - 1), \quad (23)$$

kde $P(c_i|s, D)$ určí klasifikátor maximální entropie. V závěru je použit Viterbiho algoritmus k nalezení sekvence tříd s největší pravděpodobností [18].

3.1.5. Neorientované pravděpodobnostní grafické modely

Základem aplikací, které se zabývají zpracováním signálu, obrazu, biologických dat a přirozeného textu apod. je schopnost předpovídat neviděné či neznámé veličiny a proměnné, které jsou na sobě určitým způsobem závislé [15]. Hledá se tak výstupní vektor $y = \{y_0, \dots, y_t\}$ (v případě IE jednotlivé prvky přísluší jednotlivým tokenům) na základě vektoru příznaků. Příznakem pak může být pozice slova, přítomnost velkého písmena apod. [15].

Je-li *x* množina *n* náhodných proměnných, potom *P(x)* je sdružená pravděpodobnost všech těchto proměnných. Existují-li dvě podmnožiny x: $x_A a\ x_B$, které jsou na sobě při daném x_C podmíněně nezávislé, pravděpodobnost dodržuje tuto podmíněnou nezávislost, pokud platí výrok [15]:

$$P(x_A|x_B, x_C) = P(x_A|x_C), \quad (24)$$

resp.

$$P(x_A|x_B, x_C) = \frac{P(x_A, x_B, x_C)}{P(x_C)} = \frac{P(x_A|x_B, x_C)P(x_B, x_C)}{P(x_C)} = P(x_A|x_C)P(x_B|x_C). \quad (25)$$

Někdy je používán zkrácený zápis: $x_A \perp x_B|x_C$. (26)

Existuje-li *x* a seznam výroků o podmíněné nezávislosti, je cílem nalézt skupinu pravděpodobnostních rozdělení nad *x*, které odpovídají takovým výrokům. Nechť je daný neorientovaný graf $G = (X, E)$, jehož uzly odpovídají množině náhodných proměnných. Potom pro hrany musí platit, že pokud se odstraní všechny uzly náležících množině x_C, odstraní se všechny cesty z x_A do x_B, platí $x_A \perp x_B|x_C$ [15]. Jedním ze způsobů, jak reprezentovat vzájemné vztahy mezi výstupními proměnnými, je tvorba grafického modelu, jehož příkladem můžou být Markovské sítě [15].

3.1.6. Podmíněná náhodná pole

Podmíněná náhodná pole (Conditional random fields, CRF) označuje neorientovaný grafický model, který se využívá k určení podmíněných pravděpodobností výstupních uzlů na základě uzlů vstupních [22]. Každý uzel, který v tomto případě představuje slovo, popisuje náhodnou proměnou. CRF lze jinak popsat jako model s konečným počtem stavů a nenormalizovanými pravděpodobnostmi přechodů. CRF se liší od HMM tím, že nemá přesně

určené hodnoty přechodů mezi stavy a disponuje možností mnohonásobných funkcí generujících příznaky. Tento model dosahuje lepších výsledků při zpracování dat se závislostmi vyššího řádu, které lépe odpovídají reálnému modelu [15].

Formálně lze CRF definovat takto [22]: nechť je X náhodnou proměnnou dat, v tomto případě vět a Y náhodnou proměnnou k X příslušných značek. Všechny prvky Y_i z Y náleží abecedě x, která tedy obsahuje množinu všech použitých značek. Vytvoří se pak model $P(X|Y)$ z párů pozorovaných hodnot a sekvencí značek [15].

CRF pak lze vyjádřit následovně: nechť $G = (V, E)$ je grafem, u kterého platí $Y = (Y_v), v \in V$ tak, že Y je indexováno body vektoru G. Poté (X, Y) vyjadřuje CRF v případě, kdy náhodné proměnné Y_v jsou podmíněné X a vyhovují Markovské vlastnosti s ohledem na graf, který je vyjádřený jako [22]:

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w{\sim}v), \tag{27}$$

kde $w{\sim}v$ označuje, že prvky w a v spolu ve vektoru G sousedí.

V tomto případě lze G, X i Y považovat za jednoduché řetězce [22]:

$$G = (V = \{1,2, \dots, m\}, \tag{28}$$

$$E = \{(i, i+1)\}, X = (X_1, X_2, \dots, X_n), \tag{29}$$

$$Y = (Y_1, Y_2, \dots, Y_n). \tag{30}$$

3.1.7. Lineárně řetězená CRF

Tato metoda je variantou metody CRF, která je neorientovaným grafickým modelem HMM. Lineárně řetězená CRF lze pak definovat takto [38]: nechť je dáno podmíněné rozdělení $P(y|x)$, které je odvozeno od sdruženého rozdělení HMM $P(y|x)$, což je ve skutečnosti CRF s jednou vybranou konkrétní skupinou příznakových funkcí, což jsou funkce, které na vstupu očekávají větu s, pozici i slova ve větě, třídu l_i současného slova, třídu l_{i-1} předchozího slova a na výstupu poskytuje reálné číslo (velmi často však pouze 0 či 1).

Bude-li se používat koncept příznakových funkcí, předchozí rovnici pak lze zapsat ve stručnější podobě. Každá takováto funkce má podobu $f_k(y_t, y_{t-1}, x_t)$. Nechť je daný příznak $f_{ij}(y, y', x) = 1_{\{y=1\}}1_{\{y'=j\}}$ pro každý přechod (i, j) a jeden příznak $f_{io}(y, y', x) = 1_{\{y=1\}}1_{\{x=0\}}$ pro každou dvojici pozorování (i, o). Potom lze HMM zapsat jako [38]:

$$P(y|x) = \frac{1}{Z(x)} exp \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t), \tag{31}$$

$$Z(x) = \sum_y \prod_{t=1}^{T} exp \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t), \quad (32)$$

kde *Z(x)* je normalizační konstanta, *X, Y* jsou náhodné vektory, $\theta = \{\theta_k\} \in R^K$ je parametr vektoru, $F = \{f_k(y, y',\ x_t)\}_{k=1}^{K}$ je soubor skutečných hodnot funkcí.

3.2. Metriky systémů pro extrakci informací

Systémy určené pro extrakci informací a zpracování přirozeného jazyka fungují správně jen do určité míry. Výkonost těchto systémů je potřeba měřit ať už z důvodu jejich porovnávání či ladění jednotlivých systémů. Z těchto důvodů je potřebné zavést způsob, jak úspěšnost systémů měřit. Vytvořené systémy se hodnotí podle následujících metrik, pro jejichž výpočet je zásadní, že výstup jakéhokoliv systému na extrakci informací lze ohodnotit jako pozitivní (*P*) a negativní (*N*). V tomto případě výstup *P* signalizuje, že byla entita extraktorem nalezena a výstup *N*, že entita nalezena nebyla [6], [14], [5]:

- *precision (přesnost)* – udává, jak velká část objektů zařazených do *P* je skutečně z *P*,
- *recall (úplnost)* – značí, jak velká část z celkového skutečného *P* byla zařazena do *P*,
- *F-measure (F-míra)* – dána harmonickým průměrem předchozích dvou,
- *Accuracy (Úspěšnost)* – udává podíl všech správně vyextrahovaných entit a celkový počet entit,
- *Senzitivita a specificita* – jedná se o pojmy, které byly převzaty z medicíny, kdy v případě nasazení nějakého nového léku je sledováno, u kolika nemocných pacientů lék zabere (senzitivita), a zda lék zabírá pouze na danou chorobu (specificita). V tomto případě *senzitivita* představuje schopnost extraktoru označit entitu, která je v textu obsažená. V případě *specificity* se jedná o schopnost extraktoru neoznačit entitu, která v textu obsažená není.

Při klasifikaci jednotlivých objektů do tříd *P* a *N* se rozlišují následující třídy výsledků [2]:

- *skutečně pozitivní (TP)* – správné zařazení objektu do třídy *P*,
- *skutečně negativní (TN)* – správné zařazení objektu do třídy *N*,
- *chybně pozitivní (FP)* – špatné zařazení objektu do třídy *P* (objekt patří do třídy *N*),
- *chybně negativní (FN)* – špatné zařazení objektu do třídy *N* (objekt patří do třídy *P*). Zařazení do jednotlivých tříd je dobře patrné z Obrázku 6.

Na základě těchto skupin lze definovat přesnost, úplnost a F-míru následovně:

$$Přesnost = \frac{TP}{TP+FP}, \tag{33}$$

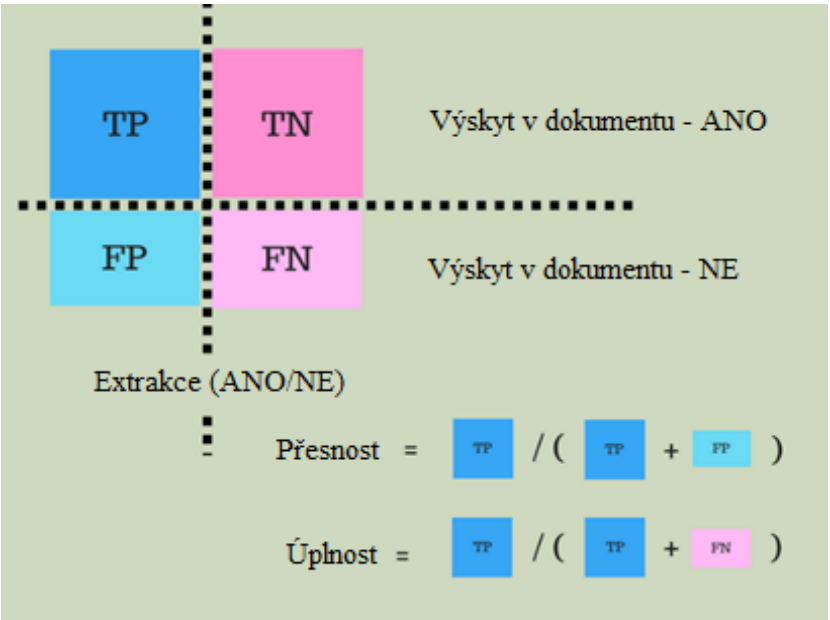
$$Úplnost = \frac{TP}{TP+FN}, \tag{34}$$

$$F - míra = \frac{2TP}{2TP+FP+FN}, \tag{35}$$

$$Úspěšnost = \frac{TP+TN}{TP+FN+FP+TN}, \tag{36}$$

$$Senzitivita = \frac{TP}{TP+FN}, \tag{37}$$

$$Specificita = \frac{TN}{FP+TN}. \tag{38}$$



Obrázek 6: Přesnost a úplnost

Zdroj: Upraveno dle [6]

Přesnost a úplnost pak představují dva různé pohledy na výsledek klasifikace. Tyto se navzájem ovlivňují a to tak, že pokud se zvýší hodnota jednoho, sníží se hodnota druhého a naopak, jejich harmonický průměr tak dává celkový pohled na výsledky nehledě na to, která z metrik má lepší výsledek. Na základě výše uvedeného se systémy soustředí podle konkrétní úlohy na jednu z metrik [14].

Pokud chce uživatel výše uvedený postup aplikovat na pojmenované entity, musí se určit mapování výsledků poskytnutých systémem na třídy *TP*, *TN*, *FP* a *FN*.

Systém se při označování jednotlivých entit v textu může dopustit následujících chyb [6]:

- označí entitu, která entitou není,

- neoznačí entitu,
- správně označí entitu, ale špatně určí třídu,
- špatně určí hranice entity,
- kombinace předchozího, kdy jsou špatně určené hranice entity a špatně určená třída.

3.2.1. MUC-6 evaluace

První metriky pro vyhodnocení výsledků systémů NER přišly společně s příchodem samotných úloh spojených s extrakcí informací. Výběr metrik byl proveden v rámci Message Understanding Conference (MUC) a je založen na způsobech měření ostatních úloh z oblasti NLP, které se do této doby používaly [14].

Dle MUC-6 je zvlášť hodnocena kvalita hledání hranic entit a jejich typu. Typ je považován za správný v případě, pokud souhlasí s originálním typem a pokud je jím označena část textu zahrnující celou původní entitu. Text je považován za správný, když se shoduje s originálním textem entity. Porovnáváni textu zahrnuje i jeho možné modifikace, a proto i entita, která má nestejnou, upravenou podobu je považována za správnou [14].

Pro span i typ jsou následně počítány [14]:

- správné odpovědi systému,
- počet uživatelem nalezených entit,
- počet nalezených entit.

Výkon celého systému je pak určen jako součet těchto čísel, z něhož následně klasickým způsobem určujeme přesnost, úplnost a F-míru.

3.2.2. CoNLL evaluace

Dle Conference on Natural Language Learning (CoNLL) je definován způsob evaluace jako striktní metoda, která považuje ze správnou odpověď pouze tu, která je bezchybná – musí být zcela správně určeny jak hranice, tak i třída. Pokud je například originální entita „The United States“ a systémem je označeno „United States“, systém je penalizován celkem dvakrát jednou za chybu FP za „United States“ a podruhé za FN „The United States“ [34].

3.2.3. ACE evaluace

Automatic Content Extraction program (ACE) reprezentuje následovníka MUC. Tento program se zaměřuje na dvě hlavní úlohy v oblasti NER: detekci entit a vztahů a rozpoznávání a normalizaci časových údajů, kdy obě tyto úlohy rozšiřují standardní definici NER [10].

Tento typ evaluace nevyužívá standardní metriky, evaluace je založena na speciálním skórovací systému, kdy každý typ chyby a entity má různou váhu [10]. Výhodu této evaluace je možnost zhodnotit výkon systému podle přesně daných kritérií. Na druhé straně je ale nutné dodržet stejný způsob hodnocení pro porovnání dvou systémů.

3.2.4. Lenient evaluace

Přílišnou striktnost při posuzování výsledků evaluace CoNLL se snaží řešit evaluace Lenient. U evaluace CoNLL může docházet k tomu, že odpověď, která je téměř správná je penalizována dokonce hned dvakrát. Oproti tomu Lenient evaluace považuje za správný výsledek i ten, kde se entita nalézá v identifikovaných hranicích a byl určen správný typ. Pro ohodnocení správnosti odpovědi systému se musí zavést následující třídy [39]:

- *not marked (nm)* – entita nebyla systémem nalezena,
- *not correct (nc)* – označen úsek textu, který není entitou,
- *partially marked (pm)* – správně určený typ entity, ale špatně určené hranice (posuzováno pro každou entitu),
- *partially correct (pc)* – správný typ entity, ale nepřesně určené hranice (posuzováno pro každou rozpoznanou entitu),
- *correct (c)* – správně určený typ i hranice entity.

Výpočet jednotlivých metrik je pak upraven následujícím způsobem [39]:

$$Přesnost = \frac{c+p}{c+nc+pc}, \tag{39}$$

$$Úplnost = \frac{c+pm}{c+pm+nm}, \tag{40}$$

$$F-míra = \frac{2 \cdot Přesnost \cdot Úplnost}{Přesnost+Úplnost}. \tag{41}$$

3.2.5. Strict evaluace

Strict evaluace představuje drobnou úpravu evaluace Lenient, kdy se od sebe liší ve výpočtu jednotlivých metrik [39]:

$$Přesnost = \frac{c}{c+nc+pc}, \tag{42}$$

$$Úplnost = \frac{c}{c+pm+nm}, \tag{43}$$

$$F - míra = \frac{2 \cdot c}{2 \cdot c+nc+nm+pc+pm}. \tag{44}$$

Evaluace Strict a Lenient dohromady představují dolní a horní mez výkonu systému, kdy dolní mez je prezentována evaluací Strict a horní pak evaluací Lenient.

3.2.6. Mikro- a Makro-průměr

Nevýhodou metriky, která je určena pro celý systém, je možná ignorace chyby pro dílčí třídy. Tuto nevýhodu je možné odstranit tím, že se zavede metrika, která bude brát v potaz výsledky pro jednotlivé třídy. Pro tyto potřeby se používají výše uvedené, modifikované metriky mikro- a makro-průměr. V angličtině se pro ně používá označení *Micro-averaged* (B_{micro}) a *Macro-averaged* (B_{macro}) přesnost/úplnost/F-míra. Nejčastěji je používáno v binární klasifikaci *B(tp, tn, fp, fn)*. Výpočet se pak provede následovně:

Nechť $L = \{\lambda_j : j = 1 \dots q\}$ je množina všech tříd pojmenovaných entit a nechť $tp_\lambda, fp_\lambda, tn_\lambda \; a \; fn_\lambda$ jsou četnosti jednotlivých kategorií výsledků klasifikace pro třídu λ, pak platí následující vztahy [39]:

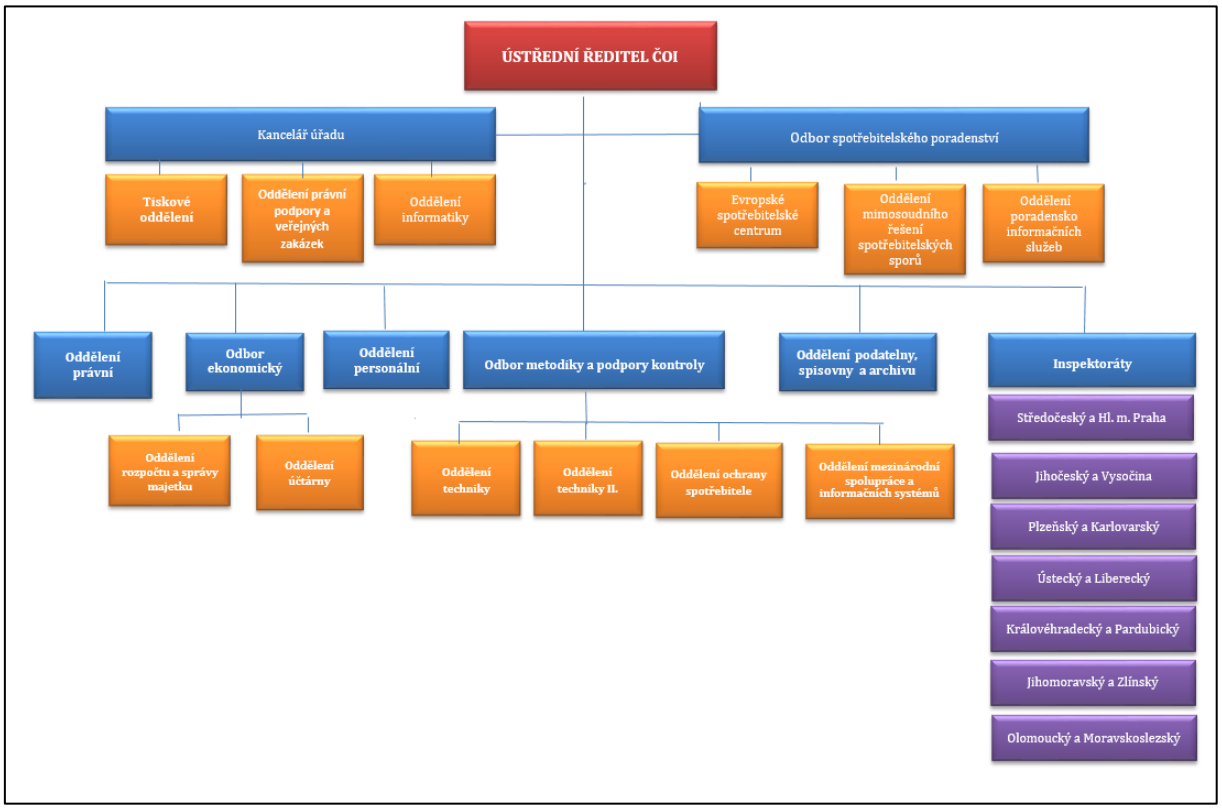
$$B_{macro} = \frac{1}{q} \sum_{\lambda=1}^{q} B(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda), \tag{45}$$

$$B_{micro} = B\left(\sum_{\lambda=1}^{q} tp_\lambda, \sum_{\lambda=1}^{q} fp_\lambda, \sum_{\lambda=1}^{q} tn_\lambda, \sum_{\lambda=1}^{q} fn_\lambda\right). \tag{46}$$

Rozdílem mezi těmito mírami je ten, že zatímco B_{macro} dává všem třídám stejnou váhu, B_{micro} dává stejnou váhu všem rozhodnutím systému napříč textem. Z výše uvedeného vyplývá, že pro měření efektivity na větších třídách je vhodnější B_{micro}, zatímco na menších třídách je to metoda B_{macro}.

4. ČESKÁ OBCHODNÍ INSPEKCE

Jelikož je tato práce zaměřená na extrakci informací z textových dokumentů pro potřeby České obchodní inspekce, je nutné zde uvést základní charakteristiku orgánu Česká obchodní inspekce (dále ČOI). ČOI je orgánem státní správy, který je podřízený Ministerstvu průmyslu a obchodu ČR. V čele ČOI je ústřední ředitel, kterého jmenuje ministr průmyslu a obchodu. ČOI se dále člení na ústřední inspektorát, kterému podléhá sedm oblastních inspektorátorů se sídly a pobočkami v krajských městech. Organizační struktur ČOI je vyobrazena na Obrázku 7 [45].



Obrázek 7: Organizační struktura ČOI

Zdroj:[8]

Dle zákona č. 64/1986 sb. Se ČOI zaměřuje především na kontrolu fyzických a právnických osob, které nabízejí, prodávají, dodávají nebo uvádějí na trh výrobky, nabízejí nebo poskytují služby nebo vyvíjejí jinou činnost dle tohoto zákona. ČOI dále [8], [45]:

a) Ukládá pokuty a jiná opatření dle výše zmíněného zákona nebo podle zvláštního právního předpisu.

b) Zobecňuje poznatky z výkonu kontroly a zveřejňuje výsledky kontrol s cílem předcházet porušování právních předpisů.

c) Provádí rozbory nebo zajišťuje jejich provedení pro účely kontroly. Provedení těchto rozborů zajišťuje u příslušných orgánů nebo osob. Na náklady kontrolovaných osob provádí rozbory nebo zajišťuje jejich provedení jen tehdy, byly-li rozborem zjištěny vlastnosti výrobků, které neodpovídají právnímu předpisu nebo jsou v rozporu s deklarací o vlastnostech výrobků uvedených zejména v prohlášení nebo v obchodním sdělení.

Internetové obchodování patří mezi nejsledovanější oblasti kontrol ČOI. Mezi nejsledovanější oblasti se řadí nejen z důvodu vysoké četnosti provozovaných e-shopů, ale také z důvodu vysokého počtu stížností spotřebitelů [7]. Mezi nejčastější stížnosti patří klamavé, nekorektní a v některých případech i podvodné jednání obchodníků.

V roce 2015 bylo ČOI provedeno celkem 1 194 inspekcí obchodů, které nabízejí zboží a služby. Z výše uvedeného počtu inspekcí bylo u 990 (82,9 %) případů zjištěno pochybení proti zákonným povinnostem. Následně bylo uloženo 825 pokut převyšujících celkovou částku 4 mil. Kč [7]. Výsledky kontrol, dle jednotlivých inspektorátorů, jsou dobře patrné z Tabulky 2.

Tabulka 2: Výsledky kontrol dle inspektorátorů

Internetové obchodování – rok 2015			
Inspektorát	**počet kontrol**	**kontroly se zjištěním**	**zjištěná porušení v %**
Středočeský a Hl. město Praha	175	114	65,1 %
Jihočeský a Vysočina	112	100	89,3 %
Plzeňský a Karlovarský	132	107	81,1 %
Ústecký a Liberecký	181	154	85,1 %
Královéhradecký a Pardubický	172	145	84,3 %
Jihomoravský a Zlínský	175	146	83,4 %
Moravskoslezský a Olomoucký	247	224	90,7 %
Celkem	**1 194**	**990**	**82,9 %**

Zdroj: [7]

Celkem bylo v roce 2015 Českou obchodní inspekcí přijato 5 178 stížností, kdy mezi nejčastější stížnosti spotřebitelů patřilo především [7]:

- způsob vyřízení reklamace,
- nedodání zaplaceného výrobku,
- dodání jiného zboží,
- dodání poškozeného, nekompletního nebo použitého zboží,
- nevrácení peněz při odstoupení od smlouvy.

ČOI při svých kontrolách internetových obchodů zjistila nejčastěji prohřešky proti zákonu č. 634/1992 Sb., o ochraně spotřebitele. Jmenovitě se jednalo o [7]:

- používání klamavých obchodních praktik spojených s neposkytnutím zákonem stanovených informací (823 případů),
- neposkytnutí řádných informací spotřebiteli o rozsahu, podmínkách a způsobu uplatnění práva z vadného plnění § 13 (497 případů),
- formální náležitosti přijetí a vyřízení reklamace (155 případů),
- neposkytnutí řádné informace o ceně nabízených výrobků nebo služeb spotřebiteli (57 případů),
- klamavé obchodní praktiky, související s nabízením nebo prodejem výrobků porušujících některá práva duševního vlastnictví § 5 odst. 2 (27 případů).

Internetové obchodování se mezi spotřebiteli těší stále větší oblíbenost, avšak z výše uvedeného je patrné, že mezi obchodníky je mnoho těch, kteří využívají různé klamavé či dokonce podvodné praktiky. Monitorování a kontrola internetového obchodování tak patří mezi prioritní kontrolní činnosti ČOI. V této práci bylo pracováno především s informacemi, která ČOI zveřejňuje na svých webových stránkách, které jsou dostupné na adrese https://www.coi.cz/pro-spotrebitele/rizikove-e-shopy/. Na této stránce ČOI zveřejňuje informace o e-shopech, které považuje za rizikové. Důvody, proč tyto e-shopy považuje za rizikové, zde zveřejňuje v krátkých popiscích e-shopů. Tyto popisky byly následně analyzovány v další části práce a na základě výsledků těchto analýz byly vytvořeny modely, pro usnadnění monitorování a kontroly e-shopů.

5. EXTRAKCE INFORMACE PRO POTŘEBY ČOI

V této části práce bylo přistoupeno k řešení extrakce informace ze stránek e-shopů s využitelností pro ČOI. Prvním úkolem, který bylo potřeba vyřešit, bylo stanovení problematiky, která bude pomocí extrakce informace řešena. Vzhledem k výsledkům tiskové zprávy, které byly prezentovány v předchozí kapitole, představuje kontrola internetového obchodování pro ČOI oblast, jež patří mezi prioritní. A právě z tohoto důvodu bylo rozhodnuto, že práce bude zaměřena na rizikové e-shopy. Rizikové e-shopy ČOI zveřejňuje na svých webových stránkách na adrese https://www.coi.cz/pro-spotrebitele/rizikove-e-shopy/. Na těchto stránkách je zobrazen výčet rizikových e-shopů společně s komentářem, kde je zdůvodněno, z jakého důvodu ČOI považuje e-shop jako rizikový. Komentáře, které popisují rizikovost e-shopů, se stanou první bází dat, která bude tato práce zpracovávat. Cílem bylo na základě četnosti jednotlivých výrazů, které popisky obsahují, stanovit, co je nejčastějším prohřeškem a tomu věnovat hlavní pozornost ve zbylé části práce. Na základě frekvenční analýzy výše uvedených komentářů, byla následně provedena extrakce jmenných entit ze stránek obchodních podmínek e-shopů. V samotném závěru práce bylo provedeno vyhodnocení vytvořeného modelu.

5.1. Výběr vhodného programového prostředí

Po stanovení problematiky, kterou bude práce řešit, bylo třeba rozhodnout, který SW bude pro extrakci informace použit. Jako nejvhodnější nástroj pro zpracování této práce byl zvolen SW Rapidminer (dále RM), který je poskytován zdarma pod licencí General public license GNU. Jedná se o SW, který poskytuje prostředí pro procesy DM, kdy pomocí modulárních operátorů umožňuje tvorbu zřetězených modelů pro řešení velkého množství problémů.

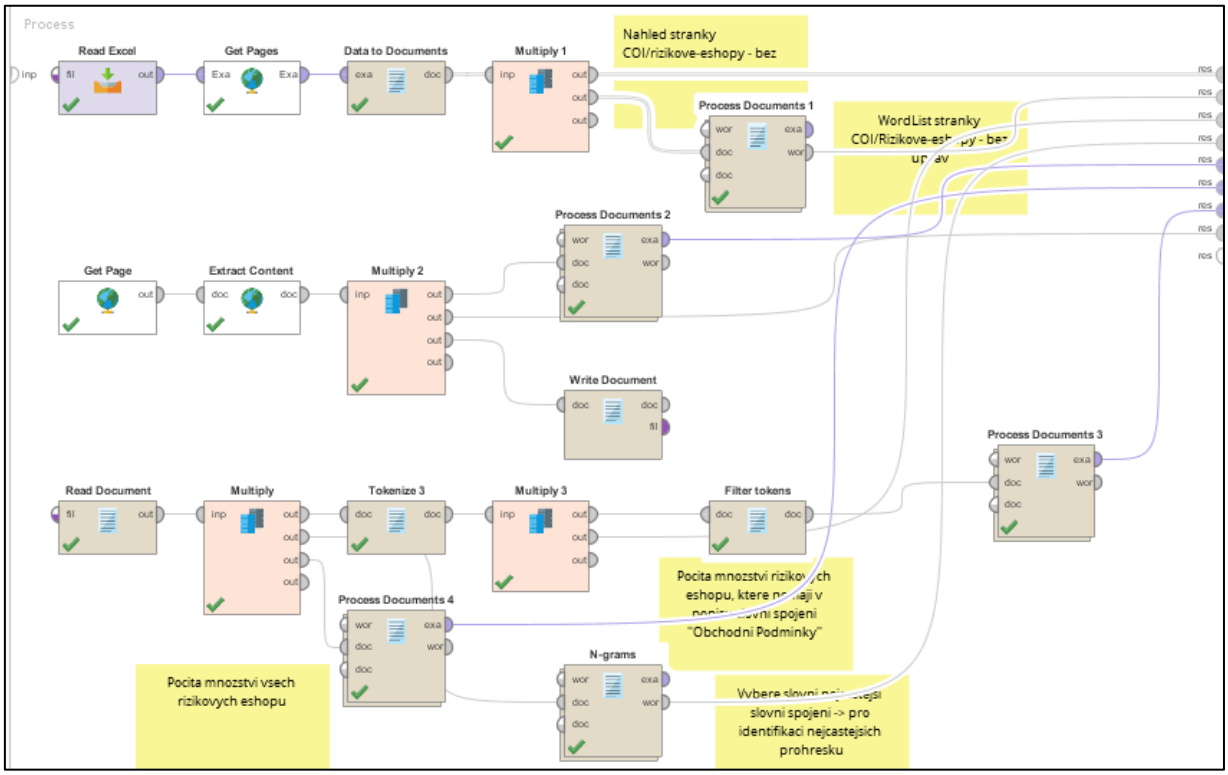
Hlavním kritériem na programové prostředí byly jeho poskytované funkcionality s ohledem na řešení text miningových úloh. Zejména se jednalo funkcionality spojené s předzpracováním textu, extrakcí informací a asociací entit. Další kritérium, které musel vybraný SW splňovat, je jeho dostupnost. Nástroj RM je po zaregistrování (čímž uživatel automaticky získá licenci) poskytován ve své trial verzi zcela zdarma. Tato verze není nikterak omezena a SW poskytuje své služby zcela neomezeně. Po uplynutí 30 dnů RM přejde do trial verze, která je již některými parametry omezena. Omezen je zde počet datových řádků (omezeno na 10 000 řádků), se kterými bude RM pracovat a počet logických procesorů, které bude využívat, na jeden procesor. Po úvaze, co budou tato omezení znamenat pro vypracování této práce, bylo vyhodnoceno, že i přes to lze SW zcela bez problémů využít. K prvnímu omezení v počtu zpracovávaných

datových řádků by vzhledem k rozsahu této práce nemělo dojít. Druhé omezení představovalo delší zpracovávání jednotlivých procesů, což se v průběhu práce projevilo větší časovou náročností při zpracovávání jednotlivých úloh [31].

Dalším kritériem bylo, aby nástroj umožňoval zpracovávat úlohy extrakce informací. Samotný nástroj RM zpracování textových úloh nepodporuje, ale je zde možnost využít některého z rozšíření, která jsou možná bezplatně doinstalovat z RM Marketplace. V této práci byly používány především rozšíření Text processing 8.1.0 pod licencí RM_EULA a Rosette Text Analytics pod licencí Vendor Specific. Bližší popis těchto rozšíření a jejich funkcionalit bude proveden v kapitole 5.4.2 Nástroj Rosette Text Analytics [31].

5.2. Frekvenční analýza rizikových e-shopů

Prvním úkolem této práce bylo vytvoření modelu, který bude ukazovat četnost slov a slovních spojení v popiscích rizikových e-shopů. Na základě této četnosti bylo posléze rozhodnuto o oblastech pro následnou extrakci informací. Výsledkem této části práce je model, jehož náhled je znázorněn na Obrázku 8. Tento model je součástí této práce na přiloženém DVD jako příloha pod názvem Model1_rizikove_eshopy_wordlist.



Obrázek 8: Model 1 – četnosti slov v rizikových e-shopech

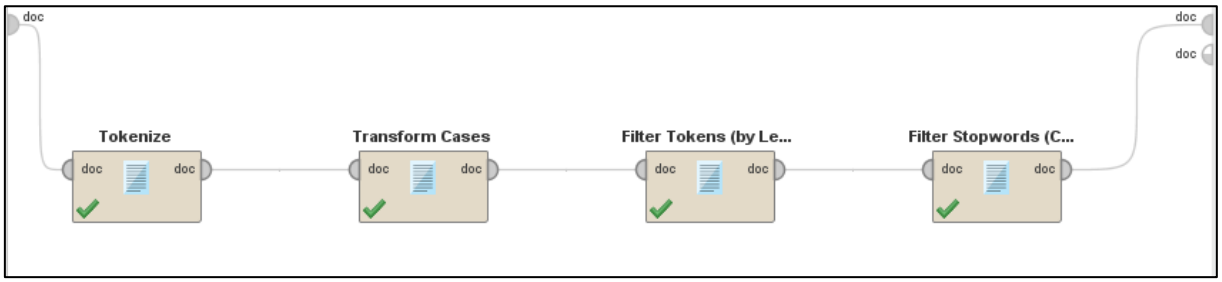
Zdroj: Vlastní zpracování

Nejprve bylo nutné načíst stránku s popisky rizikových e-shopů do programového prostředí RM. Toho bylo dosaženo pomocí modulu *Get Pages*, kdy adresa do tohoto modulu je načítána ze souboru typu xls, který byl do modelu načten pomocí modulu *Read Excel*. V modulu *Read Excel* byla nastavena cesta k souboru link_COI_rizikove_eshopy (součástí přílohy) pomocí nástroje *Import Configuration Wizard*. Tímto nástrojem nebyla nastavena jen cesta ale i rozsah dat, který bude načten, typ dat a označení názvu atributu (v tomto případě *Link*). Následně byl nastaven parametr modulu *GetPages* link attribute na: Link, ostatní parametry tohoto modulu byly ponechány v původním nastavení. Stránka byla následně získána pomocí požadavku GET a uložena v novém atributu. Výstupem tohoto modulu je tedy datová tabulka, což nebylo příliš vhodné, jelikož následující kroky vyžadovaly ke zpracování textové dokumenty. Z toho důvodu bylo následně využito modulu *Data to Documents*. Tento modul transformuje získaná data na dokumenty, kdy výstup z tohoto modulu je zobrazen na Obrázku 9.

Obrázek 9: Html kód stránky s rizikovými e-shopy

Zdroj: Vlastní zpracování

Ke splnění cíle této kapitoly, jímž bylo vytvoření frekvenční analýzy slov a slovních spojeních v popiscích rizikových e-shopů, bylo dále nutné vytvořit výčet slov, která se nachází v tomto dokumentu. K tomu byl využit modul *Process Documents 1*, jenž na svém výstupu *wor* poskytuje vektor všech slov v dokumentu. Aby se modul choval tak, jak bylo popsáno, bylo nutné do něj vnořit některé další moduly pro předzpracování textu. Ty jsou patrné z Obrázku 10.



Obrázek 10: Předzpracování dokumentu 1

Zdroj: Vlastní zpracování

Modulem *Tokenize* byla provedena tokenizace podle bílých znaků (*non letters*). Následně byly modulem Transform Cases převedeny všechny znaky v textu na malá písmena (*lower cases*). Takto vzniklé tokeny byly následně filtrovány podle své délky modulem *Filter Tokens (by lenght)*, kdy byl vybrány pouze tokeny o minimální délce 4 a maximální 25 znaků. Následně byl již využit pouze modul *Filter Stopwords (Czech)*, který měl za úkol z dokumentu odstranit česká stopslova. Výsledek této části je zobrazen v Tabulce 3.

Tento výsledek se pro další analýzu rizikových e-shopů ukázal jako nepřehledný. Bylo zde obsaženo velké množství slov, která se používají pro tvorbu webových stránek. Nejčastěji se v dokumentu vyskytovala slova *class* (1329 případů), *span* (886 případů) a *div* (842 případů).

Tabulka 3: Seznam slov 1

Word	Attribute Name	Total Occurences ↓	Document Occurences
class	class	1329	1
span	span	886	1
div	div	842	1
post	post	805	1
content	content	772	1
article	article	738	1
entry	entry	737	1
obchodní	obchodní	423	1
menu	menu	413	1
item	item	398	1
list	list	377	1
row	row	377	1
informati...	information	368	1
titles	titles	368	1
stránkách	stránkách	333	1

Zdroj: Vlastní zpracování

Dalším úkolem tedy bylo očistit tento dokument od html tagů, které tyto výrazy obsahují a vyextrahování částí textů, které obsahují popisky rizikových e-shopů. Toho bylo docíleno použitím modulu *Extract Content*, který stránky očistí od html tagů a vyextrahuje z nich pouze obsah. Tento modul zpracovával dokument s rizikovými e-shopy, který byl načten pomocí modulu *Get Page*, kde byla zadána url adresa stránky s rizikovými e-shopy https://www.coi.cz/pro-spotrebitele/rizikove-e-shopy/. Stránka byla získána opět pomocí funkce GET. Výstup modulu *Extract Content* je pak vyobrazen na Obrázku 11. Tento dokument byl pomocí modulu *Write Document* uložen jako textový dokument rizikove_eshopy_popisky.txt.

Obrázek 11: Popisky rizikových e-shopů

Zdroj: Vlastní zpracování

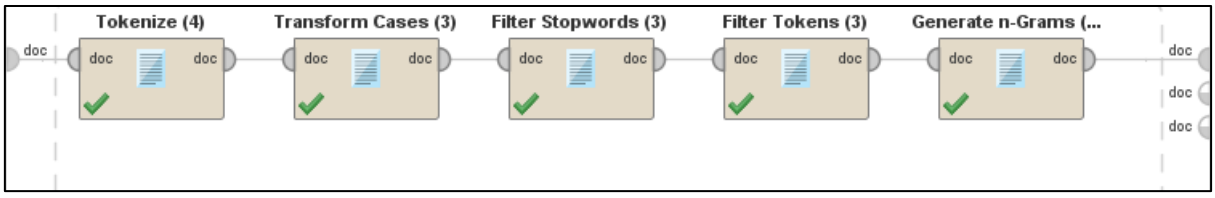
Na tomto výstupu byla, stejně jako v první případě, provedena frekvenční analýza slov, kdy bylo opět využito modulu *Process Documents*. V tomto modulu byly opět vnořené moduly *Tokenize*, *Transform Cases*, *Filter Tokens* a *Filter Stopwords*. Parametry těchto modulů byly nastaveny shodně jako v prvním případě. Výsledkem této frekvenční analýzy je Tabulka 4. Z této tabulky je již patrné, že nejčastěji použitým slovem ve vyextrahovaném textu je slovo *obchodní* (421 případů). Ani na základě tohoto výsledku však nebylo možné zcela s jistotou říci, co je nečastějším prohřeškem rizikových e-shopů. Z toho důvodu bylo dále přikročeno k vytvoření frekvenční analýzy nejčastěji užívaných slovních spojení (*n*-gramy).

Tabulka 4: Seznam slov 2

Word	Attribute Name	Total Occurences ↓	Document Occurences
obchodní	obchodní	421	1
stránkách	stránkách	333	1
stránky	stránky	265	1
zcela	zcela	263	1
česká	česká	261	1
inspekce	inspekce	260	1
není	není	236	1
spotřebitel	spotřebitel	214	1
může	může	206	1
práva	práva	199	1
nárokovat	nárokovat	198	1
smlouvu	smlouvu	198	1
uzavírá	uzavírá	198	1
vůči	vůči	198	1
komu	komu	196	1

Zdroj: Vlastní zpracování

Nejčastěji užívaná slovní spojení byla získána pomocí modulu *Process Documents*, kdy byly vnořeny moduly *Tokenize*, *Transform Cases*, *Filter Stopwords*, *Filter Tokens* a *Generate n-Grams (Terms)*, viz Obrázek 12. Zde oproti předešlým modelům přibyl pouze modul *Generate n-Grams (Terms)*, kdy byla nastavena maximální délka o třech slovech.



Obrázek 12: Vnořené moduly - N-gramy

Zdroj: Vlastní zpracování

Výsledkem této části modelu byla Tabulka 5, kdy bylo zjištěno, že velmi častými slovními spojeními jsou *obchodní_podmínky* (154 případů) a *stránky_zcela_anonymní* (116 případů).

Tabulka 5: Slovní spojení

Word	Attribute Name	Total Occurenc... ↓	Docum...
nakupem_techto_strankach	nakupem_techto_strankach	154	1
obchodní_podmínky	obchodní_podmínky	154	1
stránkách_česká	stránkách_česká	154	1
stránkách_česká_obchodní	stránkách_česká_obchodní	154	1
těchto_stránkách_česká	těchto_stránkách_česká	154	1
webu	webu	144	1
nikdo	nikdo	139	1
rizikový	rizikový	135	1
chybí	chybí	134	1
zcela_anonymní	zcela_anonymní	132	1
stránky_zcela	stránky_zcela	116	1
stránky_zcela_anonymní	stránky_zcela_anonymní	116	1
informace	informace	110	1
spotřebitele	spotřebitele	109	1
anonymní_spotřebitel	anonymní_spotřebitel	107	1

Zdroj: Vlastní zpracování

Dále bylo považováno za nutné provést výpočet, kolik rizikových e-shopů je na stránkách popsáno a v kolika případech se vyskytuje slovní spojení *„obchodní podmínky“*. Při prozkoumání dokumentu s vyextrahovanými popisky bylo zjištěno, že k oddělení komentářů e-shopů autoři používají následující sadu znaků „———————————— -->“, a právě těchto znaků bylo využito při tokenizaci textu.

Nejprve bylo provedeno načtení textového dokumentu *rizikove_eshopy_popisky.txt* do programového prostředí RM pomocí modulu *Read Document*. Tento dokument byl následně zpracován modulem *Process Documents 4*, v němž byl vnořen modul *Tokenize*. V modulu *Tokenize* byl zvolen mód tokenizace „*regular expression*“ podle již zmiňované posloupnosti znaků „———————————— -->“. Počet všech tokenů, jež představovali jednotlivé popisky rizikových e-shopů, byl pak patrný z výstupu *example set* modulu *Process Documents 4*. Dle výstupu z modelu RM bylo v den vzniku této práce na stránce

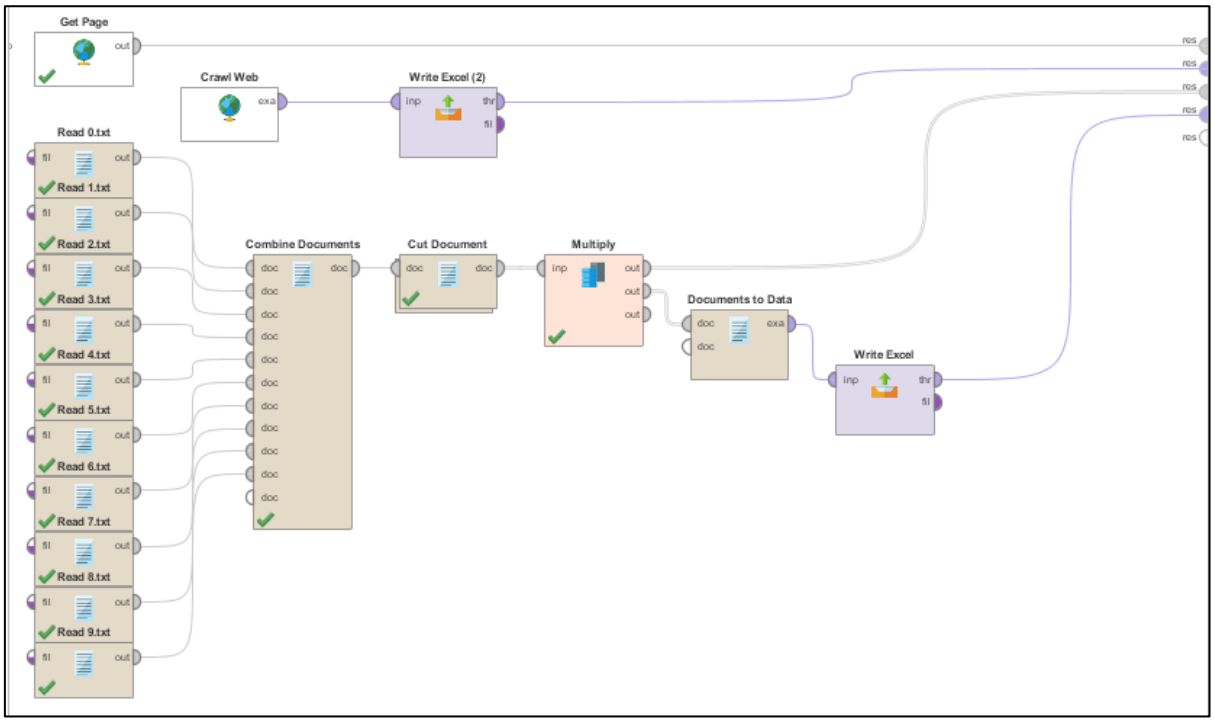
https://www.coi.cz/pro-spotrebitele/rizikove-e-shopy/ celkem 374 komentářů. Dalším úkolem bylo zjistit, v kolika z těchto komentářů se vyskytuje slovní spojení *„obchodní podmínky“*. I v tomto případě bylo k rozdělení dokumentu na jednotlivé komentáře využito modulu *Tokenize* se shodným nastavením módu tokenizace jako v předešlém případě. K vyfiltrování komentářů, které obsahují slovní spojení *„obchodní podmínky“*, bylo využito modulu *Filter Tokens (by Region)*. V tomto modulu bylo třeba nastavit parametr *condition* (což je způsob podle jakým modul tokeny filtruje) na *contains* a zadán řetězec znaků „*obchodní podmínky“*. Dle tohoto nastavení modul vyfiltroval všechny tokeny, které obsahují výše uvedené slovní spojení. Výsledkem tohoto vyfiltrování bylo, že toto spojení se nachází ve 153 případech.

Na základě přechozích výsledků bylo rozhodnuto, že práce bude zaměřena na obchodní podmínky e-shopů. Cílem bude vytvořit seznam náhodných e-shopů a pomocí funkcionalit RM určit, zda tyto e-shopy obsahují či neobsahují obchodní podmínky. Na obchodních podmínkách e-shopů bude dále provedena extrakce jmenných entit, kdy tyto entity budou voleny s ohledem na potřeby ČOI. Následně bude vyhodnocena úspěšnost nástroje RM pro extrakci informace. V poslední části pak bude poté provedena asociace jmenných entit s databází Wikidata a její vyhodnocení. Asociace bude provedena pomocí *QID* identifikátoru, který je jedinečný pro každou položku obsaženou v databázi Wikidata.

5.3. Sběr a zpracování dokumentů

Cílem této části práce bylo vytvořit sadu dokumentů, na kterých bude probíhat extrakce informace, konkrétně jmenných entit souvisejících s obchodními podmínkami. Na základě výsledků předchozí kapitoly bylo rozhodnuto, že tyto dokumenty budou představovat obchodní podmínky jednotlivých e-shopů. Pro dosažení cílů této kapitoly byl v programovém prostředí RM vytvořen model (příloha Model2_Nazvy_eshopu) uvedený na Obrázku 13.

Nejprve bylo nutné získat seznam e-shopů. K tomuto cíli byly využity webové stránky https://obchody.heureka.cz/, na kterých se nachází seznam e-shopů společně s jejich popisem. Z důvodu dalšího zpracování bylo nejprve nutné tuto stránku do programu RM načíst a blíže je analyzovat. K tomuto bylo využitu modulu *Get Page*, kdy byl zadán parametr url: https://obchody.heureka.cz/. Výsledkem tohoto modulu, bylo získání html kódu, který byl dále využit k určení pravidla, podle kterého byly vyextrahovány jednotlivé názvy e-shopů. Názvy jednotlivých e-shopů byly následně pomocí textových funkcí v programu MS Excel transformovány na webové odkazy.



Obrázek 13: Model 2 – Extrakce seznamu e-shopů

Zdroj: Vlastní zpracování

Ještě před tím, než byly vyextrahovány názvy jednotlivých e-shopů, bylo nutné získat html kódy následujících stránek, na kterých se seznam e-shopů nacházel. Bylo rozhodnuto, že bude využito celkově 10 stránek z celého seznamu, kdy na každé stránce se nacházelo 20 e-shopů. V práci se tedy dál bude pracovat s prvními 200 e-shopy, které jsou na těchto stránkách seřazeny podle počtu recenzí. K tomu bylo využito modulu *Crawl Web*, který umožňuje procházet webové stránky a ukládat je na místní uložiště. Aby byly procházeny pouze stránky, které byly požadovány, bylo nutné nastavit následující parametry:

- url: https://obchody.heureka.cz/ – počáteční adresa,
- crawling rules – pravidla podle kterých jsou stránky prohledávány a ukládány. V tomto případě bylo postačující nastavit pouze jedno pravidlo, kdy jeho aplikace byla nastavena na: follow_link_with_matching_url a jeho tvar na: „.+f=[0-9]“,
- max crawl depth: 10 – hloubka prohledávání,
- retrieve as html: true – bude stažen html kód stránek,
- write pages to disk: true – jednotlivé stránky budou uloženy do zvoleného adresáře na disku,
- max pages: 10 – bude staženo maximálně 10 stránek.

Výsledkem tohoto kroku bylo stažení deseti dokumentů html kódu, kde na každé je popisováno 20 e-shopů. Tyto stránky byly zároveň uloženy na disk jako textové soubory s příponou txt.

Dalším krokem bylo vyextrahování názvů jednotlivých e-shopů ze stažených textových souborů. K tomu bylo využito modulu *Cut Document*. Tento modul má za úkol rozdělit dokument na základě zvolených pravidel. K rozdělení dokumentu tento modul nabízí následující možnosti:

- *String matching* – je nutné určit počáteční a koncový řetězec znaků. Vše mezi tím je pak vyhodnoceno jako výsledek.
- *Regular expression* – zde se zadává seznam atributů a jejich regulárních výrazů.
- *Regular region* – atributy, pro jejichž vyhledávání se určuje počátek a konec oblasti čili dvou regulárních výrazů.
- *XPath queries* – extrakce pomocí jazyka XPath (vyžaduje kořenovou strukturu dokumentu).
- Indexed – extrakce pomocí indexu a délky shody.
- *JSON Path* – extrakce pomocí jazyka JSONPath (kořenová struktura dokumentu není vyžadována)

Před tím, než byla zvolena varianta extrakce, bylo nutné prozkoumat stažené dokumenty. Pro každý popis e-shopu byl použit obdobný html kód, část toho kódu je patrná na Obrázku 14.

```
      <td class="c-shops-table__cell c-shops-table__cell--cta">
        <div class="c-shops-table__cta o-inline-list o-inline-list--tight o-inline-list--spread">
          <a href="/profizoo-cz/recenze/overene" class="e-button e-button--simple e-button--small">Zobrazit recenze</a>
          <a href="https://www.heureka.cz/exit/profizoo-cz/?z=4" class="e-button e-button--small">Do obchodu</a>
        </div>
        <a class="c-shops-table__link" href="/profizoo-cz/recenze/overene"><svg class="e-icon" aria-label="Zobrazit recenze"><use xlink:href="#arrow--right"
></use></svg></a>
      </td>
    </tr>

    <tr class="c-shops-table__row">
      <td class="c-shops-table__cell c-shops-table__cell--logo">

          <a href="https://www.heureka.cz/exit/drmax-cz/?z=4" target="_blank"><img class="c-shops-table__logo" src="https://im9.cz/iR/importobchod-orig/6320_logo--
mm160x50.png" alt="Dr. Max lékárna"></a>

      </td>
      <th class="c-shops-table__cell c-shops-table__cell--name">
```

Obrázek 14: Html kód stránek e-shopu – příklad e-shopu profizoo.cz

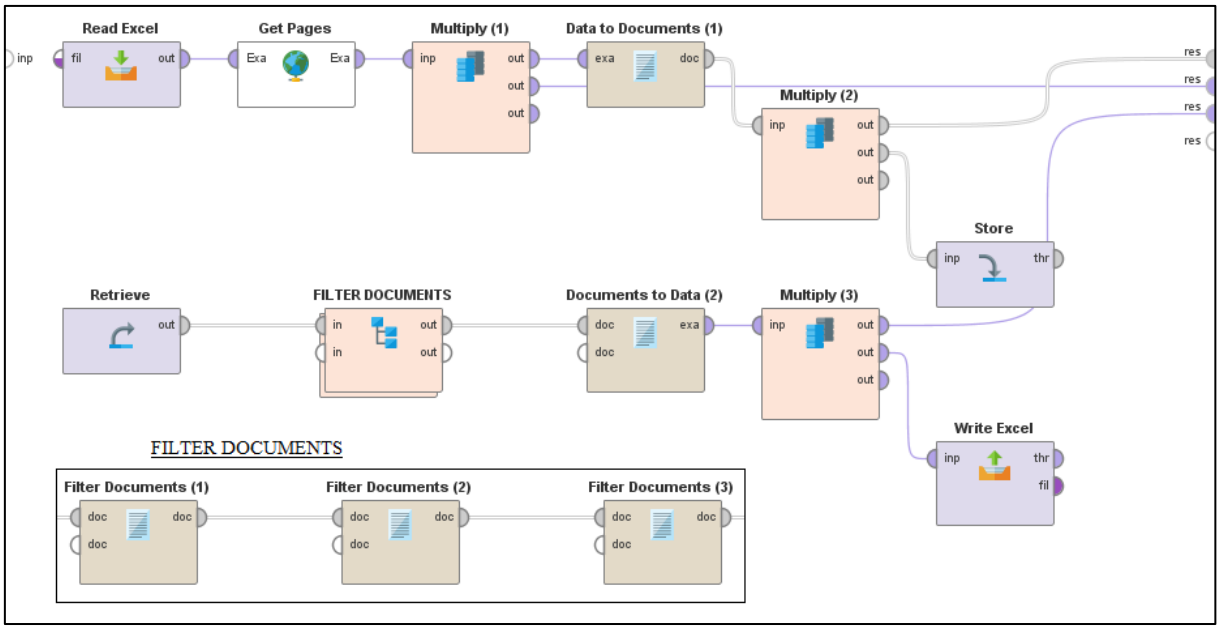
Zdroj: Vlastní zpracování

Pro další zpracování byla využita část kódu, která odkazuje na recenze e-shopu. Z této části byl následně vyextrahován název e-shopu do souboru *seznam_eshopy_HEUREKA.xlsx*. Zvolená metody extrakce tedy byla *String matching* kde:

- počáteční řetězec znaků: class="c-shops-table__link" href="/,
- koncový řetězec znaků: /recenze/overene"><svg class="e-icon".

Takto zvolený způsob extrakce zajistil, že byl prohledán celý dokument a vše mezi těmito řetězci bylo vyextrahováno. Aby nebylo nutné tuto operaci opakovat pro každý dokument zvlášť, byly jednotlivé dokumenty se seznamy e-shopů sloučeny do jednoho pomocí modulu *Combine Documents*. Výsledkem předešlého postupu bylo vyextrahování a uložení 200 názvů e-shopů (soubor *Seznam_eshopy_Heureka.txt*).

Dalším krokem, který bylo nutné provést, bylo prověření dostupnosti vyextrahovaných internetových obchodů. Nejprve bylo nutné převést názvy e-shopů na funkční odkazy. Toho bylo docíleno pomocí textových funkcí excelu. Samotné prověření dostupnosti již proběhlo v programovém prostředí RM, kde k tomu byl vytvořen model patrný z Obrázku 15.



Obrázek 15: Model 3 – Ověření dostupnosti

Zdroj: Vlastní zpracování

Nejprve byly jednotlivé odkazy načteny pomocí modulu *Read Excel*, což byl následně vstup do modulu *Get Pages*. Modul *Get Pages* měl za úkol stáhnout úvodní stránky jednotlivých internetových obchodů. Ve dnech tvorby této práce bylo tímto způsobem z výše zmíněných 200 odkazů odstraněno celkem 5 nefunkčních odkazů, což je patrné z výsledku, který je zobrazen

v Tabulce 6. Jmenovitě se jednalo o stránky: knihy-abz.cz, legenio.cz, megapixel.cz, pethome.cz a e-kondomy.cz.

Tabulka 6: Výsledek ověření dostupnosti e-shopů

Row No.	LINK	gensym1	URL	Respons...	Response-M...	Content-Type	Content-Len...
30	?	?	?	?	?	?	?
44	?	?	?	?	?	?	?
65	?	?	?	?	?	?	?
84	?	?	?	?	?	?	?
133	?	?	?	?	?	?	?
1	https://www.notino.cz	<!DOCTYPE ...	https://www.n...	200	OK	text/html; char...	233631
2	https://www.eva.cz	<!DOCTYPE ...	https://www.e...	200	OK	text/html; char...	0
3	https://www.lekarna.cz	<!DOCTYPE ...	https://www.l...	200	OK	text/html; char...	0
4	https://www.astratex.cz	<!DOCTYPE ...	https://www.a...	200	OK	text/html; char...	52412
5	https://www.parfemy-elnino.cz	<!DOCTYPE ...	https://www.p...	200	OK	text/html; char...	0
6	https://www.profizoo.cz	<!DOCTYPE ...	https://profizo...	200	OK	text/html; char...	0
7	https://www.drmax.cz	<!DOCTYPE ...	https://www.d...	200	OK	text/html;char...	0
8	https://www.kosmas.cz		https://www.k...	200	OK	text/html; char...	120539
9	https://www.altisport.cz	<!DOCTYPE ...	https://www.a...	200	OK	text/html; char...	0
10	https://www.prozdravi.cz	<!DOCTYPE ...	https://www.p...	200	OK	text/html; char...	0
11	https://www.parfemy.cz	<!DOCTYPE ...	https://www.p...	200	OK	text/html; char...	0

Zdroj: Vlastní zpracování

Jednotlivé úvodní stránky byly uloženy jako sada dokumentů pomocí modulu *Store*. Následně bylo nutné zjistit, které z e-shopů mají stránku s obchodními podmínkami. Po analyzování vzorku stránek bylo rozhodnuto, že toho bude docíleno prozkoumáním úvodních stránek, tj. zda obsahují slovní spojení „Obchodni podminky“. K této operaci bylo využito modulu *Filter Documents*. Tento modul prováděl zvolenou filtraci nad výše uvedenou sadou dokumentů, kterou bylo nejprve nutné načíst pomocí modulu *Retrieve*. V tomto modulu bylo nutné nastavit parametr *condition* na variantu *contains* a *string* na výraz „*obchodni-podminky*“, tento způsob nastavení zaručil, že modul vybere všechny dokumenty, ve kterých je obsažen výraz „*obchodni-podminky*“. Tímto způsobem bylo označeno celkem 155 dokumentů, které obsahují výraz „*obchodni-podminky*“.

Dalším krokem bylo zjistit, jakou úspěšnost má tento způsob označení e-shopů s obchodními podmínkami. Jinými slovy bylo potřeba zjistit, zda všechny e-shopy, které neobsahují slovní spojení „obchodni-podminky“ skutečně tuto stránku nemají, nebo zda pouze nebyla nalezena. Prvním krokem bylo vytvoření výpisu těchto e-shopů, k čemuž byl využit předchozí model s modulem *Filter Documents* s tím, že jeho parametr *invert condition* byl nastaven na *true.* Toto nastavení znamenalo, že modul označil všechny dokumenty, které slovní spojení „*obchodni-podminky*“ neobsahují. Výsledek tohoto kroku je zobrazen v Tabulce 7.

Tabulka 7: E-shopy bez obchodních podmínek

ExampleSet (40 examples, 0 special attributes, 10 regular attributes) Filter (40 / 40 examples): all

Row No.	obchodni_p...	LINK	URL	Response-C...	Response-M...
1	<!DOCTYPE ...	https://www.eva.cz	https://www.eva.cz	200	OK
2		https://www.kosmas.cz	https://www.kosmas.cz	200	OK
3		https://www.tsbohemia.cz	https://www.tsbohemia.cz	200	OK
4	<!DOCTYPE ...	https://www.insportline.cz	https://www.insportline.cz	200	OK
5	<!DOCTYPE ...	https://www.eobuv.cz	https://www.eobuv.cz	200	OK
6	<!DOCTYPE ...	https://www.sportobchod.cz	https://www.sportobchod.cz	200	OK
7	<!DOCTYPE ...	https://www.maxikovy-hracky.cz	https://www.maxikovy-hracky.cz	200	OK
8		https://www.exasoft.cz	https://www.exasoft.cz	200	OK
9	<!DOCTYPE ...	https://www.pneumatiky.cz	https://www.pneumatiky.cz	200	OK
10	<!DOCTYPE ...	https://www.kytary.cz	https://kytary.cz/	200	OK
11	<!DOCTYPE ...	https://www.kupkolo.cz	https://www.kupkolo.cz	200	OK
12	<!DOCTYPE ...	https://www.online-sport.cz	https://www.online-sport.cz	200	OK
13	<!DOCTYPE ...	https://www.e-pneumatiky.cz	https://www.e-pneumatiky.cz	200	OK
14	<!DOCTYPE ...	https://www.sw.cz	https://www.sw.cz	200	OK
15	<!DOCTYPE ...	https://www.xzone.cz	https://www.xzone.cz	200	OK
16	<!DOCTYPE ...	https://www.fotolab.cz	https://www.fotolab.cz	200	OK

Zdroj: Vlastní zpracování

Z důvodu dalšího zpracování byl tento výsledek uložen do souboru Seznam_eshopy_oznaceny_NEMAJI_OBCHPOD.xlsx. Následně bylo provedeno ruční prozkoumání těchto e-shopů, zda obchodní podmínky mají či nikoliv. Po důkladném prohledání jednotlivých e-shopů bylo zjištěno, že pouze jeden z označených obchodní podmínky skutečně nemá. Pomocí modulu *Filter Documents* tak bylo správně označeno 156 e-shopů z celkového počtu 195, což představovalo úspěšnost 80 %.

Z důvodu zvýšení úspěšnosti bylo po prozkoumání chybně označených stránek rozhodnuto o přidání dalších modulů *Filter Documents*. Jako první byl přidán modul, který měl z chybně označených e-shopů odstranit všechny, které obsahovaly řetězec „*obchodní podmínky*“. Po tomto kroku již bylo chybně označeno pouze 22 e-shopů, došlo tak k navýšení úspěšnosti na 88.7 %. Následně byl přidán třetí modul, který měl v parametru string zadán výraz „*o-nas*“. Těmto kroky bylo docíleno výsledku, kdy bylo chybně označeno pouze 16 e-shopů. Výsledná úspěšnost modelu tak byla 91.8 %.

Jak již bylo předesláno na začátku této kapitoly, tak samotná extrakce jmenných entit bude probíhat na obchodních podmínkách e-shopů. Zde bylo nejprve nutné vyřešit problém, který představoval modul pro extrakci jmenných entit v programovém prostředí RM. Problém spočíval v tom, že tento modul nepodporoval český jazyk. Z toho důvodu bylo rozhodnuto, že extrakce jmenných entit byla provedena na obchodních podmínkách e-shopů, které měly svou

verzi v anglickém jazyce. Po analýze výše uvedeného seznamu e-shopů, bylo zjištěno, že anglickou verzi poskytují následující stránky:

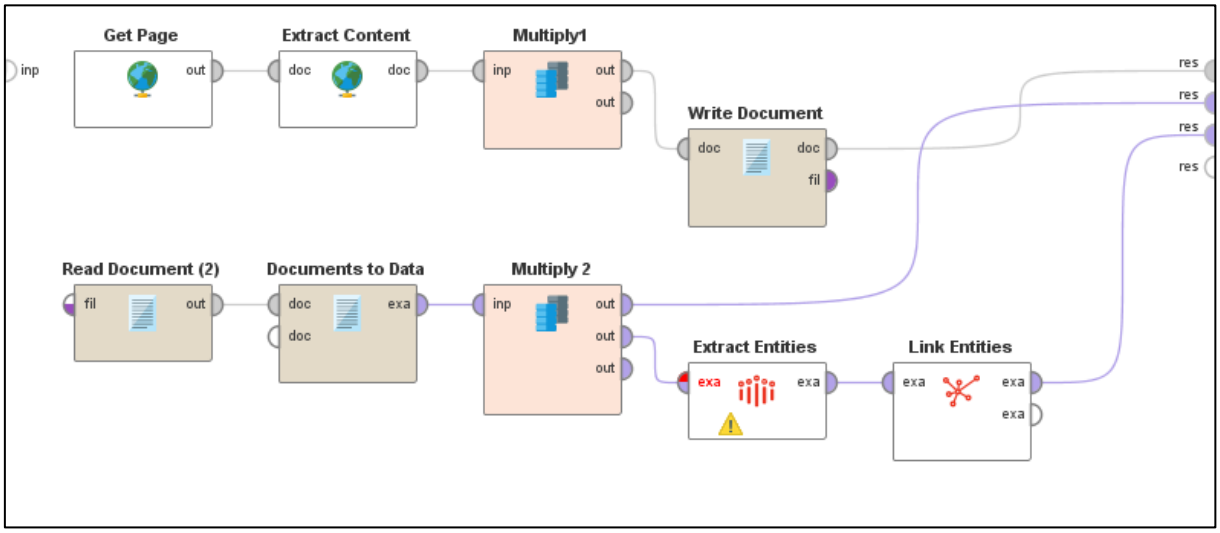
- https://www.matejovsky-bedding.com/
- https://www.insportline.eu
- https://kytary.co.uk/
- https://www.docsimon.com/
- https://www.skoda-parts.com/online-store.html
- https://www.originalky.eu
- https://www.stoklasa-eu.com/
- https://www.uni-max.co.uk/
- https://www.snowboard-zezula.com/
- https://www.ukposters.co.uk/
- https://www.fabricshouse.com/en/
- http://eshop.skoda-auto.com/cz/en/b2c

Z důvodu poměrně nízkého počtu e-shopů, které mají verzi v anglickém jazyce, bylo rozhodnuto o doplnění počtu těchto stránek na 30. Tento počet stránek by již měl být dostačující pro vyhodnocení modelu extrakce entit. Seznam byl doplněn o e-shopy, které se nacházely na seznamu stránek https://obchody.heureka.cz/ a následovaly předchozí vyextrahované e-shopy řazené dle počtu recenzí. Konkrétně bylo rozšíření provedeno o následující e-shopy:

- https://www.kasa.cz
- http://www.vltavadesign.cz
- https://www.hdt.cz
- https://www.xkko.eu
- http://www.glass-bohemia.com
- https://www.vivaco.cz
- https://www.vaprio.eu
- https://www.queens.global
- https://www.balistas.com
- https://www.dobeado.co.uk
- https://www.outfit4events.com
- https://www.kovonastroje.cz
- https://www.shopkilpi.cz
- https://www.vitalvibe.eu
- https://www.gina.cz

- https://www.indies.eu
- https://profimodel.cz
- https://www.cyklo69.cz

Pomocí modelu 4, který je na Obrázku 16, byly nejprve staženy stránky s obchodními podmínkami. Stažení těchto stránek probíhalo v následujících krocích. Nejprve byla tato stránka stažena pomocí modulu *Get Page*, kdy v tomto modulu musela být nejprve zadána url adresa stránek, ostatní parametry byly ponechány v původním nastavení. Pomocí modulu *Extract Content* byl následně ze stránek vyextrahován „čistý text“ bez veškerých html tagů. Následně byl tento obsah stránek uložen do textového souboru pomocí modulu *Write Document* (příloha složka: *Obchodni_podminky_eshopy_ENG*).



Obrázek 16: Model 4 – Extrakce entit

Zdroj: Vlastní zpracování

Po ohledání výsledků předchozího kroku bylo patrné, že se v textech nalézají výrazy s diakritikou (např. ve jménech či adresách). Toto by pro další zpracování představovalo problém (jak již bylo popsáno výše, modul pro extrakci entit nepodporoval český jazyk), a proto bylo nutné tuto diakritiku odstranit. K odstranění diakritiky byl použit program PSPad. Na každém dokumentu byla provedena úprava *Konverze-Odstranit diakritiku* a následně byl upravený dokument uložen bez diakritiky. Výsledkem předchozích kroků byla sada 30 textových dokumentů obsahujících obchodní podmínky jednotlivých e-shopů v anglickém jazyce, na kterých byla následně prováděna extrakce jmenných entit.

5.4. Extrakce zvolených entit z dokumentů

V této kapitole jsou uvedeny výsledky samotné extrakce jmenných entit. Nejprve bylo nutné rozhodnout, které entity budou z textu extrahovány. Po tomto výběru následovala již samotná extrakce v programovém prostředí RM, kdy pro tuto extrakci bylo využito rozšíření Rosette Text Analytics (bude podrobněji popsáno níže). Po extrakci jmenných entit bylo následně provedeno vyhodnocení přesnosti extrakce.

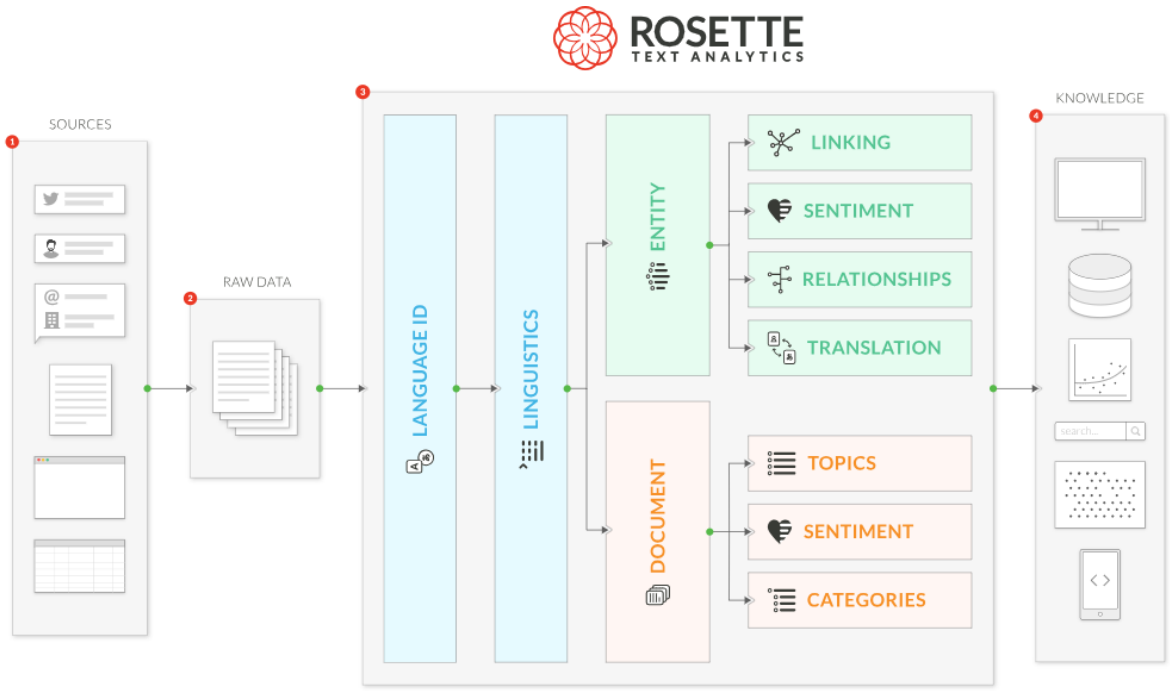
5.4.1. Výběr entit

Jak již bylo předesláno, první úkolem před samotnou extrakcí bylo rozhodnout o tom, které entity budou z textů extrahovány. Dle kapitoly 5.2 Frekvenční analýza rizikových e-shopů se ukázalo, že nejčastějšími prohřešky e-shopů jsou chybějící obchodní podmínky a jejich anonymita. Z tohoto důvodu byly zvoleny takové entity, kdy jejich výskyt v obchodních podmínkách znamená, že tento problém se dotčeného e-shopu netýká. S ohledem na výše uvedené, byly vybrány následující entity:

- Email – emailová adresa provozovatele e-shopu, kdy z textu bude extrahována konkrétní adresa.
- VAT – Value Added Tax, jedná se o ekvivalent pro české Daňové identifikační číslo.
- URL – Uniform Resource Locator, jedná se o adresu internetového obchodu, která určuje přesné umístění informací na internetu.
- Zákonná lhůta – představuje zákonem stanovenou lhůtu pro odstoupení od kupní smlouvy.
- Adresa – jedná se o adresu provozovatele e-shopu, kdy bylo rozhodnuto, že za úspěšné označení této entity bude postačující město a stát sídla organizace.
- Telefonní číslo – jedná se o kontaktní telefonní číslo.
- Název organizace – představuje název organizace provozující internetový obchod.

5.4.2. Nástroj Rosette Text Analytics

V této kapitole bude blíže představen rozšiřující nástroj Rosette Text Analytics. K analýze textu tento nástroj využívá zpracování přirozeného jazyka (NLP), statistické analýzy a strojového učení. Tento nástroj poskytuje celkem 13 nástrojů pro analýzu textu, viz Obrázek 17.



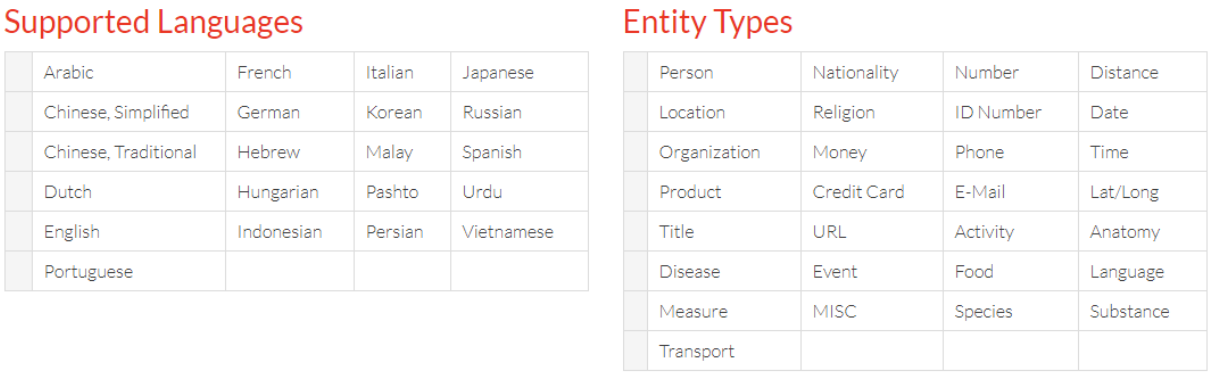
Obrázek 17: Rosette Text Analytics – přehled funkcí

Zdroj:[33]

V této práci byly využívány moduly *Extract Entities* a *Link Entitites*. Nástroj *Extract Entities* využívá hybridního přístupu pro vyvážení výkonu a přesnosti. Pro každý entitní typ je vybrán právě ten přístup, který poskytuje nejlepší výsledky. Tento nástroj kombinuje pokročilé statistické modelování, SVM a neuronové sítě (metody byly popsány v kapitole 2.1.1 a 3.1.2), které jsou doplněny o seznamy pravidelných výrazů a seznamy entit. Modely jsou učeny na vyváženém korpusu miliónů novinových článků, sociálních médiích a blogových příspěvcích. Statistické modelování hledá entity založené na kontextu, nikoliv ve shodě řetězců či vzorů. Kvalitních výsledků je tedy dosahováno pouze s velmi kvalitními trénovacími daty, i z tohoto důvodu Rosette využívá k tagování a anotování dat rodilé mluvčí. Entity, které se řídí vzorem (např. data, časy, emaily …), jsou pak označovány pomocí pravidel vyjádřených regulárními výrazy. Tento nástroj podporuje celkem 21 různých jazyků, viz Obrázek 18. V této práci byly využívány především tyto entitní typy, které zahrnují [33]:

- Location – města, státy, regiony, budovy, vodní plochy, parky, adresy …;
- Organization – korporace, firmy, instituce, agentury …;
- Person – lidský identifikátor podle jmen, přezdívek či aliasů;
- Identifier: Email – emailová adresa;
- Identifier: Phone_Number – telefonní číslo;

- Identifier: URL – webová adresa.



Supported Languages

	Arabic	French	Italian	Japanese
	Chinese, Simplified	German	Korean	Russian
	Chinese, Traditional	Hebrew	Malay	Spanish
	Dutch	Hungarian	Pashto	Urdu
	English	Indonesian	Persian	Vietnamese
	Portuguese			

Entity Types

	Person	Nationality	Number	Distance
	Location	Religion	ID Number	Date
	Organization	Money	Phone	Time
	Product	Credit Card	E-Mail	Lat/Long
	Title	URL	Activity	Anatomy
	Disease	Event	Food	Language
	Measure	MISC	Species	Substance
	Transport			

Obrázek 18: Podporované jazyky a entitní typy

Zdroj:[33]

Další nástroj, který byl v této práci využit, byl nástroj *Link Entities*. Tento nástroj využívá nejlepších osvědčených postupů textových analýz a statistického modelování k odhalení struktury a poznatků o entitách obsažených v textu. Text, který se vztahuje k entitě, se nazývá zmínka o subjektu. Zmínky o subjektech, které v reálném světě odkazují na stejnou entitu se Rosette snaží spojit. Toto spojování následně napomáhá stanovit totožnost subjektu tím, že shoduje různé názvy jako například přezdívky a formální výraz s identifikátorem entit [33], [32].

Vzhledem k povaze vstupních dokumentů bylo v této práci využito napojení Rosette na znalostní bázi Wikidata. Tato databáze je svobodná, mnohojazyčná, druhotná databáze, která shromažďuje strukturovaná data jako podporu projektů hnutí Wikimedia. Vzhledem k tomu, že se jedná o svobodný projekt, je možné data kopírovat, distribuovat, upravovat, uvádět je, včetně komerčních účelů, bez žádosti o povolení. Repozitář databáze Wikidata je tvořen položkami, které mají štítky, popisy a libovolné množství aliasů. Tyto položky jsou jednoznačně identifikovány písmenem *Q* následovaným číslem (QID identifikátor), jehož bude využito při asociaci entit. Modul *Link Entities* bude asociovat entity s položkami v databázi Wikidata. Pokud entita v databázi Wikidata existuje, poté Rosette vypíše unikátní identifikátor Wikidata QID. Tento nástroj podporuje entity, které se vyskytují v anglickém, čínském, japonském a španělském jazyce [32].

5.4.3. Extrakce entit

Po předzpracování dokumentů následovala již samotná extrakce entit, jež byla provedena pomocí modelu na Obrázku 16. Nejprve byly jednotlivé dokumenty načteny pomocí modulu *Read Document* a následně pomocí modulu *Documents to Data* převedeny do datové tabulky,

zde byl nastaven parametr *text attribute* na „*OP*“ (jedná se o označení textu). Na takto předzpracovaném dokumentu byla následně pomocí modulu *Extract Entities* provedena extrakce entit. V tomto modulu bylo nejprve nutné nastavit jeho parametry, a to především parametr *Connection*. Tento parametr zajišťuje napojení na databázi Rosette. Při prvním spuštění bylo nutné vytvořit nové připojení a získat přístupový API KEY. Ten byl získán na základě registrace na stránkách https://developer.rosette.com/. Při každém dalším spuštění již stačilo vybrat toto vytvořené připojení. Parametr *Source Language* byl nastaven na „*English*“ (obchodní podmínky e-shopů byly v anglickém jazyce) a u parametru *Attribute Selector* bylo vybráno *OP* (označení textu z modulu *Documents to Data*). Zbylé parametry byly nastaveny jako *false*. Výstup z tohoto extractoru byl přiveden na modul *Link Entities*, kde bylo provedeno nastavení parametrů obdobně jako u modulu *Extract Entities*. Parametr *Attribute Selector* byl nastaven na „*Entity*“ (asociace entit byla aplikována na vyextrahované entity). Zbylé parametry byl nastaveny na hodnotu *false*. Výsledkem pak byla tabulka, kde v sloupci *Entity* byly vypsány vyextrahované entity, v sloupci *EntityType* byl uveden typ entity a v sloupci *QID* se nacházel Wikidata identifikátor. Náhled výsledku extrakce pro obchodní podmínky e-shopu www.shopkilpi.cz je na Obrázku 19.

ExampleSet (40 examples, 2 special attributes, 4 regular attributes) Filter (40 / 40 examples): all

Row No.	InputID	OP	Entity	EntityType	QID	LinkEntity
1	1	Terms & Conditions ? ShopKilpi.cz ...	eshop@kilpi.cz	IDENTIFIER:EMAIL	T0	eshop@kilpi.cz
39	1	Terms & Conditions ? ShopKilpi.cz ...	+420 777 734 330	IDENTIFIER:PHONE_NUMBER	T0	+420 777 734 330
19	1	Terms & Conditions ? ShopKilpi.cz ...	www.sportkilpit.com	IDENTIFIER:URL	T0	www.sportkilpit.com
36	1	Terms & Conditions ? ShopKilpi.cz ...	http://ec.europa.eu/consumers/odr/	IDENTIFIER:URL	T0	http://ec.europa.eu/consumers/odr/
7	1	Terms & Conditions ? ShopKilpi.cz ...	EU	LOCATION	T0	EU
9	1	Terms & Conditions ? ShopKilpi.cz ...	Czech Republic	LOCATION	Q213	Czech Republic
32	1	Terms & Conditions ? ShopKilpi.cz ...	Prologis Park, D1 East	LOCATION	T0	Prologis Park, D1 East
38	1	Terms & Conditions ? ShopKilpi.cz ...	Ostrava - Nova Ves Czech Republic	LOCATION	Q213	Ostrava - Nova Ves Czech Republic
3	1	Terms & Conditions ? ShopKilpi.cz ...	Czech	NATIONALITY	T0	Czech
5	1	Terms & Conditions ? ShopKilpi.cz ...	English	NATIONALITY	T0	English
35	1	Terms & Conditions ? ShopKilpi.cz ...	European	NATIONALITY	T0	European
2	1	Terms & Conditions ? ShopKilpi.cz ...	VAT	ORGANIZATION	?	?
11	1	Terms & Conditions ? ShopKilpi.cz ...	Skiing Skialp Snowboard Freeride Cross	ORGANIZATION	T0	Skiing Skialp Snowboard Freeride Cross
13	1	Terms & Conditions ? ShopKilpi.cz ...	Fitness Cycling Skiing Skialp Snowboard Freeride Cross	ORGANIZATION	T0	Fitness Cycling Skiing Skialp Snowboard
15	1	Terms & Conditions ? ShopKilpi.cz ...	Skiing Snowboard Cross	ORGANIZATION	?	?
16	1	Terms & Conditions ? ShopKilpi.cz ...	Running Cycling Fitness Skiing Skialp Snowboard Freer...	ORGANIZATION	T0	Cycling Fitness Skiing Skialp Snowboard

Obrázek 19: Výsledek extrakce entit obchodních podmínek - www.shopkilpi.cz

Zdroj: Vlastní zpracování

Z náhledu výsledku extrakce obchodních podmínek e-shopu www.shopkilpi.cz jsou patrné následující entity:

- *eshop@kilpi.cz* – emailová adrese e-shopu, která byla označena jako entitní typ IDENTIFIER: EMAIL (řádek č.1),

- *+420 777 734 330* – telefonní číslo označené jako IDENTIFIER: PHONE_NUMBER (řádek č. 39),
- www.sportkilpit.com – webové stránky entitního typu IDENTIFIER: URL (řádek č. 19),
- *Ostrava – Nova Ves Czech Republic* – adresa, označená jako typ entity LOCATION,
- *VAT* – daňové organizační číslo, označené jako ORGANIZATION,
- *Ponature s.r.o* – název společnosti pod typem entity ORGANIZATION,
- *14 days* – Zákonná lhůta pro odstoupení od kupní smlouvy, typ entity TEMPORAL: TIME.

Obdobným způsobem byla provedena extrakce entit pro všech 30 dokumentů obchodních podmínek. Celkové zhodnocení výsledků a vyhodnocení přesnosti extrakce je provedeno v následující kapitole.

5.4.4. Vyhodnocení extrakce entit

V následující části práce bylo provedeno vyhodnocení extrakce pro jednotlivé typy entit. U každé entity byla vyhotovena přehledná tabulka s výsledky, kterých bylo dosaženo (Tabulka 8 - 14). Následně byly provedeny výpočty pro přesnost, úplnost, F-míru, úspěšnost, senzitivita a specificita.

Email – tato entita představuje emailovou adresu internetového obchodu. Tato entita je modulem *Extract Entities* extrahována pomocí pravidel vyjádřených regulárními výrazy. Vzhledem k tomu byla očekávána vysoká úspěšnost extrakce.

Tabulka 8: Výsledek extrakce – EMAIL

EMAIL		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	26	1
	NE	0	3

Zdroj: Vlastní zpracování

$$Přesnost = \frac{26}{26+0} \cdot 100 = 100\ \%,$$

$$Úplnost = \frac{26}{26+1} \cdot 100 = 96.3\ \%,$$

$$F - míra = \frac{2 \times 26}{2 \times 26 + 0 + 1} \cdot 100 = 98.1\ \%,$$

$$Úspěšnost = \frac{26+3}{26+1+0+3} \cdot 100 = 96.6\ \%,$$

$$Senzitivita = \frac{26}{26+1} \cdot 100 = 96.3\ \%,$$

$$Specificita = \frac{3}{0+3} \cdot 100 = 100\ \%.$$

Z výše uvedených výpočtů je patrné, že u této entity bylo dosaženo velmi dobrých výsledků. U přesnosti bylo dosaženo hodnoty 100 %, což znamená, že z 26 nalezených případů, které byly extraktorem nalezeny, se v obchodních podmínkách email nacházel skutečně 26krát. Hodnota úplnosti (senzitivity) 96.3 % představuje, že z 27 možných případů extraktor identifikoval entitu Email v 26 případech. Hodnota specificity dosáhla hodnoty 100 % (entita *Email* se v obchodních podmínkách nenacházela ve 3 případech, a právě tolikrát nebyla extraktorem nalezena). Úspěšnost extrakce entity *Email* byla 96.6 %. Vysokou úspěšnost lze vyvodit z toho, že extraktor Rosette vyhledává entitní typ Email pomocí pravidlového přístupu, kdy je využíváno regulárních výrazů.

VAT – jedná se o zkratku z anglického Value Added Tax, což představuje ekvivalent pro české daňové identifikační číslo. Vzhledem k tomu, že je předpoklad, že se tento výraz může objevovat v textu obchodních podmínek nejen k označení DIČ, nebyla očekávána příliš vysoká úspěšnost extrakce.

Tabulka 9: Výsledek extrakce – VAT

VAT		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	12	8
	NE	7	3

Zdroj: Vlastní zpracování

$$Přesnost = \frac{12}{12+7} \cdot 100 = 63.2\ \%,$$

$$Úplnost = \frac{12}{12+8} \cdot 100 = 60\ \%,$$

$$F - míra = \frac{2\times12}{2\times12+7+8} \cdot 100 = 61.5\ \%,$$

$$Úspěšnost = \frac{12+3}{12+8+7+3} \cdot 100 = 50\ \%,$$

$$Senzitivita = \frac{TP}{TP+FN} \cdot 100 = \frac{12}{12+8} \cdot 100 = 60\ \%,$$

$$Specificita = \frac{3}{7+3} \cdot 100 = 30\ \%.$$

Z výsledků je na první pohled patrné, že u entity *VAT* nebylo dosaženo vysoké úspěšnosti. Původní předpoklad, kdy extraktor nalezl entitu *VAT*, aniž by se jednalo o DIČ, se naplnil a výsledky tím byly značně ovlivněny. Přesnost extrakce byla pouze 63.2 %, což znamená, že z 19 případů (nalezených extraktorem) se ve skutečnosti v obchodních podmínkách nacházelo pouze 12. U úplnosti (senzitivity) bylo dosaženo hodnoty 60 % (z 20-ti možných případů nalezení entity *VAT* bylo nalezeno pouze 12). Specificita dosáhla hodnoty pouze 30 %. Takto nízká hodnota je zapříčiněna chybným označením entity *VAT* u obchodních podmínek, které tuto hodnotu neobsahovaly (celkem 7 případů). Úspěšnost extrakce byla tedy 50 %.

URL – Uniform Resource Locator, jedná se o adresu internetového obchodu. *URL* adresy mají svůj specifický tvar, čehož využívá extraktor Rosette. *URL* adresy jsou tak vyhledávány pomocí pravidel a regulárních výrazů. Stejně jako tomu bylo u entity Email i zde je očekávána vysoká úspěšnost extraktoru.

Tabulka 10: Výsledek extrakce – URL

URL		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	27	1
	NE	0	2

Zdroj: Vlastní zpracování

$$Přesnost = \frac{27}{27+0} \cdot 100 = 100\ \%,$$

$$Úplnost = \frac{27}{27+1} \cdot 100 = 96.4\ \%,$$

$$F - míra = \frac{2 \times 27}{2 \times 27+0+1} \cdot 100 = 98.2\ \%,$$

$$Úspěšnost = \frac{27+2}{27+1+0+2} \cdot 100 = 96.6\ \%,$$

$$Senzitivita = \frac{27}{27+1} \cdot 100 = 96.4\ \%,$$

$$Specificita = \frac{2}{0+2} \cdot 100 = 100\ \%.$$

Z výše uvedených výsledků je patrné, že u entity *URL* dosáhl extraktor vynikajících výsledků. Přesnost dosáhla 100 %, čili z 27 nalezených případů jich skutečně 27 bylo uvedeno v obchodních podmínkách. Úplnost dosáhla 96.4 %, což v tomto případě znamenalo, že entita byla nalezena celkem 27krát z 28 možných výskytů. Hodnota specificity dosáhla úspěšnosti 100 % (entita *URL* se v obchodních podmínkách nenacházela ve 2 případech, a právě tolikrát nebyla extraktorem nalezena). Úspěšnost extrakce dosáhla hodnoty 96.6 %.

Zákonná lhůta – tato entita představuje časový údaj, který udává zákonem stanovenou lhůtu pro odstoupení od kupní smlouvy. Časové entity jsou vyhledávány pomocí vzorů a regulárních výrazů, a tudíž zde byl předpoklad dosažení vysoké úspěšnosti extrakce.

Tabulka 11: Výsledek extrakce – Zákonná lhůta

ZÁKONNÁ LHŮTA		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	22	6
	NE	0	2

Zdroj: Vlastní zpracování

$$Přesnost = \frac{22}{22+0} \cdot 100 = 100\ \%,$$

$$Úplnost = \frac{22}{22+6} \cdot 100 = 78.6\ \%,$$

$$F - míra = \frac{2 \times 22}{2 \times 22+0+6} \cdot 100 = 88\ \%,$$

$$Úspěšnost = \frac{22+2}{22+6+0+2} \cdot 100 = 80\ \%,$$

$$Senzitivita = \frac{22}{22+6} \cdot 100 = 78.6\ \%,$$

$$Specificita = \frac{2}{0+2} \cdot 100 = 100\ \%.$$

I zde bylo dosaženo poměrně uspokojivých výsledků, avšak úplnost dosáhla pouze 78.6 %. Tato hodnota znamenala, že z 28 případů výskytu entity byla označena 22 krát. To bylo způsobeno především tím, že v některých obchodních podmínkách byla číslice 14 psána slovně. Hodnota specificity dosáhla úspěšnosti 100 % (entita *Zákonná lhůta* se v obchodních podmínkách nenacházela ve 2 případech a právě tolikrát nebyla extraktorem nalezena). Úspěšnost pak dosáhla 80 %.

Adresa – tato entita představuje adresu provozovatele e-shopu. V tomto případě bylo rozhodnuto, že za úspěšné označení této entity bude postačující město a stát. Entitní typ adresa je extraktorem vyhledávána pomocí statistického přístupu, neuronových sítí a strojového učení. Jelikož jsou adresy e-shopů složeny z českých názvů měst a ulic, nedá se předpokládat příliš vysoká úspěšnost (extraktor je trénován na dokumentech v anglickém jazyce).

Tabulka 12: Výsledek extrakce – Adresa

ADRESA		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	12	16
	NE	0	2

Zdroj: Vlastní zpracování

$$Přesnost = \frac{12}{12+0} \cdot 100 = 100\ \%,$$

$$Úplnost = \frac{12}{12+16} \cdot 100 = 42.9\ \%,$$

$$F - míra = \frac{2 \times 12}{2 \times 12+0+16} \cdot 100 = 60\ \%,$$

$$Úspěšnost = \frac{12+2}{12+16+0+2} \cdot 100 = 46.6\ \%,$$

$$Senzitivita = \frac{12}{12+16} \cdot 100 = 42.9\ \%,$$

$$Specificita = \frac{2}{0+2} \cdot 100 = 100\ \%.$$

Celková úspěšnost extrakce entity *Adresa* byla pouze 47 %. Tento výsledek potvrdil původní předpoklad, kdy problém spočíval, že extraktor byl trénován na dokumentech v anglickém jazyce. Je zde patrná velmi nízká úplnost (senzitivita) 42.9 %, kdy bylo extraktorem nalezeno pouze 12 adres z 28 možných. Hodnota specificity pak dosáhla 100 % (ve dvou případech se adresa v obchodních podmínkách e-shopů nenacházela a právě tolikrát nebyla nalezena)

Telefon – jedná se o kontaktní telefonní číslo. Stejně jako entity *Email, URL* a *Zákonná lhůta* je i tato vyhledávána pomocí vzorů a regulárních výrazů. Z toho důvodu by měla extrakce dosáhnout dobrých výsledků.

Tabulka 13: Výsledek extrakce – Telefon

TELEFON		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	20	5
	NE	0	5

Zdroj: Vlastní zpracování

$$Přesnost = \frac{20}{20+0} \cdot 100 = 100\ \%,$$

$$Úplnost = \frac{20}{20+5} \cdot 100 = 80\ \%,$$

$$F - míra = \frac{2 \times 20}{2 \times 20+0+5} \cdot 100 = 88.9\ \%,$$

$$Úspěšnost = \frac{20+5}{20+5+0+5} \cdot 100 = 83.3\ \%,$$

$$Senzitivita = \frac{20}{20+5} \cdot 100 = 80\ \%,$$

$$Specificita = \frac{20}{0+20} \cdot 100 = 100\ \%.$$

Úspěšnost extrakce entity *Telefon* byla 83.3 %, což bylo méně, než jaký byl předpoklad. Celková úspěšnost byla výrazně ovlivněna úplností (senzitivitou), kdy z 25 výskytů této entity bylo extraktorem nalezeno pouze 20 případů. Specificita dosáhla hodnoty 100 % (5 případů, kdy se entita *Telefon* v obchodních podmínkách nenacházela a extraktorem nebyla správně nalezena).

Název organizace – jedná se o entitu, která představuje název organizace provozující internetový obchod.

Tabulka 14: Výsledek extrakce – Název organizace

NÁZEV ORGANIZACE		*Nalezena extraktorem (ANO/NE):*	
		ANO	NE
Výskyt v obchodních podmínkách e-shopu (ANO/NE):	ANO	18	10
	NE	0	2

Zdroj: Vlastní zpracování

$$Přesnost = \frac{18}{18+0} \cdot 100 = 100\ \%,$$

$$Úplnost = \frac{18}{18+10} \cdot 100 = 64.3\ \%,$$

$$F - míra = \frac{2\times18}{2\times18+0+10} \cdot 100 = 78.3\ \%,$$

$$Úspěšnost = \frac{18+2}{18+10+0+2} \cdot 100 = 66.7\ \%,$$

$$Senzitivita = \frac{18}{18+10} \cdot 100 = 64.3\ \%,$$

$$Specificita = \frac{18}{0+18} \cdot 100 = 100\ \%.$$

Úspěšnost extrakce entity *Název organizace* dosáhla 66.7 %, kdy extraktorem nebylo nalezeno celkem 10 názvů organizací, které se ovšem v obchodních podmínkách nacházely, což představovalo úplnost (senzitivitu) 64.3 %. Specificita extrakce pak byla 100 % (ve dvou případech se entita Název organizace v obchodních podmínkách nenacházela a extraktorem nebyla extrahována). Nižší úspěšnost extrakce byla spojena s problematikou českého jazyka,

kdy extraktor očekával jazyk anglický (zkratky jako a.s., s.r.o. nebyly ve většině případů přeloženy).

5.5. Asociace jmenných entit do relací

V následující kapitole bylo provedeno vyhodnocení asociace jmenných entit. Asociace byla provedena v programu RM pomocí modulu Link Entity, jak bylo popsáno v kapitole 5.4.2 Nástroj Rosette Text Analytics. Vyhodnocení asociace bylo provedeno na výsledcích extrakce u všech dokumentů obchodních podmínek, na entitním typu *Organization*. Cílem bylo zjistit, jaká je úspěšnost přiřazení QID ientifikátoru na databázi Wikidata k jednotlivým entitám, a tedy praktická využitelnost tohoto nástroje. Výsledky jsou vyobrazeny v následující, kdy ve sloupci *E-shop* se nachází názvy jednotlivých e-shopů, ve sloupci *Možných asociací* je udáván počet entit, které se nacházejí v databázi Wikidata a v posledním sloupci *Počet nalezených* je uveden počet nalezených asociací s databází Wikidata.

Tabulka 15: Asociace entit

E-shop	Možných asociací	Počet nalezených
docsimon.com	6	3
eshop.skoda-auto.com	7	6
fabricshouse.com	6	4
kytary.co.uk	5	3
matejovsky-bedding.com	5	1
originalky.eu	15	10
skoda-parts.com	8	2
snowboard-zezula.com	6	6
stoklasa-eu.com	2	1
europosters.eu	0	0
uni-max.co.uk	3	2
balistas.com	8	6
dobeado.co.uk	5	2
gina.cz	2	0
glass-bohemia.com	1	0
hdt.cz	2	0
indies.eu	2	2
kasa.cz	7	3
kovonastroje.cz	5	1
outfit4events.com	7	4
queens.global	12	7
shopkilpi.cz	3	1
vaprio.eu	10	5
vitalvibe.eu	3	1
vivaco.cz	2	1

insportline.eu	1	0
vltavadesign.cz	7	3
xkko.eu	1	0
profimodel.cz	5	2
cyklo69.cz	7	0
CELKEM	**153**	**76**
Celková úspěšnost	49.7 %	

Zdroj: Vlastní zpracování

Z Tabulky 15 je patrné, že extraktor vyextrahoval celkem 153 entit typu ORGANIZATION, které bylo možné nalézt v databázi Wikidata. Z tohoto počtu byla provedena asociace s databází Wikidata u 76 případů. Z výše uvedeného tak vyplývá úspěšnost asociace 49.7 %. U asociace entit nebyly vypočteny metriky přesnost, úplnost a F-míra, jelikož nebylo možné přesně stanovit počet všech entit, které se v dokumentech nacházely a extraktorem nebyly nalezeny. Aby toto bylo možné, bylo by nutné prohledat všechny dokumenty obchodních podmínek a ručně označit všechny potenciální entity entitního typu *Organization.*

ZÁVĚR

Cílem diplomové práce bylo charakterizovat současné přístupy k extrakci informací z textových dokumentů a tyto následně aplikovat na dokumentech s využitelností pro ČOI.

Jednotlivé techniky extrakce byly prezentovány v první části práce, která se těmto technikám věnovala po teoretické stránce. Jako první bylo nutné představit textová data a jednotlivé techniky předzpracování těchto dat. Byly zde prezentovány techniky jako jsou segmentace textu, tokenizace, tagování, stemming, lemmatizace a odstranění stopslov, kdy tyto metody byly následně využity v druhé části práce, ve které byly tyto metody využity při tvorbě jednotlivých modelů. Poté, co byly charakterizovány jednotlivé metody předzpracování dat, byly následně prezentovány oblasti využití text miningu jako jsou kategorizace textů, shlukování textů, analýza sentimentu, shrnutí textu, získávání informací, extrakce informací a asociace entit. A právě extrakce informací a asociace entit se staly hlavními zájmy zbylé části práce. V následující kapitole byly charakterizovány jednotlivé techniky, které se věnují extrakci informací z textu a metriky systémů pro extrakci informací.

Následující část práce již byla zaměřena na úlohy spojené s extrakcí informací z textových dokumentů s ohledem na využitelnost pro ČOI. Prvním, co bylo nutné vyřešit, bylo stanovení oblasti, které se bude práce věnovat. Vzhledem k výsledkům, které ČOI prezentuje ve své závěrečné zprávě za rok 2015 (součást této práce), patří internetové obchodování mezi nejsledovanější oblasti kontrol ČOI. Z toho důvodu bylo rozhodnuto, že práce bude dále zaměřena na internetové obchodování. Dále bylo nutné zjistit, co je nejčastějším prohřeškem provozovatelů internetových obchodů. Za tímto účelem byl vytvořen model v programovém prostředí RM, pomocí něhož byla provedena frekvenční analýza výskytu slov v popiscích rizikových e-shopů, které ČOI zveřejňuje na svých internetových stránkách. Pomocí programového prostředí RM byl získán html kód stránky s popisky rizikových e-shopů. S tímto kódem bylo následně pracováno, byly na něm uplatněny metody předzpracování textu jako tokenizace, transformace velkých písmen, filtrace tokenů, odstranění stopslov a odstranění html tagů. Výsledkem bylo, že nejčastěji vyskytujícím se slovem v očištěném textu bylo slovo *obchodní*. Následně byl vytvořen model, pomocí kterého bylo zjištěno, že slovo „*obchodní“* se v textu nejčastěji vyskytuje ve spojení se slovem „*podmínky“*. Slovní spojení „*obchodní podmínky“* se v dokumentu objevilo celkem 154krát. Dále bylo zjištěno, že celkem 116krát se v textu vyskytlo slovní spojení „*stránky zcela anonymní“*. Následně byl vytvořen model, pomocí kterého bylo zjištěno, že na stránkách s popisky rizikových e-shopů se nachází celkem 374 těchto popisků a slovní spojení „obchodní podmínky“ se vyskytuje v celkem 153

případech. Z výše uvedeného bylo rozhodnuto, že extrakce entit bude dále provedena na obchodních podmínkách e-shopů.

Vzhledem k tomu, že extrakce entit měla být realizována na obchodních podmínkách e-shopů, bylo nejprve nutné dokumenty s obchodními podmínkami získat. Tomuto úkolu byla věnována kapitola Sběr a zpracování dokumentů. K tomuto účelu byla vybrána internetová stránka *obchody.heureka.cz*, na kterých se nachází seznam e-shopů řazených sestupně dle počtu recenzí. Bylo stanoveno, že pro práci bude dále postačující pracovat s 200 e-shopy. Za tímto účelem byl vytvořen model, pomocí něhož byly staženy stránky obsahující názvy e-shopů a pomocí regulárního výrazu byly následně vyextrahovány názvy jednotlivých e-shopů.

V následující kapitole byl vytvořen model, pomocí něhož byla ověřena funkčnost odkazů na jednotlivé e-shopy a také zda jednotlivé stránky obsahují stránku s obchodními podmínkami. Tímto model bylo zjištěno, že 5 odkazů na internetové stránky obchodu je již neaktivních. Aktivní e-shopy byly filtrovány podle toho, zda obsahují stránku s obchodními podmínkami. Postupně byly přidávány filtry, které stránky filtrovaly dle regulárních výrazů, kdy výsledná úspěšnost modelu dosáhla 91.8 %.

Zbylá část práce se již věnovala extrakci entit z obchodních podmínek e-shopů. Z důvodu, že modul Rosette pro extrakci entit v programovém prostředí RM nepodporuje český jazyk, bylo nutné pracovat se stránkami, které měly verzi v anglickém jazyce. Těchto stránek bylo z daného seznam pouze 12 a proto byly doplněny o dalších 18 na celkový počet 30. Následně byly stanoveny entity, které měly být v obchodních podmínkách extraktorem nalezeny. Jednalo se o entity Email, VAT, URL, Zákonná lhůta, Adresa, Telefonní číslo a Název organizace. U extrakce těchto entit bylo následně provedeno vyhodnocení dle prezentovaných metrik. Modul dosahoval vysoké úspěšnosti především u entit, které byly vyhledávány pomocí pravidel vyjádřených regulárními výrazy u těchto bylo dosaženo úspěšnosti přes 90 %. Naopak u entity Adresa, která je vyhledávaných pomocí hlubokých neuronových sítí, bylo dosaženo úspěšnosti pouze 46.6 %. Takto nízká úspěšnost byla zapříčiněna především překladem stránek do anglického jazyka, kdy učení modulu probíhalo na anglických textech.

Poslední kapitola práce byla zaměřena na asociaci entit. Asociace entit byla součástí předchozího modelu, kdy modulem *Link Entities* byly jednotlivé entity, pomocí jedinečného QID identifikátoru, asociovány s položkami databáze Wikidata. Z důvodu velké časové náročnosti byla asociace vyhodnocena pouze pro entity entitního typu *Organization*. Z celkového počtu 153 entit entitního typu *Organization* bylo správně asociováno 76 a bylo tak dosaženo úspěšnosti 49.7 %.

Stanovené cíle práce se podařilo naplnit. Po teoretické části, kde byly charakterizovány současné přístupy k extrakci informací, byly vytvořeny celkem 4 modely, jejichž výstupy by se při mapování internetového obchodování daly prakticky využít. Využít by se daly zejména při prověřování funkčnosti odkazů na internetové obchodování dále při prověřování, zda tyto e-shopy poskytují zákazníkům povinné obchodní podmínky a zda tyto obchodní podmínky obsahují i povinné náležitosti jako jsou název provozovatele, kontakt, sídlo, daňové identifikační číslo apod.

POUŽITÁ LITERATURA

[1] AGGARWAL, C. C., ZHAI, C. *Mining text data*. New York: Springer, c2012. ISBN 978-1-4614-3222-7.

[2] AWAD, Mariette a Rahul KHANNA. *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Berkley: Apress Open, [2015]. ISBN 978-1-4302-5989-3.

[3] BALOG Krisztian. *Entity-oriented search: The information retrieval series,* vol 39. Springer, Cham, 2018. ISBN 978-3-319-93933-9.

[4] BENESTY, Jacob., M. Mohan SONDHI a Yiteng HUANG. *Springer handbook of speech processing*. London: Springer, 2008. ISBN 3540491252.

[5] BERKA, Petr. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.

[6] *Classification Performance Metrics: NLP-FOR-HACKERS* [online]. [cit. 2018-02-14]. Dostupné z: http://nlpforhackers.io/classification-performance-metrics/

[7] *COI - TISKOVÁ ZPRÁVA: E-shopy klamou a nevyřizují reklamace* [online]. [cit. 2018-11-27]. Dostupné z: https://www.coi.cz/e-shopy-klamou-a-nevyrizuji-reklamace/

[8] *COI: Česká obchodní inspekce* [online]. [cit. 2018-10-09]. Dostupné z: https://www.coi.cz/

[9] CUNNINGHAM, Hamish. *Information Extraction, Automatic* [online]. [cit. 2018-03-08]. Dostupné z: https://gate.ac.uk/sale/ell2/ie/main.pdf Department of Computer Science. University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK.

[10] DODDINGTON, G., MITCHELL, A., PRZYBOCKI, R. *The automatic content extraction (ACE) program tasks, data, and evaluation* [online]. [cit. 2018-11-28]. Dostupné z: https://pdfs.semanticscholar.org/0617/dd6924df7a3491c299772b70e90507b195dc .pdf

[11] DOSTÁL, Petr, Karel RAIS a Zdeněk SOJKA. *Pokročilé metody manažerského rozhodování: konkrétní příklady využití metod v praxi*. Praha: Grada, 2005. Expert (Grada). ISBN 80-247-1338-1.

[12] FELDMAN, Ronen a James SANGER. *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press, 2007. ISBN 978-0-521-83657-9.

[13] GELBUKH, Alexander. Computational linguistics and intelligent text processing. *8th international conference*, CICLing 2007, Mexico City, Mexico, February 18-24, 2007: proceedings. New York: Springer, c2007. ISBN 978-3-540-70938-1.

[14] GRISHMAN, Ralph a Beth SUNDHEIM. *Message understanding conference - 6: A brief history* [online]. [cit. 2018-02-14]. Dostupné z: https://pdfs.semanticscholar.org/6723/dda58e5e09089ec78ba42827b65859f030e2.pdf

[15] GUPTA, Rahul. *Conditional random fields* [online]. [cit. 2018-02-14]. Dostupné z: https://pdfs.semanticscholar.org/13cb/794cae7c42bb20b2ca41041086bc3ec45b77.pdf

[16] HITOSHI, I., *Evolutionary approach to machine learning and deep neural networks*. New York, NY: Springer Berlin Heidelberg, 2018. ISBN 978-981-13-0199-5.

[17] CHAIRS, general a HAIZHOU LI AND A. KUMARAN. --. *Proceedings of the 2009 Named Entities Workshop: August 7, 2009, Suntec, Singapore*. Morristown, N.J: Association for Computational Linguistics, 2009. ISBN 9781932432572.

[18] CHIEU, Hai Leong a Hwee Tou NG. *Named Entity Recognition with a Maximum Entropy Approach* [online]. [cit. 2018-02-14]. Dostupné z: http://www.aclweb.org/anthology/W03-0423

[19] INDURKHYA, Nitin a Fred J. DAMERAU. *Handbook of natural language processing*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2010. ISBN 9781420085938.

[20] JANÍČEK, Přemysl a Jiří MAREK. *Expertní inženýrství v systémovém pojetí*. Praha: Grada, 2013. Expert (Grada). ISBN 978-80-247-4127-7.

[21] KAO, Anne. a Stephen R. POTEET. *Natural language processing and text mining*. London: Springer, c2007. ISBN 18-462-8175-x.

[22] LAFFERTY, J., MCCALLUM, A., PEREIRA, F. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Dostupné z: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers

[23] LIU, Bing. *Sentiment analysis and opinion mining*. San Rafael: Morgan & Claypool Publishers, c2012. Synthesis lectures on human language technologies, 16. ISBN 978-1-60845-884-4.

[24] MANNING, Christopher D. a Hinrich SCHÜTZE. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press, c1999. ISBN 0262133601.

[25] MANNING, Christopher D., Prabhakar RAGHAVAN a Hinrich SCHÜTZE. *An introduction to information retrieval* [online]. [cit. 2018-02-13]. Dostupné z: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[26] MARKOWETZ, Florian. *Classification by support vector machines: Computational molecular biology* [online]. [cit. 2018-02-13]. Dostupné z: http://compdiag.molgen.mpg.de/ngfn/docs/2003/mar/SVM.pdf

[27] MINER, Gary. *Practical text mining and statistical analysis for non-structured text data applications*. Waltham, MA: Academic Press, 2012. ISBN 978-0-12-386979-1.

[28] MOENS, Marie-Francine. *Information extraction algorithms and prospects in a retrieval context*. Dordrecht: Springer, 2006. ISBN 9781402049934.

[29] *Portál|Český národní korpus* [online]. [cit. 2018-02-15]. Dostupné z: http://wiki.korpus.cz/doku.php/pojmy:korpus

[30] POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014. ISBN 978-80-87895-17-7.

[31] *RapidMiner: RapidMiner Pricing* [online]. [cit. 2018-11-12]. Dostupné z: https://rapidminer.com/pricing/#RapidMiner-Studio-Pricing

[32] *Rosette API: Features & Functions* [online]. [cit. 2018-11-11]. Dostupné z: https://developer.rosette.com/features-and-functions#entity-extraction-and-linking-entity-linking

[33] *Rosette Text Analytics Extension for Rapidminer Predictive Analtics* [online]. [cit. 2018-11-10]. Dostupné z: https://www.rosette.com/rapidminer/

[34] SANG, Erik F. Tjong Kim a Fien DE MEULDER. *Introduction to the CoNLL-2003 Shared task:: Language-independent named entity recognition* [online]. [cit. 2018-02-15]. Dostupné z: http://aclweb.org/anthology/W03-0419

[35] SEKINE, Satoshi. a Elisabete RANCHHOD. *Named entities: recognition, classification, and use*. Philadelphia: John Benjamins Pub. Company, c2009. ISBN 978-90-272-8922-3.

[36] SEYMORE, Kristie a Roni ROSENFELD. *Learning Hidden Markov Model Structure for Information Extraction* [online]. [cit. 2018-02-13]. Dostupné z: https://www.ri.cmu.edu/pub_files/pub1/seymore_kristie_1999_1/seymore_kristie_1999_1.pdf

[37] SILVA, Catarina. a Bernardete. RIBEIRO. *Inductive inference for large scale text classification: kernel approaches and techniques*. Berlin: Springer, c2010. Studies in computational intelligence, v. 255. ISBN 978-3-642-04532-5.

[38] SUTTON, Charles a Andrew. MCCALLUM. *Introduction to conditional random fields*. S.l.: World Scientific, 2012. ISBN 9781601985729. str. 268 - 330

[39] VAN ASCH, Vincent. *Macro-and micro-averaged evaluation measures [[BASIC DRAFT]]* [online]. [cit. 2018-02-15]. Dostupné z: https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf

[40] *Výzkum zpracování přirozeného jazyka: Informatika @ PEF MENDELU* [online]. [cit. 2018-11-24]. Dostupné z: https://informatika.mendelu.cz/cz/clanek/zpracovani-prirozeneho-jazyka

[41] WANG, Lipo. *Support vector machines: theory and applications*. Berlin: Springer, 2005. ISBN 978-3-540-24388-5.

[42] WEISS, Sholom M. *Text mining: predictive methods for analyzing unstructured information*. New York: Springer, 2005. ISBN 0-387-95433-3.

[43] WESTON, Jason. *Support vector machine (and statistical learning theory) tutorial: 4 independence way*. Princeton, *USA* [online]. [cit. 2018-02-13]. Dostupné z: http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf

[44] YUAN-CHAO, L., MING, L., XIAO-LONG, W., *InTechOpen: Application of self-organizing maps in text clustering: A review* [online]. [cit. 2018-02-22]. Dostupné z: https://www.intechopen.com/books/applications-of-self-organizing-maps/application-of-self-organizing-maps-in-text-clustering-a-review

[45] Zákon č. 64/1986 Sb. Zákon České národní rady o České obchodní inspekci

SEZNAM PŘÍLOH

Příloha A

Popis jednotlivých adresářů a souborů:

/ (kořenový adresář)

- link_COI_rizikove_eshopy.xlsx

- rizikove_eshopy_popisky.txt

- Seznam_eshopy_HEUREKA.xlsx

- Seznam_eshopy_HEUREKA-LINKY.xlsx

- Seznam_eshopy_oznaceny_NEMAJI_OBCHPOD.xlsx

/ Modely

- Model1_rizikove_eshopy_wordlist.rmp

- Model1_rizikove_eshopy_wordlist.proprties

- Model2_Nazvy_eshopu.rmp

- Model2_Nazvy_eshopu.properties

- Model3_Overeni_dostupnosti.rmp

- Model3_Overeni_dostupnosti.properties

- Model4_extrakce_entit.rmp

- Model4_extrakce_entit.properties

/ Obchodni_podminky_eshopy_ENG

- Soubory typu *txt* obsahující obchodní podmínky jednotlivých e-shopů

/ Seznam_eshopu

- Soubory typu *txt*, které obsahující html kódy stránek obchody.heureka.cz se seznamy e-shopů.