Mendelova univerzita v Brně
Provozně ekonomická fakulta

Dolování znalostí z vícejazyčných textových dat

Bakalářská práce

Vedoucí práce:
doc. Ing. František Dařena, Ph.D.

Luděk Svozil

Brno 2017

Čestné prohlášení

Prohlašuji, že jsem tuto práci: **Dolování znalostí z vícejazyčných textových dat**
vypracoval samostatně a veškeré použité prameny a informace jsou uvedeny v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů, a v souladu s platnou *Směrnicí o zveřejňování vysokoškolských závěrečných prací.*
Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 Autorského zákona.
Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity o tom, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

V Brně dne 3. ledna 2017 .....

Abstract

Svozil, L. Mining knowledge from multilingual text data. Brno, 2017.

This paper focuses on the use of machine translation in solving the problems of classification and organization of multilingual text data. Both positive and negative effects of the translation are demonstrated on experiments using real world data. It was confirmed that thanks to the translation the English training set can be used for other languages.

Keywords

text mining, machine translation, machine learning, classification, SVM, C5

Abstrakt

Svozil, L. Dolování znalostí z vícejazyčných textových dat. Bakalářská práce. Brno, 2017.

Práce se zaměřuje na využití strojového překladu při řešení problémů klasifikace a organizace vícejazyčných textových dat. Pomocí experimentů na datech z reálného světa jsou prezentovány pozitivní i negativní dopady překladu. Bylo potvrzeno, že díky překladu je možno použít anglickou trénovací množinu i pro jiné jazyky.

Klíčová slova

dolování znalostí z textu, strojový překlad, strojové učení, klasifikace, SVM, C5

Obsah

1 Úvod a cíl práce

1.1 Úvod do problematiky

Svět se nachází hluboko v digitální éře. Během posledních desítek let byly digitalizovány filmy, přenos televizního signálu, telefonické hovory a všeobecně většina komunikace a interakcí je nyní vedena on-line a často archivována. Data nyní netvoří jen lidé, ale i různá zařízení přenášející data a informace z reálného světa do digitálního vesmíru - kamery, různé senzory, chytré telefony. Od roku 2005 roste objem dat a informací, které lidstvo při své každodenní činnosti vytvoří či reprodukuje, zhruba o 40% každý rok (IDC, 2014).

S růstem objemu dat jde ruku v ruce i zvýšená poptávka po metodách, jež dokážou s takovým množstvím dat pracovat. Například zvýšené nároky při vyhledávání řeší disciplína získávání informací (Information retrieval) - v praxi reprezentována nespočetným množstvím internetových vyhledávačů. Dále rostou nároky na organizaci, roztřídění a často je třeba takové množství dat zpracovat nebo z nich vyvodit nějaké závěry, získat z nich znalosti užitečné pro nějakou konkrétní činnost. Tím se zabývají disciplíny *Data mining*, nebo pokud jde o textová data, *Text mining* (Weiss, Indurkhya, a Zhang, 2010).

Práce s textovými daty nabývá na důležitosti v oblasti marketingu, třeba při analýze zpětné vazby od zákazníků. Součástí *Business intelligence* systémů jsou znalostní báze na podporu rozhodování. Tyto báze je třeba nějak naplnit. Analýzou velkého korpusu textových dat na internetu lze získat určitou výhodu. Tyto data však mohou být vyjádřena v různých jazycích. Často je pro potřeby analýzy vhodnější mít data v jednom společném jazyce.

Tato práce může pomoci těm, kterým se za poslední léta hromadila data v datových skladech. Je prezentováno několik praktických příkladů, jak lze s vícejazyčnými daty zacházet při dolování znalostí z nich. Je prodiskutována možnost použití strojového překladu jako jednoduchého intuitivního řešení.

Tento problém není nijak nový. Otázkou dolování znalostí z vícejazyčných textů se zabývalo již mnoho prací. Shrnutí různých přístupů diskutuje Steinberger (2010). Původní naivní přístup přeložit data do jednoho jazyka a pak na ně aplikovat algoritmy pro dolování znalostí, jako by se jednalo o jeden jazyk, trpěl několika zásadními nedostatky. Tehdejší komerční překládací systémy produkovaly často nepřesný překlad a ten byl kvůli tomu náchylný k chybám. Statistický překlad trpěl nedostatkem kvalitních paralelních dat, takže nedosahoval dostatečné univerzálnosti, aby mohl být použit celoplošně.

Vznikala proto různá specializovaná řešení, například práce autorů Shi a kol. (2010). Ta se zaměřovala na využití jednojazyčné trénovací množiny pro klasifikaci dokumentů v různých jazycích. Navrhuje postup natrénování na jednom jazyce, překlad modelu a následné přizpůsobení modelu cílovému jazyku za pomoci nově získaných roztříděných dat.

Jiný přístup použili ve své práci Fortuna a Shawe-Taylor (2005), kteří využili strojový překlad k vytvoření dvojjazyčného paralelního korpusu, podle kterého pak převáděli dokumenty na jednotnou vektorovou reprezentaci.

Doba a technologie mezi tím pokročily a strojovému překladu bylo v uplynulé dekádě věnováno spoustu pozornosti. Díky různým projektům a novým kvalitním vícejazyčným datům neustále se rozšiřující Evropské Unie zaznamenává statistický překlad velké pokroky. Tato práce byla napsána za účelem oprášení toho nejjednoduššího a nejintuitivnějšího přístupu, tedy přeložit všechna data do jednoho univerzálního jazyka, v tomto případě do angličtiny.

1.2 Cíl práce

Tato práce si klade za cíl obohatit oblast dolování znalostí z dat přezkoumáním vlivu strojového překladu na hlavní problémy dolování znalostí z textu - klasifikaci a shlukování - a to v kontextu jazykově závislých i nezávislých metod předzpracování, jako je stematizace nebo výběr rysů. Dalším cílem je pomocí různých experimentů vyhodnotit, zdali je dnes strojový překlad na takové úrovni, aby mohl být bez další složité algoritmizace použit při dolování znalostí i mezi různými jazyky.

2 Současný stav

Dolování znalostí je komplexní disciplína, čerpající metody a *know-how* ze spoustu jiných vědních oborů. Navíc v průběhu posledních desítek let zaznamenala bouřlivý rozvoj. Proto dříve, než bude možno provést jakýkoliv experiment, je třeba seznámit se s klasickými i nejnovějšími poznatky v tomto odvětví. Tento výčet si rozhodně neklade za cíl být definitivním, spíše se zaměřuje na znalosti nezbytné k provedení pokusů, které by měly prověřit dopad automatického strojového překladu.

2.1 Data mining

Data mining je definován jako proces objevování vzorů v datech. Tento proces by měl být automatický, nebo ve většině případech alespoň poloautomatický. Objevené vzory musí být smysluplné, jejich znalost by měla poskytovat nějakou výhodu, ať už ekonomickou, nebo jinou. Dalším rysem je fakt, že zdrojová data jsou přítomna ve velkém objemu.

Užitečné vzory umožňují dělat předpovědi na nových datech. Mohou být vyjádřeny třeba jako vektory či pravděpodobnostní modely - pak jsou jako černá skříňka, do které nelze nahlédnout a jejíž obsah nelze jednoduše interpretovat. Na druhé straně mohou být vyjádřeny jako soustava pravidel nebo rozhodovacích stromů - ty jsou pro člověka čitelnější a na jejich základě lze vést diskuze, jelikož znalost jimi vyjádřená je na první pohled patrná, popřípadě graficky znázornitelná (Witten a Frank, 2005).

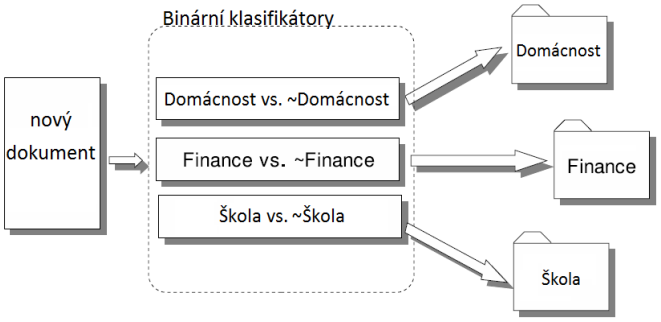
2.2 Text mining

Text mining lze obecně definovat jako proces založený na znalostech, ve kterém uživatel pracuje s kolekcí dokumentů za pomocí různých analytických nástrojů. Stejně jako při data miningu, v text miningu je cílem extrahovat užitečné informace ze zdrojových dat skrze rozpoznání a prozkoumání zajímavých vzorů. Rozdílem je ale strukturovanost dat. Nástroje data miningu očekávají data ve strukturovaném formátu a většina pozornosti při předzpracování je věnována normalizaci hodnot. Oproti tomu proces předzpracování při text miningu se zaměřuje především na identifikaci a extrakci klíčových rysů v dokumentech, které jsou psané přirozeným jazykem. Nestrukturovaná data je třeba převést do formátu, který je pro analytické algoritmy, často velmi blízké algoritmům z oblasti data miningu, srozumitelný (Feldman a Sanger, 2007).

Zatímco při dolování informací z dat je primárním vstupem tabulka s daty, víceméně kompletní, tak při dolování znalostí z textu je vstupem kolekce dokumentů. Obecně jakékoliv seskupení textových dokumentů můžeme nazvat kolekcí. Text-mining pak pracuje s kolekcemi o četnosti tisíců až desítek miliónů dokumentů. Čím větší kolekce, tím větší nároky na předzpracování (Weiss, Indurkhya, a Zhang, 2010).

2.2.1 Klasifikace

Je nazývaná také kategorizace a jedná se o jeden z typických problémů řešených v tomto odvětví. Algoritmus, který ji implementuje, se nazývá klasifikátor. V matematickém pojetí je klasifikátor funkce, která zobrazuje vstupní data na předem stanovené třídy. Jak ukazuje obrázek 1, složitější kategorizace do více tříd je často řešena pomocí soustavy binárních klasifikátorů, kdy se u nového dokumentu postupně vypočítá míra příslušnosti do každé z tříd a podle toho se pak zařadí.



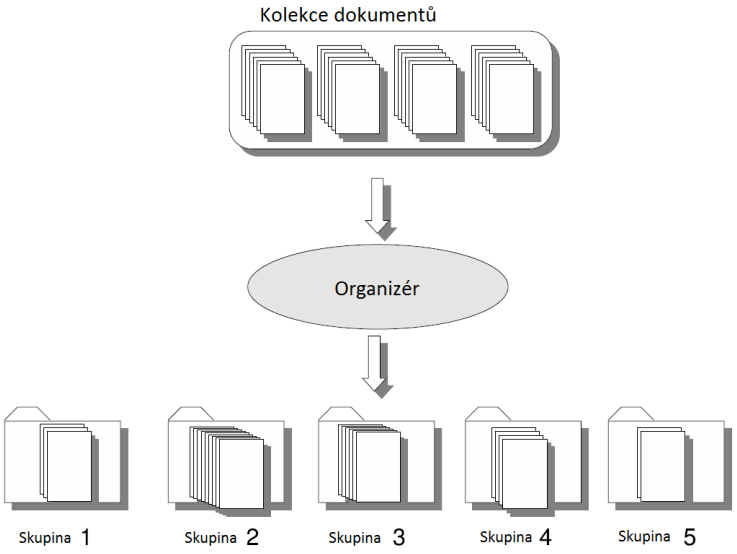
Obrázek 1: Kategorizace (Weiss, Indurkhya, a Zhang, 2010).

2.2.2 Organizace

Je podobná klasifikaci, jenže často předem není známo, do jakých tříd se rozděluje. Nemusí být známa pravidla, na základě kterých by mohlo rozdělení do tříd proběhnout, ani nemusí být k dispozici trénovací množina už roztříděných dokumentů pro použití algoritmů strojového učení. Roztřídění provede shlukovací algoritmus, a to na základě podobnosti dokumentů, znázorněno obrázkem č. 2.

2.3 Strojové učení

Zjednodušeně řečeno, cílem strojové učení je donutit počítače změnit či přizpůsobit své akce na základě nějaké předchozí zkušenosti, a to tak, aby tyto akce byly přesnější, vhodnější vzhledem k okolnostem. Taková definice ale vyvolává několik otázek. Jak stroj pozná, že jeho akce jsou vhodnější, že se zlepšuje? Na to existuje několik různých odpovědí. Můžeme mu ukázat nějakou ideální akci, jak by to mělo vypadat, a doufat, že mu jich stačí ukázat jen několik a že si získané znalosti dokáže nějak generalizovat a pak je použít. Nebo jen můžeme říci, zdali jeho akce byla správná či špatná, popřípadě udělit akci nějaké skóre a nechat jej, ať najde jiný přístup. Podle toho se dělí strojové učení do několika kategorií (Marsland, 2009).



Obrázek 2: Organizace dokumentů do skupin (Weiss, Indurkhya, a Zhang, 2010).

- **Učení s učitelem** - je poskytnuta trénovací množina se správnými odpověďmi a na jejím základě algoritmus provede generalizaci, aby odpověděl správně na všechny možné vstupy. Také se mu říká učení z příkladů.
- **Učení bez učitele** - správné odpovědi nejsou poskytnuty. Algoritmus se snaží najít podobnosti ve vstupních datech a podle toho je roztřídit.
- **Učení posilováním** - řadí se někde mezi učení s učitelem a bez učitele. Algoritmu je řečeno, zdali je jeho odpověď správná či špatná. V případě špatné odpovědi mu však není odhalena správná možnost, ale je ponechán, aby se sám napravil. Říká se mu také učení s kritikem, protože je pouze uděleno skóre, ale není navrhnuto žádné správné řešení.
- **Vývojové učení** - vzalo si inspiraci z biologické evoluce, která může být také chápána jako proces učení. Živé organismy se snaží přizpůsobit svému prostředí, aby zvýšily svoji schopnost přežít a mít potomky. Algoritmy však místo konceptu přežití používají *fitness* funkci, která jim poskytuje informaci o tom, jak se jejich aktuální řešení blíží ideálnímu.

Výpočetní náročnost metod strojového učení bude vždy jedním z klíčových aspektů, jelikož často je třeba tyto metody používat na velké sady dat. Často je rozdělována na dvě části. Na složitost při trénování, na kterou nejsou kladeny takové nároky, a na složitost při rozhodování, tedy při aplikaci natrénovaného algoritmu.

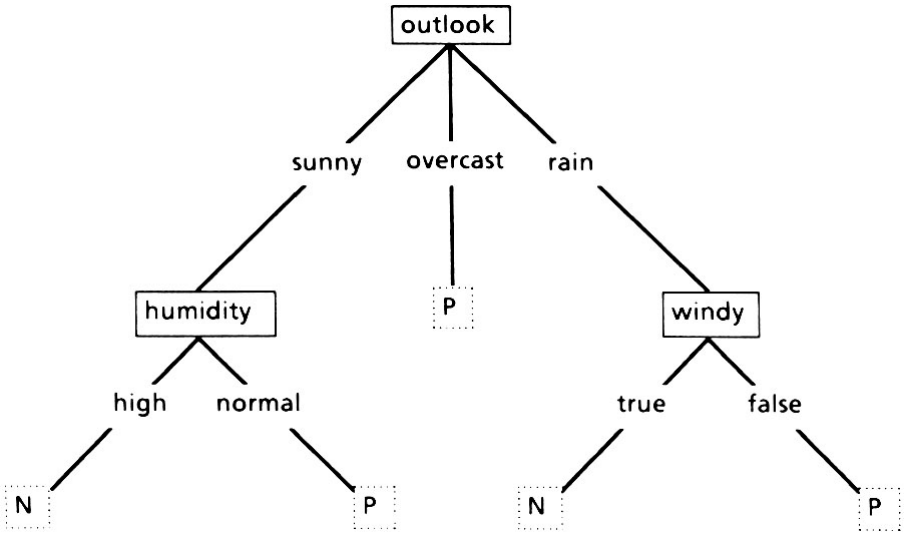
Další tři kapitoly budou pojednávat o typických algoritmech, které se při řešení problémů pomocí strojového učení běžně používají.

2.3.1 Rozhodovací stromy

Jedná se o univerzální a velmi rozšířenou metodu učení. Vytvořený model má podobu často binárního stromu, kdy rozhodovací proces začíná v kořeni a postupnými odpověďmi na otázky v uzlech se postupuje až ke konečnému rozhodnutí v listech. Jsou dobře škálovatelné a především jsou transparentní, což je jejich největší výhoda. Každé rozhodnutí je jasně zdůvodněné cestou ve stromu a menší modely jsou i velmi přehledné. Navíc algoritmy vytvářející modely ve formě rozhodovacích stromů provádí implicitně výběr rysů, takže z problému obsahujícího několik tisíc atributů vytvoří model, který je založen pouze na pár stovkách relevantních atributů (Hastie a col., 2009).

Mezi nejznámější implementace patří algoritmus *ID3* autora Rosse Quinlana. Ten postupuje tak, že v každém kroku vybere rys, jehož vybrání jako rozdělujícího prvku nejvíce sníží entropii celé množiny. Je známý fakt, že rozdělení způsobující největší úbytek entropie, poskytuje nejvíce informací (Shannon, 1948). Nejdůležitější rys je pak v kořeni stromu a postupně jsou tvořeny další uzly. Vylepšením je pak algoritmus C4.5, který navíc přidává ořezávání vzniklého stromu, aby zabránil přetrénování, a podporu pro kontinuální i diskrétní atributy (Marsland, 2009).

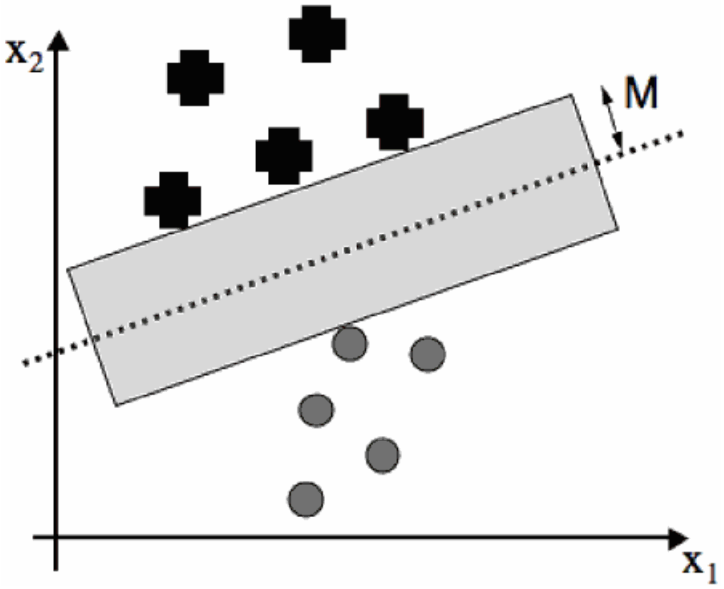
Jinou známou implementací je algoritmus *CART*. Ten funguje na podobném principu, akorát při výběru rysů používá jiné měřítko. Navíc produkuje čistě binární stromy (Marsland, 2009).



Obrázek 3: Kus rozhodovacího stromu, vytvořeného ze známého datasetu *Play tennis* (Quinlan, 1986)

2.3.2 Podpůrné vektory

Neboli *support vector machines*. Mezi hlavní vynálezce je řazen profesor Vladimir N. Vapnik. Tento algoritmus funguje na principu promítnutí rysů do vícedimenzionálního prostoru a jeho následného rozdělení nadrovinou. Body (rysy) z obou tříd nejblíže této nadrovině se pak nazývají podpůrné vektory a samy o sobě postačí k další klasifikaci nových bodů, neboť tuto rozdělující nadrovinu popisují.



Obrázek 4: Přímka rozdělující 2 dimenzionalní prostor. *M* je margin, neboli největší možná vzdálenost mezi body a rozdělující nadrovinou - v tomto případě přímkou (Marsland, 2009).

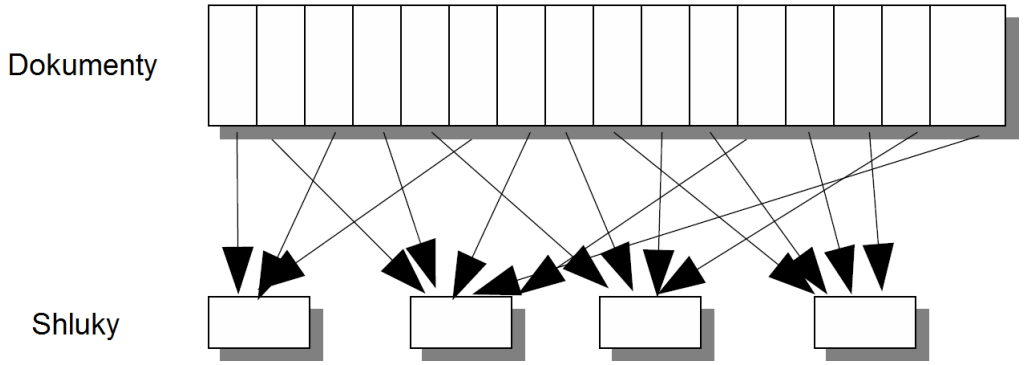
Na rozdíl od rozhodovacích stromů umí řešit i problémy, ve kterých nejsou data lineárně rozdělitelná. Umí provést mapování původních bodů do dalších rozměrů, a to tak, aby v původních dimenzích nebyly změněny, ale díky tomu jsou už lineárně rozdělitelná (Marsland, 2009).

2.3.3 Shlukování

Shlukování (také organizace) je metoda učení bez učitele. Jejím účelem je roztřídit data do předem neznámých skupin na základě podobnosti dokumentů. O nově vytvořených skupinách lze pak říci, že dokumenty v nich jsou si navzájem více podobné, než dokumenty vně skupiny. Je velmi vhodná pro analýzu velkého počtu dokumentů, o kterých je předem známo málo informací (Berry a Kogan, 2010).

Pokud každý dokument patří právě do jedné skupiny, jedná se o *pevné shlukování* (z anglického termínu *hard clustering*). Jiným přístupem je *měkké shlukování* (anglicky *soft* nebo *fuzzy clustering*), kdy jeden dokument může patřit do více skupin. Jeho míra příslušnosti ke každé ze skupin pak může být vyjádřena hodnotou v intervalu <0;1>.

Typickou implementací shlukování je *k-means* algoritmus. Ten na začátku náhodně rozdělí dokumenty do k shluků. Spočítá průměrný vektor pro každý shluk (dokumenty jsou vyjádřeny ve vektorové reprezentaci). Každý dokument pak porovná s průměrným vektorem všech shluků. Poté přesune všechny dokumenty, jejichž vektory byly blíže průměru jiného shluku než toho, ve kterém aktuálně jsou. To se opakuje tak dlouho, dokud nepřestane docházet k přesunům (Weiss, Indurkhya, a Zhang, 2010).



Obrázek 5: Schéma procesu shlukování (Weiss, Indurkhya, a Zhang, 2010).

2.4 Metody předzpracování při dolování znalostí z textu

Odhaduje se, že předzpracování může běžně zabrat okolo 80% výpočetního času a je mu tedy třeba věnovat velkou pozornost (Silva, Ribeiro, 2010).

V následujících podkapitolách budou popsány ty nejvýznamnější metody, které často bývají součástí procesu předzpracování.

2.4.1 Sjednocení velkých, malých písmen a diakritika

Základní jednoduchá úprava, obnášející převod všech písmen na jednu velikost, většinou na malá. Účelem je, aby dvě významově stejná slova nebyla zpracována jako dvě odlišná slova jen proto, že jedno má na začátku velké písmeno, zatímco to druhé ne. To však někdy může být i nežádoucí, například je-li třeba rozlišit vlastní jména. Anglický termín je *case folding*. Ze stejného důvodu také někdy bývají odstraněna diakritická znaménka a větná interpunkce (Berry a Kogan, 2010).

2.4.2 Lematizace a stematizace

Účelem lematizace a stematizace je snížit frekvenci výskytu odlišných slov standardizací různých slovních tvarů. Ať už jde o odstranění koncovek při skloňování a časování, nebo rovnou redukci slov na jejich kořeny, či dokonce převod synonym na stejný kořen, docílí se snížení počtu slov, a tedy dojde ke zjednodušení problému. To nabývá na významu zvláště u velkých úloh, jež jsou výpočetně náročné (Weiss, Indurkhya, a Zhang, 2010).

Jedná se o jazykově závislé úpravy, neboť v každém jazyce jsou pravidla ohýbání koncovek o něco jiná. Pro stematizaci je třeba pouze znát pravidla pro ohýbání slov, ale při lematizaci dochází k převodu na základní tvar slova a je třeba využít slovníku pro tyto účely.

Následuje několik příkladů originálních textů a následně textů, na něž byla aplikována stematizace.

> Špatně otočené postele vzhledem k televizní obrazovce a zvuk televize ztlumen tak, že není slyšet vůbec.
>
> spatn otocen postel vzhled k televizn obrazovk a zvuk televih ztlumen tak ze nen slyset vubec

> Grundsätzlich sind die Zimmer und Aufenthaltsräume sehr sauber, allerdings wurden nach dem Frühstüstück die Tische zwar abgeräumt aber nicht gesäubert.
>
> Grundsätzlich si di Zimm u Aufenthaltsräume sehr saub allerdings werden nach d Frühstüstück di Tisch zwar abgeräumen ab nichen säubern

> Very well maintained apartment. Lovingly teaching. The location is ideal for shopping in the city. Nice people.
>
> veri well maintain apart lovingli teach the locat i ideal for shop in the citi nice peopl

2.4.3 Výběr rysů

Neboli *feature selection*. Kolekce textových dokumentů mohou obsahovat obrovské množství slov. Většina z nich však nemusí být pro řešení daného problému, například kategorizace textu, nijak užitečná.

V každém jazyce existují seznamy slov, které pro účely dolování znalostí nemají velký význam. Jedná se často o spojky, částice, předložky nebo zájmena. Říká se jim *stopslova obecná*. Nicméně to, která slova mají nebo nemají rozlišovací potenciál, závisí na konkrétní úloze. Co může mít význam v rámci jednoho úkolu, nemusí už vůbec nic znamenat při řešení jiného úkolu. Například, je-li třeba vydolovat z textu informace ohledně polohy, nemůžeme z textu odstranit předložky jako „v, na“. Příslovce a přídavná jména jako „dobře, čistý nebo prostorný“ nebudou příliš

důležitá, ty se však v žádném obecném seznamu stopslov neobjevují. Při řešení úlohy roztřídění recenzí návštěvníků hotelu ale situace bude obrácená.

Tato přílišná obecnost se dá vyřešit sestavením vlastního seznamu slov, které pro daný úkol nejsou důležitá. Říká se jim *stopslova doménově specifická* a jsou pak vázaná na určitou doménu (oblast) jazyka, pro kterou byla vytvořena.

Některé systémy však provádí daleko agresivnější redukce, které odstraňují 90 až 99 procent rysů (slov). Aby mohlo dojít k odstranění těch nejméně užitečných rysů, je třeba definovat nějaké měřítko relevance. Jako efektivní se ukazuje počet dokumentů, ve kterých se daný rys vyskytuje (anglicky *Document Frequency*). Podle pokusů Feldmana a Sangera (2007) použití pouze 10% nejfrekventovanějších rysů nijak výrazně nesnižuje výkon klasifikátoru.

2.4.4 Reprezentace textu

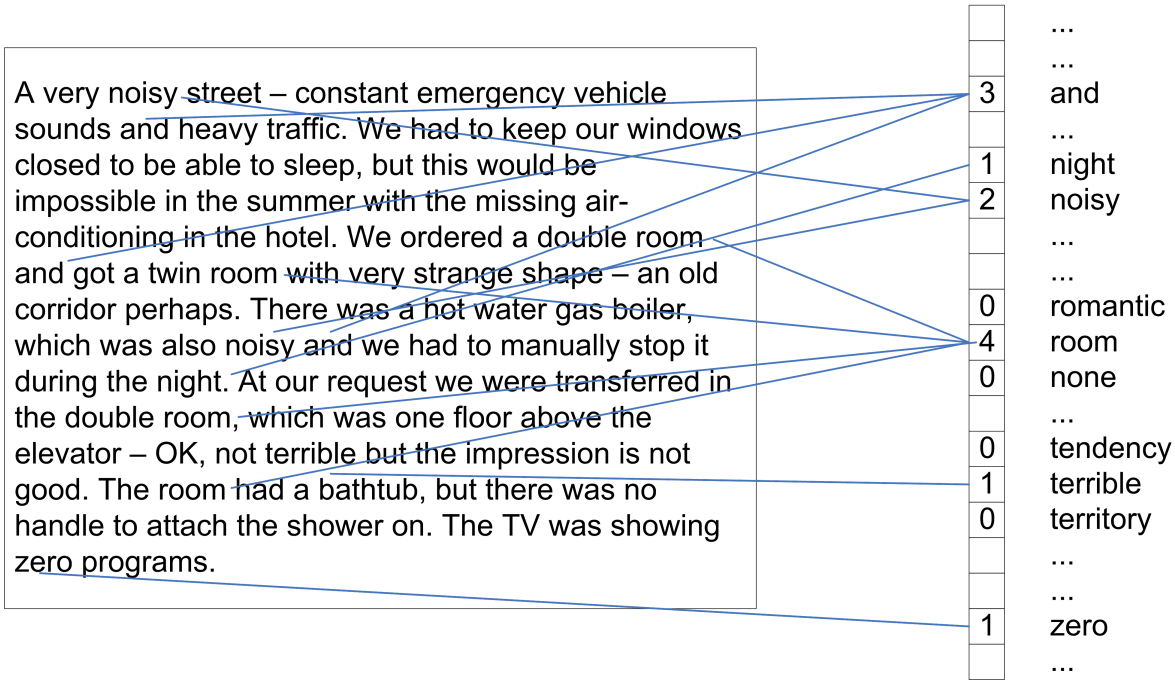
Textové dokumenty jsou nestrukturovaná data. Text, ve své podobě jako záznam přirozené lidské řeči, je pro algoritmy strojového učení, které byly vyvinuty hlavně na jasně strukturovaných úlohách data miningu, naprosto nesrozumitelný. Pro překonání tohoto problému je třeba zvolit vhodnou reprezentaci a kolekci dokumentů do ní převést. Zažitým klasickým postupem, ne však jediným, je převod do tzv. *bag-of-words* reprezentace. Jedná se o jiné pojmenování vektorového prostoru, který navrhoval již Salton (1975).

Každý rys (v nejjednodušším případě slovo) je reprezentován právě jednou dimenzí v prostoru. Počet dimenzí se tedy rovná celkovému počtu rysů. Důležitost rysů je pak vyjádřena vzdáleností od počátku. Dále každý dokument je reprezentován vektorem, který je určen vahami jednotlivých rysů. Váhy se skládají z lokální části, tedy důležitosti v rámci jednoho dokumentu a z globální části, která vyjadřuje důležitost rysu v rámci celé kolekce. (Silva, Ribeiro, 2010).

Jako lokální váha se často používá prezence termu (rysu), tedy hodnota 1 pro přítomný term v dokumentu a 0, když term v dokumentu zastoupen není. Další hojně využívanou lokální váhou je frekvence termu v dokumentu. Nejvyužívanější globální váhou je převrácená četnost slova ve všech dokumentech, neboli *Inverse document frequency*. Tato váha zvýhodňuje vzácné termy před běžnými. Čím více se totiž term vyskytuje v ostatních dokumentech, tím je jeho globální váha menší. Následující vzorce pak ukazují výpočet kombinované váhy TF-IDF a její globální složky IDF pro term j. DF(j) je počet termu j ve všech dokumentech a N celkový počet dokumentů (Weiss, Indurkhya, a Zhang, 2010).

$$TF - IDF(j) = TF(j) * IDF(j)$$

$$TF - IDF(j) = \log(\frac{N}{DF(j)})$$



Obrázek 6: Ukázka transformace textového dokumentu do vektorové reprezentace - počítání lokálních TF vah (Dařena, 2011)

2.5 Strojový překlad

Výzkum strojového překladu začal v 50. letech minulého století v USA. Hlavním převládajícím směrem byl tehdy překlad založený na pravidlech, kdy se lingvističtí experti snažili identifikovat pravidla a ty pak použít ve svých prvních překládacích systémech. Později přišel statistický překlad, jenž využíval některé metody strojového učení. K natrénování takového systému je třeba zarovnaný paralelní dvojjazyčný korpus, tj. kolekce dokumentů ve dvou jazycích, kde si pasáže v obou jazycích odpovídají a je tak možno z nich algoritmus „naučit“ překládat. Statistický překlad se postupem času stal převládajícím způsobem překladu, ačkoliv se nedá říci, že by dnes existovaly čistokrevné statistické nebo založené na pravidlech (anglicky rule-based) systémy. V praxi často jeden využívá prvky druhého (Hudík, 2015).

2.5.1 Jazykové korpusy

Jazykový korpus (latinsky corpus - tělo, těleso) je soubor mluvených či textových záznamů jazyka, sloužící lingvistickému výzkumu, tvorbě slovníků, korektorů nebo překladačů. Mohou být jednojazyčné nebo vícejazyčné. Terminologie dále rozlišuje (Chlumská, 2014):

- **Korpus paralelní** Obsahuje původní texty, čí mluvené slovo, v jazyce A a jejich překlady v jazyce B, popřípadě v dalších jazycích (korpus Europarl z roku 2012 obsahuje výstup evropského parlamentu, přeložený do 21 jazyků).

Může být jednosměrný, nebo obousměrný - ten obsahuje i překlady z jazyka B do A.

- **Korpus srovnatelný** Se skládá z částí, které byli vybrány na základě stejných kritérií. Vícejazyčný srovnatelný korpus neobsahuje překlady, ale původní texty napsané na stejné téma. Odpadá tedy vliv překladatele.

2.5.2 Různé přístupy k dolování znalostí z vícejazyčných zdrojů

V literatuře se objevuje několik různých možností, jak lze přistoupit ke klasifikaci dokumentů v různých jazycích. Je možno je rozdělit do tří kategorií:

- **Jazykově závislé:** Tyto přístupy spoléhají na využití charakteristik určitého jazyka. Jejich hlavní nevýhodou je nutnost individuálního přístupu ke každému jazyku. Například autoři (Montalvo a kol., 2006) založili svůj algoritmus na shlukovaní dokumentů ve více jazycích na práci s pojmenovanými entitami (osoba, organizace či místo). Pro každý jazyk však museli sestavit seznam těchto entit na základě pravidel určených lidskými experty.
- **Jazykově nezávislé:** Obnášejí převod dokumentů do reprezentace nezávislé na jazyce. Často jsou přitom využívány rozsáhlé slovníkové systémy jako *Babelnet* nebo *Eurovoc*(Romeo, Ienco a Tagarelli, 2015). Autoři (Gliozzo a Strapparava, 2005) zase pro změnu využili latentní sémantické analýzy (podobnost slov, na základě toho, v jakém okolí se vyskytují) na srovnatelném korpusu a nebylo potřeba žádných slovníků.
- **Využívající strojový překlad:** Do této kategorie spadají přístupy, které částečně nebo úplně překládají dokumenty do jednoho jazyka. Všechny dokumenty v kolekci například překládají autoři (Evans a Klavans, 2003) při sumarizaci novinových článků ve více jazycích.

3 Metodika

Pro vyhodnocení dopadu překladu na dolování znalostí pomocí algoritmů strojového učení je třeba navrhnout experimenty a jasně popsat, které metody předzpracování budou kdy a jak použity. Dále je třeba zvolit vhodná data. Tomu všemu se věnuje celá sekce Metodika.

3.1 Data

Jako testovací data byly použity recenze návštěvníků hotelu z webového rezervačního serveru „Booking.com“, který zajišťuje online rezervace více než 800 000 hotelů. Recenze má dvě části, pozitivní a negativní. To přímo předurčuje tuto kolekci jako materiál k testování vlivu různých technik předzpracování na klasifikaci. Recenze jsou navíc rozděleny podle země původu recenzenta, z čehož byl odvozen jazyk (Žižka a Dařena, 2011).

Jednotlivé recenze jsou uloženy tak, že na jednom řádku je vždy třída dokumentu, znak tabulátoru a text. Třída může být pozitivní nebo negativní a je značena buď písmeny P a N nebo číslicemi 1 a 2. Následuje několik příkladů.

P Odpovídá zařazení do své kategorie, vstřícný přístup zaměstnanců, dobré umístění
N V koupelně neupevněné držáky na ručník, podložky na mýdlo atp. nedaly se na odkládání použít.

1 Zimmer zur Strasse mit den Restaurants abends sehr lange laut
2 Auch für Fahrradtouristen sehr gut geeignet und äusserst freundliche Aufnahme.

1 La ropa de cama y baño debería estar incluida en el precio.
2 La ubicacion ylimpieza pero en general esta bien

1 Brak udogodnień dla małych dziecie, np. placu zabaw dla dzieci
2 Miła obsługa, wyśmienite śniadanie, spokój i blisko do centrum miasta

3.2 Zvolené algoritmy strojového učení

Různých algoritmů strojového učení je velké množství, ale tato práce si neklade za cíl je všechny porovnávat. Je zaměřena na metody předzpracování, především strojový překlad. Je nutno zdůraznit, že závěry, které byly vyvozeny z výstupů těchto algoritmů, se vážou právě na ně, a při použití jiných algoritmů strojového učení by výsledky mohli být jiné.

Pro klasifikaci byli zvoleny dva odlišné algoritmy. Prvním z nich je implementace podpůrných vektorů *SVMlight*. Byl zvolen pro svoji rychlost, přesnost a jednoduchost (myšleno jednoduchost výstupu, nikoliv principu fungování).

Jako druhý byl zvolen algoritmus *C5.0*, tedy vylepšení *ID3* implementace rozhodovacích stromů. Důvodem je snadná čitelnost vytvořených modelů (stromů). Z těch je pak patrné, které rysy jsou pro klasifikaci významné.

Na organizaci pomocí shlukování byl vybrán *Cluto*, protože měl v literatuře spoustu referencí a zároveň byly k dispozici výsledky práce z našeho ústavu, která

se zabývala nejideálnější volbou kriteriální funkce na podobných datech (Dařena, Žižka a Přichystal, 2014).

3.3 Použité metody předzpracování

Automatizace probíhala pouze pomocí skriptů psaných v Perlu. Na data byly aplikovány jazykově závislé i nezávislé způsoby předzpracování. Ty jsou v této kapitole popsány z praktického hlediska.

3.3.1 Strojový překlad

Část dat byla přeložena pomocí Google translate, nebo Power translatoru. Tito zástupci představují dva rozdílné způsoby překladu. První zmíněný je online služba běžící na serverovém cloudu opírající se o spoustu dat k tak zvanému statistickému překladu. K samotnému překladu byl využit skript, který posílá dokumenty přímo v URL adrese jako jeden z parametrů (metoda GET). Doba překladu 50 000 dokumentů trvala zhruba 5 hodin. Tento postup je však poněkud nestandardní a byl použit univerzitní proxy server. Autor však předpokládá, že v následujících letech dojde ke zvýšení dostupnosti bezplatného statistického překladu (ModernMT, 2015), ale prozatím bylo využito překladu jednoho z uznávaných komerčních poskytovatelů.

Druhý zmíněný využívá překladu založeném na pravidlech, ke svému fungování potřebuje jen osobní počítač a obsahuje slovníky umožňující offline překlad. Navíc je vybaven nástrojem pro překlad celých souborů a tímto posloužil pro tento experiment. Doba překladu 50 000 dokumentů trvala 20-40 minut, v závislosti na jazyce.

3.3.2 Převedení slov na jejich kořen (stematizace)

K převedení bylo využito Perl modulů právě pro tento účel, popsaný v kapitole 2.4.2. Lze je najít pod označením Lingua::Stem[1] a Lingua::Stem::UniNE[2]. Operace je časově nenáročná, zpracování 50 000 dokumentů se pohybuje v řádu sekund.

3.3.3 Redukce počtu slov pomocí výběru rysů

Vzhledem k vysokému počtu a náročnosti testů, které bylo třeba provést, bylo nutno značně redukovat velikost prostoru.

V praxi to znamená odstranit slova, která nemají k jednotlivým třídám žádný vztah. Jako metoda byla zvolena *The Chí Statistic*, a to na základě výsledků práce věnující se tomuto tématu (Krupník, 2014). Následující vzorec ukazuje výpočet hodnoty míry závislosti mezi termem t a třídou c (Li et. al., 2009).

[1] `http://search.cpan.org/~snowhare/Lingua-Stem/lib/Lingua/Stem.pod`

[2] `http://search.cpan.org/~patch/Lingua-Stem-UniNE-0.08/lib/Lingua/Stem/UniNE.pm`

$$CHI(t_i; c_k) = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

A – počet dokumentů v kategorii c obsahující term c.
B – počet dokumentů, které nejsou v kategorii c a obsahují term c.
C – počet dokumentů v kategorii c neobsahující term c.
D – počet dokumentů, které nejsou v kategorii c a neobsahují term c.
N – počet všech dokumentů.

Slova v trénovacích množinách byla ohodnocena touto metodou, byla seřazena od největšího po nejmenší a 75% slov[3] ze spodní části, tedy těch nejméně relevantních, bylo odebráno. Totožný seznam slov pak byl odebrán i z testovacích množin.

Dřívější experimentální práce ukázaly, že při využití podobné metody, lze snížit počet slov až o 99% bez vážnějšího vlivu na výkon klasifikátoru (Yang a Pedersen, 1997).

3.3.4 Konverze kolekce do bag-of-words reprezentace

Bylo využito nástroje[4] vzniklého na ústavu informatiky (Žižka a Dařena, 2010). Ještě před převodem byla z kolekce odstraněna slova, jejichž výskyt ve všech dokumentech nedosahoval počtu 2 a slova, která měla méně než 2 znaky, protože se často jedná o překlepy, které zvyšují výpočetní náročnost neúměrně tomu, kolik informací poskytují. Během převodu byla všechna velká písmena nahrazena malými (case folding). Vzhledem k tomu, že klasifikátor pro podpůrné vektory funguje na čistě matematické bázi, třídy dokumentu bylo třeba konvertovat na číselně vyjádřený cíl. Tedy negativní recenze byly ohodnoceny cílovou hodnotou *-1* a ty pozitivní *1*.

3.4 Vyhodnocení

Aby bylo možné odhadnout přesnost modelu v případě, kdy jej bude třeba použít na neroztřízených datech, je vhodné sestavit nějaký odhad. Prvním náznakem může být úspěšnost modelu na trénovacích datech, tedy informace, kterou algoritmy většinou ochotně na konci trénovacího procesu poskytují. Nicméně tato informace spíš poukazuje na složitost dat a množství šumu v nich. Na to, jak si povede model v případě nových neznámých dat, je nutno použít jiné metody. Některé z nich jsou uvedeny v této kapitole a jedná se o metody přímé, které rozdělují data na trénovací a testovací množiny, ze kterých lze tedy přímo určit úspěšnost klasifikátoru.

[3] Experimentálně bylo zjištěno, že v tomto případě má taková redukce zanedbatelný vliv na klasifikaci. Odstranění takového množství rysů však znamenalo nemalou úsporu výpočetního času, a tak bylo možno provést více testů.

[4] `https://akela.mendelu.cz/~darena/TextMining/`

3.4.1 Hold out

Klasickým přístupem je tak zvaná *hold out* evaluace. Tento anglický termín se dá přeložit jako „vydržet“. Znamená to, že část označených dat, u kterých je známo do které třídy patří, se ponechá stranou, čímž se vytvoří testovací množina. Tato množina se nijak neúčastní na tvorbě modelu. Až je model vytvořen, jeho přesnost se může vyhodnotit právě díky těmto zadrženým datům (Weiss, Indurkhya, a Zhang, 2010).

3.4.2 *k*-fold cross validation

Dalším jednoduchým a velmi často používaným přístupem je křížová validace. Jedná se o přesnější odhad, avšak výpočetně náročnější. Data se rozdělí na k částí, přičemž k-*1* částí se účastní tvorby modelu a na zbylé jedné části se ověří úspěšnost modelu. Tento proces se opakuje k-krát, dokud se všechny množiny nevystřídají, jak znázorňuje obrázek č. 7.

Výsledky jednotlivých validací se zprůměrují. Je třeba dát si pozor na to, aby počet množin k nebyl příliš malý, testovací část dat by pak zbytečně ubírala data na trénování a výsledný odhad by mohl být zkreslený. Opačným extrémním případem je, když se za k zvolí celkový počet příkladů a jen jeden se neúčastní tvorby modelu. Takový přístup však může být velmi výpočetně náročný - zvláště v případech, kdy je třeba znovu sestavit celý model. Literatura jako kompromis doporučuje rozdělení na 5 nebo 10 množin (Hastie a col., 2009).

Obrázek 7: Křížová validace (Hastie a col., 2009)

3.5 Popis experimentů vyhodnocující dopad překladu v rámci jednoho jazyka

Oba experimenty provedené v rámci této kapitoly mají podobnou strukturu. Struktura byla zvolena tak, aby bylo možno vyhodnotit účinky jednotlivých metod předzpracování mezi sebou navzájem i jako celku. Toho bylo docíleno rozdvojením množin v bodě, kdy se aplikovala nějaká úprava, jejíž dopad byl zkoumán. Na obrázku 8 je to graficky znázorněno. Po prvních dvou úpravách tedy z jedné původní kolekce jsou kolekce 4. Následuje jejich výčet.

- Původní dokumenty, nijak nepozměněné
- Původní dokumenty, na které byla aplikována stematizace
- Přeložené dokumenty (do anglického jazyka)
- Dokumenty přeložené a aplikována stematizace

Každá z těchto kolekcí byla v dalším kroku rozdělena na trénovací a testovací množiny. Z důvodu vyšší paměťové i časové náročnosti při sestavování rozhodovacího stromu byla každá kolekce rozdělena na 4 stejné části a teprve tyto menší části byly rozděleny v poměru 3:1 na hold out evaluaci. U jazyků, kde původní kolekce čítaly 100 000 dokumentů[5], tedy vznikly 4 trénovací množiny o velikosti kolem 18750 a 4 testovací množiny čítajíc průměrně 6250 dokumentů. Přičemž platí:

$$\bigcup_{1}^{4} trenovaci_i \cap \bigcup_{1}^{4} testovaci_i \in \emptyset$$

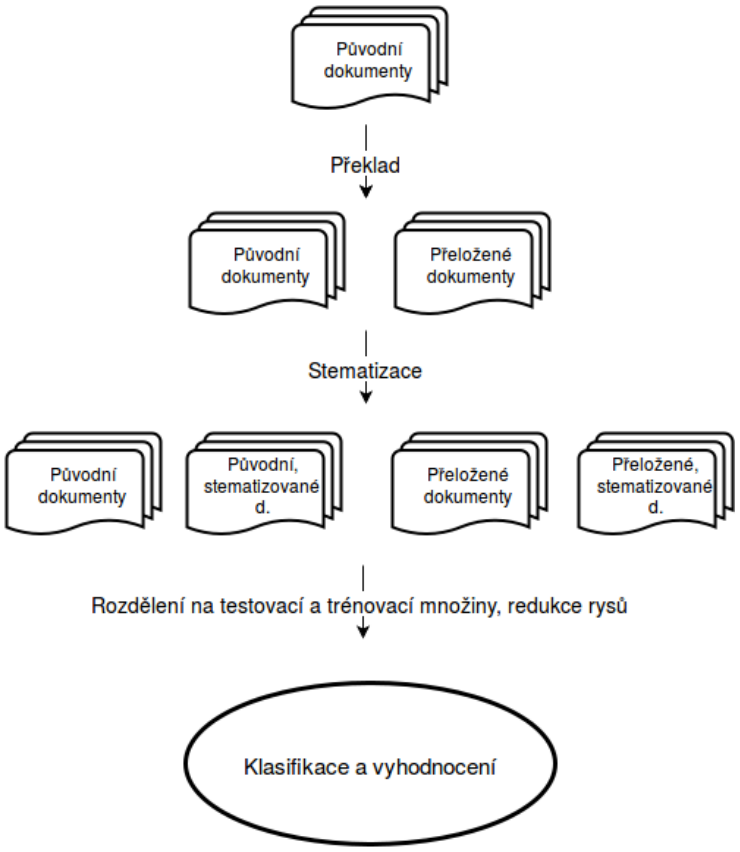
O něco jiné rozdělení probíhalo v případě testů s klasifikátorem fungujícím na principu podpůrných vektorů, který ke své činnosti zdaleka nepotřebuje tolik prostředků jako dříve zmíněné sestavování rozhodovacích stromů. Každá kolekce byla rozdělena na 5 částí. Z těchto 5 částí byly 5krát sestaveny trénovací a testovací množiny (poměr 4:1), a to tak, že každá z 5 části se stala množinou testovací právě jednou. Toto rozdělení odpovídá křížové validaci popsané v kapitole 3.4.2. U většiny jazyků, kromě češtiny, vzniklo 5 trénovacích množin o velikosti 80 000 dokumentů a 5 testovacích o velikosti 20 000.

Následně byly všechny množiny převedeny do vektorové reprezentace a z trénovacích množin byly sestaveny modely, jejich úspěšnost ověřena na testovacích množinách a výsledky byly zapsány do *csv* souboru pro vyhodnocení pomocí tabulkového procesoru.

3.5.1 Vliv strojového překladu na úspěšnost klasifikace a velikost slovníku u jednotlivých jazyků

Textová data v různých jazycích by bylo velmi výhodné skladovat v jednom společném jazyce, kterému rozumí nejvíce lidí. Kvalitou strojového překladu a jeho

[5]Jiný počet dokumentů v původní kolekci měl akorát jazyk český o velikosti 17 000



Obrázek 8: Struktura pokusů

srozumitelností se již zabývalo spoustu jiných prací. Někdy je však třeba tyto data použít dále k podpoře rozhodování, nebo jinak zpracovat pomocí algoritmů strojového učení. V takovém případě je třeba identifikovat, jaký dopad bude mít strojový překlad vybraných jazyků do angličtiny například na problém klasifikace. To se snaží adresovat následující pokus.

Kvůli časové náročnosti překladu a tvorby rozhodovacího stromu bylo náhodně vybráno 100 000 dokumentů v jazyce německém, francouzském, španělském, polském. Dále 17 000 dokumentů (více nebylo ve zdrojových datech k dispozici) v jazyce českém nebo slovenském. Ty byly přeloženy pomocí strojového překladače [6]. Dále byla na množinu přeložených i nepřeložených dokumentů aplikována jazykově závislá úprava stematizace, pomocí nástrojů uvedených v kapitole 3.3.2. Kvůli velkému počtu testů, které bylo třeba provést, bylo odstraněno 75% slov metodou uvedenou v kapitole 3.3.3. Jak dokázala jedna experimentální práce (Yang a Pedersen, 1997), podobnou metodou lze odstranit až 99% všech slov, bez vážnějšího vlivu na přesnost klasifikace. To dokazuje i složení rozhodovacího stromu ve výsledcích, který obsahuje jen zlomek všech slov. Následně došlo k převedení všech množin dokumentů do vektorové reprezentace (kapitola 3.3.4).

[6] `https://translate.google.com/`

Dále proběhlo rozdělení dokumentů do 4 množin po 25 000 (kvůli paměťové náročnosti zpracování) a tyto byly dále rozděleny v poměru 3:1 pro hold-out evaluaci. Z trénovacích množin byl vytvořen rozhodovací strom pomocí algoritmu C5.0 a na testovacích množinách byla ověřena jeho úspěšnost.

Pro vyhodnocení klasifikace pomocí podpůrných vektorů bylo využito 5-fold křížové validace, která je popsána v kapitole 3.4.2.

3.5.2 Porovnání rule-based a statistického překladu

Aby bylo možno vyhodnotit přínos statistického překladu, je třeba jej porovnat s klasičtějším způsobem překladu založeném na pravidlech. K samotnému sestavení modelu klasifikátoru je zapotřebí jen zlomek slov, které nesou informaci, díky níž je možno dokument přiřadit do jedné z tříd. Z toho vyplývá předpoklad, že překlad, který převádí slova z jednoho jazyka do druhého na základě předem sestavených a jasných pravidel bude pro účely klasifikace vhodnější.

Aby bylo možné tohle tvrzení vyvrátit nebo potvrdit, je třeba provést praktický experiment na skutečných datech z reálného světa. Z dat popsaných v úvodu metodiky bylo náhodně vybráno 100 000 dokumentů v jazyce Německém, Španělském a Francouzském. Původně měly být zařazeny i dokumenty Ruského a Polského jazyka, překlad v Power Translatoru však neustále končil předčasně chybou[7], což pro tento produkt rozhodně není dobrou vizitkou. Tyto 3 jazyky byly kromě Power translatoru přeloženy také pomocí online rozhraní překladače od společnosti Google, tedy zástupcem statistického překladu. Společně s originálními dokumenty tedy tvořily 3 množiny. Na ty byly aplikovány jazykové závislé úpravy jako stematizace, i jazykové nezávislé jako automatické odstranění stopslov a převedení do vektorové reprezentace.

Pro účely klasifikace pomocí rozhodovacího stromu C5.0 byly rozděleny po 25 tisících do 4 množin a ty byly dále rozděleny na trénovací a testovací množiny v poměru 3:1 - bylo tedy využito hold-out ohodnocení vzniklého klasifikátoru.

V případě klasifikace podpůrnými vektory bylo opět využito křížové validace.

3.6 Popis experimentu klasifikace dokumentů v různých jazycích na základě anglické trénovací množiny

Často se stane, že je k dispozici trénovací množina pouze v jednom jazyce. Následující experiment má ověřit, zdali je možno model vytvořený na základě této množiny použít i pro klasifikaci jiných jazyků.

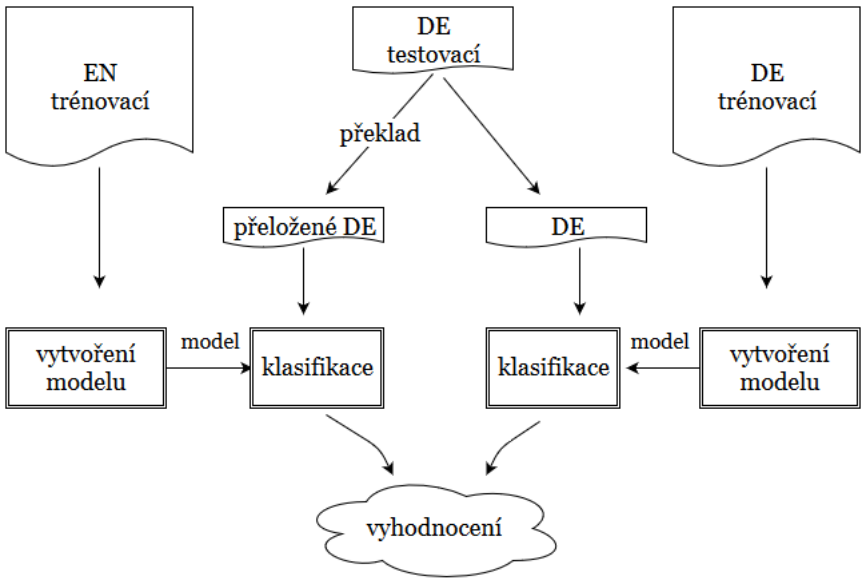
Náhodně bylo vybráno 90 000 dokumentů v anglickém jazyce. Tato množina sloužila jako trénovací k sestavení modelu pomocí SVMlight. Následně byly náhodně vybrány testovací množiny o velikosti 10 000 dokumentů v jazyce německém, francouzském, španělském a polském, od každého jazyka po třech množinách. Tyto testovací množiny byly přeloženy pomocí Google Translate do angličtiny a převedeny

[7] Autor se domnívá, že šlo o nedokonalou implementaci práce s různými kódováními.

do vektorové reprezentace tak, aby byly kompatibilní s modelem vytvořeným na základě trénovací množiny původních anglických dokumentů (pořadí atributů musí být stejné). Následně byla provedena klasifikace podpůrnými vektory.

Analogicky, aby bylo možno výsledky porovnat oproti nějaké základní hodnotě, tak k testovacím množinám v původních jazycích o mocnosti 10 000 tisíc dokumentů byl vytvořen model ze zbývajících 90 000 dokumentů téhož jazyka.

Obrázek 9 zachycuje strukturu tohoto pokusu pro jednu množinu jednoho jazyka. V každém jazyce byly postupně vybrány 3 testovací množiny.



Obrázek 9: Struktura pokusu 3.6

3.7 Popis experimentu organizace dokumentů pomocí shlukování

Shlukování může posloužit jako analytická metoda při prozkoumávání struktury kolekce. V tomto případě však bude sledována průměrná čistota a entropie vzniklých shluků.

Na přeložené i nepřeložené dokumenty byla aplikována stematizace a následně byly převedeny do vektorové reprezentace. Pomocí programu *Cluto* byla vytvořena řešení se 2,10,20,50 a 100 shluky. Na základě výsledků jiných prací byla jako kriteriální funkce zvolena H2.

4 Výsledky a diskuze

4.1 Vliv strojového překladu na úspěšnost klasifikace a velikost slovníku u jednotlivých jazyků

K vytvoření podkladů pro tento bod bylo provedeno 68 testů [8] pomocí rozhodovacího stromu a 100 testů[9] klasifikace podpůrnými vektory na původních 417 000 dokumentech v pěti jazycích.

Co se týká výsledků klasifikace rozhodovacím stromem, tak nejpozorovatelnější vliv měl překlad na německé dokumenty. Výsledky zachycuje tabulka 1. Došlo k mohutné redukci počtu slov ve slovníku a přitom ke snížení klasifikační chyby. Tento efekt je nejspíš způsoben specifikem německé gramatiky - odvozováním a skládáním slov.

Snížení počtu unikátních slov můžeme pozorovat u všech testovaných jazyků. Kromě němčiny je tento jev výrazný také u slovanských jazyků - češtiny a polštiny, jak ukazuje obrázek 10.

Nejméně příznivý vliv měl překlad na dokumenty francouzské, jak lze vidět v tabulce 2. Zmenšení slovníku bylo minimální, za to došlo ke zvýšení klasifikační chyby.

Tabulka 1: Němčina - klasifikace rozhodovacím stromem.

	prům. počet atributů	průměrná úspěšnost	prům. velikost stromu
původní	**8221**	**87,76**	**226**
bez stematizace	8791	87,53	233
se stematizací	7650	88,00	218
přeložené	**2906**	**88,38**	**238**
bez stematizace	3852	88,13	234
se stematizací	1960	88,63	241

Že tento jev je pro klasifikaci na bázi rozhodovacího stromu pozitivní, ukazují výsledky v grafu na obrázku 11. Tyto výsledky se však váží k tomuto konkrétnímu algoritmu (C5.0).

[8] (1 původní množina + 1 přeložená) * 2 aplikace stematizace * 4 rozdělení na menší množiny kvůli výkonu * 4 jazyky + 4 za češtinu (ta nebyla třeba rozdělovat na menší množiny)

[9] (1 původní množina + 1 přeložená) * 2 aplikace stematizace * 5 křížová validace * 5 jazyků

Tabulka 2: Francouzština - klasifikace rozhodovacím stromem.

	prům. počet atributů	průměrná úspěšnost	prům. velikost stromu
původní	**2919**	**87,71**	**220**
bez stematizace	3912	86,98	214
se stematizací	1926	88,45	225
přeložené	**2707**	**86,81**	**249**
bez stematizace	3531	86,18	246
se stematizací	1883	87,45	251

Tabulka 3: Španělština - klasifikace rozhodovacím stromem.

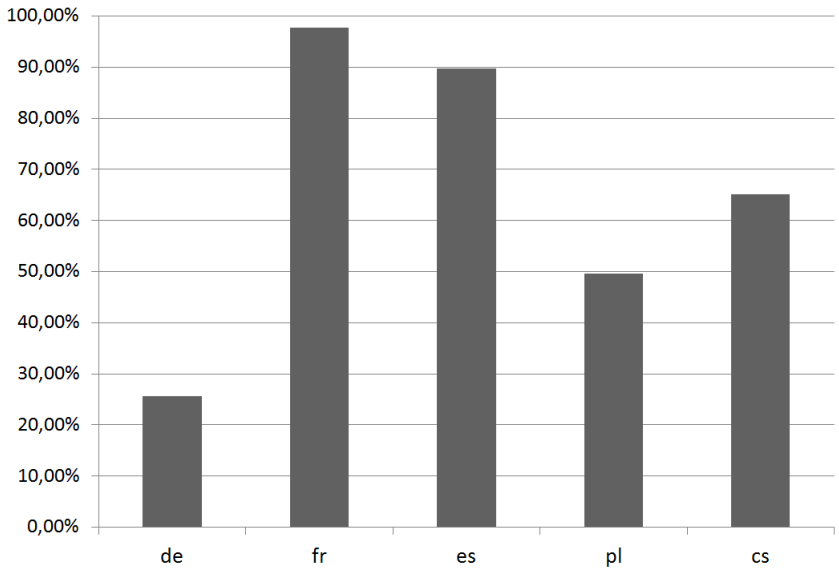
	prům. počet atributů	průměrná úspěšnost	prům. velikost stromu
původní	**3119**	**86,65**	**229**
bez stematizace	4287	85,77	247
se stematizací	1950	87,52	211
přeložené	**2527**	**86,76**	**239**
bez stematizace	3304	86,45	243
se stematizací	1750	87,07	234

Tabulka 4: Polština - klasifikace rozhodovacím stromem.

	prům. počet atributů	průměrná úspěšnost	prům. velikost stromu
původní	**6212**	**88,20**	**230**
bez stematizace	7947	87,75	222
se stematizací	4477	88,65	238
přeložené	**3168**	**89,20**	**235**
bez stematizace	4119	89,00	237
se stematizací	2218	89,40	234

Tabulka 5: Čeština - klasifikace rozhodovacím stromem.

	prům. počet atributů	průměrná úspěšnost	prům. velikost stromu
původní	**3620**	**85,00**	**180**
bez stematizace	5002	82,50	182
se stematizací	2237	87,50	178
přeložené	**2106**	**86,55**	**157**
bez stematizace	2754	86,00	153
se stematizací	1458	87,10	161

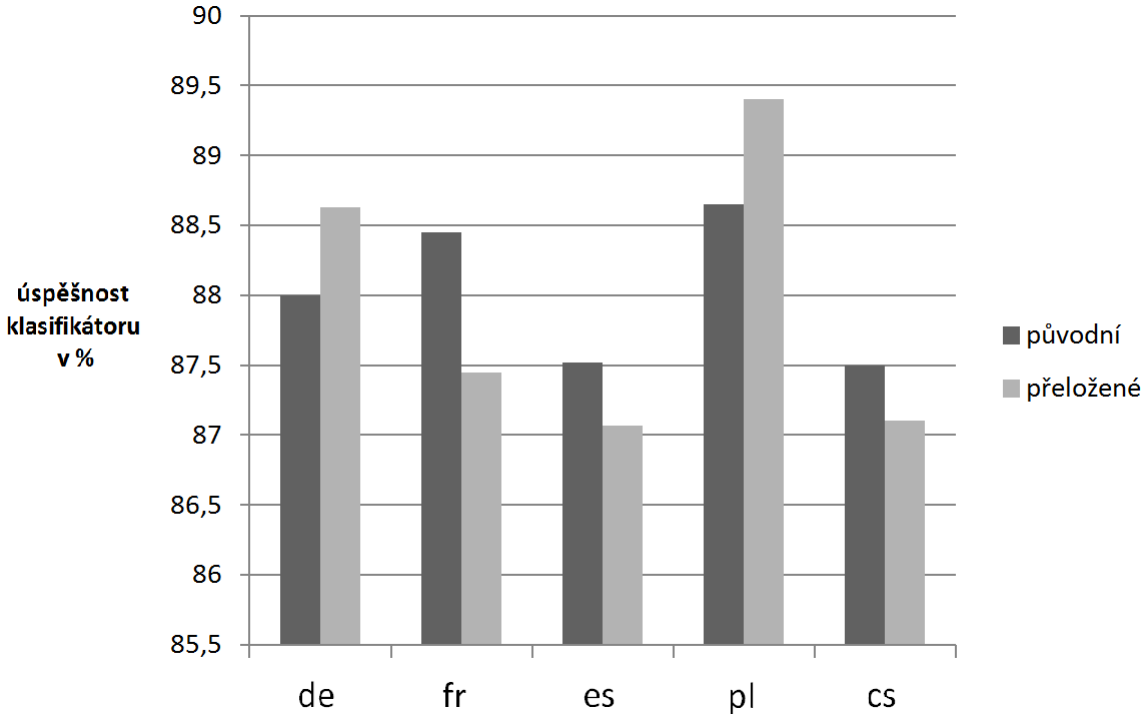


Obrázek 10: Počet unikátních slov v přeložených dokumentech, vyjádřený v % ku počtu unikátních slov v původních dokumentech (stematizace byla aplikována)

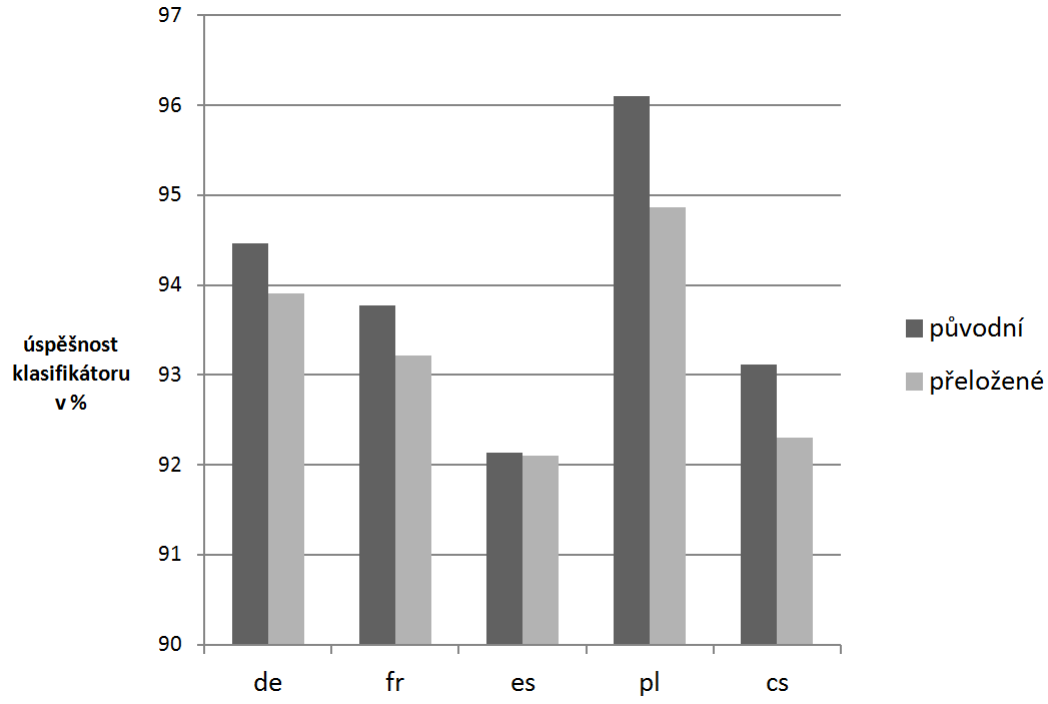
U podpůrných vektorů je situace odlišná. Generalizační schopnosti klasifikátoru na principu podpůrných vektorů totiž nejsou vysokou dimensionalitou problému příliš ovlivněny (Vapnik, 1998). Z toho vyplývá, že jev redukce počtu slov, a tím tedy i dimensionality, nebude mít na klasifikaci SVM takový pozitivní vliv, jako byl pozorován u rozhodovacích stromů. To je v souladu i s naměřenými hodnotami, které ukazuje tabulka 6 a graf na obrázku 12. Na těchto výsledcích jednoznačně převládá negativní vliv nepřesností a chyb při překladu.

Tabulka 6: Úspěšnost klasifikace SVM

	de	fr	es	pl	cs
původní	**94,47**	**93,77**	**92,13**	**96,10**	**93,11**
bez stematizace	94,51	93,74	92,12	96,35	92,38
se stematizací	94,43	93,81	92,14	95,86	93,85
přeloženo	**93,91**	**93,22**	**92,10**	**94,87**	**92,30**
bez stematizace	93,89	93,19	92,00	95,18	92,12
se stematizací	93,92	93,25	92,20	94,55	92,49



Obrázek 11: Úspěšnost klasifikace rozhodovacím stromem C5.0



Obrázek 12: Úspěšnost klasifikace SVM

4.2 Porovnání rule-based a statistického překladu z hlediska úspěšnosti klasifikace a složitosti vytvořeného modelu

V rámci tohoto pokusu, blíže popsaného v kapitole 3.5.2, bylo provedeno 72 testů[10] pomocí rozhodovacího stromu a 90 testů[11] podpůrnými vektory na celkem 300 000 dokumentech.

Tabulka 7 ukazuje výsledky klasifikace rozhodovacím stromem, seřazené podle ohodnocení úspěšnosti vytvořeného klasifikátoru na testovacích datech. Lze vidět, že úspěšnost se liší podle jazyka. Zatímco u němčiny byl úspěšnější statistický překlad, u francouzštiny a španělštiny byl o něco lepší překlad založený na pravidlech.

U obou přeložených množin pak lze pozorovat signifikantní snížení počtu atributů oproti originálům. To se jistě pozitivně projeví především na době zpracování.

Za povšimnutí dále stojí velikost vytvořeného stromu, která byla u rule-based překladu největší. Příčinou by mohlo být „roztříštění“ informací do více jazyků kvůli nedokonalému překladu.

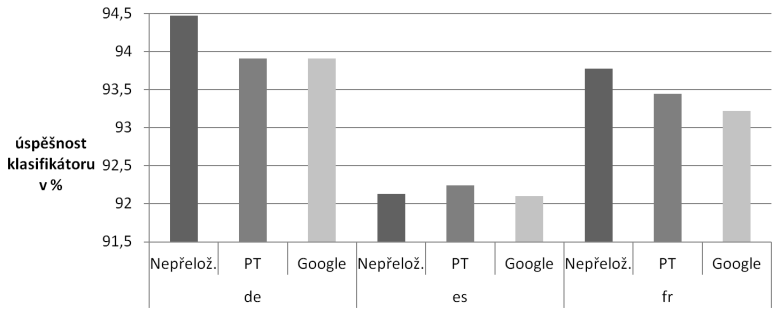
Tabulka 7: Shrnutí výsledků vlivu jednotlivých překladačů na klasifikaci rozhodovacím stromem C5.0

	prům. vel. stromu	průměrná úspěšnost	prům. počet atributů
Nepřeložené	**224**	**87,30**	**4763**
de	228	87,90	8228
es	222	86,34	3140
fr	220	87,66	2922
Power Translator	**256**	**87,17**	**2634**
de	257	87,82	3098
es	258	86,69	2332
fr	253	87,00	2471
Google	**236**	**87,24**	**2713**
de	227	88,37	2907
es	243	86,49	2530
fr	240	86,85	2702

Klasifikace podpůrnými vektory již více podpořila domněnku, že překlad založený na bázi pravidel je pro klasifikaci vhodnější. Graf na obrázku 13 ukazuje průměrné úspěšnosti klasifikátoru na jednotlivých množinách.

[10](1 původní množina + 2 přeložené) * 2 aplikace stematizace * 4 rozdělení na menší množiny kvůli výkonu * 3 jazyky

[11](1 původní množina + 2 přeložené) * 2 aplikace stematizace * 5 křížová validace * 3 jazyky



Obrázek 13: Shrnující graf výsledků klasifikace podpůrnými vektory SVMlight porovnávající jednotlivé způsoby překladu

4.3 Klasifikace dokumentů v různých jazycích na základě jedné trénovací množiny

V rámci tohoto pokusu, popsaného v kapitole 3.6, bylo provedeno celkem 24 testů[12] pomocí podpůrných vektorů. Strojový překlad testovacích množin se zde ukázal jako zajímavou jednoduchou alternativou v případě, že je k dispozici pouze trénovací množina v jednom jazyce a je třeba klasifikovat jazyk jiný. V průměru došlo ke ztrátě pouze *5,3* bodů F-score oproti ideální situaci, kdy byly dokumenty klasifikovány pomocí modelu natrénovaného na stejném jazyce. Pro srovnání, v práci, která se cross-language klasifikací přímo zabývala, dospěli podobnou metodou ke ztrátě 10 bodů z F-score (Shi, Mihalcea a Tian, 2010). Nutno podotkout, že autoři použili jiný korpus a o něco složitější klasifikaci do 7 kategorií.

Tabulka 8: Výsledky klasifikace SVMlight (F-score) s anglickou trénovací množinou oproti natrénování na množinách v jednotlivých jazycích

	de	es	fr	pl	průměr
původní	95,656	93,371	94,985	97,237	95,312
EN + přeložené	91,464	89,048	89,418	90,012	89,986
ztráta	**4,192**	**4,323**	**5,567**	**7,224**	**5,327**

4.4 Organizace dokumentů pomocí shlukování

Bylo provedeno celkem 50 testů. V tabulkách 9 a 10 je vidět, že u slovanských jazyků a němčiny se čistota snížila a entropie zvýšila. U francouzštiny a španělštiny se naopak objevil trend opačný.

[12](1 anglická množina + 1 původní) * 3 výběry * 4 jazyky

Tabulka 9: Celková čistota shluků (purity)

	2	10	20	50	100	**průměr**
cs						
původní	0,856	0,856	0,878	0,887	0,887	**0,873**
přeložené	0,845	0,854	0,865	0,867	0,868	**0,860**
de						
původní	0,844	0,868	0,868	0,869	0,873	**0,864**
přeložené	0,84	0,84	0,844	0,859	0,861	**0,849**
es						
původní	0,852	0,852	0,852	0,852	0,853	**0,852**
přeložené	0,852	0,852	0,852	0,853	0,856	**0,853**
fr						
původní	0,869	0,869	0,869	0,869	0,869	**0,869**
přeložené	0,871	0,871	0,871	0,871	0,871	**0,871**
pl						
původní	0,86	0,86	0,86	0,864	0,867	**0,862**
přeložené	0,843	0,843	0,843	0,849	0,862	**0,848**
průměr	0,853	0,857	0,860	0,864	0,866	

Tabulka 10: Celková entropie

	2	10	20	50	100	**průměr**
cs						
původní	0,493	0,461	0,432	0,413	0,404	**0,441**
přeložené	0,530	0,486	0,469	0,459	0,448	**0,478**
de						
původní	0,525	0,459	0,455	0,440	0,433	**0,462**
přeložené	0,538	0,500	0,483	0,471	0,462	**0,491**
es						
původní	0,550	0,534	0,529	0,521	0,510	**0,529**
přeložené	0,546	0,527	0,519	0,511	0,502	**0,521**
fr						
původní	0,497	0,474	0,471	0,462	0,456	**0,472**
přeložené	0,488	0,468	0,466	0,460	0,453	**0,467**
pl						
původní	0,493	0,459	0,457	0,442	0,434	**0,457**
přeložené	0,532	0,499	0,493	0,483	0,468	**0,495**
průměr	0,519	0,487	0,477	0,466	0,457	

5 Závěr

V rámci této práce bylo navrhnuto a provedeno několik experimentů. Bylo zjištěno několik pozitivních efektů použití strojového překladu, ty jsou však závislé na jazyce. Nejpozitivnější vliv byl pozorován při překladu němčiny do angličtiny, kdy došlo k redukci celkového počtu slov a v případě klasifikace rozhodovacím stromem i k lepšímu výkonu klasifikátoru. Podobný jev se vyskytl i u slovanských jazyků. Na základě těchto výsledků je možno doporučit zařazení překladu mezi ostatní metody předzpracování, pokud se jedná o klasifikaci německého, či slovanských jazyků za použití C5.0 rozhodovacího stromu.

Při porovnávání výsledků klasifikace dokumentů přeložených pomocí dvou různých typů překladačů, tedy převážně statistického online Google Translate a převážně založeného na pravidlech Power Translatoru, bylo zjištěno, že použití druhého jmenovaného (PT) je z hlediska přesnosti klasifikace výhodnější. Proti jeho nasazení však mluví nedostatečná všestrannost. Přeložení polských a ruských dokumentů často končilo předčasně chybou (nejspíše nedokonalá implementace různých kódování). Navíc velikost vytvořeného modelu (stromu) byla vyšší, což může poukazovat na nekonzistentní překlad. Tato práce tedy spíš doporučuje použití statistického překladu, ať už jde o univerzální řešení, nebo ještě lépe specifická, kde jsou překladače trénovány na datech z konkrétní oblasti.

Experimentálně bylo zjištěno, že díky překladu je možno použít model natrénovaný na jednom jazyce i pro jiný jazyk. V této práci bylo využito anglické trénovací množiny pro vytvoření univerzálního modelu a následné klasifikace jiných jazyků (před klasifikací byli přeloženy do angličtiny). Došlo jen k malé ztrátě přesnosti. V případě, kdy jsou k dispozici trénovací data v jednom jazyce a je třeba klasifikovat dokumenty v jiných jazycích, zařazení překladu dává smysl a na základě výsledků to lze doporučit.

Všechny provedené experimenty také poukázaly na fakt, že vliv překladu se liší v závislosti na jazyku. Výsledky však byly v rámci jazykových skupin konzistentní, tedy že například slovanské jazyky vykazovaly podobné trendy.

Využitelnost strojového překladu při dolování znalostí v praxi závisí také na cenové dostupnosti překladu. V této práci bylo využito komerčního překladače, který si v případě neakademického využití účtuje poplatky podle počtu přeložených znaků. To by při překladu kolekce o několika desítkách tisíc dokumentů znamenalo neúměrné vysoké náklady. Proto dalším možným rozšířením této práce by bylo sestavení vlastního statistického překladače z dat specifických pro tuto konkrétní oblast. Hlubší průzkum problematiky strojového překladu by byl nutný, jelikož se jedná o komplexní disciplínu.

6 Reference

[Berry a Kogan, 2010] Berry, M. W. & Kogan, J. *Textmining: Applications and theory*. Hoboken, NJ: John Wiley & Sons, 2010. DOI:10.1002/9780470689646.

[Dařena, 2011] Dařena, F. *Vybrané přístupy k dolování znalostí ze sociálních médií*. Habilitační práce. Brno: MENDELU, 2011.

[Dařena, Žižka a Přichystal, 2014] Dařena, F. & Žižka, J. & Přichystal, J. *Clients' freely written assessment as the source of automatically mined opinions*. Enterprise and the Competitive Environment 2014 conference. Brno, 2014.

[Evans a Klavans, 2003] Evans, D. K. & Klavans, J. L. *A Platform for Multilingual News Summarization* Columbia University Academic Commons. 2003.

[Feldman a Sanger, 2007] Feldman, R. & Sanger, J. *The text mining handbook*. Cambridge: Cambridge University Press, 2007. ISBN:978-0-511-33507-5.

[Fortuna a Shawe-Taylor, 2005] Fortuna, B. & Shawe-Taylor, J. *The use of machine translation tools for cross-lingual text mining*. 2005.

[Gliozzo a Strapparava, 2005] Gliozzo, A. & Strapparava, C. *Cross language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora.* Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, 2005, s. 9–16.

[Hastie a col., 2009] Hastie, T. & Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009. ISBN-10:0387848576.

[Hudík, 2015] Hudík, T. *Machine Translation Within Commercial Companies* Modern Computational Models of Semantic Discovery in Natural Language. 2015. s. 256-272.

[Chlumská, 2014] Chlumská, L. *Není korpus jako korpus: korpusy v kontrastní lingvistice a translatologii.* Časopis pro moderní filologii. 2014, č. 2. ISSN:0008 7386.

[IDC, 2014] International Data Corporation. *Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives* The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, 2014. [Online]. Dostupné z: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. [Accessed: 03- Nov- 2015].

[Krupník, 2014] Krupník, J. *Automatizace generování stopslov.* Diplomová práce, MendelU, 2014.

[Li et. al., 2009] Li, S., Xia, R., Huang, C. R., Zong, C. *A Framework of Feature Selection Methods for Text Categorization.* Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Singapore: ACL and AFNLP, 2009. s. 692–700.

[Marsland, 2009] Marsland, S. *Machine Learning.* An Algorithmic Perspective. Boca Raton: CRC Press, 2009. ISBN:978-1420067187.

[ModernMT, 2015] *ModernMT - Next Generation Machine Translation. [online] Dostupné z: http://www.modernmt.eu/ [Accessed 27 Nov. 2015].*

[Montalvo a kol., 2006] Montalvo, S. & Martínez, R. & Casillas, A. & Fresno, V. *Multilingual News Document Clustering: Two Algorithms Based on Cognate Named Entities* Text, Speech and Dialogue. Lecture Notes in Computer Science, 2006, sv. 4188, s. 165-172. DOI:10.1007/11846406_21.

[Quinlan, 1986] Quinlan R. J. *Induction of Decison Trees.* Machine Learning. 1986. s. 81-106.

[Romeo, Ienco a Tagarelli, 2015] Romeo, S. & Ienco, D. & Tagarelli, A. *Multilingual document classification via transductive learning* 6th Italian Information Retrieval Workshop, May 2015, Cagliari, Italy. CEUR Workshop Proceedings, sv. 1404, 2015.

[Salton, Wong a Yang, 1975] Salton, G. & Wong, A. & Yang, C. S. *A Vector Space Model For Automatic Indexing.* Communications of the ACM 18.11, 1975, s. 613-620.

[Shannon, 1948] Shannon, C. E.*A Mathematical Theory Of Communication.* Bell System Technical Journal 27.4, 1948. s. 623-656.

[Shi, Mihalcea a Tian, 2010] Shi, L. & Mihalcea, R. & Tian, M. *Cross Language Text Classification by Model Translation and Semi-Supervised Learning.* Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, MIT, Massachusetts, USA, 2010. s. 1057-1067.

[Silva, Ribeiro, 2010] Silva, C. & Ribeiro, B. *Inductive inference for large scale text classification.* Berlin: Springer. 2010. ISBN:978-3-642-04533-2.

[Steinberger, 2010] Steinberger, R. *Challenges and methods for multilingual text mining* Proceedings of Language Resources and Evaluation Conference. Malta. 2010. Dostupné z: http://www.lrec-conf.org/proceedings/lrec2010/keynotes/keynote2.pdf.

[Vapnik, 1998] Vapnik, V. *Statistical learning theory.* New York: Wiley. 1998. ISBN: 978-0-471-03003-4.

[Weiss, Indurkhya, a Zhang, 2010] Weiss, S., Indurkhya, N. & Zhang, T. *Fundamentals of Predictive Text Mining*. London: Springer-Verlag London Limited, 2010. ISBN:978-1-84996-225-4.

[Witten a Frank, 2005] Witten, I. & Frank, E. *Data Mining*. San Diego: Elsevier Science & Technology Books, 2005.

[Žižka a Dařena, 2010] Žižka, J. & Dařena, F. *Automatic Sentiment Analysis Using the Textual Pattern Content Similarity in Natural Language*. Lecture Notes in Artificial Intelligence, 2010, sv. 6231, s. 224-231. ISSN 0302-9743.

[Yang a Pedersen, 1997] Yang, Y. & Pedersen, J. O. *A Comparative Study on Feature Selection in Text Categorization*. In Proceedings of ICML-97, 14th International Conference on Machine Learning. D. H. Fisher. Nashville, TN, Morgan Kaufmann Publishers, San Francisco. 1997. s. 412–420.

[Žižka a Dařena, 2011] Žižka, J. & Dařena, F. *Mining Significant Words from Customer Opinions Written in Diferent Natural Languages*. Lecture Notes in Artificial Intelligence, 2011, sv. 6836, s. 211-218. ISSN 0302-9743.

[Žižka, Burda a Dařena] Žižka, J. & Burda, K. & Dařena, F. *Grouping of Customer Opinions Written in Natural Language Using Unsupervised Machine Learning*. Proceedings of the 14th International Symposium on Symbolic and Numeric Algorithms for scientific Computing. Timisoara: SYNASC, Romania, 2014. s. 265–270.