Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Automatická klasifikace vícejazyčných dokumentů

Plzeň 2016 Ladislav Hlom

Poděkování

Děkuji Doc. Ing. Pavlu Královi, Ph.D vedoucímu této diplomové práce za jeho ochotu, vedení, připomínky a čas, který mi věnoval.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 10. května 2016

Ladislav Hlom

Abstract

Česky

Automatická klasifikace dokumentů je úloha, ve které dokumenty zařazujeme do určitých kategorií dle jejich obsahu (např. politika, sport, ...). V práci je řešena především více třídní klasifikace, ve které může dokument patřit do více kategorií. Cílem práce bylo prozkoumat možnosti vícejazyčné klasifikace dokumentů. V rámci řešení je porovnávána metoda LDA s klasifikací po strojovém překladu do cílového jazyka. Použity jsou klasifikační metody maximální entropie a metoda podpůrných vektorů. K překladu textu jsou použity statistické systémy pro strojový překlad Moses a Google translate. Pro testování byly vybrány 3 rozdílné kolekce. První kolekce byla dodána od České tiskové kanceláře, zatímco zbylé dvě byly nalezeny na internetu. Provedené experimenty ukázaly, že varianta se strojovým překladem poskytuje solidní výsledky. Zatímco klasifikování za použití metody LDA dosahovalo nižších výsledků a nelze ho pro úlohu doporučit. Dále bylo ukázáno jak kvalita překladu ovlivňuje výslednou klasifikaci.

Klíčová slova: klasifikace, více třídní, SVM, maximální entropie, naivní bayes, LDA, klasifikace vícejazyčných dokumentů, strojový překlad, SMT

English

Automatic classification of documents is a task, where each document is classified into some categories based on its content (e.g politics, sport, etc.). The thesis is primarily focused on multi-label classification, where each document can belong to more than one category. The main aim of the thesis is a multilingual document classification. LDA method is compared with a classification after machine translation into a target language. Maximum entropy and vector machines are used as classification methods. Statistical machine translation systems Moses and Google Translate are used for the text translation. For testing three different collections were selected. The first collection was delivered from the Czech News Agency, while the other two were found on the Internet. The experiments that were done showed that the machine translation provides good-quality results. On the other hand, classification with LDA method achieved worse results and cannot be recommended for the task. Furthermore, it was shown how the quality of the translation affects the final classification.

Keywords: classification, multi-label, SVM, maximum entropy, naive bayes, LDA, multilingual document classification, machine translation, SMT

Obsah

1 Úvod **1**

2 Strojový překlad **2**
2.1 Lingvistická teorie překladu 2
2.2 Architektury systémů 3
2.2.1 Pravidlový strojový překlad 3
2.2.2 Statistický strojový překlad 3
2.2.3 Hybridní strojový překlad 4
2.2.4 Celkové Porovnání 4
2.3 Použité nástroje 5
2.3.1 Moses 5
2.3.2 Google Translate 5
2.4 Evaluační metriky 5
2.4.1 Lidská evaluace 6
2.4.2 Automatická Evaluace 6

3 Reprezentace dokumentu pomocí metody LDA **8**
3.1 Latentní Dirichletova alokace 8
3.2 Vyhodnocení podobnosti 9
3.2.1 Kosinová podobnost 9
3.2.2 Euklidovská vzdálenost 9

4 Automatická klasifikace **10**
4.1 Popis problému 10
4.2 Typy klasifikačních problémů 10
4.2.1 Binární klasifikace 10
4.2.2 Klasifikace do více oddělených tříd 11
4.2.3 Více třídní klasifikace 11
4.3 Klasifikační metody 11
4.3.1 Typy učení 11
4.3.2 Naivní Bayesův klasifikátor 12
4.3.3 Maximální entropie 13
4.3.4 Metoda podpůrných vektorů 15
4.4 Natrénování klasifikátoru a klasifikace 17
4.4.1 Nástroje ke klasifikaci 17
4.4.2 Evaluační metriky 18

5 Analýza datových kolekcí **21**
5.1 ČTK české dokumenty . . . 21
5.2 ČTK anglické dokumenty . . . 22
5.3 Reuters . . . 23
5.4 BBC . . . 24

6 Programové řešení **26**
6.1 Klasifikace . . . 26
6.1.1 Více třídní klasifikace . . . 26
6.1.2 Křížová validace . . . 27
6.1.3 Příznaky . . . 27
6.2 Testovací program . . . 27
6.3 Trénování systému pro překlad . . . 30
6.4 Paralelní korpusy . . . 30
6.5 Trénování systému pro překlad . . . 31
6.6 Překlad systémem . . . 32
6.6.1 Překlad pomocí systému Moses . . . 32
6.6.2 Překlad pomocí Google translate . . . 32
6.7 Klasifikace použitím LDA . . . 33

7 Dosažené výsledky **34**
7.1 Klasifikace českých dokumentů od ČTK . . . 34
7.2 Ohodnocení kvality překladu systémů . . . 37
7.3 Klasifikace anglických dokumentů od ČTK . . . 37
7.3.1 Testování pro jednu kolekci . . . 38
7.3.2 Kategorie namapované na českou kolekci . . . 39
7.4 Klasifikace anglických dokumentů od Reuters . . . 45
7.5 Klasifikace anglických dokumentů od BBC . . . 47
7.6 Zhodnocení výsledků . . . 51

8 Závěr **52**

Použitá literatura a zdroje

A Systémy pro překlad
A.1 Moses . . .
A.1.1 Instalace a trénování . . .
A.1.2 Překlad . . .
A.1.3 Ohodnocení kvality překladu . . .
A.2 Google translate . . .
A.2.1 Překlad . . .
A.2.2 Ukázková konfigurace . . .

B Aplikace pro klasifikaci
B.1 Konfigurace . . .
B.2 Ukázková konfigurace . . .

Seznam obrázků

2.1 Překladový trojúhelnk [23] 2

4.1 Optimální nadrovina 15
4.2 Průběh klasifikace [13] 17
4.3 Komponenty knihovny Brainy [20] 18
4.4 Schéma klasifikační úlohy [20] 18

5.1 Distribuce délek dokumentů 21
5.2 Distribuce počtu kategorií pro dokumenty 22
5.3 Distribuce délek dokumentů 22
5.4 Distribuce kategorií pro dokumenty 23
5.5 Distribuce délek dokumentů 23
5.6 Distribuce délek dokumentů 24
5.7 Distribuce kategorií pro dokumenty 25

6.1 Diagram testovacího programu 28
6.2 Diagram metody klasifikace 29
6.3 Google translate - ukázka formátu výstupu 32
6.4 Klasifikace metodou LDA 33

7.1 Vliv minimálního počtu trénovacích dokumentů pro kategorii na klasifikaci binárním sjednocením. 35
7.2 F-míra při klasifikaci českých dokumentů prahováním 36
7.3 Varianta s překladem dokumentů a následnou klasifikací (pro jednu kolekci) 38
7.4 Varianta s použitím metody LDA (pro jednu kolekci) 38
7.5 Varianta s překladem dokumentů a následnou klasifikací (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů) 40
7.6 Varianta s použitím metody LDA (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů) . . 40
7.7 F-míra při klasifikaci prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 43

7.8 F-míra při klasifikaci prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 45
7.9 Varianta s překladem dokumentů a následnou klasifikací (pro jednu kolekci) 45
7.10 Varianta s použitím metody LDA (pro jednu kolekci) 46
7.11 Varianta s překladem dokumentů a následnou klasifikací (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů) 48
7.12 Varianta s použitím metody LDA (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů) . . 48

Seznam tabulek

2.1 Srovnání RBMT a SMT systémů [3] 4

4.1 Srovnání Multinomiálního a Bernoulliho modelu [1] 13
4.2 Kontingenční tabulka 19

7.1 Klasifikace českých dokumentů binárním sjednocením (vliv minimálního počtu trénovacích dokumentů pro kategorii) 35
7.2 Klasifikace českých dokumentů prahováním 36
7.3 Porovnání systémů pro překlad metrikou BLEU 37
7.4 Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA 39
7.5 Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA 39
7.6 Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 41
7.7 Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 42
7.8 Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 43
7.9 Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 44
7.10 Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA 46
7.11 Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA 47
7.12 Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 49
7.13 Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 49

7.14 Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 50
7.15 Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí) 50

1 Úvod

Objem textových dat na internetu neustále stoupá. Data pochází z celého světa, proto jsou k dispozici v mnoha různých jazycích. Aby byla data použitelná, je nezbytné jejich třídění na základě obsahu. Například článek může být zařazen pod kategorie politika, sport nebo kultura. Články často lze zařadit pod více kategorií, které mohou být značně specifické. Například po zařazení článku do kategorie sport může být dále tříděn podle typu sportu (fotbal, hokej). Po takovém rozdělení článků může uživatel jednoduše získat přístup přesně k tomu, co ho zajímá.

Ke správnému zařazení dokumentu je nejprve nutné provést analýzu jeho obsahu, po které je možné dokument zařadit. Při současném objemu dat není v lidských silách všechna data ručně třídit. Proto je vhodné tuto činnost automatizovat, což je použitím existujících metod proveditelné. Současné metody pracují na podobném principu, který spočívá v nalezení určitých vzorů a podobností, na jejichž základě mohou být dokumenty tříděny.

Tato práce se zabývá kategorizací dokumentu podle obsahu do jedné či více kategorií. V několika jazycích jsou zkoumány dva rozdílné přístupy. Prvním přístupem je pomocí překladu dokumentů do cílového jazyka s následnou klasifikací přeloženého. Druhým přístupem je pomocí obecné reprezentace, ve které probíhá klasifikace porovnáváním podobností mezi dokumenty. Pro variantu s překladem je testováno, jak moc ovlivňuje kvalita překladu výslednou klasifikaci.

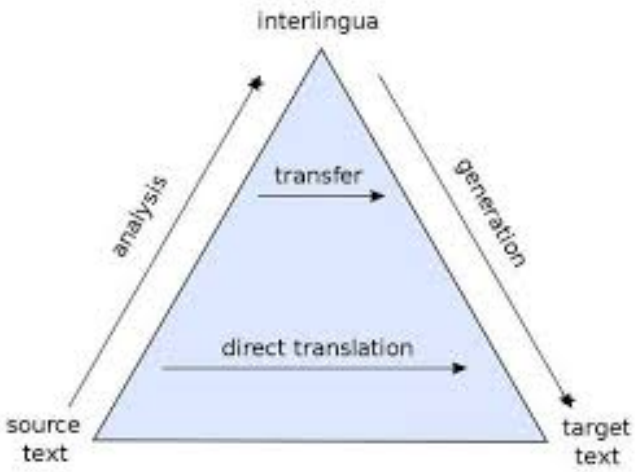
Po přečtení práce by měl čtenář porozumět současným přístupům ve strojovém překladu, problému klasifikace dokumentů a možnostem klasifikace dokumentů v cizím jazyce. Dále by měl být schopen realizovat překlad z libovolného jazyka, konfigurovat vytvořenou aplikaci a provádět upravené experimenty.

2 Strojový překlad

Strojová překlad je proces automatického překladu z jednoho jazyka do jiného pomocí počítače. Jedná se o komplikovanou úlohu, které je věnováno značné úsilí. V současné době jsou k dispozici systémy, jejichž výstup je dostatečně kvalitní pro použití v mnoha oblastech. V této kapitole budou rozebrány současné přístupy k řešení a budou zvoleny vhodné nástroje. Na závěr budou popsány možnosti ohodnocení kvality překladu.

2.1 Lingvistická teorie překladu

Jedná se o obecnou teorii, jak přeložit text z jednoho jazyka do druhého. V lingvistické teorii jsou 3 různé přístupy k překladu textu [11]. Náročnost jednotlivých přístupů je znázorněna pomocí překladového trojúhelníku (Vauquois triangle) viz obrázek 2.1.



Obrázek 2.1: Překladový trojúhelnk [23]

Direct translation model

Přeloží a přeuspořádá slova nebo n-gramy.

Transfer model

Používá znalosti o jazykových odlišnostech. Jednotlivé fáze překladu:

- analyzuje lexikální a syntaktickou strukturu zdrojové věty
- převádí strukturu ze zdrojového do cílového jazyka
- generuje odpovídající větu v cílovém jazyku

Interlingua model

Zjistí význam a vyjádří ho v cílovém jazyku. Jednotlivé fáze překladu:

- analyzuje lexikální, syntaktickou a sémantickou strukturu zdrojové věty
- interpretuje význam do kanonického umělého jazyka
- generuje větu v cílovém jazyku z kanonického jazyka

2.2 Architektury systémů

Pro strojový překlad je k dispozici několik rozdílných přístupů [3, 11]. V následující sekci budou popsány používané přístupy systémů a bude zvolena optimální varianta.

2.2.1 Pravidlový strojový překlad

Nejstarším přístupem ke strojovému překladu je pravidlový přístup, obvykle označovaný jako RBMT (Rule-based Machine Translation) [3, 11]. Rule-Based MT systémy používají lingvistický překladový model. Systém obsahuje sadu překladových pravidel a slovník. Pomocí pravidel je analyzována zdrojová věta, která je transformována do cílového jazyka. Tato pravidla musí být vytvořena lingvisty, což je primárním problémem těchto systému. Protože je tvorba těchto pravidel časově velmi náročná a vyžaduje specialisty, vývoj systému je velmi nákladnou a časově náročnou úlohou. Přidání každého dalšího pravidla navíc nezaručuje, že systém bude pracovat stejně nebo lépe. Použití tohoto přístupu může ale být jediným řešením pro jazyky, které nemají k dispozici dostatečné množství paralelního textu (korpusů) k natrénování. Použití může být vhodné, pokud se jedná o jednoduché jazyky.

2.2.2 Statistický strojový překlad

Zvyšování výpočetní síly počítačů umožnilo nástup statistických metod [3, 11]. To vytvořilo alternativu k RBMT a umožnilo eliminovat dosavadní problémy se strojovým překladem. Znalosti jsou implementovány matematickými modely, které systém sám definuje z paralelních korpusů. Následkem toho je natrénování systému pro nové jazyky mnohem snazší. Systému stačí poskytnout dostatečné množství paralelních korpusů pro trénink (čím více dat je k dispozici, tím je překlad lepší). V sou-

časné době existují dvě varianty systémů: Statistical Machine Translation (SMT) a Example-Based Machine Translation (EBMT).

2.2.3 Hybridní strojový překlad

Aktuálně se výzkum strojového překladu zaměřuje na hybridní překlad [11], který kombinuje RBMT a SMT systémy. Je navržen pro práci s méně daty k trénování. Informace získává učením z dat a použitím implementovaných překladových pravidel. Hlavní myšlenkou je automaticky se naučit překladová pravidla z omezeného množství dat.

2.2.4 Celkové Porovnání

V následující tabulce 2.1 jsou porovnány výhody a nevýhody RBMT a SMT systémů:

Výhody	**Nevýhody**
RBMT	
Založen na lingvistické teorii	Vyžaduje lingvistická pravidla a slovník
Dostačující pro jednoduché jazyky	Jazykově závislý
Nenáročný na výpočetní požadavky	Drahé vytvoření a rozšíření systému
Jednoduché odhalení chyb	Doba vytvoření systému
SMT	
Nevyžaduje lingvistické znalosti	Vyžaduje paralelní texty
Redukuje cenu lidských zdrojů	Vyžaduje velké výpočetní požadavky
Trénován lidskými překlady	Náročné odhalení chyb
Snadné vytvoření systému	Bez lingvistického pozadí

Tabulka 2.1: Srovnání RBMT a SMT systémů [3]

Jedním z požadavků je umožnit překlad z více různých jazyků. SMT systémy umožňují rychlé a snadné vytvoření překladu pro nový jazyk pomocí paralelních textů. Paralelní texty jsou dostupné pro mnoho jazyků a jsou zpravidla volně dostupné. Na základě výše uvedených důvodů bude pro překlad použit SMT systém.

2.3 Použité nástroje

V práci budou použity systémy, které jsou implementací SMT překladu. Pro strojový překlad jsou v současné době nejčastěji používány SMT systémy. V následující sekci budou zvoleny konkrétní systémy pro překlad. Dále budou probrány evaluační metriky, které slouží pro ohodnocení kvality překladu.

2.3.1 Moses

Moses je implementací statistického řešení strojového překladu [4]. Jedná se o open-source projekt pod licencí LGPL. Mezi výhody systému, který musí být natrénován je to, že ho můžeme natrénovat na takových datech, která jsou blízko typu překládaných dat. Tím může být ve výsledku dosaženo kvalitnějších výsledků, než běžné již natrénované systémy.

Moses sestává ze dvou hlavních komponent: training pipeline a dekodér. Training pipeline je kolekce nástrojů (primárně v perlu spolu s C++), které transformují data na strojový překladový model. Dekodér je aplikace v C++, která z natrénovaného překladového modelu a zdrojových vět přeloží větu ze zdrojového do cílového jazyka. Vzhledem k volné licenci a všem potřebným nástrojům je Moses nejpoužívanější open-source SMT systém.

2.3.2 Google Translate

Google Translate je stejně jako Moses implementací statistického řešení strojového překladu [24]. V současné době umožňuje překlad do více než stovky jazyků. Disponuje funkcemi na automatické rozpoznání zdrojového jazyku. Pro překlad poskytuje webové rozhraní, které je k dispozici zdarma. V tomto rozhraní je ale omezena maximální délka textu k překladu. Pro vývojáře a překlad delších textů je k dispozici API, které je placené na základě počtu přeložených slov.

2.4 Evaluační metriky

Výsledný strojový překlad musí být možné ohodnotit, aby mohla být zjištěna kvalita vytvořeného překladu. K ohodnocení kvality se používají dva základní přístupy [5].

2.4.1 Lidská evaluace

Kvalita překladu je hodnocena trénovaným překladatelem, který ohodnocuje kvalitu dle stanovené stupnice [5]. Takové ohodnocení je velmi přesné a vypovídá o kvalitě překladu. Zároveň je ale ohodnocení subjektivní a jeho zopakování je velmi obtížné. Hlavním problémem je především doba nutná k jeho provedení, od které se odvíjí také jeho vysoká cena.

2.4.2 Automatická Evaluace

Kvalita překladu je vyhodnocena strojově, na základě porovnání vytvořeného překladu s referenčními lidskými překlady [5]. Detaily porovnání jsou závislé na zvolené metrice. Výhodou je především rychlost a cena. Problémem ale je, že výsledná metrika nemusí být příliš vypovídající o kvalitě překladu. Automatická evaluace bude použita pro ohodnocení zvolených systémů v rámci této práce. Jako nejčastěji používaná automatická evaluace je označována metrika BLEU.

BLEU

Obvykle může mít věta několik kvalitních překladu [6]. Překlady se mohou lišit v použitých slovech nebo pouze pořadím jednotlivých slov. Tato metrika vychází z předpokladu, že čím je strojový překlad bližší profesionálnímu lidskému překladu, tím je lepší. Proto metrika využívá referenční lidský překlad, který porovnává se strojovým překladem. Metrika měří, jak dobře se strojový překlad překrývá s referenčními lidskými překlady. K tomu používá statistiku pro n-gramy. N-gramová přesnost je v BLEU počítána dle vzorce 2.1.

$$p_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} Count_{clip}\ (n-gram)}{\sum_{C \in Candidates} \sum_{n-gram \in C} Count\ (n-gram)} \tag{2.1}$$

kde $Count_{clip}$ je maximální počet n-gramů vyskytujících se ve strojovém a referenčním překladu. $Count(n-gram)$ je počet n-gramů ve strojovém překladu. Pro odstranění velmi krátkých překladů, které se snaží maximalizovat jejich skóre přesnosti, BLEU vypočítává tzv. trest za stručnost (brevity penalty) dle vzorce 2.2

$$BP = \begin{Bmatrix} 1 & if|c| > |r| \\ e^{1-|r|/|c|} & \text{if } |c| \leq |r| \end{Bmatrix} \tag{2.2}$$

kde $|c|$ je délka strojového překladu a $|r|$ je délka referenčního překladu. Pro výpočet BLEU je použit vzorec 2.3.

$$BLEU = BP \times exp\,(\sum_{n=1}^{N} w_n \log p_n) \tag{2.3}$$

N je standardně nastaveno na 4 a w_n (váhový faktor) je nastaven na $1/N$.

3 Reprezentace dokumentu pomocí metody LDA

V této kapitole bude probrána možnost obecné reprezentace dokumentu za pomoci LDA, díky které nebude nutné dokumenty překládat.

3.1 Latentní Dirichletova alokace

Latentní Dirichletova alokace (LDA) je generativní pravděpodobnostní model [7, 8]. Základní myšlenkou je, že každý dokument je reprezentován jako náhodná směs skrytých témat, ve kterém je každé téma charakterizováno rozdělením slov.

Princip

V množině dokumentů má být identifikováno k témat. Nejprve je nutné vyjádřit pravděpodobnost generování kolekce dokumentů, a poté hledat takové parametry, které tuto pravděpodobnost maximalizují. Za pomoci těchto parametrů lze vyjádřit vektor témat pro každý dokument a rozdělení pravděpodobnosti slov pro každé téma. Tento postup je označovaný jako generativní model a podle [8] vypadá následovně:

- Pro každý dokument i z kolekce všech dokumentů vyber parametry multinomického rozdělení $theta(i)$ z Dirichletova rozdělení s parametry $alfa$. Obě rozdělení mají k dimenzí – tedy tolik, kolik témat chceme identifikovat. $Alfa$ je vektor k reálných čísel menších než 1, která jsou společná pro všechny dokumenty a jedná se o tzv. hyperparametr LDA modelu. Pro každý dokument tedy vybereme pravděpodobnosti pro všechna z k témat. Díky volbě parametrů Dirichletova rozdělení menších než 1 bude zajištěno, že s největší pravděpodobností bude mít pouze několik málo témat nezanedbatelnou pravděpodobnost, neříkáme však, která to budou. To odpovídá skutečnosti, kde dokument většinou pojednává pouze o několika málo tématech.
- Pro každou pozici slova j v dokumentu i vyber z Dirichletova rozdělení s parametry $theta$ (i) téma $z(i, j)$. Pro každou slovní pozici v dokumentu si tedy hodíme k-stěnnou kostkou, jejíž pravděpodobnosti jsou dány parametry $theta(i)$ a přiřadíme pozici příslušného tématu.
- Pro každou pozici (i, j) vyber slovo $w(i, j)$ z multinomického rozdělení $phi(z(i, j))$. Multinomické rozdělení phi slov pro témata je globální a podobně jako $theta$ má parametry generované z Dirichletova rozdělení, tentokrát s parametrem $beta$.

Pro nalezení parametrů LDA, a tudíž i požadovaných vektorů témat pro dokumenty, je třeba maximalizovat pravděpodobnost modelu pro poskytnutá trénovací data. Pro inferenci se v LDA využívají aproximativní metody. Mezi nejčastěji používané patří variační metody a Gibbsovo vzorkování.

3.2 Vyhodnocení podobnosti

Aby bylo možné shlukovat dokumenty patřící k sobě, je nutné vypočítat jejich vzájemnou podobnost. V následující sekci jsou popsány dvě základní metody [9].

3.2.1 Kosinová podobnost

Jednou z nejpoužívanějších metrik pro porovnávání podobnosti textových dokumentů je kosinová podobnost [9]. Když jsou dokumenty reprezentovány jako vektory, podobnost dokumentů odpovídá podobnosti vektorů. Kosinová podobnost dvou dokumentů $\vec{v_a}$ a $\vec{v_b}$ je vypočtena vzorcem 3.1.

$$SIM(\vec{v_a}, \vec{v_b}) = \frac{\vec{v_a}.\vec{v_b}}{|\vec{v_a}|.|\vec{v_b}|} \tag{3.1}$$

kde $v_{i,a}$ a $v_{i,b}$ jsou n dimenzionální vektory. Výsledná hodnota je v intervalu [0,1], přičemž hodnota 1 označuje shodné dokumenty.

3.2.2 Euklidovská vzdálenost

Euklidovská vzdálenost je standardní metrika pro řešení geometrických problémů, která se často používá pro shlukování [9]. Výpočet vzdálenosti mezi dvěma dokumenty d_a a d_b, které jsou reprezentovány vektory v_a a v_b je vypočtena vzorcem 3.2.

$$d(\vec{v_a}, \vec{v_b}) = \sqrt{\sum_{i=1}^{N} (v_{i,a} - v_{i,b})^2} \tag{3.2}$$

kde $v_{i,a}$ a $v_{i,b}$ tvoří souřadnice vektorů v_a a v_b pro dimenzi i. N je velikost dimenze.

4 Automatická klasifikace

V následující kapitole bude vysvětlena automatická klasifikace. Budou popsány typy problému a učení. Dále budou vysvětleny metody, které budou využity pro klasifikaci. Na závěr bude zvolen nástroj ke klasifikaci a popsány metriky pro vyhodnocení přesnosti klasifikace.

4.1 Popis problému

Automatickou klasifikaci lze označit za úlohu [14], ve které se každé položce v rozhodovací matici přidělí hodnota z intervalu {0,1}.

	d_1	...	...	d_j	...	...	d_n
c_1	a_{11}	...	...	a_{1j}	...	...	a_{1n}
...	...	...	...	...	...	...	...
c_i	a_{i1}	...	...	a_{ij}	...	...	a_{in}
...	...	...	...	...	...	...	...
c_m	a_{m1}	...	...	a_{mj}	...	...	a_{mn}

$C = \{ c_1, ..., c_m \}$ je kolekce předdefinovaných kategorií a $D = \{ d_1, ..., d_n \}$ je kolekce dokumentů, které mají být klasifikovány. Hodnota 1 pro a_{ij} je interpretována jako rozhodnutí pro dokument d_j, že patří pod kategorii c_i. zatímco hodnota 0 je interpretována jako rozhodnutí pro dokument d_j, že nepatří pod kategorii c_i.

4.2 Typy klasifikačních problémů

Klasifikaci dokumentů můžeme rozdělit na 3 podmnožiny [10] v závislosti na tom, do kolika kategorií dokument zařazujeme.

4.2.1 Binární klasifikace

Dokument může patřit do jedné ze dvou kategorií. Výstupem klasifikátoru je ano nebo ne. Může se například jednat o klasifikaci spamu - jedná/nejedná se o spam.

4.2.2 Klasifikace do více oddělených tříd

Dokument může patřit pouze do jedné kategorie, ale celkový počet kategorií je větší než dva. Příkladem může být klasifikace obrázků ovoce, kdy se může jednat o pomeranče, jablka nebo hrušky. Předpokladem je, že se na obrázku nemůže vyskytovat více druhů, ale vždy pouze jeden.

4.2.3 Více třídní klasifikace

Dokument může patřit do jedné nebo více kategorií. Příkladem může být novinový článek, který lze zařadit pod politiku a finance. Více třídní klasifikace je hlavním cílem této práce.

4.3 Klasifikační metody

Pro klasifikaci byly na základě studia literatury vybrány metody: Naivní Bayes, Maximální entropie a Metoda podpůrných vektorů. V následující sekci je uvedeno základní dělení metod podle typu učení [12] a jsou popsány zvolené metody.

4.3.1 Typy učení

S učitelem (supervised)

Pro trénování klasifikátor obdrží množinu trénovacích dat, které mají označenu třídu, do které patří. Z těchto dat se klasifikátor naučí pravidla pro klasifikaci do jednotlivých tříd. Tento typ učení je nejčastěji používaný a bude využit pro tuto práci.

Bez učitele (unsupervised)

Množina trénovacích dat neobsahuje informaci o třídě, do které patří. Cílem je nalézt podobnosti mezi jednotlivými dokumenty a na jejím základě rozpoznat třídy. Nalezené třídy je poté nutné ručně pojmenovat.

Částečné učení s učitelem (semi-supervised)

Jedná se o kombinaci dvou předchozích metod. K trénování se používají data anotovaná učitelem spolu s daty bez přiřazené třídy. Této metody se využívá pokud máme

malé množství označených dat, ale zároveň máme dostatek neoznačených dat.

4.3.2 Naivní Bayesův klasifikátor

Pro úlohu automatické klasifikace je v literatuře velmi rozšířen tento klasifikátor [1, 15]. Mezi jeho přednosti patří především jednoduchost a rychlost klasifikace. Předpokladem pro klasifikaci je nezávislosti atributů.

Princip

Zařazení dokumentu do třídy probíhá výpočtem podmíněných pravděpodobností pro každou třídu a výběrem třídy s maximální aposteriorní pravděpodobností. Výpočet aposteriorní pravděpodobnost (pravděpodobnost zařazení dokumentu d do třídy c) je proveden použitím Bayesovi věty (vzorec 4.1).

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{4.1}$$

Apriorní pravděpodobnost $P(c)$ je vypočtena z trénovacích dat pomocí vzorce 4.2.

$$P(c) = \frac{N_c}{N} \tag{4.2}$$

kde N_c je počet dokumentů ve třídě a N je celkový počet dokumentů. Podmíněná pravděpodobnost $P(d|c)$ je vypočtena použitím vzorce 4.3.

$$P(d|c) = \frac{n_d!}{\prod_{t=1}^{n_d} f(t,d)!} \prod_{t=1}^{n_d} P(t|c)^{f(t,d)} \tag{4.3}$$

Hodnota $f(t,d)$ určuje počet výskytů termu t v dokumentu d. Hodnota n_d určuje počet termů v dokumentu.

Podmíněnou pravděpodobnost $P(t|c)$ určíme za pomoci vzorce 4.4.

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \tag{4.4}$$

kde Tct je počet výskytů t v trénovacích dokumentech pro třídu c, dělený součtem výskytů všech t v dané třídě.

Výběr nejlepší třídy pro dokument je proveden dle vzorce 4.5.

$$c_{map} = \arg\max_{c \in C} P(c|d) = \arg\max P(c) \prod_{k=1}^{n} P(t_k|c) \tag{4.5}$$

kde n je počet slov v dokumentu.

Modely

Pro tento klasifikátor jsou používány dva modely: Multinomiální a Bernoulliho, které se liší reprezentací termů. Multinomiální model si uchovává počet výskytů termů oproti Bernoulliho modelu, který pro term uchovává pouze to, zda ho dokument obsahuje. V tabulce 4.1 je uvedeno srovnání obou modelů.

	Multinomiální model	**Bernoulliho model**
činnost modelu	generování tokenů	generování dokumentu
náhodná proměnná	X = t if t na dané pozici	U_t = t if v dokumentu
reprezentace dokumentu	d =<t_1, ..., t_k, ..., t_{n_d} >, $t_k \in V$	d =<e_1, ..., e_i, ..., e_M >, $e_i \in \{0, 1\}$
odhad parametru	$\widehat{P}(X = t\|c)$	$\widehat{P}(U_i = e\|c)$
rozhodovací pravidlo	$\widehat{P}(c) \prod_{1 \leq k \leq n_d} \widehat{P}(X = t_k\|c)$	$\widehat{P}(c) \prod_{t_i \in V} \widehat{P}(U_i = e_i\|c)$
vícenásobný výskyt	brán v úvahu	ignorován
délka dokumentu	zvládne delší dokumenty	nejlépe funguje pro krátké dokumenty
počet příznaků	zvládne větší počet	nejlépe funguje jen s několika
odhad pro příznak *the*	$\widehat{P}(X = the\|c) \approx 0.05$	$\widehat{P}(U_{the} = 1\|c) \approx 1.0$

Tabulka 4.1: Srovnání Multinomiálního a Bernoulliho modelu [1]

4.3.3 Maximální entropie

Dalším klasifikátorem je klasifikátor maximální entropie [18], který poskytuje velmi dobré výsledky pro kategorizaci textů.

Princip

Trénovací data jsou použita pro nalezení omezení pravděpodobnostního rozložení, které musí být splněno při odhadu modelu. Jako vektor příznaků označíme funkci

$f_i(d, c)$ pro dokument d a třídu c. D je množina trénovacích dat. Cílem je nalezení modelu, který bude tomuto rozložení odpovídat. Hledaná podmíněná pravděpodobnost $P(c|d)$ zařazení dokumentu d do třídy c musí splňovat vlastnost 4.6.

$$\frac{1}{|D|}\sum_{d\in D} f_i(d, c(d)) = \sum_d P(d)\sum_c P(c|d)f_i(d, c) \tag{4.6}$$

$P(D)$ označuje pravděpodobnostní rozdělení dokumentů, které ale neznáme. Proto je za použití trénovacích dat vytvořena aproximace této hodnoty viz vzorec 4.7.

$$\frac{1}{|D|}\sum_{d\in D} f_i(d, c(d)) = \frac{1}{|D|}\sum_{d\in D}\sum_c P(c|d)f_i(d, c) \tag{4.7}$$

Dalším krokem je nalezení příznakových funkcí z trénovacích dat, které budou vhodné pro klasifikaci. Poté je nutné pro každý příznak vypočítat očekávanou hodnotu, která bude použita jako omezení modelu.

Parametrická forma

Po vypočtení všech omezení je zaručeno, že nalezený model má maximální entropii. Podle [19] může být ukázáno, že rozložení má vždy exponenciální formu viz vzorec 4.8.

$$P(c|d) = \frac{1}{Z(d)} exp\ (\sum_i \lambda_i f_i(d, c)) \tag{4.8}$$

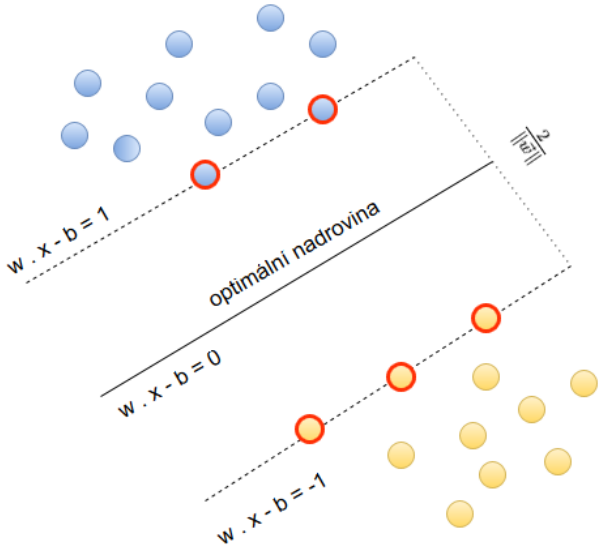
kde $f_i(d, c)$ je model příznaku, λ_i je parametr, který má být odhadnut a $Z(d)$ je normalizační faktor pro zajištění správné hodnoty pravděpodobnosti vypočtený podle vzorce 4.9.

$$Z(d) = \sum_c exp\ (\sum_i \lambda_i f_i(d, c)) \tag{4.9}$$

Po vyčíslení všech omezení z označených trénovacích dat je řešení problému maximální entropie také řešením duálního věrohodnostního problému (maximum likehood problem) pro modely, které mají stejnou exponenciální formu. Navíc je zaručeno, že konvexní funkce má jedno globální maximum a žádná lokální maxima. Proto existuje jedno možné řešení pro nalezení maximální entropie, za použití Gradientního (Hill climbing) algoritmu. Z libovolného počátečního odhadu exponenciálního rozdělení konverguje k řešení věrohodnostního problému, které je zároveň globálním řešením maximální entropie.

4.3.4 Metoda podpůrných vektorů

Další zvolenou metodou klasifikace je metoda podpůrných vektorů dále označovaná jako SVM (Support Vector Machine) [1, 16, 17]. Pro jednoduchost se omezíme na binární klasifikátor. Metoda hledá nadrovinu, která v prostoru příznaků optimálně rozděluje trénovací data. Optimální nadrovina je taková, ve které je minimální vzdálenost bodů od nadroviny pro obě množiny co největší. Grafické znázornění metody se nachází na obrázku 4.1.



Obrázek 4.1: Optimální nadrovina

Červeně označené body na obrázku znázorňují podpůrné vektory (support vectors), z kterých je určena nadrovina viz rovnice 4.10

$$w \,.\, x - b = 0 \tag{4.10}$$

Podle nadrovin je provedena klasifikace pro jednotlivé třídy. Všechny body splňující podmínku 4.11

$$w \,.\, x - b \geq 1 \tag{4.11}$$

jsou zařazeny do třídy ”modrých koleček” (standardně označované jako y=+1). Všechny body splňující podmínku 4.12

$$w \,.\, x - b \leq 1 \tag{4.12}$$

jsou naopak zařazeny do třídy "oranžových koleček" (standardně označované jako y=-1). Pro uvedené podmínky je hledání nadroviny následujícím optimalizačním problémem 4.13.

$$(w, b) = argmax \frac{1}{2}||w||^2 \tag{4.13}$$

K řešení tohoto optimalizačního problému se používá tzv. duální problém.

Duální problém

Pro tento případ se hledají parametry α_i, které jsou řešením rovnice 4.14.

$$\vec{a} = argmax\ (\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=j}^{n} \alpha_i \alpha_j y_i y_j\ (x_i, x_j)) \tag{4.14}$$

za podmínek

$$\alpha_i \leq 0, i = 1, 2..na \sum_{i=1}^{n} \alpha_i y_i = 0$$

Za pomoci nalezených parametrů α vypočítáme hodnotu w podle vzorce 4.15

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{4.15}$$

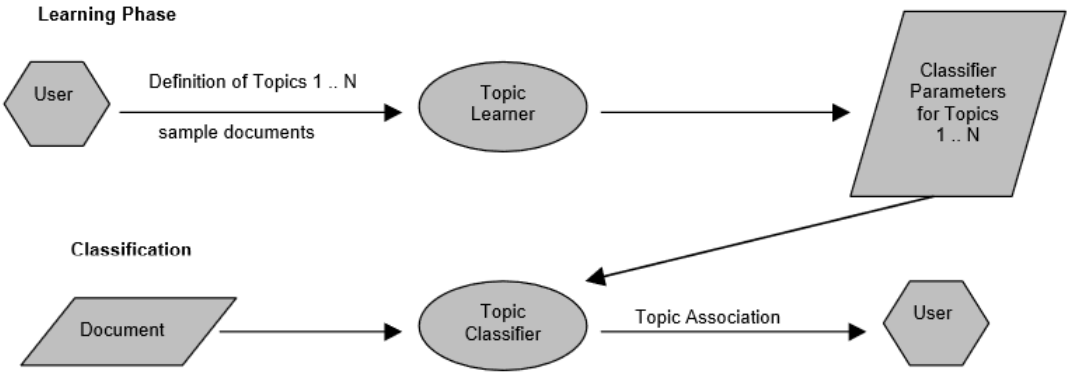
Nelineární klasifikace

Doposud jsme se zabývaly jednoduchým případem, ve kterém jsou data snadno lineárně rozdělitelná. Při reálném použití ovšem data obvykle nejsou lineárně rozdělitelná. Proto se jako řešení používá převedení dat do vyšší dimenze, které převede úlohu na lineárně separovatelnou. K převedení existuje jednoduchý způsob, tzv. trik (kernel trick), který pro určení nadroviny používá skalární součiny jádrové funkce

$$K = \ (\vec{x_i}, \vec{x_j}) = \vec{x_i}^T, \vec{x_j} \tag{4.16}$$

4.4 Natrénování klasifikátoru a klasifikace

Proces klasifikace se skládá ze dvou fází [13]. První fází je učící fáze (learnig phase), ve které se klasifikátor trénuje na rozpoznání jednotlivých tříd. Z trénovacích dat nalezne příznaky, sloužící ke klasifikaci. Druhou fází je klasifikace (classification), ve které je pro příchozí dokument použit natrénovaný klasifikátor, který může dokument zařadit do dané kategorie. Celý průběh je znázorněn na obrázku 4.2.



Obrázek 4.2: Průběh klasifikace [13]

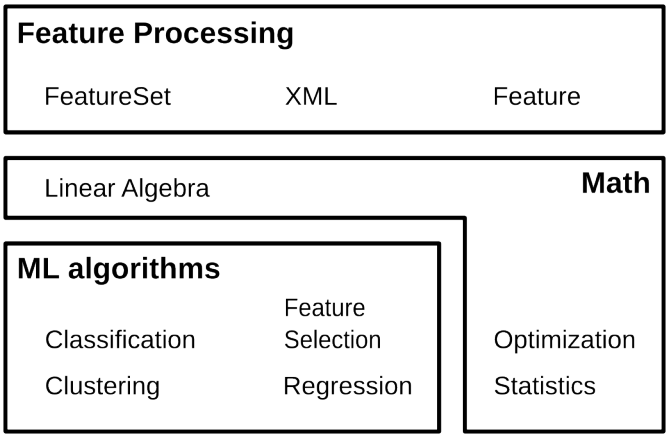
4.4.1 Nástroje ke klasifikaci

Na základě doporučení byla pro klasifikaci zvolena knihovna Brainy [20]. Brainy obsahuje všechny typy klasifikátoru, které mají být použity a je k dispozici zdarma. Proto byla zvolena jako optimální možnost, bez nutnosti hledat další nástroje.

Brainy

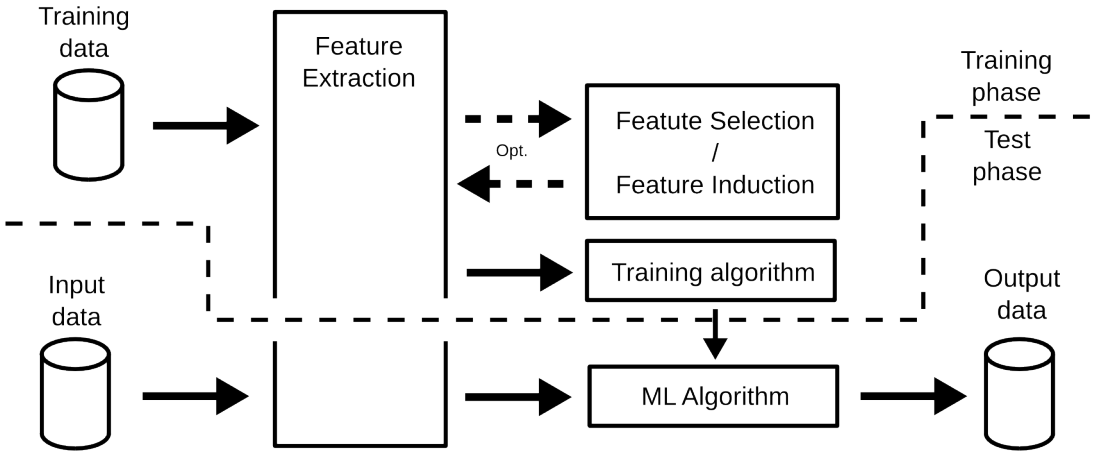
Jedná se o knihovnu [20] pro klasifikaci napsanou v JAVĚ, která pochází ze ZČU. Brainy se skládá ze tří hlavních komponent viz obrázek 4.3. Primární komponentou je machine learning, která poskytuje rozhraní pro klasifikaci nebo clustering. Další komponentou je ML, která obsahuje efektivní implementaci lineárně algebraických struktur a algoritmů. Poslední komponentou je feature processing, která má na starosti tvorbu matic a vektorů z objektů poskytnutých uživatelem.

Použití knihovny sestává z několika fází. První fází je extrakce příznaků, ve které jsou uživatelské objekty převedeny na matice a vektory. Druhá fáze je závislá na použitém typu úlohy, pro naší úlohu je zajímavá úloha typu učení s učitelem. Pro tuto



Obrázek 4.3: Komponenty knihovny Brainy [20]

úlohu uživatel nejprve inicializuje trénovací třídu implementující zvolenou metodu. Trénovací třída poté zpracuje trénovací data a vytvoří klasifikátor. Po natrénování je klasifikátor připraven k použití. Schéma celého procesu je znázorněno na obrázku 4.4



Obrázek 4.4: Schéma klasifikační úlohy [20]

4.4.2 Evaluační metriky

Pro určení přesnosti klasifikace se používá několik evaluačních metrik [1]. Pro vyhodnocení toho, zda byl dokument správně zařazen je potřeba znát třídu, do které dokument patří. Na základě toho lze určit konfuzní matici, která je nutná pro vyčíslení většiny metrik. Matice a její jednotlivé prvky jsou vysvětleny tabulkou 4.2.

	Relevantní	Nerelevantní
Obdržené	tp	fp
Neobdržené	fn	tn

Tabulka 4.2: Kontingenční tabulka

Vysvětlení jednotlivých buňek:

- tp (true positive) - predikovaná kategorie se shoduje se správnou kategorií
- fp (false positive) - predikovaná kategorie se neshoduje se správnou kategorií
- fn (false negative) - klasifikátor nezařadil dokument do kategorie, do které patří
- tn (true negative) - klasifikátor správně nezařadil dokument do kategorie

Přesnost

Přesnost (precision) je poměr z obdržených výsledků, které jsou relevantní. Výpočet je určen pomocí vzorce 4.17.

$$P = \frac{tp}{tp + fp} \tag{4.17}$$

Úplnost

Úplnost (recall) je poměr z relevantních výsledků, které byly obdrženy. Výpočet je určen pomocí vzorce 4.18.

$$R = \frac{tp}{tp + fn} \tag{4.18}$$

F-míra

F-míra patří k nejpoužívanějším metrikám pro ohodnocení kvality klasifikace. Je založena na předpokladu, že požadujeme co nejvyšší úplnost, přičemž tolerujeme jen určité procento nerelevantních výsledků. Výpočet je proveden za pomoci přesnosti a úplnosti, kdy se jedná o jejich harmonický průměr viz vzorec 4.19.

$$F = \frac{1}{\alpha\frac{1}{P} + \ (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad kde \quad \beta^2 = \frac{1-\alpha)}{\alpha} \tag{4.19}$$

kde $\alpha \in [0, 1]$ a tedy $\beta^2 \in [0, \infty]$.

Standardně používané hodnoty jsou $\alpha = 1/2$ *nebo* $\beta = 1$. Při jejichž použití se metrika označuje jako F_1 a vzorec 4.19 se zjednoduší na vzorec 4.20.

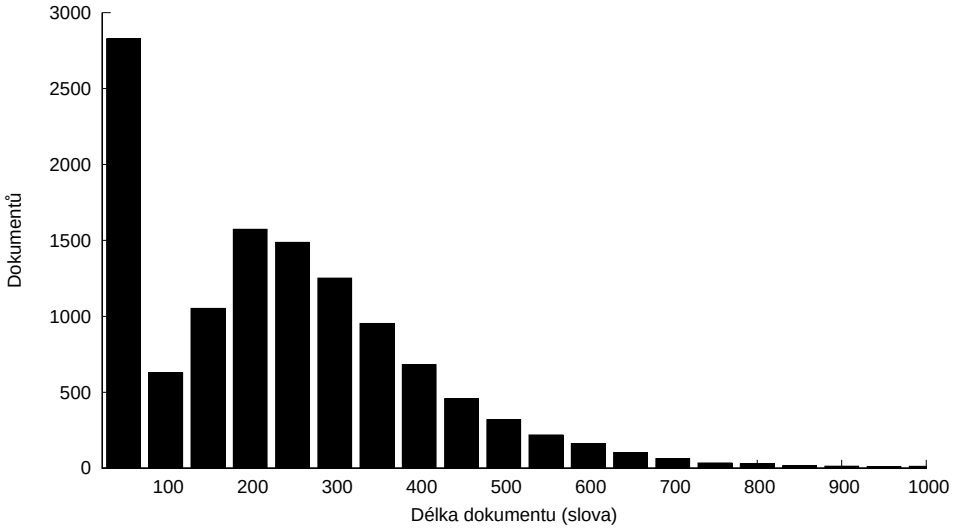
$$F_{\beta=1} = \frac{2PR}{P + R} \tag{4.20}$$

5 Analýza datových kolekcí

Data pro trénování a testování klasifikátorů poskytla Česká tisková kancelář (dále jen ČTK). Dodané kolekce byly dvou typů. První kolekce obsahovala české dokumenty, uložené jako textové soubory. Druhá kolekce obsahovala cizojazyčné zprávy, uložené v XML souboru. Kolekce byly dle zadání doplněny o dvě další volně dostupné kolekce v cizím jazyce. V následující kapitole jsou všechny kolekce detailně analyzovány.

5.1 ČTK české dokumenty

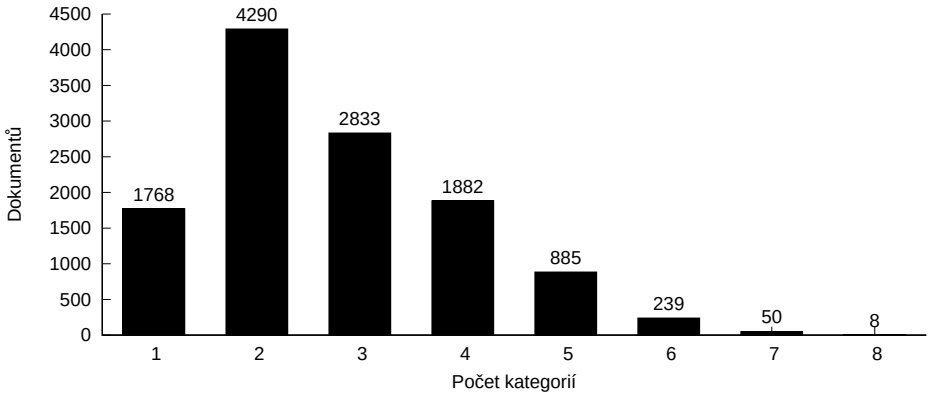
Tato kolekce je souhrnem českých dokumentů získaných od ČTK[1]. Byla dodána jako kolekce textových souborů, které ve svém názvu obsahují všechny kategorie, do kterých patří. Kolekce obsahuje celkem 11 955 dokumentů, které obsahují 2 922 319 slov. Distribuce délek dokumentů je znázorněna na obrázku 5.1.



Obrázek 5.1: Distribuce délek dokumentů

V kolekci je k dispozici celkem 60 kategorií. Každý dokument může patřit do více kategorií. Distribuce počtů kategorií pro dokumenty je znázorněna na obrázku 5.2.

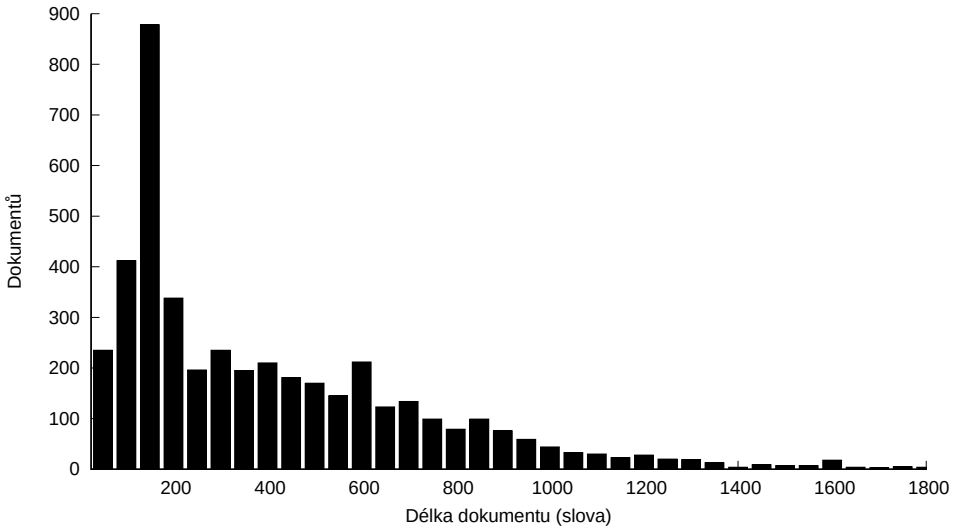
[1] http://home.zcu.cz/~pkral/sw/



Obrázek 5.2: Distribuce počtu kategorií pro dokumenty

5.2 ČTK anglické dokumenty

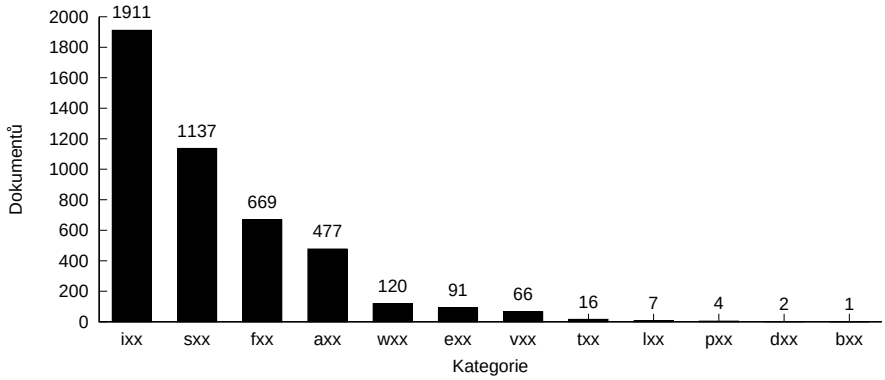
Tato kolekce je souhrnem cizojazyčných dokumentů získaných od ČTK. Byla dodána jako XML soubor, který obsahoval velké množství atributů. Mezi důležité atributy patří zpráva, agentura (z agentury lze určit jazyk, ve kterém je dokument) a kategorie. Kolekce obsahuje 4 501 dokumentů v anglickém jazyce, které obsahují 2 230 239 slov. Distribuce délek dokumentů je znázorněna na obrázku 5.3.



Obrázek 5.3: Distribuce délek dokumentů

V kolekci je k dispozici celkem 12 kategorií, přičemž každý dokument patří pouze do jedné kategorie. Celkový počet kategorií je nízký a proto je možné zobrazit počet

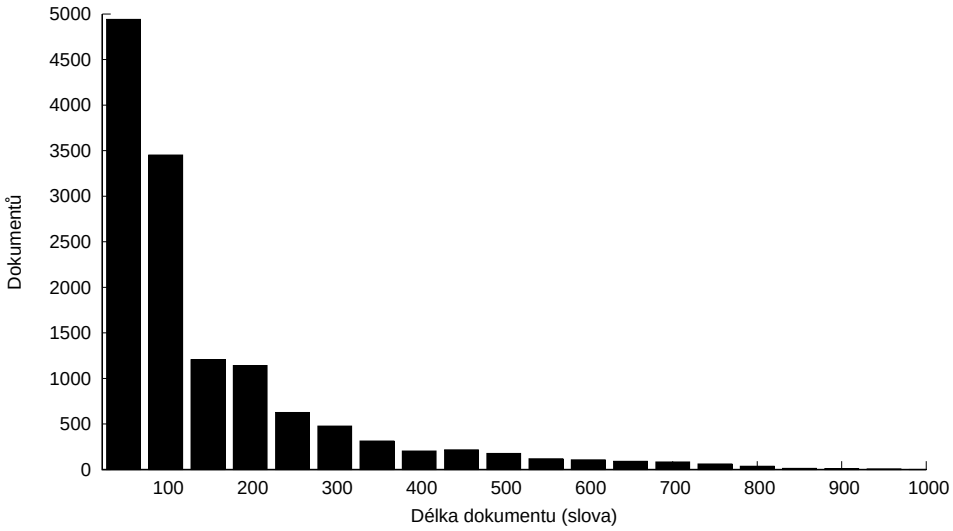
dokumentů pro všechny kategorie. Kategorie jsou velmi obecné, ale některé lze přibližně namapovat na českou kolekci od ČTK. Distribuce kategorií pro dokumenty je znázorněna na obrázku 5.4.



Obrázek 5.4: Distribuce kategorií pro dokumenty

5.3 Reuters

Tato kolekce je souhrnem anglických dokumentů od Reuters[2]. Je k dispozici jako kolekce textových souborů, které jsou rozděleny do kategorií podle složek, ve kterých se nachází. Kolekce je rozdělena na kolekci pro trénování a testování. Obsahuje celkem 3 744 testovacích a 9 583 trénovacích dokumentů, které obsahují 2 022 177 slov. Distribuce délek dokumentů je znázorněna na obrázku 5.5.



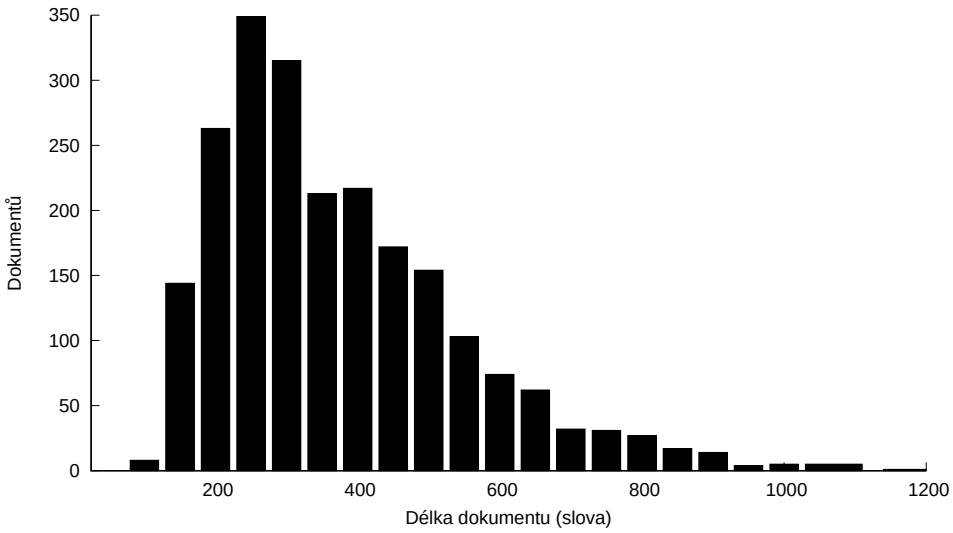
Obrázek 5.5: Distribuce délek dokumentů

[2] http://disi.unitn.it/moschitti/corpora.htm

V kolekci je k dispozici celkem 90 kategorií, přičemž každý dokument patří pouze do jedné kategorie. Kategorie jsou velmi specifické, a proto je nelze namapovat na českou kolekci od ČTK.

5.4 BBC

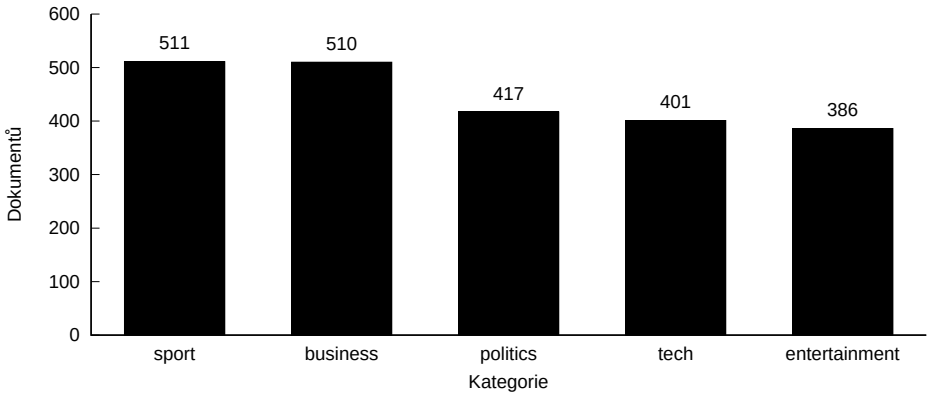
Tato kolekce je souhrnem anglických dokumentů, které pocházejí od BBC[3]. Konkrétně se jedná o internetového zpravodajství za období 2004-2005. Je k dispozici jako kolekce textových souborů, které jsou podle složek rozděleny do kategorií, do kterých patří. Kolekce obsahuje celkem 2225 dokumentů, které obsahují 854 490 slov. Distribuce délek dokumentů je znázorněna na obrázku 5.5.



Obrázek 5.6: Distribuce délek dokumentů

V kolekci je k dispozici celkem 5 kategorií, přičemž každý dokument patří pouze do jedné kategorie. Celkový počet kategorií je nízký, a proto je možné zobrazit počet dokumentů pro všechny kategorie. Distribuce kategorií pro dokumenty je znázorněna na obrázku 5.4. Kategorie jsou obecné a lze je namapovat na českou kolekci od ČTK.

[3] http://mlg.ucd.ie/datasets/bbc.html



Obrázek 5.7: Distribuce kategorií pro dokumenty

6 Programové řešení

V následující kapitole je popsáno, jakým způsobem bylo vytvořeno programové řešení. Nejprve jsou rozebrány detaily použité při klasifikaci a vytvořený klasifikační program. Dále je vysvětlen způsob trénování systému pro překlad a způsob, jakým jsou systémy použity.

6.1 Klasifikace

Pro klasifikaci byla zvolena možnost trénování s učitelem. Z klasifikačních metod byly vybrány následující - Naive Bayes, Maximální entropie a SVM. Uvedené metody fungují pro binární klasifikátor, ale pro testování je nutné převést úlohu na více třídní klasifikaci. Pro převod na tento typ úlohy se používají postupy, které jsou popsané v části více třídní klasifikace. Poté je popsána křížová validace, která se používá k důkladnému ověření výsledků.

6.1.1 Více třídní klasifikace

Binární sjednocení

Pro každou kategorii je natrénován binární klasifikátor. Trénovací data jsou pro každý klasifikátor rozdělena na dvě množiny. První množinou jsou dokumenty, které obsahují danou kategorii. Druhou jsou všechny dokumenty, které neobsahují danou kategorii. Při více třídní klasifikaci je klasifikováno postupně každým binárním klasifikátorem a dokument je zařazen do dané kategorie, pokud je pravděpodobnost zařazení do první třídy větší než do druhé. Nevýhodou této metody je především časová náročnost trénování a klasifikace.

Prahování

Pro všechny kategorie je natrénován jeden společný klasifikátor. Při klasifikaci je dokument zařazen pod každou kategorii, pro kterou je pravděpodobnost, že patří do dané kategorie větší, než stanovená mez. Výhodou této metody je především rychlost, oproti binárnímu sjednocení. Problémem je ale nalezení vhodné meze.

6.1.2 Křížová validace

Vstupní množina dat je rozdělena na podmnožiny. Jedna množina je určena jako testovací, zatímco zbylé slouží jako trénovací množiny. Tento proces se několikrát opakuje, pokaždé s jinou podmnožinou pro trénovací a testovací množinu, až je množina postupně otestována celá.

6.1.3 Příznaky

Pro klasifikaci je nutné zvolit příznaky, které budou použity pro klasifikaci. V aplikaci jsou používány dva modely, které jsou dále popsány.

Model - Bag of words

Model je natrénován pomocí trénovacích dokumentů, ze kterých jsou uložena všechna slova s počtem výskytů. Každý dokument ke klasifikaci je reprezentován indexy slov.

Model pro LDA

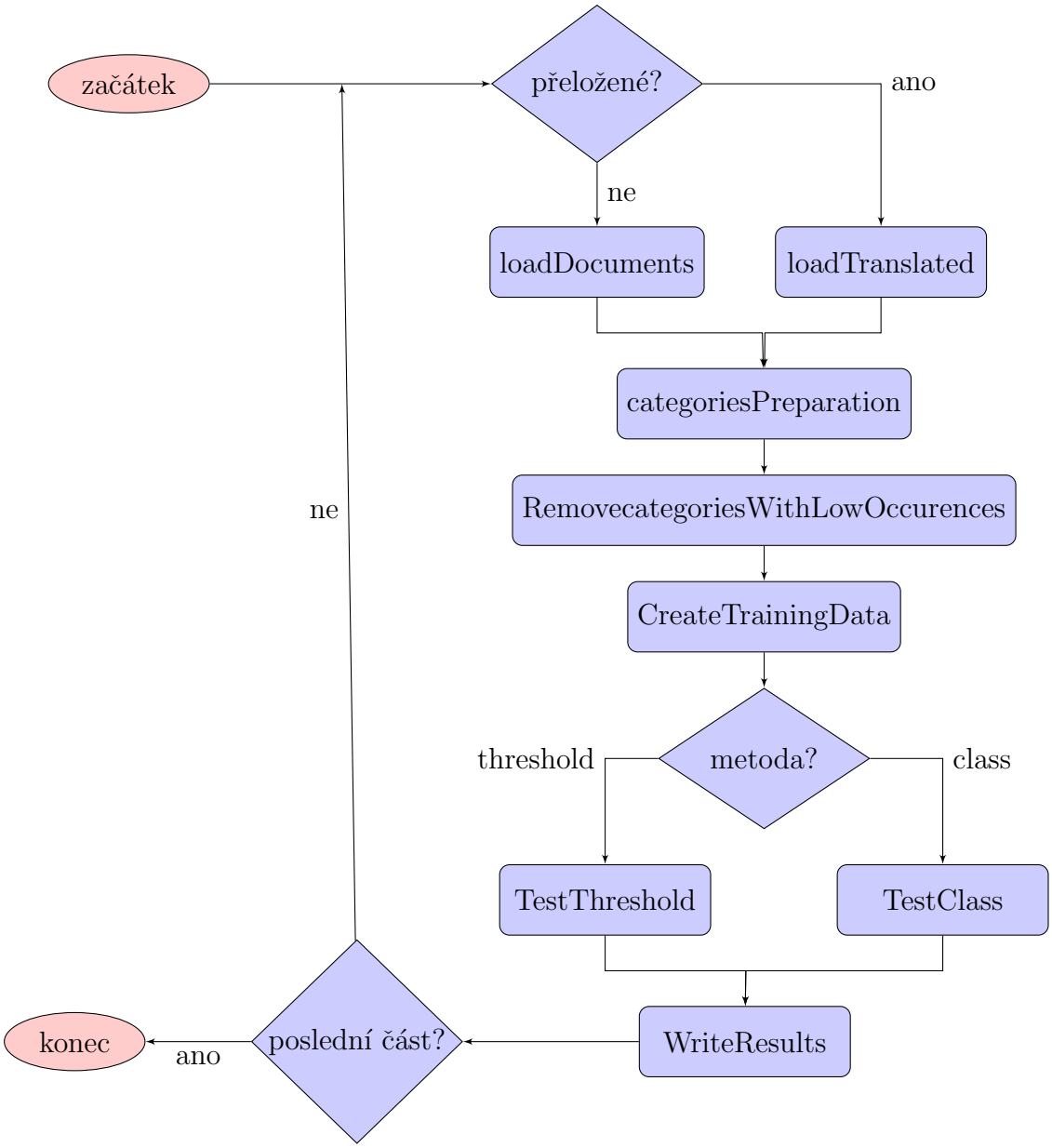
Model si uloží vektory všech trénovacích dokumentů. Při klasifikaci je vektor dokumentu porovnáván se všemi vektory trénovacích dokumentů. Podobnost dokumentů je vyhodnocována pomocí kosinové podobnosti.

6.2 Testovací program

Pro experimenty byl vytvořen program, umožňující konfiguraci a spouštění testů. Program poskytuje statistiky pro testovaný korpus, umožňuje spouštění testů pro jednu kolekci, která je testována. Dále lze testy provádět pro dvě kolekce, kdy jako trénovací jsou použity české dokumenty od ČTK a testovací je zvolená kolekce v cizím jazyce. Diagram průběhu klasifikace se nachází na obrázku 6.1. Jednotlivé metody jsou dále stručně popsány.

LoadAllDocuments - Metoda načte všechny soubory a uloží je do struktury pro další zpracování.

LoadTranslatedDocuments - přeložené zprávy se nahrají ze souboru a aktualizují se záznamy jednotlivých dokumentů. Metoda nahrává soubory nebo soubor podle nastavené kolekce a typu překladového systému.



Obrázek 6.1: Diagram testovacího programu

CategoriesPreparation - z načtených trénovacích dokumentů se získají všechny kategorie.

RemoveCategoriesWithLowOccurences - odstraní se kategorie, které obsahují menší počet trénovacích dokumentů než je stanovená mez. Tyto kategorie jsou odstraněny také ze všech dokumentů a pokud dokument již nepatří do žádné další kategorie, pak je vyhozen.

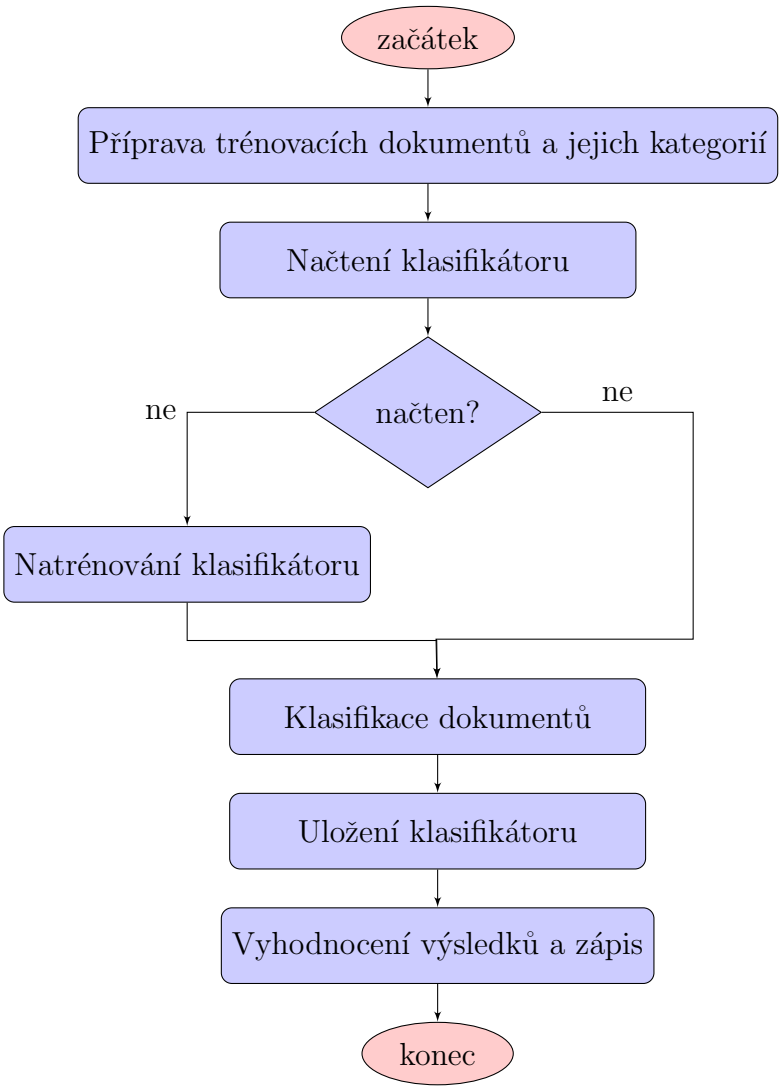
CreateTrainingData - připraví data pro trénování. Odstraní ze zpráv html znaky a zprávy jsou tokenizovány.

TestThreshold - natrénuje a otestuje klasifikátor prahováním.

TestClass - natrénuje a otestuje klasifikátor binárním sjednocením.

WriteResults - zapíše výsledky metrik (přesnost, úplnost, f-míra) do souboru.

Metody TestThreshold a TestClass pracují podle algoritmu, který je popsán na obrázku 6.2. Metody se liší v nepatrných detailech, které jsou dále popsány.



Obrázek 6.2: Diagram metody klasifikace

TestThreshold při vyhodnocení výsledků pro dokument určí pravděpodobnost zařazení do každé kategorie. Dokument je do kategorie zařazen, pokud je daná hodnota větší než stanovená prahovací mez. Alternativně může dokument zařadit pouze do třídy s největší pravděpodobností. Při této klasifikaci je použit pouze jeden klasifikátor.

TestClass naproti tomu obsahuje tolik klasifikátorů, kolik je kategorií. Při vyhodnocení výsledků je dokument zařazen do kategorie, pokud je pravděpodobnost že do ní patří větší než 0,5. Klasifikace je prováděna v cyklu podle počtu kategorií. TestClass je tedy výrazně pomalejší, ale oproti TestThreshold není třeba určovat vhodnou prahovací mez.

Načtení klasifikátoru - je metoda, která zjišťuje zda již nebyl klasifikátor natrénován. Pokud byl natrénován, pak je uložen na disku a lze ho načíst. Toto je umožněno využitím serializace, která umožňuje uložit celou třídu a také ji jednoduše načíst. Tento postup ušetří velké množství času, neboť trénování klasifikátoru je pro velké množství dat časově náročnou úlohou.

6.3 Trénování systému pro překlad

Pro překlad byly zvoleny systémy používající statistický strojový překlad. Ty musí být nejprve natrénovány pomocí paralelních korpusů. V následující sekci je popsán způsob výběru korpusů a je popsáno, jak byly systémy trénovány.

6.4 Paralelní korpusy

Jako výchozí jazyk pro testování byla zvolena angličtina, která bude překládána do češtiny. Následující sekce obsahuje popis nalezených korpusů a důvody pro zvolení použitého trénovacího korpusu.

Europarl

Europarl[1] je korpus obsahující záznamy z jednání Evropského parlamentu [21]. Tyto záznamy jsou překládány do rodných jazyků všech členských zemí Evropské Unie. Velikost korpusu je závislá na datu vstupu, kdy byla země do Unie přijata. Česko-anglický korpus obsahuje 646 605 paralelních vět.

CzEng 1.0

CzEng 1.0[2] je čtvrté vydání paralelního česko-anglického korpusu od Ústavu formální a aplikované lingvistiky (UFAL), který je volně k dispozici pro nekomerční účely [22].

[1] http://www.statmt.org/europarl/
[2] http://ufal.mff.cuni.cz/czeng

Korpus obsahuje 15 136 126 paralelních vět, ze sedmi různých zdrojů (například z evropské legislativy, filmových titulků, technické dokumentace).

Volba korpusu

Výsledná kvalita překladu statistických strojových systému je závislá na množství dat, které mají k dispozici pro trénování. Pro natrénování systému Moses je nutné zvolit vhodný korpus. Použití CzEng 1.0 korpusu by bylo vhodnější, neboť obsahuje podstatně více dat a kvalita překladu roste s počtem trénovacích dat. Jedním z cílů této práce ale je, porovnat vliv kvality překladu na klasifikaci. Proto je vhodné, aby jeden systém poskytoval horší překlad. Europarl korpus je podstatně menší než CzEng 1.0 korpus a tak při natrénování systému Moses korpusem Europarl dostaneme znatelně horší překlad. Google translate nemusí být trénován a poskytuje vysoce kvalitní překlad, proto je vhodné, aby Moses nebyl natrénován příliš velkým korpusem. Další výhodou Europarl korpusu je, že je k dispozici pro všechny jazyky členských zemí EU. Systém tak může být natrénován pro více jazyků z podobných dat. Z těchto důvodů bude pro natrénování systému pro překlad použit Europarl korpus.

6.5 Trénování systému pro překlad

Statistické strojové systémy musí být natrénovány za pomoci paralelních korpusů. V testovací aplikaci jsou využívány systémy Moses a Google translate. Google translate poskytuje překlad prostřednictvím API služby. Moses je k dispozici ve formě systému, který musí být nejprve natrénován paralelním korpusem.

Moses

Moses je natrénován za pomoci Europarl korpusu. Tento korpus musí být před trénováním podroben následujícím krokům:

- tokenizace - Mezi slova a interpunkci je vložena mezera
- truecasing - Počáteční slova v každé větě jsou konvertována (velikost písmen) na jejich nejpravděpodobnější variantu
- cleaning - Prázdné nebo dlouhé věty jsou odstraněny, dále jsou odstraněny nesprávně zarovnané věty.

Poté je nutné vytvořit jazykový model pro cílový jazyk. Po tomto kroku je k dispozici vše potřebné a může proběhnout trénování systému. Když je dokončeno trénování, pak je již možné překládat text. Dále lze provést ladění systému, které

vylepšuje kvalitu překladu. Tento krok byl při trénování přeskočen, aby bylo možné zjistit vliv kvality překladu na klasifikaci.

6.6 Překlad systémem

V následující sekci je popsán způsob, jakým jsou oba systémy pro překlad používány. Vytvořený program ke klasifikaci připraví soubor se všemi zprávami a jejich identifikátorem pro překlad. Tento soubor je za pomoci systému přeložen a poté si program ke klasifikaci načte přeložený soubor.

6.6.1 Překlad pomocí systému Moses

Překlad pomocí systému Moses je jednoduchý, ale musí být natrénován paralelním korpusem. Při překladu stačí nejprve tokenizovat soubor pro překlad. Poté se jedním příkazem spustí překlad, po jehož dokončení mohou být přeložené zprávy nahrány zpět do programu ke klasifikaci.

6.6.2 Překlad pomocí Google translate

Během prohlížení webu byla nalezena možnost bezplatného použití systému Google translate. Google poskytuje adresu k překladové API, které má několik základních omezení. Jedním z nich je maximální velikost textu pro překlad, který je omezen na 2000 znaků. Kvůli tomuto omezení je text k překladu nutné vhodně rozdělovat na části, které poté lze přeložit. Dalším problémem je výstupní formát překladu, který je nutné zpracovat pomocí regulárních výrazů. Ukázka formátu výstupu při překladu dvou anglických vět viz obrázek 6.3.

```
[ [
  ["Dnes je pondělí. ","Today is monday.",,,1],
  ["To je jen test.","This is just a test.",,,0]
  ],,"en"]
```

Obrázek 6.3: Google translate - ukázka formátu výstupu

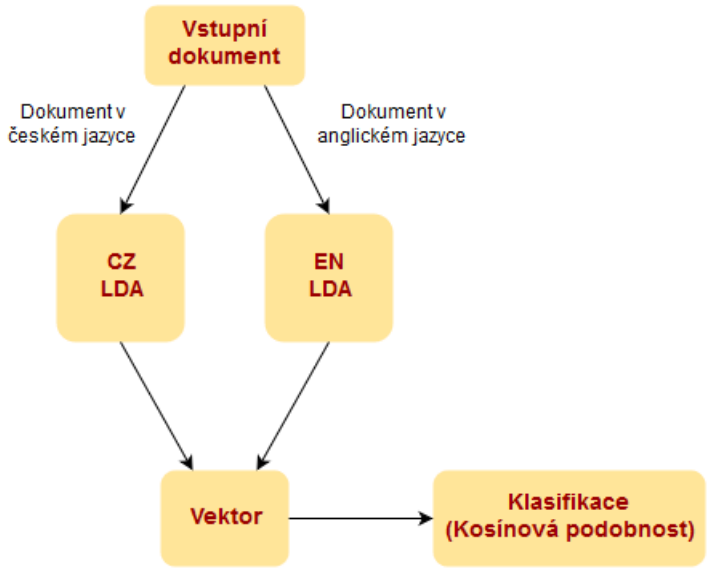
Pro tuto úlohu byla vytvořena samostatná aplikace, která načte soubor k přeložení. Z tohoto souboru načte celý řádek, podle jehož začátku provede operaci. V závislosti na nastavení pak může být proveden překlad. Tento řádek je načten a rozdělen tak, aby byl menší než 2000 znaků. Poté je přeložen pomocí Google

translate API, které vrací výsledek ve výše uvedeném formátu. Výsledek je zpracován regulárními výrazy a uložen. Řádek je zpracováván do té doby, než je celý přeložen. Při rozdělování řádku tak, aby měl méně než 2000 znaků musí brát zřetel na to, aby věta dávala smysl a Google ji mohl kvalitně přeložit. Proto je řádek rozdělován do vět, které jsou postupně přidávány k překladu až do povolené velikosti. Takto je postupně řádek překládán po větách, než je celý přeložen. Tento postup se opakuje do přeložení celého souboru.

6.7 Klasifikace použitím LDA

LDA reprezentuje dokument jako vektor pravděpodobností pro všechna definovaná témata. Pro úlohu klasifikace bylo zvoleno LDA se 100 tématy. K umožnění klasifikace z jednoho jazyka do druhého je nutné mít dva modely reprezentace dokumentu. Jeden pro český jazyk a druhý pro cizí jazyk. Témata musí být pro oba modely totožná a tudíž musí být LDA natrénováno ze shodně anotovaných dat. Pro porovnání pak lze použít kosínovou metriku, která může použitím dvou modelů porovnat dokumenty v rozdílných jazycích a určit jejich podobnost. Výsledná podobnost je v intervalu <0,1>, přičemž výsledek 1 znamená, že jsou dokumenty identické.

Celý proces je zobrazen na obrázku 6.4. Vstupní dokument je dle jazyka zpracován příslušným LDA modelem, které jej transformuje na vektor. Tento vektor může být klasifikován použitím kosinové podobnosti. Příznaky pro klasifikovaný dokument jsou jeho podobnosti se všemi trénovacími dokumenty. Implementace metody LDA k použití v testovacím program byla obdržena od vedoucího práce spolu s natrénovanými modely pro český a anglický jazyk.



Obrázek 6.4: Klasifikace metodou LDA

7 Dosažené výsledky

V této kapitole budou provedeny testy jednotlivých metod a parametrů. Testy budou provedeny na čtyřech různých datových kolekcí. První kolekce je od ČTK, která obsahuje dokumenty v českém jazyce. Zbylé tři kolekce obsahují dokumenty v anglickém jazyce, přičemž se jedná o kolekce od ČTK, Reuters a BBC.

Pro klasifikaci byly vybrány metody zvolené v předchozích kapitolách, tedy metoda maximální entropie, SVM a Naive bayes. Pro klasifikaci do více tříd budou použity dva přístupy k více třídní klasifikaci - binární sjednocení a prahování. Kolekce v českém jazyce bude podrobena testu, jak minimální počet trénovacích dokumentů pro kategorii ovlivňuje klasifikaci binárním sjednocením. Dále standardním testům na zjištění přesnosti klasifikace prahováním.

Kolekce v anglickém jazyce budou podrobeny dvěma rozdílným testům. První test bude proveden pro zjištění, jak kvalita překladu ovlivňuje výslednou klasifikaci. Pro tento test je připraven překlad dvěma různorodými nástroji - Google translate a Moses. Kvalita překladu obou nástrojů bude vyhodnocena metrikou BLEU.

Dalším testem bude klasifikace bez překladu, za pomoci obecné reprezentace dokumentu pomocí LDA. Poté, pokud to bude proveditelné, budou kategorie dané kolekce namapovány na českou kolekci a budou provedeny testy na natrénovaném modelu (CZ) pro testovací dokumenty z anglického jazyka.

Některé kolekce mohou být rozděleny na trénovací a testovací data. V takovém případě je respektováno toto rozdělení a klasifikace je provedena s daným rozdělením. Obvykle ale data takto rozdělena nejsou a není jich k dispozici velké množství. V takovém případě testování probíhá za pomoci křížové validace s parametrem 5. To znamená, že se kolekce rozdělí na 5 stejných částí, kdy pro trénování jsou použity 4/5 a pro testování zbývající část. Klasifikace pro jednu kolekci je tak spuštěna celkem pětkrát a to pro všechny možné kombinace. Získané výsledky jsou následně zprůměrovány. Úspěšnost klasifikace je vyhodnocena za pomoci metriky f-míra.

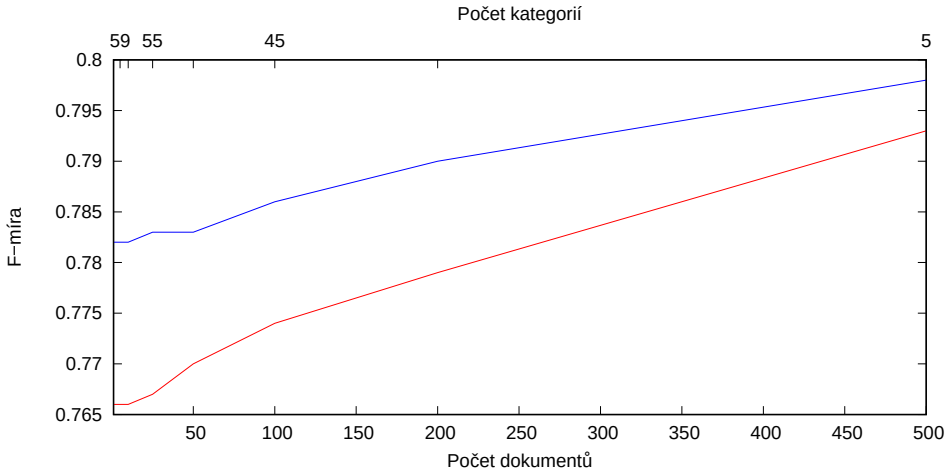
7.1 Klasifikace českých dokumentů od ČTK

Klasifikace na českých dokumentech je provedena použitím binárního sjednocení a prahování, které jsou následně porovnány. Pro klasifikaci binárním sjednocením je proveden test, jak minimální počet trénovacích dokumentů pro kategorii ovlivňuje výsledky klasifikace viz tabulka 7.1.

Počet dokumentů	Počet kategorií	ME			SVM		
		Přesnost	Úplnost	F-míra	Přesnost	Úplnost	F-míra
1	60	0,899	0,693	0,782	0,905	0,663	0,766
5	59	0,899	0,693	0,782	0,905	0,663	0,766
10	59	0,899	0,693	0,782	0,905	0,663	0,766
25	55	0,899	0,693	0,783	0,905	0,666	0,767
50	48	0,898	0,694	0,783	0,905	0,670	0,770
100	45	0,898	0,698	0,786	0,905	0,675	0,774
200	5	0,897	0,705	0,790	0,905	0,685	0,779
500	5	0,892	0,722	0,798	0,899	0,709	0,793

Tabulka 7.1: Klasifikace českých dokumentů binárním sjednocením (vliv minimálního počtu trénovacích dokumentů pro kategorii)

Výsledky f-míry pro vliv minimálního počtu trénovacích dokumentů pro kategorii, při použití klasifikace binárním sjednocením jsou zobrazeny v grafu 7.1. Výsledky jsou volbou minimálního počtu trénovacích dokumentů ovlivněny minimálně nejspíše proto, že při malém množství trénovacích dokumentů je jich také málo pro testování. Proto příliš neovlivňují výsledky metrik. Ze získaných výsledků lze nicméně usoudit, že minimální počet trénovacích dokumentů pro kategorii nemá příliš významný vliv, a proto už nebude dále zkoumán. Jako výchozí počet pro další testy bude zvolen minimální počet trénovacích dokumentů na 100.

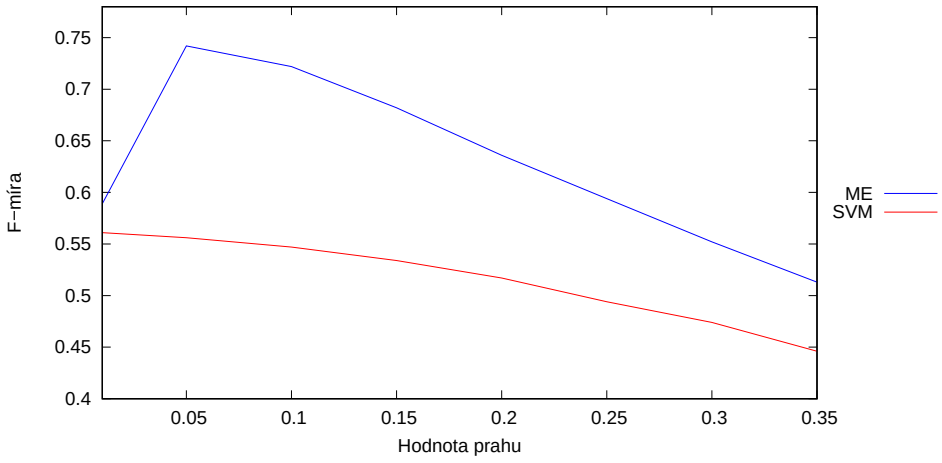


Obrázek 7.1: Vliv minimálního počtu trénovacích dokumentů pro kategorii na klasifikaci binárním sjednocením.

Dalším testem je klasifikace prahováním viz tabulka 7.2. Pro práh byly vybrány hodnoty v intervalu $[0, 1]$, v tabulce jsou ale zobrazeny pouze nejlepší výsledky. Je tedy zobrazeno celkem 8 hodnot, s krokem 0,05 až do hodnoty 0,35. Pro klasifikaci prahováním jsou výsledky f-míry zobrazeny také v grafu 7.2.

Metoda	Metrika	Hodnota meze pro prahování							
		0,01	0,05	0,1	0,15	0,2	0,25	0,3	0,35
ME	Přesnost	0,453	0,769	0,864	0,900	0,917	0,928	0,935	0,940
	Úplnost	0,851	0,716	0,620	0,548	0,487	0,437	0,392	0,353
	F-míra	0,589	0,742	0,722	0,682	0,636	0,594	0,552	0,513
SVM	Přesnost	0,613	0,690	0,735	0,764	0,783	0,794	0,799	0,802
	Úplnost	0,517	0,466	0,436	0,411	0,386	0,359	0,337	0,309
	F-míra	0,561	0,556	0,547	0,534	0,517	0,494	0,474	0,446

Tabulka 7.2: Klasifikace českých dokumentů prahováním



Obrázek 7.2: F-míra při klasifikaci českých dokumentů prahováním

Experimenty ukázaly jako optimální hodnotu meze pro prahování 0,05. Při zjemnění intervalu kolem nalezené hodnoty by se mohlo podařit najít nalézt lepší výsledky, zlepšení by ale bylo minimální.

7.2 Ohodnocení kvality překladu systémů

Jedním z naplánovaných testů je zjištění vlivu kvality překladu na výsledky klasifikace. Proto musí být ohodnoceno, jak kvalitní překlad systémy poskytují. Překlad bude získáván ze dvou systémů - Google translate a Moses. Výstup obou systému bude ohodnocen použitím BLEU metriky.

Pro vyhodnocení metriky BLEU musí být k dispozici text k přeložení (v anglickém jazyce) a referenční lidský překlad (v českém jazyce). K porovnání obou systému pro překlad byly zvoleny dva testovací korpusy[1]. Výstup obou systému byl vyhodnocen metrikou BLEU viz tabulka 7.3.

Systém pro překlad	nc_devtest	nc-dev
Google translate	27,09 %	37,69 %
Moses	7,63 %	13,28 %

Tabulka 7.3: Porovnání systémů pro překlad metrikou BLEU

7.3 Klasifikace anglických dokumentů od ČTK

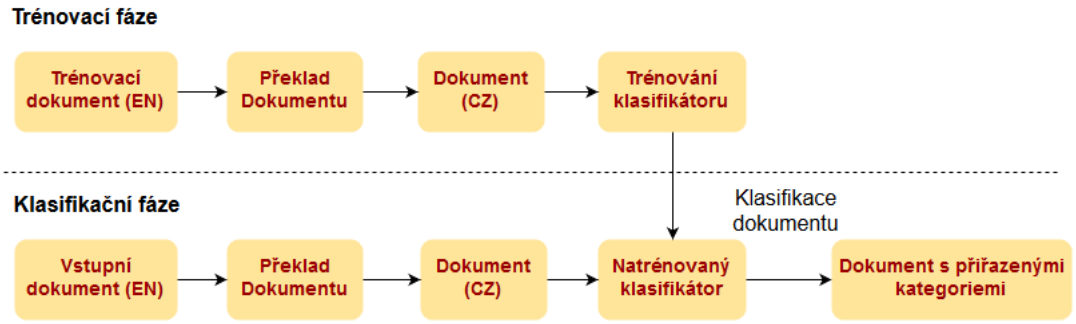
Pro klasifikaci anglických dokumentů od ČTK jsou opět zvoleny metody binárního sjednocení a prahování, které jsou vzájemně porovnávány. K porovnání jsou zvoleny dva rozdílné testy:

- V první části se porovnávají dva odlišné přístupy ke klasifikaci na jedné kolekci. Prvním je překlad dokumentu do češtiny a jeho následná klasifikace, která je znázorněna na obrázku 7.3. Druhým je obecná reprezentace dokumentu pomocí LDA, která nevyžaduje překlad, ale zkoumá podobnost mezi dokumenty pomocí kosinové podobnosti. Průběh klasifikace metodou LDA je znázorněn na obrázku 7.4.
- V druhé části se porovnávají dva odlišné přístupy ke klasifikaci pro dvě odlišné kolekce. Kolekce anglických dokumentů je použita jako testovací, zatímco model je natrénován českou kolekcí od ČTK. Prvním přístupem je překlad dokumentu do češtiny a jeho následná klasifikace, která je znázorněna na obrázku 7.5. Druhým je obecná reprezentace dokumentu pomocí LDA, která nevyžaduje překlad, ale zkoumá podobnost mezi dokumenty pomocí kosinové podobnosti. Průběh klasifikace metodou LDA je znázorněn na obrázku 7.6.

[1] http://www.statmt.org/wmt07/shared-task.html

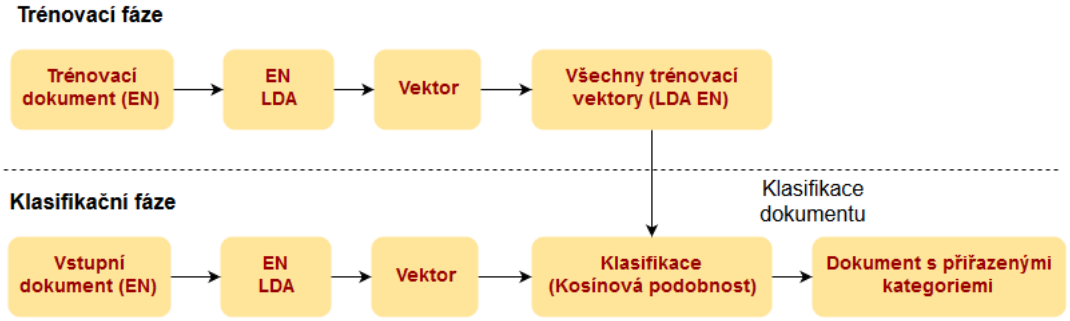
7.3.1 Testování pro jednu kolekci

Na obrázku 7.3 je znázorněn průběh klasifikace, při použití varianty s překladem a následnou klasifikací. Anglická kolekce slouží jako trénovací a zároveň testovací.



Obrázek 7.3: Varianta s překladem dokumentů a následnou klasifikací (pro jednu kolekci)

Na obrázku 7.4 je znázorněn průběh klasifikace, při použití varianty s metodou LDA. V této variantě anglická kolekce slouží jako trénovací a zároveň testovací. Trénovací fáze spočívá v uložení všech vektorů transformovaných z trénovacích dokumentů metodou LDA při použití anglického modelu. Při klasifikaci jsou příznaky získány porovnáním vektoru vstupního dokumentu se všemi vektory trénovacích dokumentů. Vstupní dokument je převeden na vektor použitím LDA anglického modelu.



Obrázek 7.4: Varianta s použitím metody LDA (pro jednu kolekci)

Následující klasifikace je provedena binárním sjednocením viz tabulka 7.4. Pro tuto klasifikaci byl minimální počet trénovacích dokumentů pro kategorii stanoven na 50 a to vzhledem k celkovému počtu dokumentů, který je oproti kolekci s českými dokumenty poloviční. Při této klasifikaci bylo k dispozici celkem 7 kategorií.

Překlad	Metrika	Metoda	
		ME	SVM
Google	Přesnost	0,893	0,899
	Úplnost	0,820	0,801
	F-míra	0,855	0,847
Moses	Přesnost	0,880	0,895
	Úplnost	0,820	0,796
	F-míra	0,849	0,843
LDA	Přesnost	0,740	0,754
	Úplnost	0,739	0,730
	F-míra	0,739	0,742

Tabulka 7.4: Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA

Vzhledem k tomu, že dokumenty v této kolekci patří vždy jen do jedné kategorie, při klasifikaci prahováním není třeba určovat vhodnou hodnotu. Místo toho stačí dokument zařadit do kategorie, která má nejvyšší pravděpodobnost. Při takové klasifikaci bude přesnost a úplnost vždy stejná. Neboť F-míra je pouze harmonický průměr přesnosti a úplnosti. Proto má smysl uvádět při klasifikaci prahováním pouze hodnotu f-míry. Klasifikaci prahováním zobrazuje tabulka 7.5.

Překlad	Metoda	
	ME	SVM
Google	0,866	0,850
Moses	0,861	0,846
LDA	0,749	0,739

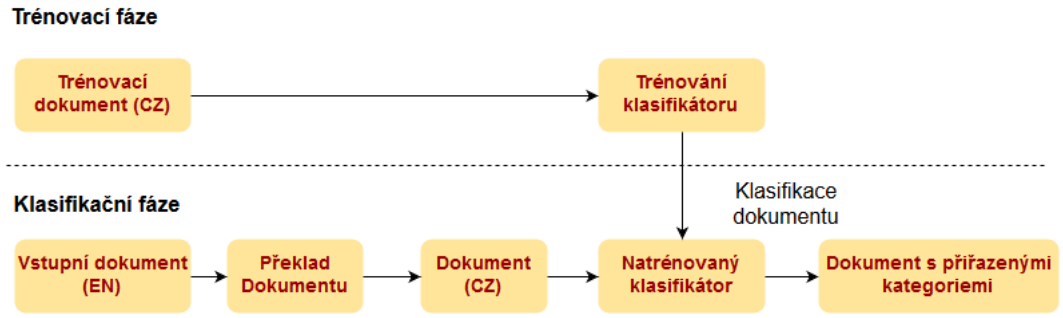
Tabulka 7.5: Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA

7.3.2 Kategorie namapované na českou kolekci

Následující testy se snaží porovnat, jak kvalita překladu ovlivňuje klasifikaci a zda není lepší reprezentace dokumentu pomocí metody LDA. Pro tento test byly katego-

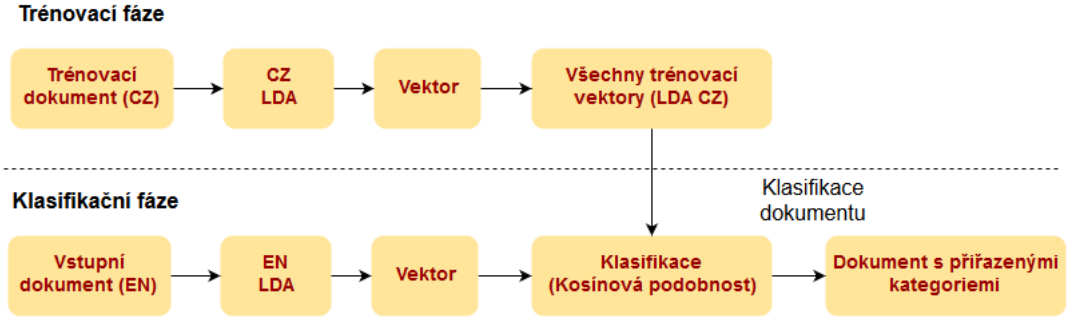
rie z anglické kolekce použité v předchozích experimentech namapovány na kategorie v české kolekci od ČTK, aby si co nejpřesněji odpovídaly. Kategorie v anglické kolekci jsou ale dosti obecné na rozdíl od českých, které jsou velmi specifické, proto nemusí být namapování úplně přesné. Pro získání přibližných výsledků k vytyčeným cílům je ale namapování dostačující.

Na obrázku 7.5 je znázorněn průběh klasifikace, při použití varianty s překladem a následnou klasifikací. Pro tento test byla celá česká kolekce od ČTK použita jako trénovací a byly na ní natrénovány klasifikátory. Anglické dokumenty sloužily pouze jako testovací.



Obrázek 7.5: Varianta s překladem dokumentů a následnou klasifikací (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů)

Na obrázku 7.6 je znázorněn průběh klasifikace, při použití varianty s metodou LDA. Česká kolekce od ČTK slouží jako trénovací, zatímco kolekce anglických dokumentů jako testovací. Trénovací fáze spočívá v uložení všech vektorů transformovaných z trénovacích dokumentů metodou LDA při použití českého modelu. Při klasifikaci jsou příznaky získány porovnáním vektoru vstupního dokumentu se všemi vektory trénovacích dokumentů. Vstupní dokument je převeden na vektor použitím LDA anglického modelu.



Obrázek 7.6: Varianta s použitím metody LDA (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů)

Pro tyto testy byly zvoleny dva rozdílné scénáře:

- V prvním byly všechny kategorie, které se nevyskytovaly v testovacích datech ponechány.
- Zatímco ve druhém scénáři byly kategorie, které se nevyskytovaly v testovacích datech odstraněny.

První scénář - všechny kategorie ponechány

V tomto scénáři testu byly ponechány i kategorie, které se nevyskytují v testovacích datech. Klasifikace tak bude znatelně ovlivněna a jsou očekávány horší výsledky. Nejprve je provedena klasifikace binárním sjednocením viz tabulka 7.6.

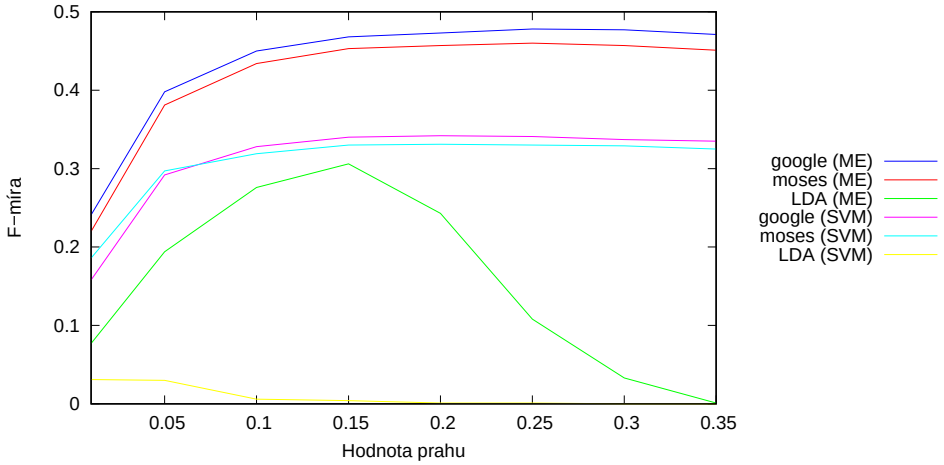
Překlad	Metrika	Metoda	
		ME	SVM
Google	Přesnost	0,385	0,510
	Úplnost	0,436	0,380
	F-míra	0,409	0,436
Moses	Přesnost	0,382	0,546
	Úplnost	0,401	0,352
	F-míra	0,391	0,428
LDA	Přesnost	0,018	0,088
	Úplnost	0,041	0,291
	F-míra	0,025	0,135

Tabulka 7.6: Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

Dále je klasifikace provedena prahováním viz tabulka 7.7. Výsledky f-míry pro klasifikaci jsou zobrazeny v grafu 7.7.

Metoda	Metrika	Hodnota meze pro prahování							
		0,010	0,050	0,100	0,150	0,200	0,250	0,300	0,350
Google (ME)	Přesnost	0,147	0,303	0,387	0,434	0,464	0,494	0,515	0,528
	Úplnost	0,660	0,580	0,538	0,508	0,482	0,463	0,445	0,425
	F-míra	0,241	0,398	0,450	0,468	0,473	0,478	0,477	0,471
Moses (ME)	Přesnost	0,131	0,282	0,366	0,418	0,450	0,481	0,502	0,517
	Úplnost	0,677	0,587	0,532	0,494	0,464	0,441	0,420	0,400
	F-míra	0,220	0,381	0,434	0,453	0,457	0,460	0,457	0,451
LDA (ME)	Přesnost	0,040	0,120	0,234	0,370	0,513	0,597	0,699	0,500
	Úplnost	0,859	0,500	0,339	0,261	0,159	0,059	0,017	0,000
	F-míra	0,077	0,194	0,276	0,306	0,243	0,108	0,033	0,001
Google (SVM)	Přesnost	0,092	0,224	0,295	0,333	0,357	0,373	0,382	0,407
	Úplnost	0,556	0,418	0,371	0,347	0,327	0,313	0,302	0,284
	F-míra	0,158	0,292	0,328	0,340	0,342	0,341	0,337	0,335
Moses (SVM)	Přesnost	0,115	0,239	0,289	0,320	0,337	0,353	0,361	0,375
	Úplnost	0,486	0,392	0,357	0,341	0,324	0,310	0,301	0,286
	F-míra	0,186	0,297	0,319	0,330	0,331	0,330	0,329	0,325
LDA (SVM)	Přesnost	0,019	0,019	0,004	0,003	0,001	0,001	0,000	0,001
	Úplnost	0,085	0,076	0,009	0,005	0,001	0,001	0,000	0,000
	F-míra	0,031	0,030	0,006	0,004	0,001	0,001	0,000	0,000

Tabulka 7.7: Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)



Obrázek 7.7: F-míra při klasifikaci prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

Druhý scénář - kategorie nevyskytující se v testovacích datech nepoužity

Následují testy pro druhý scénář testu, ve kterém nebyly použity kategorie, které se nevyskytují v testovacích datech. Klasifikace tak nebude ovlivněna a jsou očekávány lepší výsledky, než v předchozím scénáři. Klasifikaci binárním sjednocením zobrazuje tabulka 7.8.

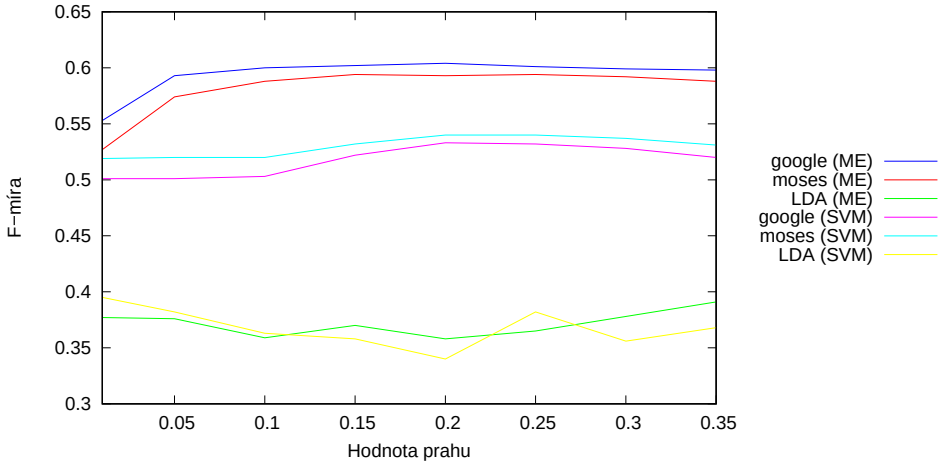
Překlad	Metrika	Metoda	
		ME	SVM
Google	Přesnost	0,635	0,669
	Úplnost	0,498	0,451
	F-míra	0,558	0,539
Moses	Přesnost	0,625	0,682
	Úplnost	0,458	0,435
	F-míra	0,529	0,531
LDA	Přesnost	0,086	0,200
	Úplnost	0,071	0,081
	F-míra	0,078	0,115

Tabulka 7.8: Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

Klasifikace pro druhý scénář je provedena také prahováním viz tabulka 7.9. Výsledky f-míry pro klasifikaci jsou zobrazeny v grafu 7.8.

Metoda	Metrika	Hodnota meze pro prahování							
		0,010	0,050	0,100	0,150	0,200	0,250	0,300	0,350
Google (ME)	Přesnost	0,450	0,536	0,572	0,592	0,607	0,617	0,626	0,634
	Úplnost	0,718	0,663	0,632	0,612	0,600	0,586	0,575	0,565
	F-míra	0,553	0,593	0,600	0,602	0,604	0,601	0,599	0,598
Moses (ME)	Přesnost	0,413	0,504	0,549	0,573	0,588	0,602	0,613	0,620
	Úplnost	0,728	0,665	0,633	0,616	0,599	0,586	0,572	0,559
	F-míra	0,527	0,574	0,588	0,594	0,593	0,594	0,592	0,588
LDA (ME)	Přesnost	0,407	0,409	0,391	0,402	0,389	0,397	0,411	0,426
	Úplnost	0,351	0,348	0,332	0,342	0,331	0,337	0,349	0,362
	F-míra	0,377	0,376	0,359	0,370	0,358	0,365	0,378	0,391
Google (SVM)	Přesnost	0,401	0,408	0,414	0,471	0,517	0,541	0,554	0,578
	Úplnost	0,668	0,649	0,642	0,585	0,549	0,523	0,505	0,473
	F-míra	0,501	0,501	0,503	0,522	0,533	0,532	0,528	0,520
Moses (SVM)	Přesnost	0,445	0,456	0,461	0,502	0,538	0,557	0,564	0,580
	Úplnost	0,622	0,605	0,596	0,566	0,542	0,524	0,512	0,491
	F-míra	0,519	0,520	0,520	0,532	0,540	0,540	0,537	0,531
LDA (SVM)	Přesnost	0,346	0,331	0,330	0,342	0,316	0,383	0,387	0,398
	Úplnost	0,459	0,452	0,403	0,377	0,368	0,381	0,329	0,342
	F-míra	0,395	0,382	0,363	0,358	0,340	0,382	0,356	0,368

Tabulka 7.9: Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

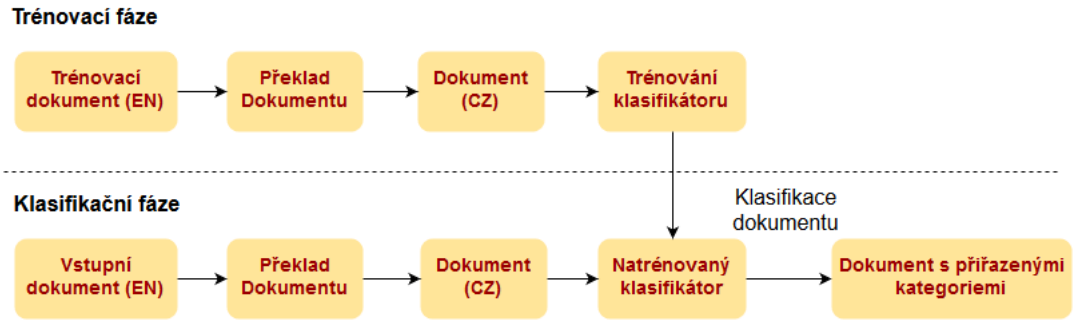


Obrázek 7.8: F-míra při klasifikaci prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

7.4 Klasifikace anglických dokumentů od Reuters

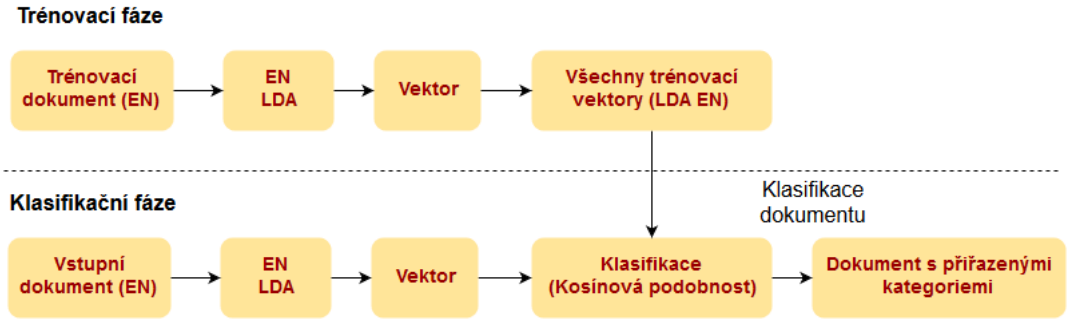
Pro klasifikaci anglických dokumentů z kolekce Reuters jsou opět zvoleny metody binárního sjednocení a prahování, které jsou vzájemně porovnávány. Během těchto testů se porovnávají dva odlišné přístupy ke klasifikaci na jedné kolekci. Prvním je překlad dokumentu do češtiny a jeho následná klasifikace. Druhým je obecná reprezentace dokumentu pomocí LDA, která nevyžaduje překlad, ale zkoumá podobnost mezi dokumenty pomocí kosinové podobnosti.

Tuto kolekci se nepodařilo namapovat na českou kolekci, a proto byly provedeny testy pouze na jedné kolekci. Na obrázku 7.9 je znázorněn průběh klasifikace, při použití varianty s překladem a následnou klasifikací. Anglická kolekce slouží jako trénovací a zároveň testovací.



Obrázek 7.9: Varianta s překladem dokumentů a následnou klasifikací (pro jednu kolekci)

Na obrázku 7.10 je znázorněn průběh klasifikace, při použití varianty s metodou LDA. Anglická kolekce slouží jako trénovací a zároveň testovací. Trénovací fáze spočívá v uložení všech vektorů transformovaných z trénovacích dokumentů metodou LDA při použití anglického modelu. Při klasifikaci jsou příznaky získány porovnáním vektoru vstupního dokumentu se všemi vektory trénovacích dokumentů. Vstupní dokument je převeden na vektor použitím LDA anglického modelu.



Obrázek 7.10: Varianta s použitím metody LDA (pro jednu kolekci)

Klasifikaci binárním sjednocením zobrazuje tabulka 7.10. Minimální počet trénovacích dokumentů pro kategorii byl stanoven na 100. Při této kvalifikaci bylo k dispozici celkem 16 kategorií.

Překlad	Metrika	Metoda	
		ME	SVM
Google	Přesnost	0,880	0,891
	Úplnost	0,789	0,759
	F-míra	0,832	0,820
Moses	Přesnost	0,873	0,895
	Úplnost	0,794	0,754
	F-míra	0,832	0,818
LDA	Přesnost	0,591	0,598
	Úplnost	0,594	0,599
	F-míra	0,592	0,599

Tabulka 7.10: Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA

Tato kolekce obsahuje dokumenty, které mohou patřit pouze do jedné kategorie. Vzhledem k tomu, že dokumenty v této kolekci patří vždy jen do jedné kategorie, při klasifikaci prahováním není třeba určovat vhodnou hodnotu. Místo toho přiřadíme dokument kategorii, která má nejvyšší pravděpodobnost. Klasifikaci prahováním zobrazuje tabulka 7.11.

Překlad	Metoda	
	ME	SVM
Google	0,841	0,823
Moses	0,848	0,818
LDA	0,594	0,562

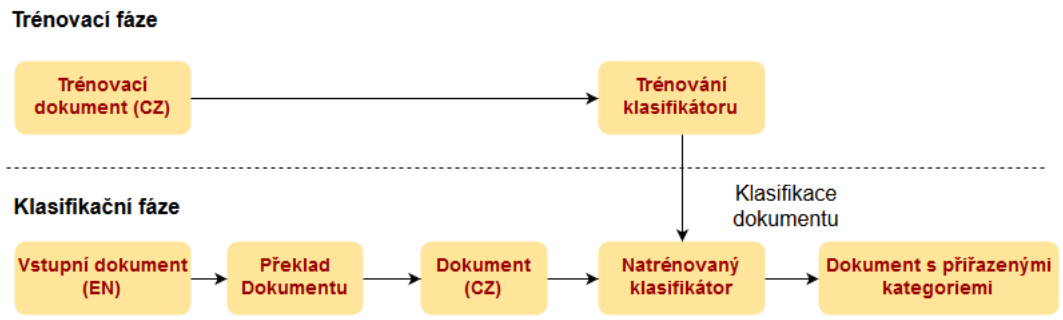
Tabulka 7.11: Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA

7.5 Klasifikace anglických dokumentů od BBC

Pro klasifikaci anglických dokumentů z kolekce BBC jsou opět zvoleny metody binárního sjednocení a prahování, které jsou vzájemně porovnávány. Během těchto testů se porovnávají dva odlišné přístupy ke klasifikaci na dvou rozdílných kolekcích. Prvním je překlad dokumentu do češtiny a jeho následná klasifikace. Druhým je obecná reprezentace dokumentu pomocí LDA, která nevyžaduje překlad, ale zkoumá podobnost mezi dokumenty pomocí kosinové podobnosti.

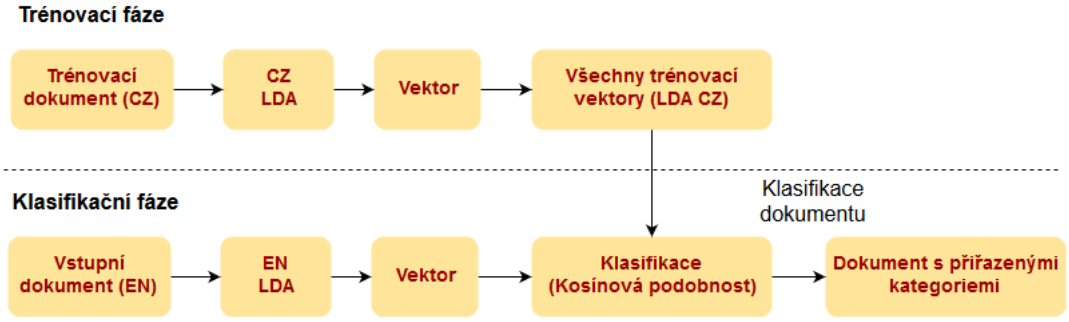
Díky úspěšnému namapování kategorií na české dokumenty od ČTK se bude jednat o testy pro naučený model z českých dat a testy pouze pro jednu kolekci nebudou provedeny. Dokumenty od BBC mají obecné kategorie, kterých je celkem pět. Namapováno bylo všech pět kategorií, přičemž každá byla namapována na jednu kategorii z českých dat.

Pro testy byla česká kolekce od ČTK použita jako trénovací a byly na ní natrénovány klasifikátory. Anglické kolekce sloužila pouze jako testovací. Na obrázku 7.11 je znázorněn průběh klasifikace, při použití varianty s překladem a následnou klasifikací.



Obrázek 7.11: Varianta s překladem dokumentů a následnou klasifikací (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů)

Na obrázku 7.12 je znázorněn průběh klasifikace, při použití varianty s metodou LDA. Česká kolekce od ČTK slouží jako trénovací, zatímco kolekce anglických dokumentů jako testovací. Trénovací fáze spočívá v uložení všech vektorů transformovaných z trénovacích dokumentů metodou LDA při použití českého modelu. Při klasifikaci jsou příznaky získány porovnáním vektoru vstupního dokumentu se všemi vektory trénovacích dokumentů. Vstupní dokument je převeden na vektor použitím LDA anglického modelu.



Obrázek 7.12: Varianta s použitím metody LDA (trénovací kolekce - české dokumenty od ČTK, testovací kolekce - kolekce anglických dokumentů)

Pro tyto testy byly připraveny dva testovací scénáře:

- V prvním byly všechny kategorie, které se nevyskytovaly v testovacích datech ponechány.
- Zatímco ve druhém scénáři byly kategorie, které se nevyskytovaly v testovacích datech odstraněny.

První scénář - všechny kategorie ponechány

V tomto scénáři testu byly ponechány i kategorie, které se nevyskytují v testovacích datech. Klasifikace tak bude znatelně ovlivněna a jsou očekávány horší výsledky. Nejprve je provedena klasifikace binárním sjednocením viz tabulka 7.12.

Překlad	Metrika	Metoda	
		ME	SVM
Google	Přesnost	0,390	0,500
	Úplnost	0,567	0,526
	F-míra	0,462	0,512
Moses	Přesnost	0,379	0,516
	Úplnost	0,480	0,445
	F-míra	0,423	0,478
LDA	Přesnost	0,032	0,048
	Úplnost	0,102	0,351
	F-míra	0,049	0,085

Tabulka 7.12: Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

Dále je klasifikace provedena prahováním viz tabulka 7.13.

Překlad	Metoda	
	ME	SVM
Google	0,519	0,373
Moses	0,480	0,299
LDA	0,226	0

Tabulka 7.13: Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

Druhý scénář - kategorie nevyskytující se v testovacích datech vyhozeny

Následují testy pro druhý scénář testu, ve kterém nebyly použity kategorie, které se nevyskytují v testovacích datech. Klasifikace tak nebude ovlivněna a jsou očekávány lepší výsledky, než v předchozím scénáři. Klasifikace binárním sjednocením viz tabulka 7.14.

Překlad	Metrika	Metoda	
		ME	SVM
Google	Přesnost	0,682	0,770
	Úplnost	0,633	0,629
	F-míra	0,656	0,692
Moses	Přesnost	0,614	0,701
	Úplnost	0,573	0,524
	F-míra	0,593	0,600
LDA	Přesnost	0,295	0,223
	Úplnost	0,136	0,320
	F-míra	0,186	0,263

Tabulka 7.14: Srovnání klasifikace binárním sjednocením pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

Klasifikace je také provedena prahováním viz tabulka 7.15.

Překlad	Metoda	
	ME	SVM
Google	0,707	0,701
Moses	0,648	0,650
LDA	0,432	0,380

Tabulka 7.15: Srovnání klasifikace prahováním pro metodu s překladem anglických dokumentů a metody LDA (model dokumentů natrénován českou kolekcí)

7.6 Zhodnocení výsledků

Během testování se neosvědčila klasifikační metoda Naive Bayes, která poskytovala velmi špatné výsledky. Výsledky této metody proto nejsou uvedeny v této práci. Proto testování proběhlo využitím metod ME a SVM. Tyto metody během testů poskytovaly uspokojivé výsledky, které se ale v některých případech lišily poměrně výrazně.

Z variant pro klasifikaci cizojazyčných dokumentů předčilo klasifikování po strojovém překladu klasifikaci použitím LDA. LDA dosahovala horších výsledků a nelze ji pro tuto úlohu doporučit. Metoda klasifikace po strojovém překladu dosahovala solidních výsledků, které lze považovat za vhodné k aplikačnímu nasazení. Při použití klasifikace binárním sjednocením dosahovala lepších výsledků metoda SVM, pro kterou byl největší naměřený rozdíl f-míry 0,05. Naproti tomu při klasifikaci prahováním dosahovala lepších výsledků metoda Maximální entropie. Která dosáhla největšího rozdílu f-míry 0,181.

Kvalita překladu na klasifikaci měla očekávaný vliv a kvalitnější překlad dosahoval lepších výsledků, rozdíly ale nebyly příliš velké. Maximální dosažené rozdíly mezi výstupy byly pro f-míru 0,074. Z toho lze usoudit, že uspokojivých výsledků při klasifikaci mezi rozdílnými jazyky lze dosáhnout i za použití horšího překladu.

8 Závěr

Cílem této práce bylo prozkoumat možnosti automatické klasifikace cizojazyčných dokumentů. V práci byly rozebrány dva rozdílné přístupy ke klasifikaci. Prvním přístupem byl překlad za pomoci statistického strojového systému, pro který byly zvoleny dva rozdílné systémy. Dalším cílem pro tuto variantu bylo prozkoumat vliv kvality překladu na klasifikaci. Druhým přístupem bylo reprezentování dokumentu za pomoci obecné reprezentace využitím metody LDA, která dokáže porovnat podobnost dokumentů v různých jazycích.

V práci bylo provedeno velké množství experimentů. Nejprve byly otestovány metody použité pro klasifikaci na dokumentech v jednom jazyce. Tyto testy ukázaly, že metoda LDA s porovnáváním podobností dokumentů může sloužit pro klasifikaci, ačkoliv nedosahuje tak kvalitních výsledků, jako klasifikace po překladu.

Pro další experiment byly cizojazyčné kolekce namapovány na českou kolekci od ČTK, která sloužila jako trénovací kolekce. To bylo problematické, neboť kategorie kolekcí byly dosti obecné oproti české kolekci od ČTK, která měla velmi specifické kategorie. Poté následovaly testy pro oba přístupy, ve kterých LDA dosahovalo horších výsledků. Oproti tomu klasifikace po překladu poskytovala solidní výsledky. Testy dále ukázaly, že při klasifikaci prahováním dosahuje lepších výsledků metoda Maximální entropie. Zatímco při použití binárního sjednocení dosahuje lepších výsledku klasifikátor SVM.

Při klasifikaci s překladem byl dále testován vliv kvality překladu na klasifikaci, který potvrdil očekávání, neboť kvalitnější překlad dosahoval lepších výsledků. Nicméně i nepříliš kvalitní překlad dosahoval překvapivě dobrých výsledků. Pro klasifikaci cizojazyčných dokumentů tak lze doporučit variantu s překladem.

Pro další experimenty by bylo vhodné vytvořit kolekci v cizím jazyce, která by obsahovala dokumenty se stejnými kategoriemi jako trénovací dokumenty v českém jazyce. Ideálně tak, aby byly dokumenty do kategorií zařazené kvalifikovaným expertem. Takové kolekce, které by odpovídaly kategoriemi dokumentům od ČTK nebyly na internetu nalezeny.

Seznam zkratek

RBMT - Rule-based Machine Translation

SMT - Statistical Machine Translation

LDA - Latentní Dirichletova alokace (Latent Dirichlet allocation)

ME - Maximální entropie

SVM - Metoda podpůrných vektorů (Support Vector Machine)

API - Application Programming Interface

ČTK - Česká tisková kancelář

BBC - British Broadcasting Corporation

Literatura

[1] MANNING, Christopher D., et al. Introduction to information retrieval. Cambridge: Cambridge university press, 2008.

[2] Statistical Machine Translation [on-line]
Dostupné z URL: http://www.coli.uni-saarland.de/courses/late2-14/Slides/LT2%20-%20SMT%201%20Intro.pdf

[3] Costa-Jussa, Marta R., et al. "Study and comparison of rule-based and statistical catalan-spanish machine translation systems."Computing and Informatics 31.2 (2012): 245-270.

[4] Moses [on-line]
Dostupné z URL: http://www.statmt.org/moses/?n=Moses.Overview

[5] LAVIE, Alon. Evaluating the output of machine translation systems. AMTA Tutorial, 2010.

[6] LIN, Chin-Yew; HOVY, Eduard. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003. p. 71-78.

[7] BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. the Journal of machine Learning research, 2003, 3: 993-1022.

[8] LDA [on-line]
Dostupné z URL: http://fulltext.sblog.cz/2011/12/22/semanticka-analyza-textu-6/

[9] HUANG, Anna. Similarity measures for text document clustering. In: Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008. p. 49-56.

[10] TSOUMAKAS, Grigorios; KATAKIS, Ioannis. Multi-label classification: An overview. Dept. of Informatics, Aristotle University of Thessaloniki, Greece, 2006.

[11] CHÉRAGUI, Mohamed Amine. Theoretical Overview of Machine Translation. Proceedings ICWIT, 2012, 160.

[12] MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. Foundations of machine learning. MIT press, 2012.

[13] GOLLER, Christoph, et al. Automatic Document Classification-A thorough Evaluation of various Methods. ISI, 2000, 2000: 145-162.

[14] SEBASTIANI, Fabrizio. A tutorial on automated text categorisation. In: Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence. Buenos Aires, AR, 1999. p. 7-35.

[15] SHIMODAIRA, Hiroshi. Text Classification using Naive Bayes. Learning and Data Note, 2014, 7.

[16] SVM [on-line]
Dostupné z URL: http://www.iicm.tugraz.at/about/Homepages/cguetl/courses/isr/old/opt/classification/Support_Vector_Machine.html

[17] ZRIGUI, Mounir, et al. Arabic text classification framework based on latent dirichlet allocation. CIT. Journal of Computing and Information Technology, 2012, 20.2: 125-140.

[18] NIGAM, Kamal; LAFFERTY, John; MCCALLUM, Andrew. Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering. 1999. p. 61-67.

[19] PIETRA, Stephen Della; PIETRA, Vincent Della; LAFFERTY, John. Inducing features of random fields. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997, 19.4: 380-393.

[20] Brainy [on-line]
Dostupné z URL: http://home.zcu.cz/~konkol/Brainy/tutorial/documentation.html

[21] KOEHN, Philipp. Europarl: A parallel corpus for statistical machine translation. In: MT summit. 2005. p. 79-86.

[22] DUŠEK, Ondrej, et al. The Joy of Parallelism with CzEng 1.0.

[23] Vauquois Triangle [on-line]
Dostupné z URL: http://www.kennislink.nl/publicaties/automatisch-vertalen

[24] Google translate [on-line]
Dostupné z URL: https://cloud.google.com/translate/

A Systémy pro překlad

A.1 Moses

A.1.1 Instalace a trénování

Popis instalace a natrénování systému se nachází na stránce http://www.statmt.org/moses/?n=Moses.Baseline. Doporučena je instalace na operačním systému Linux. Systém byl testován na operačním systému Debian 8 v 64-bit verzi.

A.1.2 Překlad

Při překladu je nejprve nutné překládaný text připravit. Pro tyto kroky má Moses připravené skripty. Prvním krokem je tokenizace textu:

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en \
< dev/newstest2011.en > newstest2011.tok.en
```

Dalším krokem je truecasing, který vyžaduje jazykový model:

```
~/mosesdecoder/scripts/recaser/truecase.perl --model truecase-model.en \
< newstest2011.tok.en > newstest2011.true.en
```

Poté lze přeložit soubor:

```
~/mosesdecoder/scripts/training/filter-model-given-input.pl  \
filtered-newstest2011 mert-work/moses.ini ~/corpus/newstest2011.true.en \
-Binarizer ~/mosesdecoder/bin/processPhraseTableMin
```

A.1.3 Ohodnocení kvality překladu

Kvalita překladu je ohodnocena za použití metriky BLEU. Skripty implementující tuto metriku jsou součástí Moses instalace. K dispozici musí být přeložený text k ohodnocení a referenční lidský překlad. Oba soubory musí být tokenizovány a následně podrobeny truecasingu. Poté lze provést ohodnocení kvality následujícím příkazem:

```
~/mosesdecoder/scripts/generic/multi-bleu.perl \
-lc ~/corpus/newstest2011.true.cs              \
< ~/working/newstest2011.translated.cs
```

A.2 Google translate

Aplikace je vytvořena v jazyku C#, jako konzolová aplikace. Je snadno konfigurovatelná pomocí properties.file.

A.2.1 Překlad

Aplikace je navržena tak, aby překládala postupně jeden textový soubor. Funguje ve dvou režimech, prvním režimem je překlad řádek po řádku. Druhý režim je pro překlad dokumentů, pro který je v textovém souboru jeho id a zpráva. Pokud se překlad nepodaří, což se může výjimečně stát například kvůli timeoutu. Pak ukládá do souboru s chybami číslo dokumentu, pro který překlad selhal. Po dokončení je výsledkem soubor s překladem a soubor s chybami. Aplikace je konfigurovatelná za pomoci properties.file:

- mode - Označuje mód, který bude použit pro překlad. Povolené hodnoty jsou 1 a 2. 1 značí překlad textů bez indexů. 2 značí překlad dokumentů, které jsou rozlišeny indexem.
- inputFile - Soubor k překladu
- outputFile - Výstupní soubor obsahující přeložený text
- errorFile - Soubor obsahující indexy dokumentů, jejichž překlad se nezdařil
- targetLanguage - zkratka jazyka, do kterého má být překlad proveden
- sourceLanguage - zkratka jazyka, ve kterém je soubor k překladu

A.2.2 Ukázková konfigurace

```
<?xml version="1.0" encoding="utf-8" ?>
<configuration>
   <startup>
       <supportedRuntime version="v4.0" sku=".NETFramework,Version=v4.5" />
   </startup>
  <appSettings>
    <add key="mode" value="1" />
    <add key="inputFile" value="input.txt" />
    <add key="outputFile" value="translated.txt" />
```

```
    <add key="errorFile" value="errors.txt" />
    <add key="targetLanguage" value="cs" />
    <add key="sourceLanguage" value="en" />
  </appSettings>
</configuration>
```

B Aplikace pro klasifikaci

Aplikace pro klasifikaci je vytvořena v jazyku JAVA. Aplikace poskytuje všechny metody potřebné pro klasifikaci. Konfigurace experimentů je nastavitelná za pomoci properties.file. Nastavení pro několik experimentů je součástí odevzdaného archivu.

B.1 Konfigurace

- path - cesta do které se ukládají výsledky a odkud jsou načítány klasifikátory
- MIN_DOCS_FOR_TRAINING - minimální počet trénovacích dokumentů pro kategorii. Pokud kategorie neobsahuje daný počet, je trénovacím a testovacím dokumentům kategorie odstraněna.
- DIVIDE_SIZE - parametr pro křížovou validaci. Pro parametr 5 je provedeno 5 testů a data rozdělena na 5 částí.
- translate - určuje zda se jedná o klasifikaci z cizího jazyka. Možné hodnoty jsou true/false.
- google - určuje zda byl dokument přeložen Googlem. V případe false byl přeložen pomocí Moses. Tato hodnota je nastavována z důvodu, že google překládá vždy jen jeden soubor, zatímco Moses překládá dokumenty soubor po souboru. Což je nutné zohlednit při načítání. Možné hodnoty jsou true/false.
- loadTranslated - určuje, zda se mají načíst přeložené soubory při klasifikaci z cizího jazyka. Možné hodnoty jsou true/false.
- method - určuje typ klasifikace, která má být použita. Možné hodnoty jsou threshold, class, lda.
- classificatorType - určuje metodu, která bude použita při klasifikaci. Možné hodnoty jsou ME, SVM, bayes.
- thresholdValue - určuje hodnotu meze při klasifikaci prahováním.
- testsForOneCollection - určuje zda jsou testy provedeny pouze pro jednu kolekci. Možné hodnoty jsou true/false. Při true je k testování a trénování použita pouze jedna kolekce. Při false je k trénování použita kolekce českých dokumentů od ČTK a pro testování je vybrána kolekce zvolená dle parametru collection. Pro testovací kolekci jsou kategorie namapovány na kolekci trénovací.
- removeOtherCategories - určuje zda při testech pro více kolekcí mají být odstraněny kategorie, které se nevyskytují v testovacích datech.
- collection - určuje kolekci, která má být použita. Možné hodnoty jsou ctk, reuters, bbc.
- threshold_only_best - určuje zda má být při klasifikaci prahováním dokument

zařazen pouze do kategorie, která má nejvyšší pravděpodobnost. Možné hodnoty jsou true/false. Při false je dokument zařazován pod kategorii, pokud je pravděpodobnost výskytu větší než stanovená prahovací mez.

- ldaClass - Bere se v potaz pouze při nastaveném parametru method LDA. Označuje zda pro LDA použít typ klasifikace binárního sjednocení. Při false je použit typ klasifikace prahováním. Možné hodnoty jsou true/false.
- ldaFile - cesta k natrénovanému česko-anglickému modelu

Spuštění klasifikace probíhá vhodným nastavením výše uvedených parametrů. Poté lze spustit jar soubor aplikace:

```
java -Xmx7256M -jar classification.jar
```

Důležitým parametrem je nastavení velikost paměti, kterou má aplikace k dispozici. To je provedeno použitím parametru -Xmx. Při trénování klasifikátoru pro LDA je například nutné až 17 GB paměti. Po proběhnutí klasifikace jsou výsledky dostupné dle definovaného parametru path v konfiguračním souboru, ve složce vypis. Po proběhnutí klasifikace jsou ukládány klasifikátory, které lze při opakované klasifikaci načíst bez nutnosti klasifikátor znovu trénovat. Trénování některých klasifikátorů je při velkém množství dat časově náročnou úlohou.

B.2 Ukázková konfigurace

- path=test/
- MIN_DOCS_FOR_TRAINING=100
- DIVIDE_SIZE=5
- translate=true
- google=false
- loadTranslated=false
- method=lda
- classificatorType=SVM
- thresholdValue=0.5
- cosine=true
- testsForOneCollection=false
- removeOtherCategories=false
- collection=ctk
- threshold_only_best=false
- ldaClass=false

- ldaFile=data/LDA_cz-en_100topics.bin