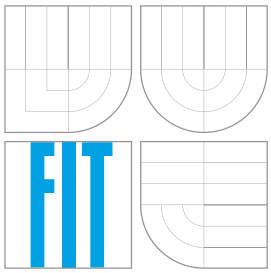
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

STROJOVÉ UČENÍ V OBLASTI STYLOMETRIE A URČOVÁNÍ AUTORSTVÍ

MACHINE LEARNING IN THE DOMAIN OF STYLOMETRY AND AUTHORSHIP ATTRIBUTION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

KAREL DRÁPELA

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2016

Abstrakt

Práce se zabývá identifikací autorů anglických internetových komentářů. Popisuje aktuální stav v oboru určování autorství na sociálních sítích. Vysvětluje fungování a strukturu vytvořeného systému na určování autorství, který funguje na základě výběru nejinformativnějších příznaků z převážně písmemnných n-gramů a slovních druhů. Prezentuje výsledky testování systému na internetových službách Quora a Twitter.

Abstract

Thesis deals with authorship attribution of english internet comments. It describes state of art in authorship attribution on social networks. It decsribes how the new system created during the work on this thesis functions. System is based on selection of most informative characteristics mostly from character n-grams and part of speech tags. It presents results of testing on comments from social networks Quora and Twitter.

Klíčová slova

určování autorství, strojové učení, výběr příznaků, quora, twitter

Keywords

authorship attribution, machine learning, feature selection, quora, twitter

Citace

DRÁPELA, Karel. *Strojové učení v oblasti stylometrie a určování autorství.* Brno, 2016. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Smrž Pavel.

Strojové učení v oblasti stylometrie a určování autorství

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana pana doc. RNDr. Pavla Smrže Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Karel Drápela
17. května 2016

Poděkování

Děkuji panu doc. RNDr. Smržovi Ph.D. za odbornou pomoc a vedení této bakalářské práce.

Obsah

Kapitola 1

Úvod

Určování autorství se dlouhodobě používá ve forenzní lingvistice k odhalování pachatelů trestných činů. V počátcích se jednalo hlavně o rukou psané listiny a dopisy. S rozvojem počítačů a internetu se otevřelo nové pole působnosti. Začaly se objevovat případy vydírání, kyberšikany a případy vytváření falešných účtů, kde všude je možné využít metody z oboru určování autorství pro identifikaci pachatele.

Cílem této práce je vytvořit systém na určování autorství pro textové komentáře na sociálních sítích. Sociální sítě představují jednu z nejrychleji se rozvíjejících oblastí internetu. Příspěvky na těchto službách bývají často velmi krátké, a to v řádu desítek či stovek znaků, což znamená nutnost využití jiných metod než například pro knihy nebo články.

Nejčastější úlohou systému určování autorství vytvořeného v rámci této práce je stanovení, která osoba z předem předvybrané skupiny autorů napsala analyzovaný text, což může např. posloužit příslušníkům bezpečnostních složek během vyšetřování trestných činů. Velikost množiny možných autorů se předpokládá několik desítek, případně stovek autorů. Na tuto velikost se v případě vyšších velikostí musí množina podezřelých uživatelů redukovat jinými prostředky (datum založení účtu, typ provinění...).

Text bakalářské práce je rozdělen do sedmi kapitol. Následující kapitola rozebírá vývoj určování autorství, kde je stručně shrnuta historie a současný stav oboru. Dále je zde popsána metoda strojového učení používaná v této práci a následuje popis vybraných metod pro výběr příznaků. Ve třetí kapitole je jsou popsány v současné době používané příznaky pro určování autorství. U každého typu příznaku jsou uvedeny princip metody, výhody, nevýhody a použití v jiných studiích o určování autorství.

Čtvrtá kapitola se zabývá systémem pro určování autorství, který byl vytvořen v rámci této práce. Je popsána architektura systému a používané typy příznaků. Dále je vysvětleno řešení několika problémů v této oblasti, jako je práce s nestejně dlouhými texty a s texty různých typů nebo závislost některých typů příznaků na tématu textu oproti čisté závislosti na stylu. Následuje pátá kapitola, kde je systém otestován na datové sadě. Je zde ukázáno, jak se vyvíjí přesnost v závislosti na množství dostupných textů a počtu kandidátních autorů. Dále je zde rozebráno, jak jsou jednotlivé typy charakteristik důležité pro funkčnost systému.

V závěrečné kapitole jsou shrnuty dosažené výsledky a zjištění prezentované v této práci a jsou navrženy další možné směry výzkumu a také vývoje systému.

Kapitola 2

Vývoj určování autorství

Určování autorství je obor, který se zabývá zjišťováním identity autora. Jedním z hlavních podoborů je identifikace autorství přirozeného textu. Dále v této práci se určováním autorství rozumí určování autorství textu. Určování autorství se snaží identifikovat autory, shlukovat je podle určitých kritérií nebo verifikovat, zda bylo určité umělecké dílo vytvořeno konkrétním člověkem.

2.1 Historie

Prvotní motivací pro rozvoj oboru určování autorství byla snaha zjistit autorství literárních děl, u kterých byla totožnost autora sporná nebo úplně neznámá. První pokusy začaly už v 19. století, kdy Menhendall pracoval s díly Bacona, Marlowa a Shakespeara (1887) a Mascol zkoumal autorství částí Nového zákona (1888). V počátcích základní myšlenkou byla snaha o nalezení jednoho příznaku, který dokáže samostatně charakterizovat autora. Tedy příznaku, který pro texty jednoho autora zůstává přibližně stejný, ale jeho hodnoty se liší u textů jiných autorů. Yule navrhl v roce 1944 délku vět, ale nakonec se tato metoda ukázala jako nespolehlivá. Byly navrženy i jiné příznaky jako bohatost slovní zásoby nebo délka slov. Žádný z těchto příznaků ale nebyl dostatečně diskriminativní, aby mohl sloužit jako jednotný invariant a v současnosti se tento přístup nepoužívá. [17]

V roce 1964 navrhli Mosteller and Wallace novou metodu pro určování autorství Listů federalistů, která kombinovala několik příznaků (jednola se o stovky funkčních slov). Klasifikace probíhala pomocí Naivního Bayesova klasifikátoru. Touto prací se ukázalo, že navrženou metodou jde spolehlivě a nezávisle na obsahu určovat autorství. Principem nového přístupu je reprezentace dokumentů jako bodů v prostoru a následně použití nějaké metody na měření vzdálenosti pro zařazení neznámého dokumentu k nejbližšímu autorovi.

V devadesátých letech nastal rozvoj strojového učení a to se projevilo i v této oblasti. Začaly se ve větší míře používat neuronové sítě, support vector machines a další podobné metody. Zde jsou jednotlivé texty reprezentovány číselnými vektory a učící algoritmus se snaží najít hranice mezi jednolivými autory. Následoval rozvoj metod, které sice generují velké množství příznaků jako n-gramy nebo sekvence slovních druhů, ale pomocí metod strojového učení je možné z nich vybrat diskriminativní složky a s nimi dále pracovat [24].

2.2 Určování autorství v současnosti

Postupem času se také posunula hlavní aplikace určování autorství. Většina literárních děl už byla důkladně prozkoumána a byly udělány závěry o jejich pravděpodobných autorech. Následně největší využití mělo určování autorství ve forenzní lingvistice, kde se posudky, zda určitý text psal nebo nepsal podezřelý, používají k odhalení pachatele. Původně toto dělali lidští odborníci, ale se zdokonalováním metod určování autorství v informatice se stále více využívá možností počítačů.

Příbuzným oborem je potom detekce plagiátorství. Nejedná se striktně o určování autorství, ale má s tímto oborem dost rysů společných. Význam má hlavně v akademickém světě, kde je důležité zajistit, aby jednotlivé práce byly originální a vědci neopisovali pasáže z jiných děl bez citace. Obecně se používají jednodušší a rychlejší algoritmy než pro určování autorství, protože často je potřeba aktuální text porovnat s velkým množstvím jiných prací v reálném čase. Většina algoritmů funguje na hledání podobných pasáží v jiných pracech a počítání míry jejich podobnosti.

V posledním desetiletí nastal velký rozvoj sociálních sítí, hlavně zásluhou Facebooku, a počty uživatelů těchto služeb se počítají ve stovkách milionů. Proto je potřeba řešit nové problémy, a to zejména spamboty [5], falešné účty, které je si možné koupit pro šíření nějaké agendy [8], nebo uživatele páchající trestnou činnost [11]. Důsledkem je potřeba určování autorství stále menších a menších úseků textu, které jsou typické pro vyjadřování u tohto typu služeb. Na jedné z nejpopulárnějších komunikačních (sociálních) služeb na internetu Twitteru je maximální délka příspěvku 140 znaků. Dalšími častými problémy se kterými se musíme vypořádat při klasifikaci textů na intenetu, jsou překlepy, gramatické chyby a časté používání zkratek.

Do oblasti určování autorství lze zařadit i profilování uživatelů. Zde není cílem zjistit, kdo daný text napsal, ale jaké jsou vlastnosti autora. Tento obor by se dal rozdělit na dvě skupiny. Prvním jsou psychologické atributy, kde se zkoumají povahové vlastnosti autora. Například zde se jedná o Jungovo dělení na introverta/extroverta nebo pětifaktorový model osobnosti známý pod názvem Velká pětka (otevřenost, svědomitost, extraverze, přívětivost, neuroticismus) [21]. Druhou skupinou jsou další údaje o člověku jako například věk, pohlaví nebo země původu [15].

2.3 Strojové učení

Před použitím metod strojového učení byl počet příznaků obvykle v jednotkách nebo desítkách. V dnešní době existují systémy, které používají příznaky v řádech desítek tisíc [4]. Obvykle se ale pohybují počty příznaků v desítkách až stovkách. Zvýšení dimenzionality reprezentace textu má za následek, že každý jednotlivý příznak nemusí mít vysokou diskriminační sílu pro konkrétní text. Metoda strojového učení sama vybere, které příznaky se nejlépe hodí pro tento konrétní případ a podle nich provede klasifikaci. Čím je příznaků více, tím je vyšší pravděpodobnost, že v ní budou i příznaky, které se hodí na konkrétní korpus, na kterém je prováděna klasifikace. Samozřejmě je nutné brát ohledy i na výpočetní možnosti dnešních počítačů. Při příliš vysokém množství příznaků se jejich extrakce a následná klasifikace stává příliš náročnou a nepraktickou.

2.3.1 Support vector machines

Support vector machines (dále SVM) je jednou z nejpoužívanějších metod strojového učení. Data reprezentuje jako body v n-rozměrném prostoru a snaží se najít nadrovinu v tomto prostoru, která optimálně rozděluje testovací data. Nadrovina data optimálně rozděluje v případě dvou tříd, pokud body z různých tříd leží v opačných poloprostorech a nadrovina je co nejdál od nejbližších bodů obou poloprostorů. [20]

Důležitý je výbět kernelu (česky "jádra", dále bude používán výraz "kernel"). Kernel umožňuje převést původně lineárně neseparovatelnou úlohu na úlohu lineárně separovatelnou převodem do vyšší dimenze. Výhodou kernelu je, že nemusí počítat přesnou pozici bodů v novém prostoru, a převod je proto relativně rychlý. [20] Základní typy kernelů jsou lineární a RBF (Gaussovský). Oba kernely jsou porovnány v kapitole 5.6. Předpokládá se, že bude vhodnější použít lineární kernel, protože při vyšším počtu příznaků je pravděpodobné, že problém je v onom vysoce dimenzionálním prostoru lineárně separovatelný a přesnost by tedy měla být srovnatelná s RBF. Lineární kernel je také obecně rychlejší.

2.4 Výběr příznaků

Jak již bylo zmíněno metody strojového učení často pracují s velkým množstvím příznaků. Konkrétně například u určování autorství použití n-gramů generuje spoustu příznaků. Podobně u klasifkace obrazových dat se často používá metoda, kde je každý pixel obrazu reprezentován jedním nebo více příznaky a výsledný vektor je tedy obsáhlý. Toto ale přináší i několik problémů. Hlavním problémem je přeučení (anglicky overfitting). Jedná se o stav, kdy se klasifikátor příliš dobře naučí učící sadu příkladů a tedy ji zvládne přiřazovat s velmi vysokou přesností, ale na úkor obecnosti modelu a to má za následek sníženou přesnost na testovací sadě (neviděných vzorcích dat). Klasifikátor poté není schopen zobecnit učící příklady na další neznámé vzorky a tedy nefunguje správně. Přeučení může nastat, pokud probíhá učení na trénovací sadě příliš dlouho nebo pokud není dostatek dat pro dostatečně komplexní natrénování klasifikátoru.

Právě při velkém množství příznaků (tisíce, desetitisíce..) je výsledný stavový prostor obrovský a proto většinou není k dispozici dostatek příkladů na naučení klasifikátoru a tedy dojde k přeučení.

Z důvodu výše zmíněných problémů byly vyvinuty různé metody pro výběr příznaků, aby se snížil počet potřebných příznaků na úspěšné naučení klasifikátoru. Tyto metody se dělí na filtrační (filters) a obalovací (wrappers). Filtrační jsou založené na statistických metodách a pomocí nich se snaží vybrat ty příznaky, které mají největší vliv na výsledné zařazení do tříd. Obalovací používají učící algoritmy pro zjištění, které příznaky jsou nejdůležitější pro správnou klasifikaci.

V následujícím textu budou vysvětleny principy čtyř metod výběru příznaků používaných v systému pro určování autorství, který bude představen v kapitole 5.1:

- Informační přínos (Information gain): Tato metoda měří změnu entropie, pokud se sada příkladů S rozdělí podle atributu A. Počítá se podle vzorce 2.1, kde $H(S)$ je entropie S, T tvoří podmnožiny t vzniklé rozdělením S podle A, $p(t)$ je podíl příkladů v t oproti původnímu počtu v S a $H(t)$ je entropie podmnožiny t. [16]

$$IG(A, S) = H(S) - \sum_{t \in T} p(t)H(t) \tag{2.1}$$

- Giniho nečistota (Gini impurity): Říká, jak často by byl vybraný text klasifikován nesprávně, pokud by byl výsledek klasifikace náhodně vybrán na základě rozložení tříd v trénovací sadě. Podobně jako metoda informační přínos rozděluje rozděluje příznaky podle atributů. [3]

- Chí-kvadrát (Chi-square): Snaží se eliminovat příznaky, které jsou nezávislé na jednotlivých třídách. Vzorec pro výpočet je 2.2. Chí-kvadrát měří nezávislost mezi příznakem t a kategorií c. A je počet příkladů kde se vyskytuje t a c současně. B je počet příkladů, kde se t vyskytuje bez c. C je počet příkladů, kde se c vyskytuje bez t. N je celkový počet příkladů. [25]

$$\chi^2(t,c) = \frac{N * (AD - CB)^2}{(A + C) * (B + D) * (A\ + B) * (C + D)} \tag{2.2}$$

- Rekurzivní výběr příznaků (Recursive feature selection): Funguje na principu výběru příznaků pomocí křížové validace. Klasifikátor se naučí se všemi příznaky. Poté se odstraní ty s nejmenším přínosem pro klasifikátor a znovu se provede učení. Toto se opakuje, dokud se nedosáhne cílového počtu příznaků.

Kapitola 3

Charakteristiky textu

Charakteristika (příznak) textu je hodnota, která nese nějakou informaci o daném textu. Právě správná kombinace charakteristik je základem k úspěšné klasifikaci textů. V následující kapitole budou představeny v současné době nejpoužívanější typy charakteristik. U každé bude popsán princip a poté výhody a nevýhody, které k ní vztahují.

3.1 N-gramy

N-gram je sled n po sobě jdoucích prvků z dané posloupnosti. N je kladné celé číslo. Pro $n = 2$ se používá speciální označení bigram a pro $n = 3$ trigram.

V počátcích se používaly pro určování autorství pouze frekvence výskytu předem daných slov nebo znaků. Nevýhodou tohoto postupu je, že nebere v úvahu kontext, ve kterém se slovo nebo znak nachází. Na tento problém odpovídají n-gramy, jejichž výhodou je, že nejen mohou sledovat frekvence sledovaných řetězců, ale přidávají i informaci, s jakými okolními řetězci se slovo nachází v kombinaci, tedy informaci o kontextu. Další výhodou n-gramů je jednoduchost jejich extrakce. Pro většinu komplexních stylistických ukazatelů, jako jsou například slovních druhy, jsou nezbytné speciální programy. Ty jednak nejsou dostupné pro všechny jazyky a také bývají oproti extrakci n-gramů z textu pomalé. Pro extrakci n-gramů stačí několik řádků kódu a extrakce je rychlá.

N-gramy jsou závislé na obsahu textu a tedy zachycují nejen autorův styl psaní, ale i téma, o kterém píše. Důležitým úkolem při použití n-gramů je minimalizovat vliv tématu ve vybrané sadě příznaků, protože téma textu se nijak netýká autorova stylu psaní. Příznaky, které jsou výsledkem využití n-gramů, jsou typické svou řídkostí (vysoké množství nulových hodnot). N-gramy také více než jednotlivé frekvence zachycují redundantní informace, protože každé místo v textu je zaznamenáno v několika n-gramech. Proto je důležité pomocí efektivních metod výběru příznaků vybrat pouze ty užitečné.

3.1.1 Písmenné n-gramy

Písmenné n-gramy nepracují s celými slovy, ale s jednotlivými písmeny v textu. Oproti slovním n-gramům zachycují hlavně krátká slova (například stopslova mívají obvykle krátkou délku), gramatické a typografické chyby, překlepy, interpunkci a další speciální znaky. Z těchto důvodů jsou písmenné n-gramy významným prostředkem pro určování autorství textů na internetu, které jsou typické právě krátkou délkou, větším množstvím chyb oproti jiným médiím a větším poměrem speciálních znaků k alfanumerickým znakům.

Písemné n-gramy jsou jednou z nejpoužívanějších metod pro určování autorství. Layton [14] úspěšně použil metodu SCAP založenou na písmenných n-gamech pro určování autorství příspěvků na Twitteru. Giraud [9] aplikoval písmenné n-gramy v kombinaci s dalšími charakteristikami pro zjištění autorství článků na internetových fórech. Koppel [17] zjistil, že písmené n-gramy mají nejvyšší úspěšnost na řadě typů textů (blogy, knihy, emaily). Sapkota [19] se domnívá, že vysoká úspěšnost písmenných n-gramů je způsobena jejich schopností zachytit stylistické i morfologické znaky textu.

3.1.2 Slovní n-gramy

Slovní n-gramy se používají pro zachycení slov a slovních spojení, které autor často používá. Jejich hlavním problémem je vysoká dimenzionalita výsledných příznaků. Z tohoto důvodu se nepoužívají obvykle slovní n-gramy pro $n > 3$. Pro vyšší hodnoty n je příznaků jednak příliš velké množství a také jsou jednotlivé příznaky příliš specifické a hrozí tedy, že nebudou fungovat na neznámým textech. Slovní n-gramy mají význam především u delších textů, jako jsou například knihy, kde je dostatek dat pro vysoké množství příznaků generovaných slovními n-gramy.

3.2 Stopslova

Stopslova jsou slova, která nenesou žádnou významnou sémantickou informaci a mají zpravidla pouze syntaktický význam. Typicky se jedná o spojky a předložky. Stopslova jsou důležitá pro konstrukci věty a spojení obsahových slov. Také jsou významným zdrojem příznaků v oblasti určování autorství. Jejich výhodou je, že se vyskytují v jakémkoliv typu textu o prakticky libovolné délce nebo tématu. To vyplývá z jejich nezbytnosti při stavbě věty. Počet různých druhů stopslov je v porovnání s obsahovými slovy velmi malý. Průměrný anglický rodilý mluvčí zná více než 100 000 slov, ale pouze 400 z nich jsou stopslova. Tato malá část slovníku ale tvoří významnou část každé věty [12].

Další výhodou je, že autoři i čtenáři věnují jejich použití menší pozornost než u obsahových slov [12]. Toto je důležité v případě, když chceme určit autora, který se snaží skýt, že je skutečným autorem textu. Pokud se snaží pachatel svůj styl psaní změnit, je u stopslov větší šance, že zde byl jeho styl psaní dotčen minimálně. Například Schindler v roce 1978 na tuto skutečnost ukázal jednoduchým eperimentem. Zjistil, že lidé často nenajdou všechny písmena „f“ v následujícím textu:

Finished files are the result of years of scientific study combined with the experience of many years.

Celkem jich je 6 a důvodem neúspěchu většiny lidí je právě slovo of, které je jedno z nejčastějších slov v anglickém jazyce, nenese žádný obsahový význam, a proto ho lidé podvědomě přeskakují. To je v souladu s dalšími výzkumy, kdy se například zjistilo, že lidé mají problém nalézt pravopisné chyby ve stopslovech [6]. V neposlední řadě je jejich výhodou právě jejich nízký obsahový význam. Díky tomu sa dají použít pro určování autorství nezávisle na tématu, o kterém autor píše.

3.3 Obsahová slova

Obsahová slova jsou definována tím, že mají informační hodnotu pro obsah textu. Jsou to podstatná jména, většina sloves a přídavných jmen. Při jejich použití se musí postupovat opatrně právě kvůli jejich závislosti na tématu textu. Při přílišné závislosti na obsahu

je poté systém pro určování autorství nepoužitelný, případně má výrazně horší výsledky v oblastech odlišných od té, na které byl systém natrénován. Samostatně se pro klasifikaci textů nepoužívají a uvádím je zde jako doplnění stopslov. Stále ale mají svůj význam, a to zejména u úplných (čistých) synonym, slov totožného významu, kde má autor volbu, které ze slov použije a jeho volba závisí právě na jeho stylu psaní. Například se jedná o dvojice pěkný a hezký nebo odvážný a statečný.

3.4 Slovní druhy

Slovní druhy jsou další populárním typem příznaků. Ty podobně jako slovní n-gramy pracuje se slovy, které převedou na odpovídající slovní druh. Tyto sekvence pak ukazují, jaké typy frází autor často používá. Pracují na obecnejší úrovni než slovní n-gramy. Výhodou je podobně jako u stopslov nezávislost na obsahu textu. Slovní druhy byly použity s poměrně vysokou úspěšností v mnoha studiích jak pro články[1], tak pro knihy [26].

Sledování frekvencí nebo n-gramů slovních druhů ve větě je jedním z nejjednodušších analýz syntaxe. Extrakce je oproti jiným metodám jako například tvorba derivačního stromu věty poměrně rychlá a v dnešní době existuje řada, nejenom anglických, parserů, které extrakci slovních druhů umožňují.

3.5 Další charakteristiky

V následující kapitole budou popsány charakteristiky, které se používají pro určování autorství jako podpůrné příznaky, ale sami o sobě nestačí pro klasifikaci s dostatečně vysokou úspěšností, aby mohly být používany samostatně.

3.5.1 Průměrná délka slov a vět

U průměrné délky slov jsou používány dva základní přístupy. První spočítá průměrnou délku slov v textu jako podíl počtu znaků a počtu slov. Druhým přístupem je vytvoření vektoru V o délce k a do každého prvku vektoru V_k se ukládá relativní četnost slov o počtu znaků k. Podle J. Grieva [10] je druhý způspob lepší pro určování autorství. Je to důsledek faktu, že první metoda zaznamená jen průměrnou délku, zatímco druhý způsob ukládá i histogram použití slov jednotlivých délek. Průměrná délka vět je velmi podobná průměrné délce slov. Jediný rozdíl je, že délku věty můžeme měřit buď počtem znaků nebo počtem slov. Zde bývá častější použít znaky, protože to dává přesnější výsledky oproti slovům, které mohou mít různou délku. Také je užitečné měřit počet slov ve větě.

3.5.2 Bohatost slovní zásoby

Tento příznak se snaží o vyjádření bohatosti slovní zásoby autora textu. Nejjednoduší způsob výpočtu je podíl počtu unikátních slov a počtu slov celkem. Tento způsob ale není považován za nejlepší, protože počet unikátních slov výrazně závisí na délce textu. Tato závislost ale není lineární: při krátkých textech roste počet unikátních slov rychleji než u delších textů [24]. Bylo navrženo velké množství funkcí, které počítají bohatost slovní zásoby. Žádná z nich není uznávána jako nezpochybnitelně nejlepší. V práci J. Grieva [10] dosahuje nejlepších výsledků vzorec 3.1, kde p_v je relativní četnost v-tého nejčastějšího slova.

$$Entropy = -100 * \sum p_v * \log p_v \tag{3.1}$$

3.5.3 Interpunkce

Interpunkce má ve stylometrii velký význam, protože frekvence používání jednotlivých interpunkčních znaků ukazuje na různé rysy stylu autora. Například pokud bude autor s vysokou frekvencí používat závorky, tak to může napovídat, že je často využívá na vysvětlení některých pojmů místo například vložené věty. Dále uvozovky jsou často užívané na vyjádření sarkasmu nebo velká frekvence čárek zase ukazuje na zálibu autora v dlouhých souvětích. [18]

3.5.4 Gramatické a typografické chyby

I jediná chyba, které se autor pravidelně dopouští, může mít velký vliv na přesnost určení textu, jelikož chyby jsou významné idiosynkrazie. Na principu hledání idiosynkrazií pracují i lidští experti při snaze určit autora neznámého díla [13]. Výhodou této skupiny příznaků je také skutečnost, že nezávisí na obsahu nebo typu textu. Gramatické chyby také umožňují identifikovat osoby, které nepíší ve svém mateřském jazyce.

Na druhou stranu je u těchto příznaků největší pravděpodobnost, že autor se je aktivně snaží odstranit, a proto chyba, podle které bylo možné určit autora před několika lety, nemusí být v současné době vůbec relevantní. Také mohou také vznikat nové chyby a proto je důležité mít k dispozici aktuální texty [7]. Také je nutno mít na paměti fakt, že i když se autor pravidelně dopouští určitého typu chyby, tak text může být natolik krátký, že se tato chyba ani neprojeví.

3.5.5 Specializované příznaky

Existuje řada dalších pomocných charakteristik, které mohou zpřesnit výsledný model. Často jsou specifické pro danou aplikační doménu – např. “průměrný počet hashtagů na tweet” pro určování autorství na Twitteru.

Kapitola 4

Výběr a zpracování testovacích dat

Důležitým krokem byl výběr zdroje textů do datové sady. Bylo nutné vybrat služby, kde je možné identifikovat účty určitého autora a stáhnout dostatečné množství textů. Dalším kritériem bylo, aby typ textů těchto služeb nebyl úplně stejný. V následujícím textu budou jednotlivé služby představeny.

4.1 Quora

Quora je internetová služba, která slouží k pokládání a zodpovídání otázek. Byla založena v roce 2009 dvěma bývalými pracovníky Facebooku a od té doby zažila prudký vzestup v popularitě. Rozsah témat je sice neomezený, ale převažují otázky na téma IT, internetu a startupů. Uživatelé mohou hlasovat pro odpovědi, o kterých si myslí, že jsou nejlepší. Příspěvky na Quoře mají velmi variabilní délku od jednoslovných či jednovětných odpovědí po celé články s vysvětlujícími obrázky a podpořené referencemi na odbornou literaturu. Výrazná většina uživatelů používá spisovný jazyk a vyjadřuje se v celých větách. Oblíbené internetové zkratky jako „LOL“ nebo „OMG“ se prakticky vůbec nevyskytují. Zkratky se používají většinou pouze pro odborné termíny jako například „SVM“ (support vector machines) nebo „NN“ (neural networks).

4.2 Twitter

Twitter je sociální síť a mikroblogovací služba, který umožňuje uživatelům posílat a číst příspěvky zaslané jinými uživateli, známé jako tweety. Tweet je textový příspěvek s maximální délkou 140 znaků. Uživatelé si mohou vybrat jiné uživatelé, které budou sledovat (follow) a poté se jejich tweety zobrazují na osobní stránce uživatele. Twitter je v současné době velmi důležitý marketingový nástroj pro řadu firem, osobností a celebrit. Umožňuje jednoduchou a přímočarou komunikaci s fanoušky nebo potenciálními zákazníky. Je jasné, že na Twiteru se vyskytují pouze krátké texty a z toho plynou typické charakteristiky textů na Twitteru:

- velké množství zkratek, a to i u normálně nezkracovaných slov kvůli úspoře místa v tweetu,
- jelikož uživatelé často používají twitter pro vyjádření své nálady, tak se často používají emotikony,

- velmi vysoké procento tweetů obsahuje URL odkaz buď na pokračování textu nebo na stránku o které referuje tweet,
- Speciální znaky: Twitter obsahuje 2 znaky, které mají na této službě specifický význam,
  - Hashtag (#téma) se používá pro shlukování nebo vyhledávání tweetů o stejném tématu,
  - Mention (@jméno) se používá pro komunikaci mezi uživateli.

4.3 Stažení a filtrace

Sběr dat proběhl v únoru 2016. Autoři byli vybráni ze seznamu nejsledovanějších uživatelů za rok 2013 na Quoře[1], protože u těchto uživatelů je pravděpodobnější, že budou mít větší množství příspěvků na obou službách než u průměrného uživatele. Následně byli vyfiltrováni ti, co neměli účet na Twitteru nebo ho dostatečně nepoužívali. Celkem bylo staženo 35000 příspěvků na Quoře a 50000 na Twitteru pro celkem 150 uživatelů. Každý vybraný uživatel měl minimálně 100 příspěvků na Quoře a 200 na Twitteru. Následně byly jednotlivé texty příspěvků předfiltrovány. Byly odstraněny opakující se konstrukce jako datum a čas příspěvku u textů z Quory nebo retweety (příspěvky od jiných uživatelů než je majitel účtu, který je sdílí na své osobní stránce) z Twitteru.

[1] https://quorabot.quora.com/The-most-followed-Top-Writers

Kapitola 5

Systém pro určování autorství

V následující kapitole bude vysvětlena architektura a postup práce systému pro určování autorství. Rozhodnutí o vybraných metodách a algoritmech budou ilustrována na experimentech, které porovnají jejich efektivitu.

5.1 Architektura systému

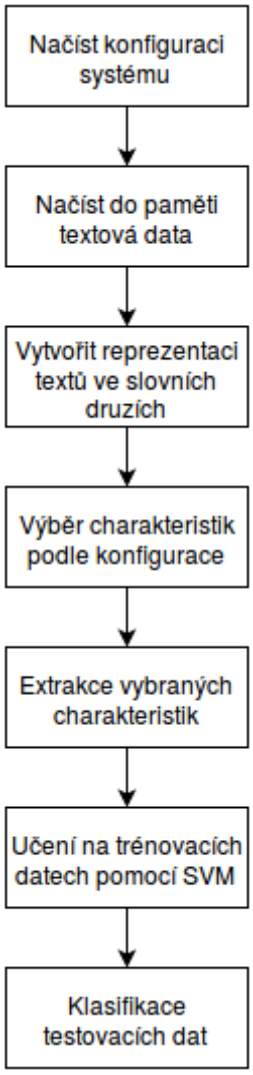
Na obrázku 5.1 je možno vidět schéma celého systému včetně načtení a vyhodnocení dat. Nejdříve je z konfiguračního souboru načteno nastavení. Následně se načtou data z příslušné složky, kde jsou uloženy texty jednotlivých autorů. Vytvoří se struktura v paměti, kde jsou uloženy všechny texty, rozdělené podle autora a služby, kde byly napsány. Text v kódování ASCII je relativně úsporný, a proto nebyl problém mít všechny texty uložené v paměti. Ovšem při příliš velkém množství textů by mohly nastat problémy a bylo by potřeba systém upravit, aby texty načítal postupně, například po autorovi. Každopádně i při klasifikaci pro 150 autorů, kde se používaly desetitisíce textů, nebyl s nedostatekem paměti žádný problém.

Po načtení do paměti je potřeba pro každé slovo najít jeho odpovídající slovní druh. Je to potřeba, jelikož některé kategorie charakteristik slovní druhy využívají. Převedená data jsou uložené do podobné struktury jako původní data.

Poté se provede výběr vhodných charakteristik k extrakci. Podrobněji bude tento proces popsán v kapitole 5.3. Vybrané charakteristiky jsou extrahovány z textu, spojí se do jednoho dokumentu a dají se na vstup SVM klasifikátoru. Ten se na datech natrénuje a následně klasifikuje testovací data. Poté jsou do výstupního souboru vypsány statistiky a výsledky pro jednotlivé texty v testovacích datech.

5.2 Použité knihovny

V průběhu tvorby systému jsem se snažil znovu neimplementovat funkce, které už jsou k dispozici ve veřejně dostupných knihovnách. Extrakce charakteristik je založena na knihovně Scipy, která poskytuje podporu pro vědecké výpočty v Pythonu a knihovně Scikit-learn což je knihovna pro programy využívající strojové učení. Knihovna Scipy umožňuje použití řídkých matic, které jsou optimalizovány pro ukládání řídkých vektorů. Díky tomu je paměťová náročnost relativně nízká. Pro převod textu na jeho reprezentaci ve slovních druzích byl použit Natural Language Toolkit (NLTK).



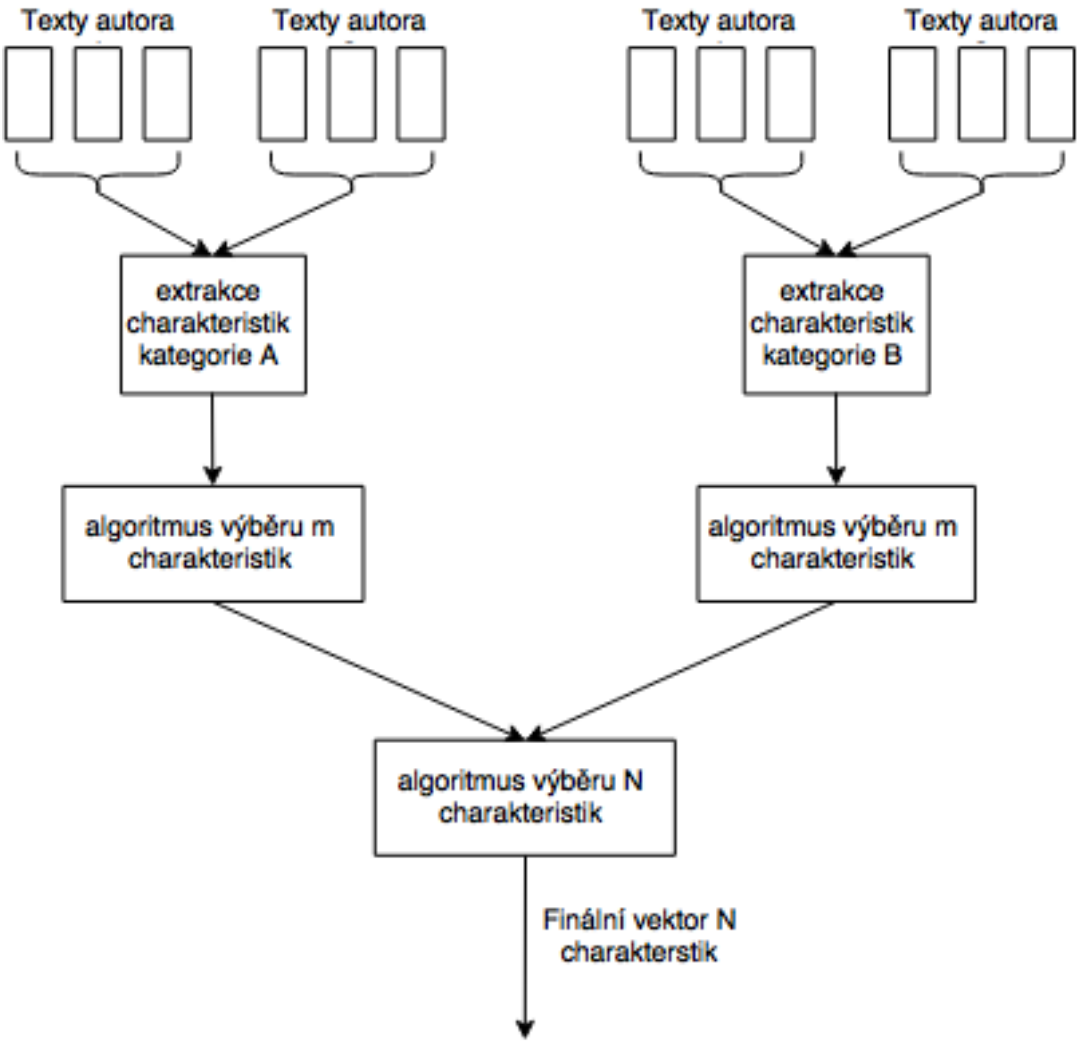
Obrázek 5.1: Postup práce systému

Pro stažení dat z Quory byl použit Selenium webdriver s driverem PhantomJS. Pro stažení dat z Twitteru byl vytvořen program postavený na API Twitteru a knihovně Tweepy, která usnaďnuje použití zmíněné API.

5.3 Systém výběru charakteristik

Jak už bylo zmíněno v kapitole 2.4, zejména v případě využití n-gramů není možné a ani efektivní použít všechny příznaky z důvodů přeučení a dlouhého běhu systému. Z tohoto důvodu jsem vyvinul dvojvrstvý systém filtrace charakteristik, který by měl zajistit, že se výsledný algoritmus strojového učení bude učit na těch nejinformativnějších příznacích.

Na obrázku 5.2 je znázorněno schéma algoritmu pro výběr charakteristik. Nejdříve se vezmou jednotlivé texty od všech autorů, kteří se mají klasifikovat a z jejich textů se extrahují příznaky náležející do jednotlivých kategorií. Používané kategorie pro tuto práci budou podrobněji popsány v následující kapitole. Schéma je ale obecné a tyto kategorie mohou představovat prakticky libovolnou třídu charakteristik. Také mohou být přidány či ode-



Obrázek 5.2: Výběr charakterstik

brány další kategorie. Po extrakci následuje filtrace, kde se vyberou pouze charakteristiky s minimálně L výskyty. Tento krok se provádí, aby se redukoval počet vstupních charakteristik do další fáze. Filtrace se provádí, protože velké množství charakteristik u n-gramů má počet výskytů v jednotkách nebo desítkách, což pro větší soubory dat má nízkou vypovídací hodnotu. Poté se pro každou kategorii vybere M nejpřínosnějších příznaků. M má obvykle pro každou kategoii jinou hodnotu a závisí na tom, kolik příznaků prošlo předchozí filtrací. Vybírá se přibližně 10% vstupních příznaků. Následně se spojí tyto vybrané charakteristiky dohromady a provede se další výběr, ale tentokrát už jen jeden a to pro všechny kategorie dohromady. Zde by se měly odstranit případné závislé příznaky mezi kategoriemi.

První výběr charakteristik má za úkol zúžit počet kandidátních příznaků, aby potom druhý výběr netrval příliš dlouhou dobu, jelikož i přes filtraci nejméně častých charakteristik jich často zbyde velké množství, pokud hodnota L není nastavena vysoko. Pokud je nastavena vysoko, tak hrozí, že přijdeme o cenné informace, jelikož filtrace pouze na základě pouze počtu výskytů je velmi hrubá metoda výběru a její hlavní výhoda je rychlost a jednoduchost.

5.4 Kategorie charakteristik

V následující podkapitole budou představeny typy charakteristik (kategorie), které byly použity klasifikaci. Systém pochopitelně umožňuje v závislosti na typu a množství klasifikovaných textů kategorie přidávat, odebírat nebo měnit.

5.4.1 Písmenné n-gramy

Písmenné n-gramy byly použity, protože bylo ve velkém množství porovnávacích studií prokázáno, že jsou jedním z nejefektivnějším typů charakteristik. [17] [24] [14] Dalším důvodem jejich použití byla skutečnost, že se velmi dobře hodí pro texty z Twitteru, které tvoří značnou část testovaných textů. Umí zachytit často používané zkratky, smajlíky a překlepy, které jsou typické pro komunikaci na Twitteru. Hodnota $n = 4$ bylo vybrána experimentálně. Alternativně bylo možné použít $n = 3$, kde byla lehce nižší úspěšnost klasifikace. N-gramy pro $n > 4$ byly moc specifické, generovaly velmi vysoké množství n-gramů s nízkou průměrnou frekvencí výskytu. Pro $n < 3$ byly zase příliš obecné. Tento poznatek je v souladu s dalšími pracemi. Například Stamatos [24] zjistil, že nejvhodnějším typem n-gramů pro anglické texty jsou tetragramy (4-gramy).

Jak už bylo zmíněno v kapitole 3.1.1, písmenné n-gramy jsou do určité míry závislé na obsahu textu. Pokud by byla tato závislost příliš velká, byl byl systém využívající tyto charakteristiky nepoužitelný. Proto byl proveden následující experiment, který má za úkol zjistit, jak velká tato závislost je a jestli je rozumné písmenné n-gramy použít pro určování autorství. Na stejných textech jsem provedl dva typy testů. První s nezměněnými texty s normálními písmennými n-gramy. V druhém testu byla odstraněna všechna obsahová slova (podstatná jména, slovesa, přídavná jména) a byla nahrazena reprezentací jejich slovního druhu. Rozdíl přesnosti klasifikace mezi těmito dvěma testy by měl ukázat jak významná je závislost na obsahových slovech pro klasifikaci.

Tabulka 5.1: Závislost písemných n-gramů na obsahu textu

	Přesnost klasifikace	**Autor je mezi pěti nejlepšími**
S nezměněným textem	60,3%	76%
S odstraněnými obsahovými slovy	41,1%	64%

Experiment je shrnut v tabulce 5.1. Je vidět, že písemné n-gramy jsou u tohoto typu textů závislé na obsahu z méně než jedné třetiny. Tento rozdíl byl posouzen jako dostatečně nízká závislost, aby bylo možné použít tento typ charakteristik.

5.4.2 Slovní druhy

Samotné frekvence slovních druhů poskytují jen omezenou informaci, protože není jasné jak se slov ze tvoří věty a fráze. Tento nedostatek je možné odstranit využitím n-gramů slovních druhů. Charakteristiky z této kategorie mají za úkol zachytit syntaktickou stavbu věty autora a jednoduché fráze, které autor často využívá.

5.4.3 Kombinace slovních druhů a stopslov

Tato kategorie má za účel kombinovat výhody slovních druhů a slovních n-gramů. Výhodou slovních druhů je nezávislost na obsahu textu, což je naopak problém u slovních n-gramů, kde určitá závislost je. Problémem slovních druhů je ale přílišná obecnost, kvůli které někdy nejde od sebe odlišit určité autory. U tohoto typu charakteristiky nechávám stopslova v textu a obsahová slova jsou nahrazeny prvním písmenem z anglického označení jeho slovního druhu (takže například „pes“ se nahradí „n“ – noun). Tato metoda zachovává nezávislost na obsahu, což je výhoda n-gramů slovních druhů, ale umožňuje být specifičtější na základě použitých stopslov. Takže je možné rozdělit dva autory, kteří používají stejnou frázi v rámci slovních druhů, ale jeden v ní používá jedno stopslovo a druhý jiné.

5.4.4 Ostatní

Tato kategorie není tak homogení jako předchozí. Byly do ní zařazeny všechny ostatní příznaky, které nejsou dostatečně významné, aby byly samostatně popsány. Vždy je nejdříve název charakteristiky a poté případně krátké vysvětlení, co daná charakteristika zachycuje a proč byla vybrána.

Obecné stylometrické příznaky:

- Frekvence alfanumerických znaků: podíl prostého textu.
- Frekvence nealfanumerických znaků: podíl speciálních znaků. Například pokud autor často používá smajlíky, tak tato hodnota bude vyšší než normálně.
- Frekvence číslic: ze stylometrického hlediska je důležité hlavně to, jestli autor používá. slovní vyjádření číslic nebo přímo číslici napíše. V budoucnu by se dal tento příznak vylepšit detekcí typu čísla (datum, částka..) a v jakém formátu je autor obvykle píše.
- Počet odstavců: jde o to zjistit, jaký má autor sklon dělit text na odstavce a jestli dělá spíše dlouhé odstavce nebo krátké.
- Frekvence velkých písmen: jednak se tímto příznakem zjistí, jestli autor používá celé věty (velké první písmeno) a také případné použití klávesy CAPS LOCK.
- Průměrná délka citací: pozorováním bylo zjištěno, že různí autoři mají sklon dělat různě dlouhé citace. Někteří tam vloží pouze jedno slovo, někteří zase spíše skupinu slov.
- Frekvence interpunkčních znamének: jak už bylo zmíněno v kapitole 3.5.3, interpunkční znaménka jsou prověřená charakteristika, která podává dobré výsledky.

Příznaky vět a slov: Následující tři charakteristiky mají za úkol zjistit, jak komplexní a dlouhé věty a slova autor tvoří a používá:

- průměrný počet slov ve větě,
- průměrná délka věty,
- průměrná délka slova.

Stopslova: Sleduje frekvenci 200 nejpoužívanějších stopslov v anglickém jazyce. Jsou zde jako doplňkové příznaky pro kategorii Kombinace slovních druhů a stopslov. Většina z těcto stopslov nebude ve výsledném vektoru, ale pokud autor používá nějaké stopslovo výrazně častěji než obvykle, tak to tento příznak nejlépe zachytí, protože n-gramy by mohli rozprostřít použití toho stopslova do mnoha příznaků a tato informace by zmizela nebo by její informační hodnota klesla.

Příznaky chyb: Následující charakteristiky se zaměřují na chyby (převážně typografické). Většinu chyb je obtížné zjistit prostým prohlížením textu a je nutné mít speciální parser nebo program na hledání chyb, proto jich je zde využito pouze několik, i když chyby jsou velmi užitečným příznakem:

- jak často dělá autor mezeru před čárkou,
- jak často dělá autor mezeru před tečkou,
- skupina teček o velikosti jiné než 1 nebo 3,
- nesprávné použití citace,
- opakované alfanumerické znaky,
- opakované nealfanumerické znaky.

Příznaky pro elektronické texty: Tyto příznaky jsou specifické pro texty na internetu a například pro knihy by neměly význam. Frekvence hashtagů a mentions jsou určené pro Twitter a pro klasifikaci na jiných službách mají mnohem menší význam, ale na Twitteru je informace, jak často používají tyto fráze, velmi užitečná pro určení autora:

- frekvence textových smajlíků,
- frekvence internetových zkratek: ze seznamu 50 nejčastěji používaných,
- frekvence hashtagů.

5.5 Délka textů

V datové sadě jsou texty ze dvou různých služeb, které se liší tím, jaké příspěvky na ně uživatelé píší. Patrné je to i u délky textů, kde příspěvky na Twitteru mají maximální délku 140 znaků a na Quoře často bývají příspěvky o délce stovek znaků. Toto je problém pro strojové učení, protože příspěvky od stejného autora budou mít výrazně odlišné hodnoty na textech z odlišných služeb a vede to k nesprávné klasifikaci. Řešením je poskytnout algoritmu informaci o délce textu. Byly vyzkoušeny následující metody pro řešení tohoto problému:

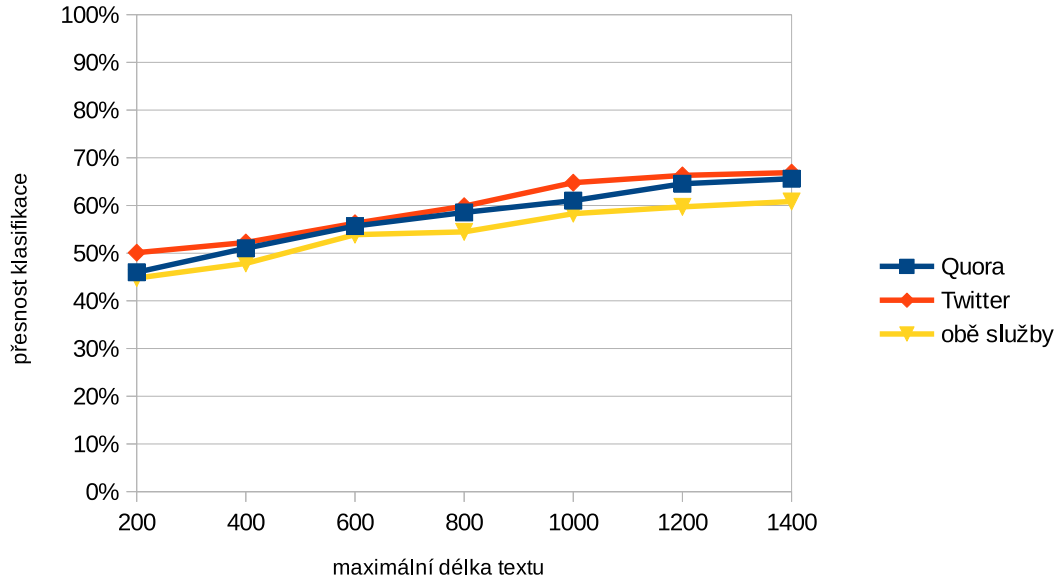
- Frekvence: jedná se o jednu z nejčastěji používanou metod. Každý příznak se vydělí celkovou délkou textu a z počtu výskytu příznaků se stane frekvence. Nevýhodou je, že se každý příznak ještě musí dělit a poté obvykle normalizovat do rozmezí (0, 1), což zabere určité kvantum strojového času.
- Spojení textů: jednotlivé texty se spojí do přibližně stejně velkých celků a následně se s nimi pracuje jako s jedním. Výhodou je zvýšení rychlosti. Jednak bude méně

textů, tedy je potřeba méně vektorů příznaků pro jejich klasifikaci, což vyústí v rychlejší naučení klasifikátoru a klasifikaci. Nevýhodou je potřeba větší množství textu, abychom měli k dispozici reprezentativní vzorek dat, na kterém je možné natrénovat klasifikátor.

- Stejná délka textů: všechny texty od stejného autora se spojí v jeden dokument a poté se tento dokument rozdělí na stejně velké části o určité délce (třeba 1000 znaků). Výhodou je, že všechny texty budou mít stejnou délku, takže hodnoty příznaků jsou přímo porovnatelné bez nutnosti dělení délkou textu nebo jinou normalizací. Nevýhodou je, že texty jsou rozděleny na nepřirozených místech, třeba i v půlce věty.

Normalizace délkou textu se ukázala jako neefektivní. Zbylé dvě metody jsou porovnány na datové sadě, kde na každého autora bylo 100 textů z Quory a 400 z Twitteru, aby se vyrovnal mnohem vyšší počet znaků při textech Quory a obě služby měli přibližně srovnatelnou reprezentaci. Přesnost klasifikace bez zásahů do původních dat je 45%.

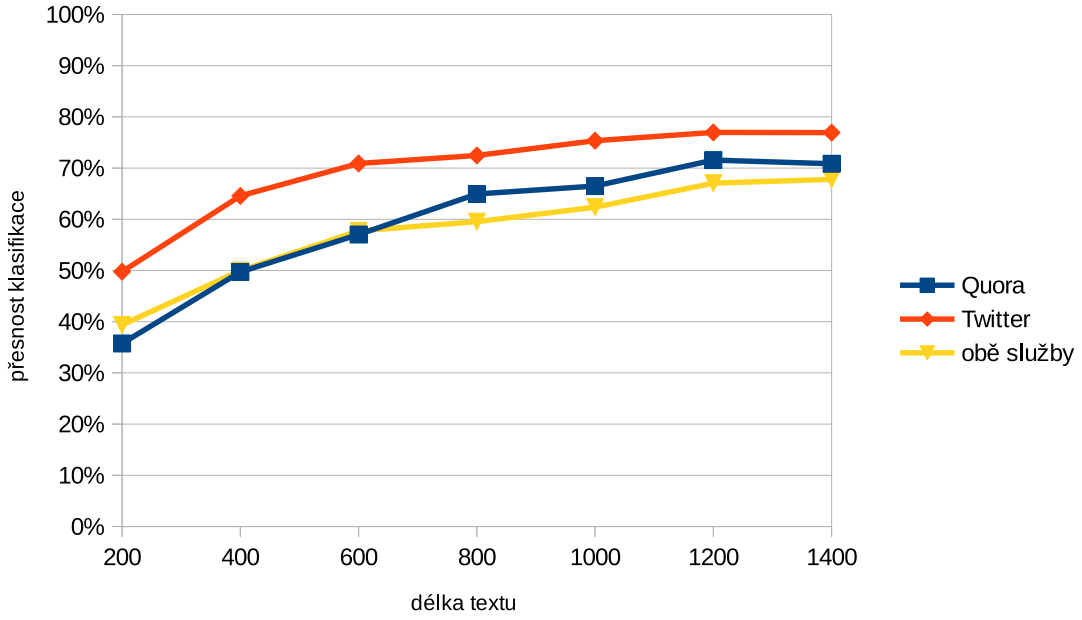
Na obrázku 5.3, je vidět jak se zvyšuje úspěšnost při spojování textů do větších celků. Při této metodě texty nebyly rozdělovány, takže všechny zůstaly celistvé. Texty byly spojovány, dokud nebyla dosažena určitá hranice maximálního počtu znaků. Hodnoty této hraniční veličiny jsou zobrazeny na ose x. Z grafu je vidět, že přesnost klasifikace se postupně mírně zvyšuje, ale nejedná se o žádný výrazný rozdíl. Twitter má vyšší úspěšnost, protože je k dispozici více příspěvků.



Obrázek 5.3: Spojení textů

Na obrázku 5.4 jsou znázorněny výsledky druhé metody. Všechny texty od jednoho autora byly spojeny do jednoho dokumentu a tento dokument byl poté rozdělen na přesně stejně velké části. Velikost jedné části je zobrazena na ose x. Při použití této metody se přesnost klasifikace výrazně zvyšuje a předčila předchozí metodu spojování textů a to už od délky textu rovné 400 znakům. Metoda je také výrazně lepší než přesnost při klasifikaci původních, nezměněných textů. Nevadí ani, že jeden původní příspěvek může být rozdělen na několik části a třeba i v půlce věty nebo slova. Nevýhodou je potřeba mít větší množ-

ství textů, aby bylo možné jich několik spojit a stále mít dostatek příkladů pro naučení klasifikátoru.



Obrázek 5.4: Stejná délka textů

Pro použití v systému jsem tedy zvolil rozdělení na stejně velké úseky o délce 1200 znaků. Z grafu je patrné, že po této hodnotě se už přesnost výrazně nezvyšuje.

5.6 Strojové učení

Pro učení z extrahovaných příznaků a následnou klasifikaci jsem zvolil SVM, jehož princip je popsán v kapitole 2.3. Vyzoušel jsem i jiné metody, jako například neuronové sítě, které ale vykazovaly lehce nižší přesnost klasifikace a vyšší čas potřebný pro naučení, nebo algoritmus k-NN, který měl výrazně nižší přesnost. Další výhodou SVM je jednoduchost použití a fakt, že není nutné dlouho optimilizovat jednotlivé parametry pro dobré výsledky.

Nejdůležitějšími parametry SVM je hodnota prametru C a zvolený kernel. Parametr C určuje poměr mezi snahou algoritmu minimalizovat chybu klasifikace a snahou zobecnit model. Pro lineární kernel byla zvolena hodnota $c = 500$, od které se už přesnost dále nezvyšovala. Na základě provedených testů je možné konstatovat, že výsledky přesnosti klasifikace jsou srovnatelné, ale RBF kernel má mnohem delší čas učení.

Kandidátní metody výběru příznaků byly představeny v kapitole 2.4. Nejlepší výsledky dávala metoda *informační přínos*. Proto je ve finální verzi systému používana tato metoda. Nicméně systém umožňuje použít všechny metody.

5.7 Ovládání a konfigurace

Program nemá GUI a veškerá komunikace s uživatelem probíhá přes terminál a konfigurační soubor. Konfigurační soubor je v hlavní složce s programem a jmenuje se *config*. Konfigurační soubor jsem se oproti argumentům programu v terminálu rozhodl využít z důvodu

vysokého počtu možných parametrů, které je možno nastavit a upravit tak chování systému. Příklad nastavení je možno vidět na obrázku 5.5. V první sekci jsou obecné parametry systému jako počet autorů a podíl textů, který se použije pro testování. Poté následují parametry pro jednotlivé kategorie charakteristik.

```
text_folder = ./data_1500/
text_unknown = ./unknown/
authors = 50
number_of_services = Twitter, Quora
train_test_split = 0.2
texts_to_extract = 400, 100
text_split_method = normal
LEN = 1200, 1200
feature_count = 1000
k_value = 0.75
selection_method = entropy
top = 5
repeat = 5

# Nasleduji parametry jednotlivych kategorii charakteristik.
pos, n=3, l=20, m=0.1
nva, n=3, l=20, m=0.1
char, n=4, l=20, m=0.1
misc
```

Obrázek 5.5: Konfigurační soubor

Průběh práce programu je potom možné sledovat na výpisech v terminálu. Následně se výsledky vypíší do výstupních souborů. Do každé složky s autorem, který byl klasifikován, se vypíše podrobný výpis, jakým způsobem byl každý text v testovací sadě vyhodnocen. Poté se ještě vypíší obecné statistiky jako celková přesnost klasifikace, čas běhu programu a další. Podrobnější popis je možno nalézt v příloze B.1.

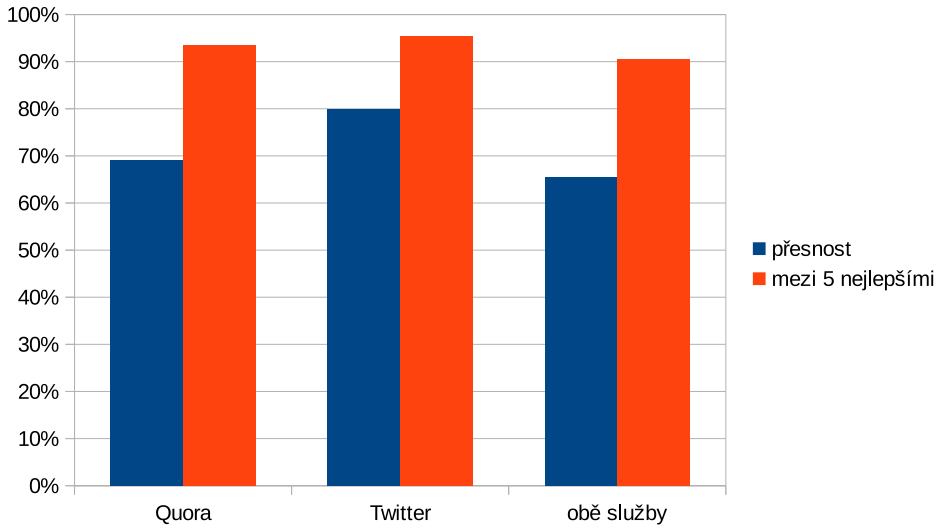
Kapitola 6

Výsledky a testování

V této kapitole budou ukázány a vyhodnoceny nejdůležitější vlastnosti systému. Nejdřív bude popsán experiment, dále budou v tabulce či grafu zobrazeny výsledky. Následovat bude komentář a vyhodnocení výsledků. Výsledkem klasifikace je žebříček kandidátních autorů s přiřazeným procentuálním ohodnocením jak je pravděpodobné, že daný autor je skutečným autorem textu. Součet všech ohodnocení je 100%. Přesnost značí poměr testovacích případů kdy byl skutečný autor určený správně (získal nejvyšší ohodnocení z kandidátních autorů). "Mezi X nejlepšími", kde X je číslo, značí jak často byl skutečný autor vybrán mezi X nejlépe ohodnocenými autory. Pokrytím se rozumí podíl klasifikovaných textů k celkovému počtu textů. Pokud není řečeno jinak, tak nastavení systému bylo následující: 50 autorů, maximálně 100 textů z Quory a 400 z Twitteru, spojení textů na stejnou délku 1200 znaků, 80% dat použito pro učení a zbytek pro testování.

6.1 Přesnost klasifikace

V grafu 6.1 je zobrazena přesnost systému pro jednotlivé služby a také pro obě současně při klasifikaci padesáti autorů.



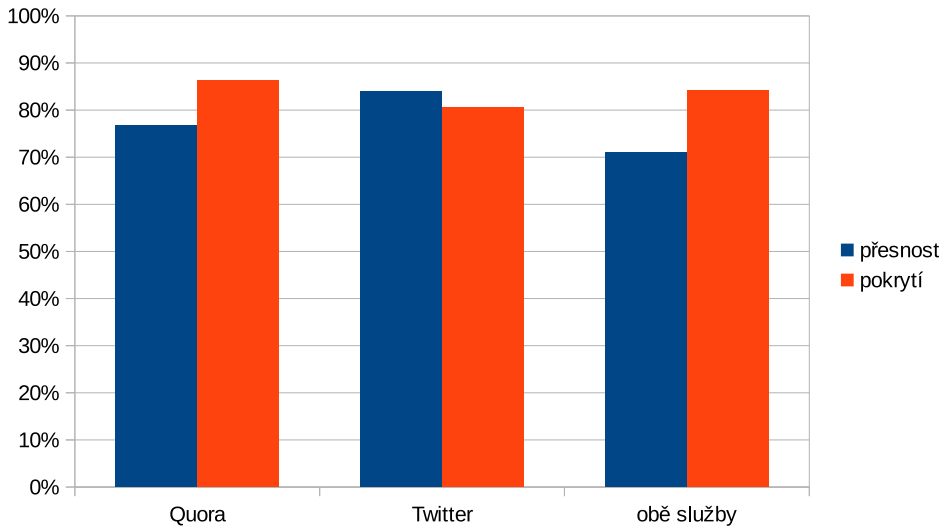
Obrázek 6.1: Přesnost klasifikace

Často se stává, že skutečný autor mezi kandidátními autory vůbec není. V takovém případě by byl systém nucen k nesprávné klasifikaci textu. Také v určitých případech systém nedokáže s dostatečnou mírou jistoty říci, který autor text napsal. Systémy pro určování autorství proto často mají možnost se rozhodnout text neklasifikovat.

Tuto funkcionalita byla implementována i v tomto systému. Výstup SVM sestává z vektoru procentuálních hodnot, které značí, jak je pravděpodobné, že daný text napsal příslušný autor. Samotná výše procentuální hodnoty závisí hlavně na počtu autorů. Proto jsem se rozhodl, kritériem (mírou jistoty), že vybraný autor je skutečným autorem analyzovaného textu, je porovnání nejpravděpodobnějších možností. Zjistí se to porovnání skóre autora s nejvyšší pravděpodobností (značen A) a druhou nejvyšším pravděpodobností (značen B). Text se nebude klasifikovat, pokud platí rovnice 6.1.

$$A * k < B. \tag{6.1}$$

Parametr k může nabývat hodnot v intervalu $(0, 1>$ a při hodnotě $k = 1$ budou všechny texty klasifikovány. Jako základní hodnotu pro systém byla vybrána hodnota $k = 0.75$, která byla experimentálně zvolena pro dobrou rovnováhu mezi přesností klasifikace a podílem klasifikovaných textů. Na grafu 6.2 je zobrazena přesnost se základní hodnotou $k = 0,75$.

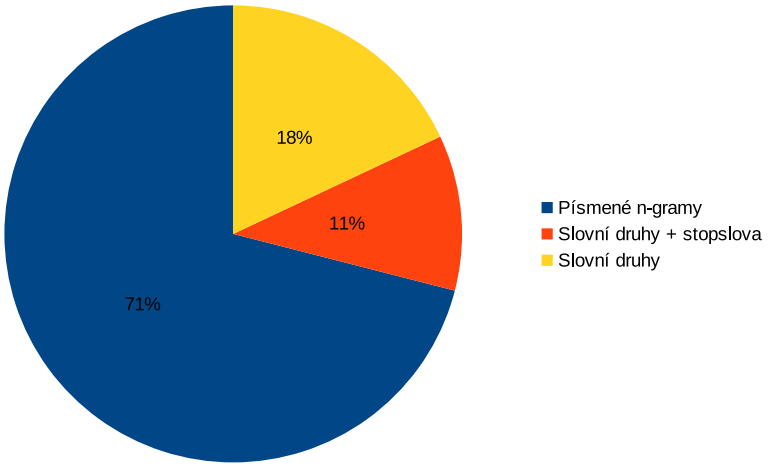


Obrázek 6.2: Přesnost klasifikace

6.2 Vybrané charakteristiky

Graf 6.3 zobrazuje, v jakém poměru jsou jednotlivé katogorie charakteristik vybírány do finální skupiny příznaků. Největší podíl mají podle očekávání písmenné n-gramy. Důvodem je hlavně vysoké množství informací, které zaznamenávají. Dále také určitě přispěla určitá závislost na tématu textu, která zvyšuje přesnost klasifikace tohoto typu charakteristik zhruba o třetinu.

Zbylé kategorie charakteristik plní úlohu spíše pomocného charakteru, ale obě mají stále dostatečně velký podíl, aby se vyplatilo je zařadit do výběru. Samotné slovní druhy mají vyšší podíl než kombinace se stopslovy, což je pro mě trochu překvapivé. Zřejmě je důvodem této vyšší úspěšnosti slovních druhů jejich větší obecnost a větší zaměření na

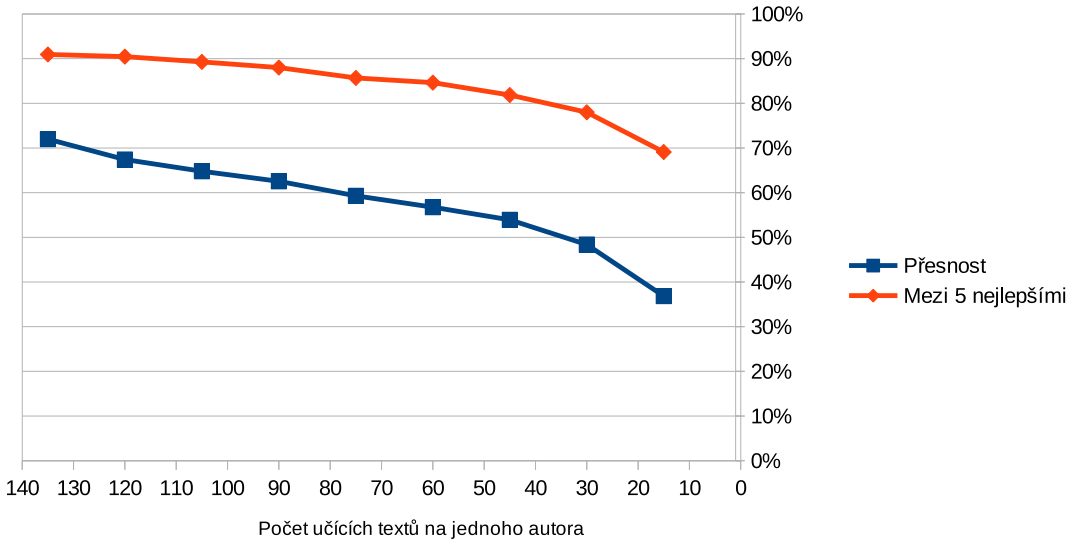


Obrázek 6.3: Podíly vybraných charakteristik

syntaktickou stránku textu, jelikož na lexikální stránku jsou zaměřeny písemné n-gramy. Vybrané charakteristiky pro texty z Quory a Twitteru se nijak výrazně neliší. Twitter má lehce vyšší podíl písemných n-gramů.

6.3 Poměr učících a testovacích dat

Systém rozdělí texty každého autora na učící a testovací sadu. Jedním z parametrů systému je hodnota, která určuje poměr tohoto rozdělení. V základní verzi jsou data rozdělována v poměru 80% na učení a 20% na klasifikaci. Tyto texty byly rozděleny do úseků textu o 1200 znacích. Celkem jich bylo 150 na jednoho autora.



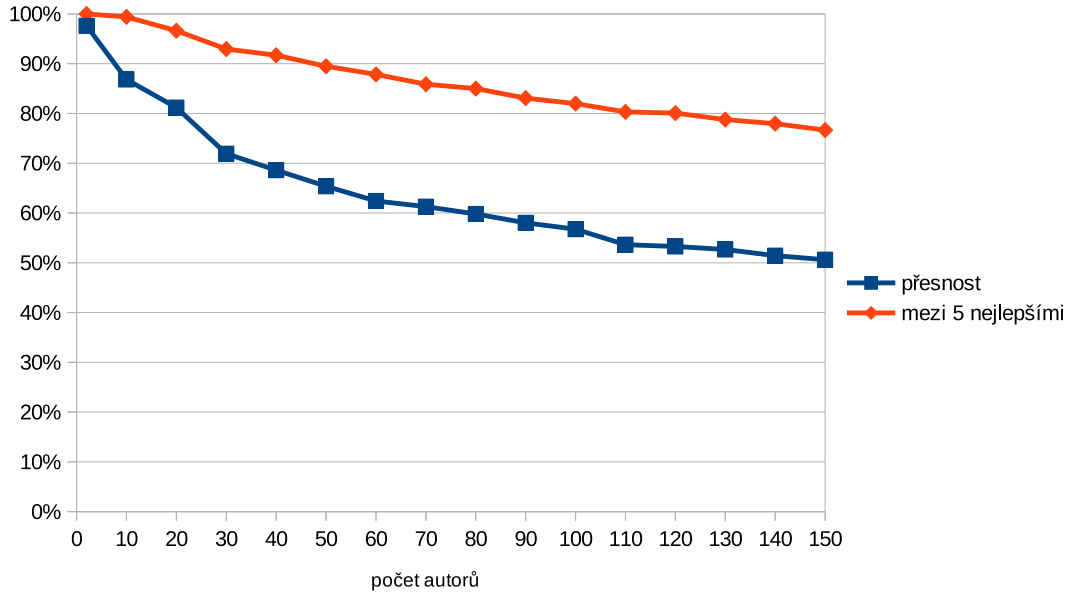
Obrázek 6.4: Poměr učících a testovacích dat

Z grafu je vidět, že přesnost určení se se snižujícím se počtem učících textů snižuje stále rychleji. U 15 textů na autora je přesnost určení pouze 45%. Tento systém tedy

není ideální pro určování autorství na velmi malých souborech dat, kde se počet textů na jednoho autora počítá v jednotkách, protože už nedokáže toto malé množství dat dostatečně zobecnit a naučit se styl autora. Případně by se musely udělat nějaké úpravy nastavení, aby byl systém pro malé množství textů lépe optimalizován.

6.4 Počet autorů

Čím více je možných autorů, mezi kterými se musí systém rozhodovat, tím je pochopitelně přesnost klasifikace nižší. Graf 6.5 zobrazuje závislost mezi přesností klasifikace a počtem možných autorů. Snižování přesnosti na této datové sadě se s přidáváním dalších autorů zpomaluje.



Obrázek 6.5: Přesnost klasifikace v závislosti na počtu autorů

6.5 Srovnání s jinými pracemi

V této sekci se pokusím srovnat přesnost mého systému se systémy v jiných studiích. Toto je obtížný úkol, protože každý studie si zvolila jinou datovou sadu, počet autorů, množství klasifikovaného textu. Vybral jsem tedy několik, které jsou s podmínkami mé datové sady srovnatelné. Srovávat jde pouze určování autorství tweetů, protože jsem nenašel žádné práce, které by prováděly klasifikaci na Quoře. V tabulce 6.1 je přehledné srovnání nejdůležitějších parametrů a výsledků prací, které jsem našel.

Vybral jsem studie, kde se pracovalo se stejným počtem autorů a přibližně stejným počtem tweetů na autora. Neexistuje žádný populární korpus pro Twitter (jako existuje pro knihy nebo novinové články), takže prakticky každá studie používá vlastní stažená data. Srovnání vychází z předpokladu, že náhodně stahované tweety v dostatečném množství jsou přibližně stejně náročné na klasifikaci, i když každý používá tweety jiné. Pro přesnost mého systému je použita přesnost při spojení více tweetů dohromady na délku 1200 znaků. Z

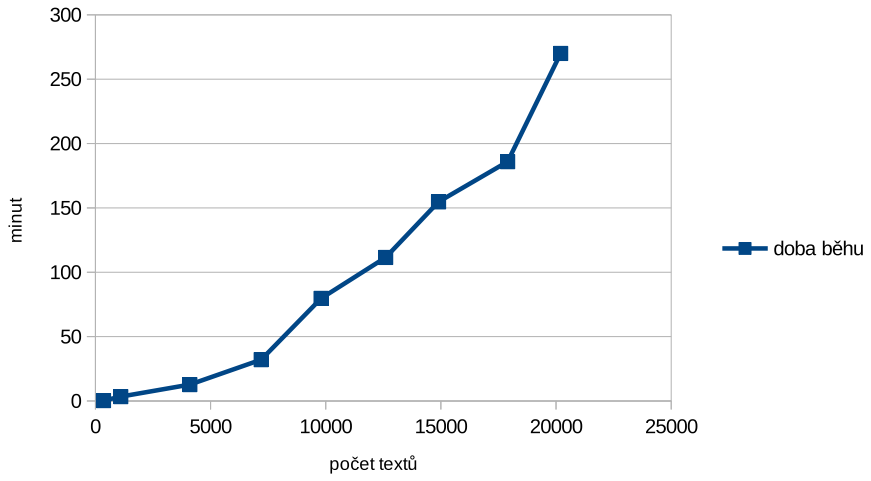
Tabulka 6.1: Srovnání s jinými pracemi

Autor (zdroj)	Přesnost	Počet autorů	Max. počet tweetů na autora
Můj systém	79%	50	400
Layton [14]	72%	50	200
Silva [23]	61%	50	250
Cavalcante [4]	55%	50	400
Schwartz [22]	60%	50	400
Boutwell [2]	48%	50	190

porovnání je vidět, že si systém z hlediska přesnosti vede dobře a to i přesto, že potřebuje 400 tweetů na jednoho autora, což je horní hranice v rámci porovnávaných systémů.

6.6 Škálovatelnost

U každého programu je také důležité, jak rychle dokáže předat výsledek. U zde prezentovaného systému tento čas závisí na počtu textů, které zpracovává, a velikost vektoru příznaků, kterým je každý jednotlivý text reprezentován. Na grafu 6.6 je zobrazena závislost počtu textů na době běhu systému. Pod pojmem "doba běhu systémumi" se rozumí čas od spuštění programu do doby než jsou vypsány výsledky. Každý text je v tomto případě reprezentován jako vektor o velikosti 1000. Každý text má konstatní délku 1200 znaků, tedy graf není zkreslen různými délkami. Jeden text tak odpovídá délce 2/3 normostrany. Testy byly prováděny na počítači s procesorem Intel Core i7-2670QM s frekvencí 2,2 GHz a kapacitou operační paměti 6 GB.



Obrázek 6.6: Závislost doby běhu systému na počtu zpracovávaných textů

Závislost je přibližně kubická, takže čas zpracování roste stále rychleji. To je způsobeno hlavně použitou SVM metodou, ve které s přibývající komplexností klasifikačního problému roste čas, který potřebuje na zpracování. Systém by se dal v případě potřeby zrychlit použitím méně příznaků, ale výrazně vyšší počet textů by buď přineslo větší snížení přesnosti klasifikace nebo příliš dlouhý běh systému.

Kapitola 7

Závěr

Tato práce se zabývá určováním autorství přirozeného textu v anglickém jazyce na sociálních sítích. V teoretické části práce jsou představeny nejpoužívanější typy charakteristik v současnosti a pro úplnost několik historických. Dále jsou popsány metody výběru příznaků používané v systému pro určování autorství a metoda strojového učení SVM, což je jedna z nejpoužívanějších metod pro klasifikaci textů.

Byly navrženy dvě metody spojení textů více komentářů do jednoho textu za účelem zvýšení přesnosti klasifikace. První je spojení jednotlivých textů do určité maximální délky. Druhá je spojení textů jednoho autora na přesně stejnou délku ve znacích. Obě metody překonaly v přesnosti klasifikace původní nezměněné texty. První metoda byla úspěšnější do délky jednoho textu 400 znaků, poté už byla druhá metoda lepší. Druhá metoda je tedy obecně lepší, ale vyžaduje větší množství dat.

Cílem práce bylo vytvořit systém pro určování autorství internetových komentářů. Systém funguje na principu výběru charakteristik pomocí metody "informační přínos" a následné klasifikace pomocí SVM. Typy charakteristik jsou převážně písemné n-gramy, stopslova a slovní druhy. Internetové komentáře byly staženy ze služeb Quora a Twitter. Tyto služby byly vybrány kvůli rozdílné délce a typu textů, které se na nich vyskytují.

Přesnost klasifikace na Quoře a Twitteru zároveň pro 50 autorů je 65% a pokud se systému povolí některé příklady neklasifikovat, tak při 84% pokrytí je přesnost 71%. Přesnost na Twitteru je při 400 textech na autora 79%, což je lepší výsledek než ve srovnávaných studiích v kapitole 6.5. Pro porovnání klasifikace na Quoře nebyly nalezeny žádné srovnatelné práce. Mezi pěti nejpravděpodobnějšími autory byl pro 50 možných autorů skutečný pisatel textu přítomen ve více než v 90% případů.

Systém je tedy možné použít pro identifikaci uživatelů na sociálních sítích. A v případě vhodného nastavení parametru k systému, aby byla dostatečná jistota, že výsledné zařazení je správné, i pro identifikaci osob podezřelých z trestné činnosti.

Dalšími možnostmi vývoje je přidání dalších typů charakteristik nebo převedení slov na základní tvar pomocí lemmatizéru. Dále by bylo užitečné stáhnout texty z dalších služeb (Facebook, LinkedIn, Wikipedia) a porovnat, jak se od sebe liší a jaké příznaky je nejlépe charakterizují.

Literatura

[1] Argamon-Engelson, S.; Koppel, M.; Avneri, G.: Style-based text categorization: What newspaper am I reading. In *Proc. of the AAAI Workshop on Text Categorization*, 1998, s. 1–4.

[2] Boutwell, S. R.: *Authorship attribution of short messages using multimodal features*. Diplomová práce, Naval Postgraduate School Monterey CA, University Circle 1, Monterey, Spojené státy americké, March 2011 [cit. 10.5. 2016], dostupné z: http://oai.dtic.mil.

[3] Breiman, L.: *Classification and regression trees*. New York, N.Y: Chapman & Hall, repr. vydání, 1993, ISBN 0412048418.

[4] Cavalcante, T.; Rocha, A.; Carvalho, A.: Large-Scale Micro-Blog Authorship Attribution: Beyond Simple Feature Engineering. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Springer, 2014, s. 399–407, doi:10.1007/978-3-319-12568-8_49.

[5] Crabb, E. S.; Mishler, A.; Paletz, S.; aj.: *HCI International 2015*, kapitola Reading Between the Lines: A Prototype Model for Detecting Twitter Sockpuppet Accounts Using Language-Agnostic Processes. Springer International Publishing, 2015, ISBN 978-3-319-21380-4, s. 656–661, doi:10.1007/978-3-319-21380-4_111.

[6] Drewnowski, A.; F. Healy, A.: Detection errors on *the* and *and*. *Memory & Cognition*, ročník vol. 5, č. issue 6, 1977: s. 636–647, ISSN 0090502x, doi:10.3758/BF03197410.

[7] Foster, D. W.: *Author unknown*. New York: Henry Holt, 2000, ISBN 0805063579.

[8] Fusilier, D. H.; Montes-y Gómez, M.; Rosso, P.; aj.: *Computational Linguistics and Intelligent Text Processing*, kapitola Detection of Opinion Spam with Character n-grams. Springer International Publishing, 2015, ISBN 978-3-319-18117-2, s. 285–294, doi:10.1007/978-3-319-18117-2_21.

[9] Giraud, F.-M.; Artières, T.: Feature Bagging for Author Attribution. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, editace P. Forner; J. Karlgren; C. Womser-Hacker, 2012, ISBN 978-88-904810-3-1, ISSN 2038-4963, s. 17–20.

[10] Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, ročník 22, č. 3, 2007: s. 251–270.

[11] Hidalgo, J. M. G.; Díaz, A. A. C.: Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification. In *CLEF 2012 Evaluation Labs and*

Workshop, Online Working Notes, 2012 [cit. 10.5.2016], s. 17–20, dostupné z: http://ceur-ws.org/.

[12] Kestemont, M.: Function Words in Authorship Attribution From Black Magic to Theory? *EACL 2014*, 2014: s. 59–66.

[13] Koppel, M.; Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, ročník 69, 2003, str. 72.

[14] Layton, R.; Watters, P.; Dazeley, R.: Authorship Attribution for Twitter in 140 Characters or Less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, IEEE, 2010, ISBN 9781424480548, s. 1–8, doi:10.1109/CTC.2010.17.

[15] López-Monroy, A. P.; Montes-y Gómez, M.; Escalante, H. J.; aj.: Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, ročník vol. 89, 2015: s. 134–147, ISSN 09507051, doi:10.1016/j.knosys.2015.06.024.

[16] Mitchell, T. M.: *Machine learning*. Boston: McGraw-Hill, 1997, ISBN 0070428077, 58 s.

[17] Moshe, K.; Jonathan, S.; Shlomo, A.: Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, ročník vol. 60, č. issue 1, 2009: s. 9–26, ISSN 15322882, doi:10.1002/asi.20961.

[18] Olsson, J.: Using groups of common textual features for authorship attribution. *Forensic Linguistics Institute*, 2006 [cit. 10.5. 2016], dostupné z: thetext.co.uk/authorship/authorship.doc.

[19] Sapkota, U.; Bethard, S.; Montes-y Gómez, M.; aj.: Not all character n-grams are created equal: A study in authorship attribution. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 2015, s. 93–102.

[20] Scholkopf, B.; Smola, A. J.: *Learning with kernels*. Cambridge, Mass.: MIT Press, 2002, ISBN 0262194759.

[21] Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; aj.: Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, ročník 8, č. 9, 09 2013: s. 1–16, doi:10.1371/journal.pone.0073791.

[22] Schwartz, R.; Tsur, O.; Rappoport, A.; aj.: Authorship Attribution of Micro-Messages. In *Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2013, s. 1880–1891.

[23] Sousa Silva, R.; Laboreiro, G.; Sarmento, L.; aj.: *Natural Language Processing and Information Systems*, kapitola 'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ISBN 978-3-642-22327-3, s. 161–168, doi:10.1007/978-3-642-22327-3_16.

[24] Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, ročník 60, č. 3, 2009: s. 538–556, ISSN 15322882, doi:10.1002/asi.v60:3.

[25] Yang, Y.; Pedersen, J. O.: A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, ročník 97, 1997, s. 412–420.

[26] Zhao, Y.; Zobel, J.: Searching With Style: Authorship Attribution in Classic Literature. In *Thirtieth Australasian Computer Science Conference (ACSC2007)*, *CRPIT*, ročník 62, editace G. Dobbie, Ballarat Australia: ACS, 2007, s. 59–68.

Přílohy

Seznam příloh

Příloha A

Obsah CD

A.1 Obsah kořenové složky

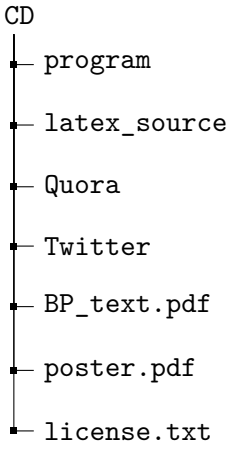
Na obrázku A.1 je zobrazen obsah CD. Nachází se zde složka se systémem pro určování autorství *program*. Obsah složky program je potom vidět na obrázku A.2 a je podrobněji popsán v kapitole A.2. Ve složce *latex_source* je zdrojový kód technické zprávy v LaTeXu. Výsledné pdf obsahuje technickou zprávu je *BP_text.pdf*. Soubor *poster.pdf* obsahuje plakát vyžadovaný zadáním. V souboru *license.txt* jsou uloženy licence použitých knihoven a odkazy na plné znění těchto licencí. Složky *Quora* a *Twitter* obsahují zdrojové kódy programů, které jsem vytvořil pro stažení komentářů z těchto služeb.

A.2 Obsah složky s programem

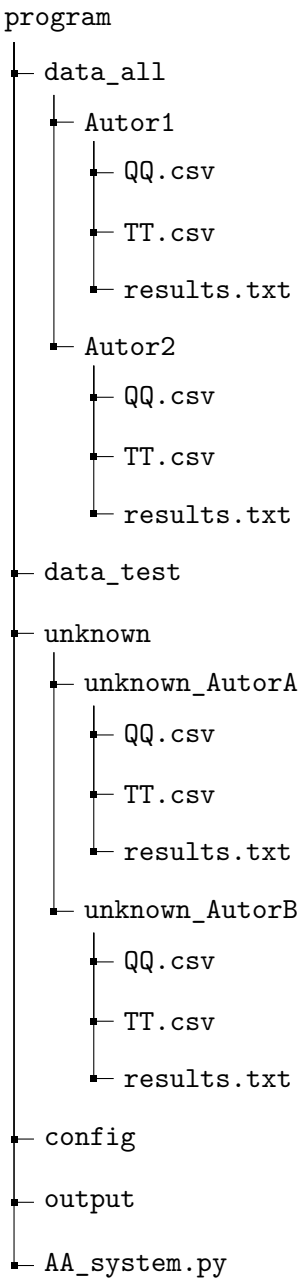
Na obrázku A.1 je znázorněna adresářová a souborová struktura systému. Zobrazeny jsou pouze prvky důležité pro běh systému. Hlavní modul programu je uložen v souboru *AA_system.py*. Soubor *config* obsahuje konfigurační soubor, který bude podrobněji popsán v následující sekci B.1.

Následují dvě složky *data_all* a *unknown*, které obsahují texty určené pro učení a klasifikaci. Jejich struktura je identická, ale účel mají jiný. Složka *data_all* obsahuje všechny texty, které jsou k dispozici které jsou k dispozici pro analýzu navrženým programem. Ze souborů zde se vybere podle nastavení programu v konfiguračním souboru určitý počet autorů a jejich texty se rozdělí na učící a testovací část. Do složky *unknown* patří neznámí autoři, jejihž totožnost chceme zjistit. Odtud nebude žádný text použit k učení systému. Všechny texty budou klasifikovány po ukončení učení a následně budou vypsány výsledky. Složka *data_test* obsahuje část dat ze složky *data_all* (texty pro 10 autorů). Slouží k jednodušššímu otestování funkcí systému. Například obsahuje i všechny autory ze složky *unknown* a je tak možné porovnat, jak systém identifikuje neznámé autory.

Nyní bude popsána struktura obsahu dvou zmíněných složek. Každá obsahuje dva soubory *QQ.csv* a *TT.csv*, které obsahují texty stažené z Quory a Twitteru ve formátu CSV. První položka CSV souboru je id v rámci daného souboru, aby se na příslušný text dalo jednoduše odkazovat ve výsledcích. Druhá položka je samotný text komentáře stažený z příslušné služby. Posledním souborem je *results.txt*, který obsahuje výsledky klasifikace. Do složky každého autora se ukládají výsledky klasifikace jeho textů.



Obrázek A.1: Obsah CD



Obrázek A.2: Složka s programem

Příloha B

Manuál

V následující kapitole bude popsáno ovládání programu vytvořeného v rámci této bakalářské práce a popsaného v kapitole 5.

B.1 Konfigurační soubor

Příklad nastavení konfiguračního souboru je možno vidět na obrázku B.1. Konfigurace je uložena v klasickém textovém souboru, kde na každém řádku je jeden záznam ve formátu "parametr=hodnota", případně "parametr=hodnota, hodnota, hodnota", pokud je explicitně povoleno více hodnot. Nyní budou jednotlivé parametry podrobně popsány.

Obecné parametry:

- **text_data**: Složka se všemi dostupnými texty, které se mají použít na učení nebo test systému.
- **text_unknown**: Složka s texty neznámých autorů, jejihž totožnost chceme zjistit.
- **authors**: Počet autorů, kteří se mají vybrat ze složky uložené v prametru **text_data**. Pokud je hodnota parametru 0, tak se vyberou všechni autoři. Povolené hodnoty: celá čísla vyšší než 2.
- **services**: Které služby se mají použít. Momentálně jsou povolené pouze dvě hodnoty "Quora", "Twitter". Je možné je zadat i obě zároveň.
- **train_test_split**: Desetinné číslo, které udává podíl načtených textů od každého autora, které budou použity pro validaci/testování systému. Povolené hodnoty: desetiné číslo v intervalu (0.0, 1.0).
- **texts_to_extract**: Počet položek musí být stejný jako počet použitých služeb na základě parametru **services**. Každá položka značí kolik textů (příspěvků) se má z každé služby načíst.
- **text_split_method**: Tento parametr udává, jaká metoda se má použít pro zpracování textů před jejich předložením učícímu algoritmu. Bližší vysvětlení je v kapitole 5.5.
  - **Frekvence**: hodnota "normal",
  - **Spojemí textů**: hodnota "join",

  - **Stejná délka textů**: hodnota “equal”.

.

- **LEN**: Počet položek musí být stejný jako počet použitých služeb na základě parametru **services**. Jedná se o rozšiřující parametr k parametru **text_split_method**. Pro hodnoty “join” a “equal” udává maximální délku jednoho textu. Povolené hodnoty: celá čísla větší než 0.
- **feature_count**: Parametr nastavuje výsledný počet vybraných charakteristik textu, které budou použity pro učení SVM. Povolené hodnoty: celá čísla větší než 0.
- **k_value**: Nastavuje parametr k. Ten udává, jak si systém musí být jistý výsledkem, aby přiřadil nejpravděpodobnějšího autora k textu jako skutečného autora. Povolené hodnoty: desetinné číslo v intervalu $(0.0, 1.0)$. Při hodnotě $k = 1.0$ je autor přiřazen vždy. Podrobnější vysvětlení je obsaženo v kapitole 6.1.
- **top**: Ovlivňuje výpis výsledků. Program může sledovat, jestli se správný autor vyskytuje mezi X nejpravděpodobněnjšími výsledky, kde X se nastavuje právě pomocí parametru **top**. Povolené hodnoty: celé číslo, vyšší než 0 a musí být nižší nebo rovno počtu autorů (parametru **authors**).
- **repeat**: Umožňuje spustit systém vícekrát se stejným nastavením systému automaticky. Tento parametr určuje, kolikrát se bude spuštění opakovat. Povolené hodnoty: celé číslo, vyšší než 0.
- **selection_method**: Výběr metody pro výběr charakteristik. Jsou k dispozici čtyři možné argumenty:
  - **entropy**: Informační přínos,
  - **gini**: Gini impurity,
  - **chi2**: Chí-kvadrát,
  - **rec_elim**: rekurzivní eliminace.

Parametry jednotlivých kategorií: Na každém řádku jsou argumenty pro jednu kategorii charakteristik. Jednotlivé parametry a jejich hodnoty jsou:

- **id**: Identifikátor charakteristiky. Musí být vždy na prvním místě, proto se píše rovnou hodnota. Možné hodnoty (bližší popis jednotlivých kategorií v kapitole 5.4):
  - **pos**: slovní druhy,
  - **char**: písemnné n-gramy,
  - **nva**: slovní druhy + stopslova,
  - **misc**: ostatní charakteristiky,
  - **word**: slovní n-gramy.
- **n**: Pouze pro kategorie založené na n-gramech. Udává hodnotu N.
- **m**: Parametr ovlivňující filtraci charakteristik. Příznaky s méně než m výskyty v korpusu budou vyřazeny před zpracováním.
- **l**: Parametr udává, kolik možných příznaků se má vybrat v první fázi výběru. Schéma výběru je na obrázku 5.2.

```
text_folder = ./data_1500/
text_unknown = ./unknown/
authors = 50
number_of_services = Twitter, Quora
train_test_split = 0.2
texts_to_extract = 400, 100
text_split_method = normal
LEN = 1200, 1200
feature_count = 1000
k_value = 0.75
selection_method = entropy
top = 5
repeat = 5

# Nasleduji parametry jednotlivych kategorii charakteristik.
pos, n=3, l=20, m=0.1
nva, n=3, l=20, m=0.1
char, n=4, l=20, m=0.1
misc
```

Obrázek B.1: Konfigurační soubor

B.2 Instalace knihoven, spuštění a výpis výsledků

V této kapitole bude popsáno, jak nainstalovat všechny nezbytné knihovny, aby bylo možné program spustit a následně typické použití programu. Návod je určen pro distribuci Ubuntu 15.04 a Python 2.7.9

Pro spuštění je potřeba nainstalovat následující knihovny:

- Numpy 1.8.2: `sudo apt-get install python-numpy`
- Scipy 0.14.1: `sudo apt-get install python-scipy`
- Sklearn 0.17: `sudo pip install sklearn` [1]
- NLTK 3.1: `sudo pip install nltk`
- Spustit program pro stáhnutí dalších dat pro NLTK ve složce s programem: `python nltk_download.py`

Předpokládá se, že je instalovananý jazyk Python verze 2.7 a spouští se příkazem `python`. Program spustit příkazem `python AA_system.py`.

Výsledky jsou vypisovány do souborů *results.txt*. V každé složce autora se nachází jeden soubor *results.txt*, kde jsou uloženy výsledky klasifikace tohoto autora. Jedná se o textový soubor, kde jsou jednotlivé položky odděleny znakem tabulátoru. První položkou je seznam ID hodnot textů v rámci jednoho ze dvou výše zmíněných souborů. Tato položka říká, o jaký text nebo texty jde. Dále následuje výsledek klasifikace. Tím je buď název složky (jméno autora), ve které byl text uložen nebo řetězec "unknown", pokud systém nedokázal v závislosti na svém nastavení přiřadit s dostatečnou jistotou některého z možných autorů. Poslední položkou je číselná hodnota, která říká, jak je si systém jistý tímto výsledkem.

[1] Instalátor Pip je možno nainstalovat příkazem `sudo apt-get install pip`

Jedná se o podíl ohodnocení (pravděpodobnosti) nejpravděpodobnějšího a druhého nejpravděpodobnějšího autora. Může nabývat hodnot mezi 1 a 0. Čím je hodnota blíže 0, tím je si systém více jistý svým zařazením.