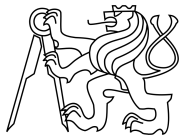
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název: Identifikace funkčního stylu dokumentu
Student: Svetlana Ekimova
Vedoucí: doc.RNDr.Ing. Marcel Jiřina, Ph.D.
Studijní program: Informatika
Studijní obor: Teoretická informatika
Katedra: Katedra teoretické informatiky
Platnost zadání: do konce letního semestru 2015/16

Pokyny pro vypracování

1. Proveďte rešerši metod, které se zabývají automatizovanou detekcí funkčního stylu dokumentu (hovorový, administrativní, vědecký, publicistický, řečnický a umělecký). Zaměřte se na česky psané dokumenty.
2. Na základě rešerše vyberte vhodné metody, případně navrhněte úpravy těchto metod, které umožní detekovat funkční styl zadaného českého textu.
3. Navržené algoritmy implementujte ve vhodném programovacím jazyce.
4. Implementované algoritmy ověřte na reálných datech a vyhodnoťte dosaženou úspěšnost.

Seznam odborné literatury

Dodá vedoucí práce.

L.S.

doc.Ing. Jan Janoušek, Ph.D.
vedoucí katedry

prof.Ing. Pavel Tvrdík, CSc.
děkan

V Praze dne 26. února 2015

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Bakalářská práce

Identifikace funkčního stylu dokumentu

Bc. Svetlana Ekimova

Vedoucí práce: doc. RNDr. Ing. Marcel Jiřina, Ph.D.

16. května 2016

Poděkování

Chtěla bych poděkovat svému vedoucímu bakalářské práce doc. RNDr. Ing. Marcelu Jiřinovi, Ph.D. za cenné rady a pomoc, které mi poskytoval při zpracování této práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 16. května 2016 .....

Odkaz na tuto práci

Ekimova, Svetlana. *Identifikace funkčního stylu dokumentu.* Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2016.

Abstrakt

Práce se zabývá identifikací funkčního stylu textových dokumentů. Úloha identifikace je řešena jako úloha klasifikační: funkční styl dokumentu se určuje pomocí metod, které bývají používány pro klasifikaci textů. Pro naučení klasifikátoru jsou textové dokumenty reprezentovány jako vektor atributů a označeny hodnotou zastupující jejich funkční styl. Následně jsou předloženy metodě strojového učení s učitelem. Úspěšnost vytvořeného tímto způsobem klasifikátoru je vyhodnocena pomocí křížové validace.

Klíčová slova text mining, zpracování přirozeného jazyka, strojové učení s učitelem, klasifikace textů, kategorizace textů, identifikace stylu, funkční styly

Abstract

The thesis deals with the identification of the functional style of text documents. The identification task is solved as a classification task: the functional style of a document is detected by means of methods which are being used for the text classification. To learn the classifier, text documents are represented as a vector of attributes and labeled by a value describing their functional style. They are then given as input to a supervised machine learning method.

The success rate of the learned classifier is evaluated by using k-fold cross-validation.

Keywords text mining, natural language processing, supervised machine learning, text classification, text categorization, style identification, functional styles

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Pojem *funkční styl* je jedním ze základních pojmů české stylistiky. Při tvorbě textu musí autor volit vhodné jazykové prostředky, aby text plnil svou funkci a nepůsobil kuriózně. Každý funkční styl je spojen s určitou funkcí, popř. funkcemi, a je definován použitím jazykových prostředků, které jsou pro danou funkci vhodné.

V dnešní době jsou mnohé texty tvořeny i uchovávány v elektronické podobě. Často potřebujeme vyhledat text nejen podle tématu, ale i podle jeho druhu: například můžeme chtít najít odborné práce o kvantových počítačích, ale nezajímají nás novinové články na toto téma. K tomu bychom potřebovali, aby vyhledávač, který za tímto účelem používáme, dokázal určit nejen témata textů, ale i jejich (funkční) styl. Jakým způsobem to lze udělat, se zabývá daná práce.

Cílem práce je návrh a implementace algoritmu, jenž bude identifikovat funkční styl textových dokumentů. Základní postup se skládá ze čtyř kroků:

1. Provedení rešerše studií, které se zabývají stejnou nebo shodnou problematikou, tj. identifikací (funkčního) stylu dokumentů.

2. Volba a případná úprava metod vhodných pro splnění cíle práce.

3. Implementace navrženého algoritmu.

4. Hodnocení úspěšnosti navrženého algoritmu.

Identifikace funkčního stylu dokumentů je pojata jako klasifikace dokumentů podle jejich funkčního stylu.

Práce se skládá z teoretické (kapitoly 1–4) a praktické (kapitoly 5–8) částí. Kapitola 1 je úvodem do problematiky funkčních stylů. V kapitole 2 je čtenář seznámen se základními pojmy z oblasti strojového zpracování přirozeného jazyka a některými metodami strojového učení. Kapitola 3 popisuje v rámci

práce provedenou rešerši metod používaných pro úlohu klasifikace textů. V kapitole 4 je uveden vlastní návrh algoritmu identifikace funkčního stylu. Kapitola 5 obsahuje popis dat, jež byla použita při implementaci a vyhodnocení navrženého algoritmu. Samotná implementace je popsána v kapitole 6. Hodnocení úspěšnosti algoritmu obsahují kapitoly 7 (nastavení parametrů metod strojového učení) a 8 (porovnání úspěšnosti jednotlivých metod). V závěru práce jsou diskutovány dosažené výsledky a navrženy možné způsoby vylepšení použitého postupu.

Kapitola 1

Funkční styly v české jazykovědě

Pojem *funkční styl* je jedním ze základních pojmů české stylistiky. Teorie funkčních stylů vznikla v první polovině 20. století a její základní myšlenkou je to, že výběr jazykových prostředků při tvorbě textu je určen funkcí, kterou daný text má plnit.

Chloupek [1, s. 38] definuje funkční styl jako „*okruh výrazových prostředků plnicích jednu a tutéž konkrétní funkci nebo funkce přibližně stejné*“.

Počet funkcí jazyka i funkčních stylů se v teorii stylistiky postupem času měnil. Tak, ještě v 90. letech 20. století se mluvilo o čtyřech základních funkčních stylech: prostě sdělovacím (hovorovém), odborném, publicistickém a uměleckém [1, s. 40]. Současná stylistika vymezuje již šest základních funkčních stylů [2]:

- styl hovorový (funkce prostě sdělná),
- styl odborný (funkce odborně sdělná a vzdělávací),
- styl administrativní (funkce direktivní, zpravovací a jednací),
- styl publicistický (funkce sdělná a persvazivní),
- styl řečnický (funkce informativní a persvazivní),
- styl umělecký (funkce esteticky sdělná).

Vedle výše uvedených funkčních stylů, kterým se také říká *primární*, existují funkční styly *sekundární*, neboli *odvozené*, ale jelikož se daná práce věnuje identifikaci pouze základních (primárních) funkčních stylů, bude v dalším textu používán termín *funkční styl* pro funkční styly základní.

Stylistika rozlišuje mezi stylem subjektivním (stylem jednotlivce, individuálním) a stylem objektivním (tzv. interindividuálním, tj. společným pro větší skupinu lidí). Funkční styl je stylem objektivním. Aby text plnil svou funkci, musí jeho tvůrce využít určitých kompozičních prostředků charakteristických

pro danou funkci. Kompozičními prostředky se rozumí nejen prostředky jazykové (slovní zásoba, morfologické a syntaktické prostředky), ale i kompoziční výstavba textu (členění a uspořádání obsahu apod.) [2, s. 20]. Právě tyto prostředky pak mohou sloužit k tomu, abychom byli schopni určit funkční styl textu, aniž bychom předem věděli, o jaký styl se jedná.

Uvažujme jako příklad funkční styl hovorový. Slovotvorným jevem poukazujícím na to, že text má být zařazen do daného funkčního stylu, může být např. záměna kořenových samohlásek *í/ý* a *é* (*okýnko* místo *okénko*, *míň* místo *méně*[1]) nebo slova s příponami *-ář*, *-ák*, *-ovka* (*prvňák*, *sodovka*) aj. Typickými sémantickými prostředky jsou tzv. intenzifikující přídavná jména: *děsný*, *šílený* apod. Ze syntaktického hlediska jsou pro hovorový funkční styl nejcharakterističtější elipsy (výpustky – věty jsou gramaticky neúplné) a parenteze (vsuvky).

Příznaky odborného funkčního stylu jsou především použití termínů, multiverbizované jazykové vyjádření (*projevit souhlas* místo *souhlasit*), časté použití trpného rodu a těsné spojení vět.

Podobné jazykové jevy jsou typické i pro funkční styl administrativní: odborná terminologie, multiverbizované vyjádření, trpný rod, ale kromě toho a časté užití instrumentálu (*rozhodnutím bylo ustanoveno*) a ustálených vazeb/opakujících se frází (*předem Vám děkujeme za laskavé vyřízení*).

V textech publicistického funkčního stylu se zpravidla vyskytují tyto jevy: frazémy, opakující se obrazná vyjádření, syntaktická kondenzace (např. hromadění genitivních vazeb: *vypracování návrhu koncepce programu protidrogové politiky*) a expresivní vyjádření.

Funkční styl řečnický je charakterizován hojným použitím tropů, stylistických figur, oslovení a citátů. Častým jevem jsou také výpovědi signalizující příští sdělení (*jistě nevíte, že…*, *bude vás určitě zajímat, že…*).

Texty uměleckého funkčního stylu typicky obsahují tzv. lexikální poetismy (archaismy, neologismy apod.), metafory a tropy.

Všechny výše popsané prostředky mohou sloužit k automatizované identifikaci funkčních stylů. Některé z nich však mohou vést k chybným výsledkům. Například odborná terminologie je typickým jevem hned dvou funkčních stylů (odborného a administrativního), ale používá se občas i v textech publicistických.

Nejproblematičtějším ze všech funkčních stylů je nepochybně styl umělecký, neboť některé vzorky krásné literatury mohou ve značné míře obsahovat rysy jiných stylů (například dílo současné prózy napodobující promluvu nějaké osobnosti k davu by stroj mohl zařadit do kategorie stylu řečnického apod.). Navíc, na rozdíl od ostatních funkčních stylů, je u funkčního stylu uměleckého poměrně těžké vyčlenit jeho charakteristické rysy.

V následujících podkapitolách je na krátkých ukázkách textů znázorněno, jaké jazykové prostředky bývají často používány u jednotlivých funkčních

[1] Tento a další příklady v této kapitole byly převzaty z [2].

stylů.

1.1 Administrativní funkční styl

„Základním předpisem odůvodňujícím vznik hlavního města Prahy je ústavní zákon č. 1/1993 Sb., Ústava České republiky, ve znění pozdějších předpisů.

Základním předpisem upravující postavení hlavního města Prahy je zákon č. 131/2000 Sb. o hlavním městě Praze, ve znění pozdějších předpisů, (dále jen „zákon“). Tento zákon upravuje postavení hlavního města Prahy jako hlavního města České republiky, kraje a obce a dále postavení městských částí.

Hlavní město Praha je veřejnoprávní korporací, která má vlastní majetek, má vlastní příjmy a hospodaří podle vlastního rozpočtu. Hlavní město Praha vystupuje v právních vztazích svým jménem a nese odpovědnost z těchto vztahů vyplývající.

Hlavní město Praha je samostatně spravováno Zastupitelstvem hlavního města Prahy; dalšími orgány hlavního města Prahy jsou Rada hlavního města Prahy, primátor hlavního města Prahy, Magistrát hlavního města Prahy, zvláštní orgány hlavního města Prahy a Městská policie hlavního města Prahy.“

(Magistrát hl. m. Prahy: *Důvod a způsob založení*)[2]

Prostředky typickými pro administrativní funkční styl v této ukázce jsou: 7. pád, trpný rod, opakující se fráze („*Základním předpisem*“), ustálené vazby („*ve znění pozdějších předpisů*“).

1.2 Hovorový funkční styl

„Tak už sem to skoukl, jako nic moc film, sem to chtěl v pulce vypnout, ale pak se to trochu rozjelo.. Ale nechápu co to říká Iveták o feminismu a zaměření proti ženám? Jestli myslí tím to chování té psycho ženské, tak nevim teda, ta baba byla fakt šiblá, ale žeby tím naznačovali něco o všech ženách to fakt nechápu..“

(Příspěvek uživatele na diskuzním fóru)[3]

Prostředky typickými pro administrativní funkční styl v této ukázce mj. jsou: intenzifikující přídavná jména, elipsy („*jako nic moc film*“ – věta bez přísudku), časté použití zájmena *ten*.

1.3 Odborný funkční styl

„Energie jednoho druhu se obecně přeměňuje v jiný druh konáním práce.

V makroskopickém popisu se však od mikroskopického působení silových interakcí zpravidla odhlíží a přeměna se může jevit jako bezprostřední (při anihilaci částice a antičástice látky v klidu) nebo se zavádějí nové veličiny fenomenologicky popisující disipaci či skrytý přenos energie a formulují se nová pravidla pro energetické děje. V termodynamice se proto zavádí teplo a přeměna energie se v termodynamickém popisu řídí prvním a druhým zákonem termodynamiky.

[2]Zdroj: `http://www.praha.eu` [cit. 10.5.2016]

[3]Zdroj: `http://pauza.zive.cz` [cit. 10.5.2016]

Jeden druh energie (přeměňované či přeměněné) lze zpravidla považovat za energii potenciální, která je "uložena"v silovém poli (polohová energie) nebo klidové hmotnosti (klidová energie) daného fyzikálního systému i v jeho relativním klidu, druhá je energií dynamickou, projevující se v časové přeměně či pohybu (kinetická energie, energie vlnění).“

(wikipedia.org: *Energie*)[4]

Prostředky typickými pro administrativní funkční styl v této ukázce jsou: odborná terminologie, trpný rod, těsné spojení vět.

1.4 Publicistický funkční styl

„Luxusní kabát za deset tisíc korun ušitý podle londýnských módních trendů, vypiplaný bezlepkový narozeninový dort i terapie pro psy, kteří upadli do deprese. Nad některými výrobky či službami určenými pro zvířecí mazlíčky nechovatelé kroutí hlavou.

[. . .]

Zahrnout pejska luxusním zbožím a výběrovými dobrotami ovšem uspokojí hlavně jeho pána, zvíře nemusí být v dobrém rozpoložení.

Psí psycholog a veterinář Alexandr Skácel hovoří o „syndromu zlaté klece“, spojovaném spíše se znuděnými a rozmařilými partnerkami bohatých mužů. „Některé nejsou šťastné, ani když mají všechno, podobné je to u psů,“ říká.“

(idnes.cz: *ČEŠI A PSI: Místo boudy zlatá klec. K mání jsou psí dorty i kabáty z tvídu*)[5]

Prostředky typickými pro administrativní funkční styl v této ukázce jsou: frazémy (*kroutí hlavou*), expresivní vyjádření (*vypiplaný dort*).

1.5 Řečnický funkční styl

„Milí spoluobčané,

dnes uplynulo deset let od okamžiku, kdy vznikla Česká republika jako samostatný stát.

Co jejímu vzniku předcházelo?

V podstatě tři roky velkého revolučního kvasu, kdy se měnil celý právní systém a obnovovala demokracie, kdy ožívaly všechny po desítiletí potlačované svobody, kdy ohromné státní vlastnictví masivně přecházelo do privátních rukou, kdy se téměř ze dne na den objevovaly celé skupiny nových politiků, rodily se nové politické strany, probouzel se opět spolkový život a kdy jsme se - jako politici vesměs začátečníci - narychlo seznamovali s velikostí problémů, které dřímaly pod povrchem předchozích poměrů a jejichž hloubku předtím málokdo z nás dokázal přesně odhadnout.“

(Novoroční projev prezidenta republiky Václava Havla (2003))[6]

Prostředky typickými pro administrativní funkční styl v této ukázce jsou: oslovení, tropy a stylistické figury.

[4] Zdroj: `https://cs.wikipedia.org` [cit. 10.5.2016]

[5] Zdroj: `http://zpravy.idnes.cz` [cit. 10.5.2016]

[6] Zdroj: `https://cs.wikisource.org` [cit. 10.5.2016]

1.6 Umělecký funkční styl

„Před třemi dny jsem přiklekl na zahrádce k rozkvetlému trsu dlužichy, abych ji očistil od plevele; měl jsem slabou závrať, ale to se mi stávalo častěji. Snad ta závrať způsobila, že se mi to místo zdálo krásnější než kdy dosud: jiskřivě rudé klásky dlužichy a za nimi bílé, chladivé laty tavolníků, – bylo to tak krásné a skoro tajemné, že mi šla hlava kolem. Na dva kroky ode mne seděla na kameni pěnkavka, hlavičku na stranu, a dívala se na mne jedním okem: Co ty vlastně jsi? Ani jsem nedýchal, bál jsem se, že ji zaplaším; cítil jsem, jak mi bouchá srdce. A najednou to přišlo. Nevím, jak bych to popsal, ale byl to strašně silný a jistý pocit smrti. Skutečně to neumím jinak vyjádřit; myslím, že jsem zápasil o dech či co, ale jediné, čeho jsem si byl vědom, byla nesmírná úzkost. Když to polevilo, klečel jsem ještě, ale měl jsem plné ruce urvaného listí. Opadlo to jako vlna a nechalo to ve mně smutek, který nebyl nepříjemný.“

(Karel Čapek: *Obyčejný život*)[7]

Prostředky typickými pro administrativní funkční styl v této ukázce jsou především metafory a tropy.

[7]Zdroj: `https://cs.wikisource.org` [cit. 10.5.2016]

Kapitola 2

Zpracování přirozeného jazyka. Strojové učení a jeho metody

Cílem práce je detekce funkčního stylu textového dokumentu[8]. To znamená, že navrhovaný algoritmus, jehož úkolem je určení funkčního stylu, bude mít na vstupu soubor s textem psaným v přirozeném (českém) jazyce. Takový text má *nestrukturovanou* podobu [3], která není vhodná pro strojové zpracování. Pro získání informací z textových dokumentů jsou proto potřeba speciální postupy, jejichž zkoumáním se zabývá disciplína zvaná *text mining*. Text mining spojuje v sobě metody *data miningu* (získávání informací ze strukturovaných dat), statistiky, vyhledávání informací v dokumentech (*information retrieval*), strojového učení a také zpracování přirozeného jazyka (*natural language processing*, dále *NLP*) [4]. Přehled základních postupů NLP je uveden v podkapitole 2.1.

Text mining se používá pro různé typy úloh, které souvisejí se zpracováním dat z nestrukturovaných dokumentů. Jako příklad lze uvést vyhledávání v textech, přiřazení klíčových slov dokumentům, určení tématu textu a další [4]. V rámci této práce je však nejzajímavější úlohou *klasifikace textů*: identifikaci funkčního stylu dokumentu lze totiž převést na přiřazení tomuto dokumentu jedné ze šesti tříd, které odpovídají šesti funkčním stylům. Přesnou definici klasifikace textů uvedeme v kapitole 3.

V moderní počítačové lingvistice je úloha klasifikace textů obvykle řešena pomocí metod strojového učení (*machine learning*). Alternativním přístupem těmto metodám je znalostní inženýrství (*knowledge engeneering*). Rozdíl mezi znalostním inženýrstvím a strojovým učením spočívá ve způsobu vytvoření pravidel pro klasifikaci: zatímco v znalostním inženýrství se pravidla vytvářejí ručně experty, u metod strojového učení je za vytvoření pravidel zodpovědný, jak je patrno z názvu, stroj. Ačkoliv znalostní inženýrství dosahuje při klasifikaci textů poměrně dobrých výsledků, je považováno za „drahé“, neboť

[8] Pojmy *text*, *dokument* a *textový dokument* se v práci používají jako synonyma.

vyžaduje práci doménových expertů a navíc není přenositelné mezi doménami [5]. Tento přístup byl populární až do 90. let 20. století, kdy byl vytlačen strojovým učením [6], jehož metody také mohou dávat dobré výsledky, ale přitom nevyžadují účast odborníků vytvářejících pravidla a jsou lépe přenositelné. Úspěšnost klasifikace za pomocí strojového učení však závisí na vstupních datech, která by měla být vysoce kvalitní [5]. Podrobněji o strojovém učení a jeho metodách lze přečíst v podkapitole 2.2.

2.1 Adaptace textových dokumentů pro strojové učení

Cílem této podkapitoly je seznámit čtenáře s pojmy z oblasti NLP, které souvisejí se zpracováním nestrukturovaného textu do strukturované podoby. Podrobněji jednotlivé procesy a jejich problematiku projednávají Manning a Schütze [7].

Prvním krokem bývá *tokenizace*, tj. rozdělení souvislého textu na jednotlivé tokeny – slova, čísla, interpunkční znaménka apod. Po tokenizaci zpravidla následují *stemming* nebo *lemmatizace*. Stemmingem se rozumí ořezávání předpon a přípon slova, tak, že ve výsledku zůstává pouze kmen. Lemmatizace je převod slova na základní tvar (1. pád jednotného čísla u podstatných jmen, neurčitý tvar u sloves apod.). Cílem těchto procesů je omezit počet tokenů sloučením různých forem jednoho lexému.

V některých případech se také provádí tzv. *POS-tagging* (*part-of-speech tagging*, označení slovních druhů). Tento proces předzpracování je vhodný, pokud nás zajímají nejen lexikální, ale i gramatické vlastnosti textů.

2.2 Strojové učení

Existují dva typy strojového učení: s učitelem (*supervised machine learning*) a bez učitele (*unsupervised machine learning*). V prvním případě jde o naučení klasifikačního algoritmu na základě kolekce již klasifikovaných dokumentů, tj. algoritmus má informace o počtu tříd a rozdělení jednotlivých dokumentů do těchto tříd. V druhém případě nejsou žádné informace o třídách a jejich počtu k dispozici, a úkolem algoritmu je tak najít nejvhodnější rozdělení pro předložená data. Zatímco strojové učení bez učitele bývá spojováno s pojmem shlukování (*clustering*), je strojové učení s učitelem často používáno v souvislosti s úlohou klasifikace [7, s. 232] [6], kterou se zabývá daná práce. Proto v dalším textu se budeme věnovat právě strojovému učení s učitelem.

Metody strojového učení pracují na základě hodnot různých *atributů* (také *příznaků*, *features*) předložených *vzorků* (v našem případě textových dokumentů), a proto pro úlohu klasifikace textů potřebují být upraveny a zahrnují tak tyto kroky (podle [5]):

1. Určení atributů, které budou sloužit k reprezentaci jednotlivých dokumentů.
   Reprezentovat dokument mohou např. v něm obsažená slova, jejich četnost, průměrná délka vět nebo jiné statistické hodnoty, atd. Podrobněji o různých způsobech výběru atributů lze přečíst v kapitolách 3 a 4.

2. Převod textových dokumentů na trénovací vzorky, tj. určení hodnot atributů z prvního kroku pro každý dokument a přiřazení dokumentům značky třídy, do které patří (*label*). Tomuto procesu se také říká indexace textů (*text indexing* [6]).
   K převodu dokumentů na trénovací vzorky bývají používány procesy popsané v podkapitole 2.1. Reprezentace dokumentu pomocí hodnot atributů se občas nazývá *vektor* dokumentu [8, 9].

3. Nalezení vzoru pro rozlišení jednotlivých tříd mezi sebou.
   Toto je úkol algoritmu vytvářejícího klasifikátor, tj. metody strojového učení.

4. Vyhodnocení výsledků a výběr nejlepšího vzoru (zpravidla z hlediska minimalizace chyby klasifikace).

V následujících sekcích budou krátce popsány vybrané metody strojového učení s učitelem[9].

2.2.1 Rozhodovací stromy

Rozhodovací stromy (*decision trees*) je jedna z nejrozšířenějších metod strojového učení. Klasifikátor je reprezentován rozhodovacím stromem, jehož uzly odpovídají atributům a listy třídám, do kterých mají být vzorky klasifikovány. Větve vedoucí z uzlu korespondují s možnými hodnotami, kterých atribut reprezentovaný daným uzlem může nabývat.

Tento klasifikátor lze také popsat jako sadu pravidel „*if – then*“ a jeho výhodou je snadná interpretovatelnost. Dalšími klady této metody jsou odolnost vůči chybám v trénovacích datech a skutečnost, že rozhodovací stromy bývají často vhodné pro řešení klasifikačních problémů (*classification problems*). Nevýhodou metody je především možné přeučení (zejména v případě, kdy nejsou k dispozici kvalitní trénovací data) a také to, že se hůř vypořádá se spojitými hodnotami atributů.

[9] Metody v sekcích 2.2.1–2.2.3 jsou popsány podle [10], sekce 2.2.4 podle [11].

2.2.2 Naivní Bayesův klasifikátor

Tato metoda je založena na Bayesově větě pro výpočet podmíněné pravděpodobnosti jevů, která je vyjádřena vzorcem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}. \tag{2.1}$$

V kontextu strojového učení se jednotlivé proměnné interpretují následujícím způsobem:

- $P(h)$ – pravděpodobnost, že platí hypotéza h, předtím, než byla použita trénovací data (také apriorní pravděpodobnost).
- $P(D)$ – pravděpodobnost, že bude pozorována množina trénovacích dat D (apriorní pravděpodobnost).
- $P(D|h)$ – pravděpodobnost, že budou pozorována data D, pokud platí hypotéza h.
- $P(h|D)$ – pravděpodobnost, že platí hypotéza h, pokud byla pozorována data D (aposteriorní pravděpodobnost).

Pokud se jedná o klasifikační úlohu, zajímá nás nalezení nejvhodnější třídy pro daný vzorek, tj. nalezení nejvíce pravděpodobné hypotézy h. Proto se používá *maximální aposteriorní pravděpodobnost*

$$h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}, \tag{2.2}$$

kde H je množina hypotéz (hypotézou h se v klasifikační úloze rozumí předpoklad, že vzorek patří do třídy h). $P(D)$ je konstanta nezávislá na h, a proto

$$h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h). \tag{2.3}$$

Pokud máme množinu tříd V a vzorek popsaný atributy $a_1, a_2, \ldots, a_n$, pak lze rovnice 2.2 a 2.3 zapsat jako

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} P(v_j|a_1, a_2, \ldots, a_n) = \\ &= \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n|v_j)P(v_j)}{P(a_1, a_2, \ldots, a_n)} = \\ &= \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n|v_j)P(v_j), \end{aligned} \tag{2.4}$$

kde v_j je třída z množiny tříd V, $P(v_j)$ je její apriorní pravděpodobnost (obvykle bývá stejná pro všechny třídy z množiny V) a $P(a_1, a_2, \ldots, a_n|v_j)$ je pravděpodobnost, že pokud vzorek patří do v_j, pak má atributy s hodnotami

$a_1, a_2, \ldots, a_n$. Naivní Bayesův klasifikátor předpokládá, že atributy jsou na sobě nezávislé, a proto lze vzorec 2.4 přepsat následujícím způsobem:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j). \tag{2.5}$$

Proměnná v_{NB} v rovnici 2.5 představuje výstup Naivního Bayesova klasifikátoru, tj. třídu, do které je zařazen předložený klasifikátoru vzorek.

Ačkoliv předpoklad, že atributy jsou na sobě nezávislé, nemusí platit, tento klasifikátor často dosahuje vysoké přesnosti a je proto jedním z nejčastěji používaných.

2.2.3 k nejbližších sousedů

k nejbližších sousedů (*k-Nearest Neighbor*, dále k-NN) je jednou z metod učení založeného na instancích (*instance-based learning*). Typické pro takové metody je to, že nevytvářejí klasifikační model, ale ukládají předložené trénovací vzorky. Výpočet probíhá až ve chvíli, kdy klasifikátoru je předložen nový vzorek, který má být zařazen do jedné ze tříd. Konkrétně v k-NN se počítá vzdálenost nového vzorku od ostatních a hledají se k nejbližších instancí. Nový vzorek pak bude zařazen do třídy, do které patří většina jeho k sousedů.

Vzdálenost lze počítat např. jako eukleidovskou vzdálenost, ale i dalšími způsoby.

2.2.4 Support Vector Machines

Metoda *Support Vector Machines* (dále SVM) se zakládá na principu minimalizace strukturálního rizika, tzn. úkolem je najít takovou hypotézu h, u níž pravděpodobnost, že náhodně zvolený testovací vzorek bude klasifikován chybně, je minimální. Základní myšlenka metody SVM spočívá v nalezení nadroviny, která by oddělovala data patřící do různých tříd. Nadrovina je zpravidla vytvořena lineární funkcí.

Výhodou SVM je nezávislost na dimenzionalitě prostoru atributů (hodí se i pro velké počty atributů) a vysoká úspěšnost v úloze klasifikace textů. Nedostatkem je to, že je určena primárně pro binární klasifikaci. Avšak existují modifikace této metody, které umožňují klasifikaci do více tříd (např. MSVM, podrobněji [12]).

Kapitola 3

Klasifikace textových dokumentů

Klasifikace textů (často se používá synonymum *kategorizace textů*, v anglicky psané literatuře se také vyskytuje pojem *topic classificatio*n nebo *topic spotting*, pokud jde o klasifikaci podle tématu textu) může být definována následujícím způsobem (dle [3, s. 256]): Nechť $\mathbb{X}$ je prostor dokumentů, $d \in \mathbb{X}$ je popis dokumentu a $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ je množina tříd pevně dané velikosti. Dále existuje trénovací množina $\mathbb{D}$, která obsahuje klasifikované dokumenty, tj. prvky $\langle d, c\rangle \in \mathbb{X} \times \mathbb{C}$. Úkolem je pak pomocí určité metody učení naučit klasifikátor neboli klasifikační funkci γ, která provádí zobrazení $\gamma : \mathbb{X} \to \mathbb{C}$. Metoda strojového učení se označuje $\Gamma(\mathbb{D}) = \gamma$.

Klasifikovat texty lze např. podle jejich tématu, autoru nebo stylu. Texty mohou být zařazovány do jedné třídy (tzv. *one-of*) nebo do více tříd najednou, popř. do žádné (*any-of*) [3]. Speciálním případem zařazení do jedné třídy je binární třídění: text patří buď do třídy c_i, nebo do jejího doplňku $\overline{c}_i$ [6].

Každá klasifikační úloha začíná určením tříd, do kterých budou rozdělovány jednotlivé vzorky. Protože se jedná o učení s učitelem, musí existovat sada vzorků, kterým již byl (často manuálně) přiřazen label (hodnota označující třídu). Jelikož se tato práce zabývá klasifikací textových dokumentů, budeme pro sadu vzorků používat, jak je v počítačové lingvistice zvykem, termín *korpus*. Korpus je „*soubor počítačově uložených textů* [...]*, který primárně slouží k jazykovému výzkumu*“ [13].

Každý dokument v korpusu je reprezentován pomocí atributů (příznaků). Výběr atributů je zásadní pro úspěšnost budoucího klasifikátoru, neboť právě na základě jejich hodnot jsou hledána pravidla pro rozlišení mezi jednotlivými třídami. V odborné literatuře věnující se klasifikaci textů jsou popsány dva rozdílné přístupy výběru atributů: na jedné straně jsou to atributy lexikální, na druhé atributy strukturální.

Lexikální atributy představují jednotlivá slova obsažená v textu, popř. skupiny těchto slov. Problémem tohoto přístupů je velká dimenzionalita výsled-

ných vektorů (každý token reprezentuje jednu dimenzi), proto se provádí tzv. výběr atributů (*feature selection*), eliminují se nevýznamná slova (*stop words*) apod. Reprezentace pomocí lexikálních atributů se používá především pro klasifikaci textů podle tématu. Oproti tomu *strukturální atributy* (popisující gramatické vlastnosti textu a jeho strukturu) slouží zpravidla ke klasifikaci podle stylu, která je podrobněji popsána v následující podkapitole.

3.1 Rešerše studií zabývajících se klasifikací textů na základě stylu

V anglicky psané literatuře se pro označení lexikálních atributů občas používá pojem *bag of words*. Finn a Kushmerick [14] označují jejích výběr pro reprezentaci textových dokumentů jako „standardní přístup" a porovnávají ho se dvěma jinými způsoby reprezentace: pomocí *statistiky slovních druhů* (POS atributy) a pomocí *textové statistiky* (průměrná délka věty, průměrná délka slova apod.). Ačkoliv autoři této studie mluví o klasifikaci dokumentů podle žánru (*genre*) a uvádějí, že žánr je spojen se stylem[10], jejich pojetí žánru je poněkud zvláštní, neboť ve svém experimentu klasifikují dokumenty podle těchto dvou kritérií:
a) zda má text subjektivní nebo objektivní charakter,
b) zda je recenze pozitivní nebo negativní.
Zejména druhé rozdělení souvisí spíše s analýzou sentimentů, nikoliv s detekcí stylu.

Karlgren [15] a Kessler [16] se pokoušejí o klasifikaci textů podle žánru v obvyklejším slova smyslu: jako třídy jsou používány literární žánry (např. romány), ale i další kategorie reprezentující publicistické i administrativní texty. Karlgren et al. volí jako atributy jak strukturální příznaky (počet znaků, počet adverbií, počet dlouhých slov), tak i některá *funkční slova*. Funkční slova jsou taková slova, která se vyskytují ve většině textů nezávisle na jejich tématu (u Karlgren jsou to například slova *I*, *me*, *that*, *which* atd.). Kessler et al. [16] poukazují na to, že výše popsány přístup vyžaduje POS-tagging a navrhují alternativní, jednodušší způsob reprezentace dokumentů: kromě strukturálních a lexikálních vlastností lze totiž využít vlastností znakových (interpunkce) a odvozených (poměry a míry odchylek). Úspěšnost klasifikace se v obou případech jeví jako obstojná, u [15] klesá s růstem počtu tříd, přičemž pro 2 třídy je úspěšnost 96%, pro 4 třídy 73% a pro 10 65%.

Argamon et al. [17] porovnávají dva typy atributů: lexikální (četnost funkčních slov) a pseudo-syntaktické (POS trigramy). Nejvyšší úspěšnosti nakonec v jejich studii dosáhla kombinace obou typů.

[10]Literární žánr a literární styl jsou samozřejmě dva odlišné pojmy, navíc je funkční styl stylem jazykovým, nikoliv literárním. I přesto jsou metody klasifikace textů podle žánru v rámci dané práce relevantní, protože jak jednotlivé žánry, tak i jednotlivé styly jsou tvořené mimo jiné pomocí jazykových prostředků.

Pustylnikov [18] se zaměřuje na klasifikaci textů pouze podle strukturálních atributů, kterým říká *bag of features*. I když klasifikační metoda v jejím experimentu představuje učení bez učitele, dosahuje úspěšnosti nad 80%. Shodný přístup (atributy získané z logické struktury textu) používají Pustylnikov a Mehler [19] pro klasifikaci na základě tématu dokumentu. Porovnávají učení s učitelem (metoda SVM) a bez učitele (shlukování) a dospívají k závěru, že klasifikátor získaný učením s učitelem dává lepší přesnost.

Stamatatos et al. [20] se naopak pokoušejí detekovat styl textu pomocí přístupu použitého v dřívějších studiích pro určení autorství, a to pomocí četnosti nejčastěji vyskytujících se slov. Doplňují a vylepšují pak tuto metodu o četnost výskytu interpunkčních znamének. Avšak i bez tohoto vylepšení dosahují až více než 90% přesnosti klasifikace (míra přesnosti závisí na počtu atributů, tj. nejčastějších slov).

Zajímavý experiment s využitím lexikálních atributů je popsán v článku Lee a Myaeng [21]. Jejich metoda je založena na statisticky zvolených atributech získaných z trénovacích dat klasifikovaných jak podle tématu, tak podle žánru. I tento přístup dosáhl vysoké míry úspěšnosti.

3.1.1 Klasifikace česky psaných textů

Určitý počet českých studií věnujících se klasifikaci textových dokumentů používá pro naučení a testování klasifikátoru korpusy anglicky psaných textů, jako je například kolekce Reuters-21578 [22, 23]. Pokud se zaměříme na práce a články popisující klasifikační experimenty s českými texty, zjistíme, že zkoumají klasifikaci textů podle témat, a proto používají metody založené na reprezentaci dokumentů pomocí lexikálních atributů.

Například Hrala a Král [8, 24] se zabývají klasifikací novinových článků do jedné i více tříd. Při vytvoření reprezentace dokumentů používají lemmatizaci a POS-tagging, přičemž v [8] uvádějí, že zatímco POS-tagging je pro úspěšnost klasifikátoru důležitý, lemmatizace nevede k výraznému zlepšení výsledků. Ke stejnému názoru dospívá i Toman [25], který zkoumá vliv normalizace slov na klasifikaci. Podle něj je lepším přístupem pouhé odstraňování stop-slov bez použití lemmatizace. Další studie [26] navrhuje pro reprezentaci textů použití jmenných, předložkových a slovesných skupin (místo nebo vedle jednotlivých slov). Král [9] popisuje další experiment s vytvářením atributů, a to pomocí tzv. *pojmenovaných entit* (*named entities*). Avšak ani jeden ze dvou posledních přístupů nevedl k významnému zvýšení úspěšnosti klasifikátoru.

Lexikální struktury textu při strojovém zpracování česky psaných dokumentů využívá studie [27] zkoumající rozpoznávání dialogových aktů (*dialogue act*). Autory navrhují nový přístup: vedle lexikálních atributů používají informace i o pozici slov ve větě. Jejich experiment ukazuje, že tyto informace zlepšují přesnost rozpoznávání dialogových aktů.

Přímo funkčním stylům v kontextu automatického zpracování jazyka se věnuje ve svém článku Panevová [28], avšak z hlediska automatické opravy

chyb, nikoliv klasifikace textů.

3.2 Hodnocení klasifikátorů

Po naučení klasifikátoru bývá zpravidla potřeba ověřit jeho přesnost. To znamená, že je nutno zjistit, jakou část z předložených mu nových dat dokáže klasifikátor zařadit do správné třídy. K tomuto účelu slouží různé metodiky, nejznámější jsou tzv. *train-and-test* a *křížová validace* (*k-fold cross validation*). Níže následuje krátké vysvětlení těchto dvou přístupů[11] a také uvedení dalších pojmů souvisejících se způsoby hodnocení klasifikátorů.

Principem *train-and-test* přístupu je rozdělení množiny vzorků, která je určena k naučení klasifikátoru, na dvě části: trénovací a testovací. Trénovací množina se používá při učení klasifikátoru; na testovací množině se pak ověřuje jeho úspěšnost. Důležité je, aby tyto množiny byly disjunktní, jinak by byly výsledky testování zkreslené. Občas bývá z trénovací množiny oddělena validační množina, která slouží k otestování nastavení parametrů klasifikátoru na různé hodnoty.

Křížová validace se zakládá na rozdělení množiny vzorků na k části, za nímž následuje vytvoření k klasifikátorů, které jsou vyhodnoceny jako *train-and-test*, přičemž pro každý klasifikátor slouží jako testovací množina jiná část původní množiny. Pro získání výsledné přesnosti musí být přesnosti dílčích klasifikátorů zprůměrovány.

Přesnost (*accuracy*) klasifikátoru je vypočítána podle vzorce

$$\frac{TP + TN}{TP + TN + FP + FN}, \tag{3.1}$$

kde TP (*true positives*) označuje počet vzorků třídy c_i, které byly klasifikovány jako c_i; TN (*true negatives*) je počet vzorků, které nepatří do třídy c_i a nebyly klasifikovány jako c_i; FP (*false positives*) je počet vzorků, které nepatří do c_i, ale byly klasifikovány jako c_i; FN (*false negatives*) označuje počet vzorků, které patří do c_i, ale nebyly klasifikovány jako c_i.

		rozhodnutí experta	
		c_i	$\overline{c}_i$
rozhodnutí klasifikátoru	c_i	TP_i	FP_i
	$\overline{c}_i$	FN_i	TN_i

Tabulka 3.1: Konfúzní matice

TP, TN, FP a FN jsou zpravidla uváděny v konfúzní matici (neboli matici záměn, tab. 3.1). Kromě přesnosti lze z těchto hodnot vypočítat i další míry popisující úspěšnost klasifikátoru:

[11] Podle [6].

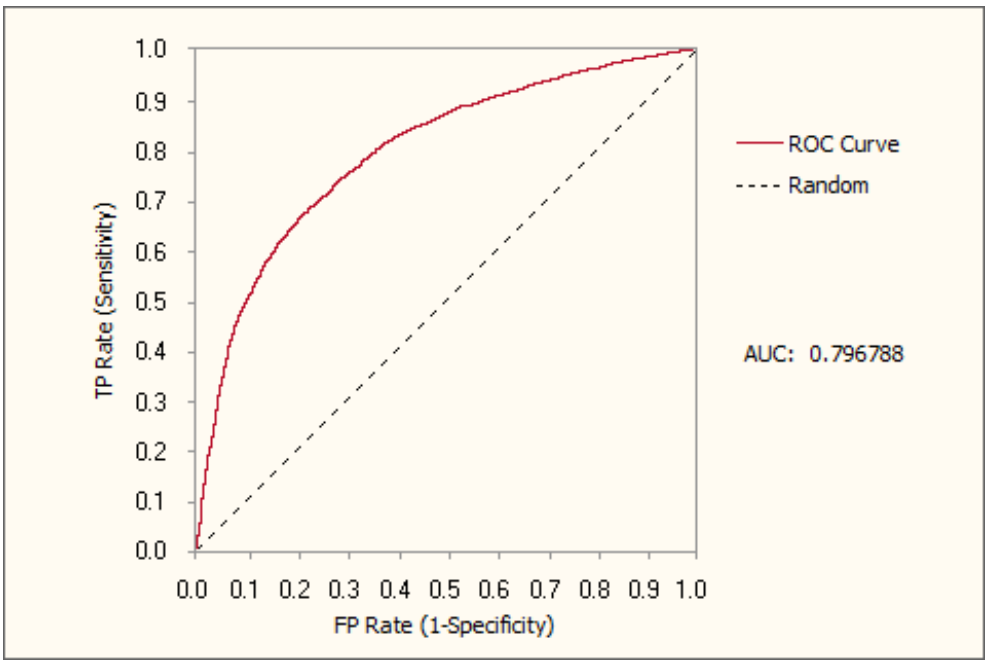
- *chybovost* (*error rate*): vypočítává se jako $ERR = 1 - ACC$, kde ERR je chybovost klasifikátoru a ACC je přesnost klasifikátoru;
- *přesnost*[12] (*precision*) je pravděpodobnost, že pokud náhodný dokument d_i byl klasifikován jako c_j, pak skutečně patří do c_j; vypočítává se jako $\frac{TP}{TP+FP}$;
- *úplnost* (*recall*, také *senzitivita*) je pravděpodobnost, že pokud náhodný dokument d_i patří do třídy c_j, pak bude klasifikován do této třídy; vypočítává se jako $\frac{TP}{TP+FN}$;
- *specificita* (*specificity*) je pravděpodobnost, že pokud náhodný dokument d_i byl klasifikován jako $\overline{c}_j$, pak skutečně nepatří do c_j; vypočítává se jako $\frac{TN}{TN+FP}$;

Úplnost bývá označována i jako *true positive rate* (TPR) a specificita jako *true negative rate*. S tím souvisí pojem *false positive rate* (FPR), který se také nazývá *fall-out* a rovná se 1 − specificita.

TPR a FPR jsou hodnoty, na jejichž základě je vytvořen nástroj používaný mimo jiné i pro ilustraci úspěšnosti klasifikátoru – *ROC křivka*. ROC křivka je tvořena všemi možnými dvojicemi (TPR,FPR) [29] a plocha pod ní, které se říká AUC (*Area Under Curve*) je dalším ukazatelem úspěšnosti klasifikátoru. AUC = 0.5 odpovídá náhodnému klasifikátoru a AUC = 1 klasifikátoru ideálnímu. Příklad ROC křivky je znázorněn na obrázku 3.1.

Je nutno podotknout, že ROC křivka a AUC se používají při hodnocení binárních klasifikátorů.

[12] *Accuracy* i *precision* se do českého jazyka překládají jako *přesnost*. V dalším textu je pojmem *přesnost* míněno *accuracy*, pokud není uvedeno jinak.



Obrázek 3.1: ROC křivka a AUC.
Zdroj: `http://www.gepsoft.com/`

Kapitola 4

Návrh vlastního řešení

Úkolem práce je navržení algoritmu, který identifikuje funkční styl textového dokumentu. Tento úkol lze vyřešit jako klasickou klasifikační úlohu: nechť d je dokument z množiny dokumentů $\mathbb{D}$ a $\mathbb{C} = \{c_1, c_2, c_3, c_4, c_5, c_6\}$ je množina tříd, kde jeden prvek odpovídá jednomu funkčnímu stylu. Pak „identifikovat funkční styl dokumentu d“ znamená přiřadit tomuto dokumentu jednu ze šesti tříd z množiny $\mathbb{C}$. Jedná se o klasifikaci typu *one-of*, tj. každému dokumentu může být přiřazena právě jedna třída.

Jak bylo ukázáno v kapitole 3, pro úlohu klasifikace textů podle stylu se často volí reprezentace dokumentů pomocí lexikálních atributů, tj. pomocí slov obsažených v textu. Studie, které staví experimenty s využitím tohoto přístupu, dosahují celkem dobré přesnosti klasifikátoru, ale v rámci této práce se pokusíme klasifikovat texty pomocí alternativního přístupu: dokumenty budou reprezentovány pomocí strukturálních atributů. Toto rozhodnutí bylo učiněno na základě stylistické teorie českého jazyka, která popisuje funkční styly právě z hlediska stavby vět i celých textů a také morfologických specifik češtiny (např. dvojích koncovek u 1. osoby j. č. a 3. osoby m. č. sloves). Dalším důvodem je ten fakt, že studie, které se zabývají klasifikací česky psaných textů, se zaměřují na reprezentaci pomocí lexikálních atributů a problémy spojené s tímto přístupem (především způsoby redukce počtu atributů)[13]. Některé z nich kombinují lexikální příznaky s informacemi syntaktického charakteru (např. pozice slov ve větě), ale o pokusech využití ryze strukturálních atributů pro klasifikaci českých textů se nepodařilo najít žádnou informaci.

Následující podkapitoly popisují jednotlivé kroky algoritmu, které korespondují s čtyřmi kroky zmíněnými v podkapitole 2.2.

[13] Musíme také podotknout, že se zpravidla jedná o klasifikaci podle tématu, nikoliv podle stylu textu.

4.1 Určení atributů

Nejprve je nutno navrhnout, jak budou reprezentovány textové dokumenty sloužící k naučení klasifikátoru. O tom, jaké atributy by mohly vhodně reprezentovat texty za účelem klasifikace podle funkčního stylu, se rozhodovalo převážně na základě popisu vlastností funkčních stylů v [2] a částečně na základě vlastního zkoumání dokumentů.

Všechny atributy lze rozdělit do následujících podskupin:

1. *Struktura textu.* Do této kategorie patří statistické hodnoty, jako je např. průměrná délka věty, a také informace o horizontálním členění textu.

2. *Interpunkce a speciální symboly.* Speciálními symboly se rozumí všechny znaky odlišné od malých a velkých písmen české abecedy, číslic a interpunkčních znamének.

3. *Slovní druhy.* Podskupina zahrnuje všechny atributy vyjadřující zastoupení jednotlivých slovních druhů v textu.

4. *Morfologické příznaky.* Veškeré příznaky, které souvisí se skloňováním a časováním (popř. i dalšími způsoby tvarosloví) se nachází v dané kategorii.

5. *Syntaktické příznaky.* Sem patří informace o výstavbě věty a větných členech.

6. *Stylistické příznaky.* Do této kategorie lze zařádit některá funkční slova, jejichž použití může být podmíněno funkcí textu (např. zájmeno *jenž* nebo tvary slov označované jako nespisovné). Ačkoliv se jedná spíše o lexikální atributy, k jejich určení není potřeba vytvářet speciální slovník, protože některé POS-tagging nástroje umožňují detekovat nespisovné tvary slov.

V tabulce 4.1 jsou shrnuty všechny atributy s uvedením jejich skupiny, datového typu a případně způsobu jejich výpočtu. Některé popisy atributů, které by mohly být čtenáři nejasné, jsou vysvětleny níže.

Zda se v textu vyskytují dialogy, je určováno následujícím způsobem: pokud jsou v textu alespoň dva sousední řádky začínající uvozovkami, počítá se, že v textu je alespoň jeden dialog (proměnná je nastavena na *true*, jinak je nastavena na *false*).

Podobně se detekuje výskyt výčtů: v textu musí být alespoň dva sousední řádky, které začínají pomlčkou nebo jednou ze čtyř kombinací dvou symbolů: prvním symbolem je písmeno nebo číslice, druhým symbolem je znak „.“ nebo „)“.

Za neúplnou větu je považována věta bez přísudku.

4.2 Převod textových dokumentů na vektory hodnot

Korpus dokumentů (uložených jako *plain text*, tj. neobsahujících žádná metadata, jako např. formátování, xml-tagy apod.) je zpracován algoritmem vypočítávajícím hodnoty navržených v prvním kroku atributů pro každý dokument zvlášť. Výstupem tohoto algoritmu je matice, jejíž řádky (vektory) představují dokumenty a sloupce (složky vektorů) jsou atributy. Kromě toho je do matice přidán sloupec obsahující značky funkčního stylu dokumentů, tj. jejich třídy (celočíselné hodnoty od 1 do 6).

K převodu dokumentů na vektory je nutno použít POS-tagging nástroj, s jehož pomocí lze spočítat hodnoty atributů ze skupin 2 – 6. Pro určení hodnot ostatních atributů budou napsány vlastní metody. Popis zvoleného POS-tagging nástroje obsahuje kapitola 6.

4.3 Vytvoření klasifikátoru

Pro vytvoření klasifikačního algoritmu bylo rozhodnuto využít metod strojového učení s učitelem. Vstupem pro ně slouží matice vzniklá v předchozím kroku. Na základě provedené rešerše bylo usouzeno, že nejvhodnějšími v rámci dané práce by mohly být následující metody učení s učitelem: Rozhodovací stromy, Naivní Bayesův klasifikátor, SVM a k-NN. Dané algoritmy bývají často voleny pro úlohu klasifikace textů [26] a zejména SVM a Naivní Bayesův klasifikátor jsou považovány za jedny z nejúspěšnějších, co se týče přesnosti naučeného klasifikátoru.

Úspěšnost těchto čtyř metod bude změřena a porovnána. Ve výsledném algoritmu identifikace funkčního stylu bude použita metoda vybraná na základě porovnání úspěšnosti klasifikátorů.

4.4 Hodnocení klasifikátoru

Měření úspěšnosti klasifikátorů proběhne pomocí křížové validace. Budou spočítány tyto ukazatele: přesnost (*accuracy*) metody a senzitivita (TPR) pro každou třídu v rámci jedné metody. Dále budou metody porovnány na základě ROC křivky a hodnoty AUC. Jelikož klasifikace podle funkčního stylu není binární klasifikací, bude nutno vytvořit ROC křivky a spočítat hodnoty AUC pro každou třídu c_j zvlášť, a to tak, že ostatních pět tříd budou chápány jako doplněk třídy c_j.

Skupina	Atribut	Datový typ	Popis
struktura textu	rozdělení na kapitoly	boolean	Text je / není dělen na kapitoly
	počet odstavců	int	
	průměrná délka odstavce	double	Počet slov ÷ Počet odstavců
	průměrná délka věty	double	Počet slov ÷ Počet vět
	dialogy	boolean	V textu jsou / nejsou dialogy
	výčty	boolean	V textu jsou / nejsou výčty
interpunkce a speciální symboly	počet vykřičníků	int	
	počet otazníků	int	
	počet středníků	int	
	průměrný počet čárek ve větě	double	Počet čárek ÷ Počet vět
	znak %	boolean	V textu je / není aspoň 1 znak %
	znak §	boolean	V textu je / není aspoň 1 znak §
	počet speciálních symbolů	int	
slovní druhy	podíl sloves	double	Počet sloves ÷ Počet slov
	podíl podstatných jmen	double	Počet podstatných jmen ÷ Počet slov
	podíl přídavných jmen	double	Počet přídavných jmen ÷ Počet slov
	podíl přísloví	double	Počet přísloví ÷ Počet slov
	podíl spojek	double	Počet spojek ÷ Počet slov
	podíl citoslovcí	double	Počet citoslovcí ÷ Počet slov
	podíl číslovek (psaných číslicemi)	double	Počet číslovek ÷ Počet slov
morfologické příznaky	výskyt podstatných jmen v 5. pádě	boolean	V textu je / není aspoň 1 substantivum v 5. pádě
	podíl plurálu 2. osoby sloves	double	Počet sloves 2. os. pl. ÷ Počet sloves 2. os.
	infinitivy končící *-ci*	boolean	V textu je / není aspoň jeden infinitiv končící *-ci*
	slovesní koncovky 1. osoby singuláru	double	Počet sloves s koncovkou *-i* ÷ Počet sloves 1. os. sg.
	slovesní koncovky 3. osoby plurálu	double	Počet sloves s koncovkou *-ou* ÷ Počet sloves 3. os. pl.
	podíl vět s nepřímým slovosledem	double	Počet vět s nepřímým slovosledem ÷ Počet vět
	podíl neúplných vět	double	Počet neúplných vět ÷ Počet vět
syntaktické příznaky	podíl pasivních tvarů sloves	double	Počet pasivních tvarů ÷ Počet sloves
	průměrný počet sloves ve větě	double	Počet sloves ÷ Počet vět
	podíl rozkazovacího způsobu	double	Počet sloves v rozk. způsobu ÷ Počet sloves
	podíl podmiňovacího způsobu	double	Počet sloves v podm. způsobu ÷ Počet sloves
stylistické příznaky	tvary zájmena *jenž*	boolean	V textu jsou / nejsou tvary zájmena *jenž*
	tvary zájmen *týž* a *tentýž*	boolean	V textu jsou / nejsou tvary zájmen *týž* a *tentýž*
	četnost výskytu tvarů zájmena *ten*	double	Počet tvarů zájmena *ten* ÷ Počet slov
	výskyt nespisovných tvarů	boolean	V textu jsou / nejsou nespisovné tvary slov

Tabulka 4.1: Popis atributů sloužících k reprezentaci dokumentů

Kapitola 5

Popis v práci použitých dat

Vstupem pro navrhovaný algoritmus je soubor textových dokumentů psaných v českém jazyce (dále korpus). Jelikož pro metody strojového učení s učitelem je velice důležitá kvalita vstupních dat, byly texty vybírány a zařazovány do jednotlivých tříd manuálně.

Každý dokument je uložen do vlastního souboru jako plain-text v kódování UTF-8. Dokumenty jsou rozděleny do šesti složek podle funkčního stylu. Tabulka 5.1 popisuje zastoupení jednotlivých funkčních stylů v korpusu. Celkový počet vzorků je 1511.

Nejproblematičtějším z hlediska výběru reprezentativních vzorků se jeví funkční styl hovorový. Tento funkční styl je typický především pro mluvený jazyk, a proto by se k jeho reprezentaci nejvíce hodily přepisy audiozáznamů mluvených projevů. Bohužel takové přepisy je velmi obtížné najít v dostatečném množství a hlavně v požadované kvalitě (mluvené texty jsou zpravidla zaznamenávány pomocí transkripce, která se liší od „tradičního“ psaného textu a proto není vhodná pro zpracování nástroji pracujícími s texty psanými v přirozeném jazyce). Z tohoto důvodu bylo rozhodnuto použít texty sice psané, ale stylisticky podobné textem mluveným: příspěvky na fórech, hodnocení zboží v internetových obchodech apod.

funkční styl	**počet textů**	**druhy textů**
administrativní	200	zákony, vyhlášky, rozhodnutí, předpisy apod.
hovorový	300	příspěvky na fórech, hodnocení produktů
odborný	208	odborné práce a články, vědecké posudky a recenze
publicistický	300	zprávy, interview, reportáže
řečnický	203	proslovy, projevy
umělecký	300	romány, povídky, pohádky, básně apod.

Tabulka 5.1: Zastoupení funkčních stylů v korpusu

Kapitola 6

Implementace

6.1 Volba programovacího jazyka a pomocného softwaru

Navržený algoritmus byl implementován v jazyce Java.

Pro lemmatizaci a POS-tagging vstupních dokumentů byl použit nástroj MorphoDiTa, který je podrobněji popsán v podkapitole 6.2. K vyhodnocení a porovnání úspěšnosti různých metod učení s učitelem sloužil nástroj RapidMiner Studio [30].

Metoda, která byla vyhodnocena jako nejúspěšnější, byla následně použita v algoritmu, a to za pomocí knihovny *Java Machine Learning Library* (Java-ML, podrobněji v podkapitole 6.4).

6.2 MorphoDiTa

MorphoDiTa (Morphological Dictionary and Tagger) [31] je nástroj pro morfologickou analýzu textů psaných v přirozeném jazyce. Je vyvíjen Ústavem formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy a vydává se pod licenci Creative Commons (CC BY-NC-SA).

MorphoDiTa je dostupný jako samostatný nástroj nebo jako knihovna. Ačkoliv je psán v jazyce C++, je distribuován společně s API pro další programovací jazyky včetně Java.

Pro použití MorphoDiTa v Java-programu je potřeba následující:

- Java 6 nebo novější;
- G++ 4.7 nebo novější;
- SWIG 2.0.5 nebo novější;
- make.

Kromě samotného nástroje je třeba mít stažený jazykový model, který je nezbytný pro provedení morfologické analýzy. Doporučuje se použití modelu *MorfFlex CZ 160310* [32].

Podrobný návod k instalaci lze najít na stránkách MorphoDiTa [33]. CD přiložené k dané práci obsahuje již přeložené soubory.

6.3 Práce s POS-tagy

V této podkapitole je popsána struktura POS-tagů neboli pozičních tagů.

POS-tag představuje řetězec skládající se z 15 znaků. Každá pozice kóduje jednu morfologickou kategorii pomocí jednoho znaku. Popis pozicí je shrnut v tabulce 6.1.

pozice	**název**	**popis**
1	POS	slovní druh
2	SubPOS	slovní poddruh
3	Gender	rod
4	Number	číslo
5	Case	pád
6	PossGender	rod vlastníka
7	PossNumber	číslo vlastníka
8	Person	osoba
9	Tense	čas
10	Grade	stupeň
11	Negation	negace
12	Voice	slovesný rod
13,14	Reserve1,2	rezervy
15	Var	varianta, styl

Tabulka 6.1: Poziční tagy

Například pozice 1 může nabývat těchto hodnot: N (podstatné jméno), A (přídavné jméno), V (sloveso), D (příslovce), P (zájmeno), C (číslovka), R (předložka), I (citoslovce), J (spojka), T (částice), X (neznámý slovní druh) a Z (interpunkce). Detailní výčet všech hodnot každé pozice lze najít v [34].

6.4 Java-ML

Java-ML je kolekce algoritmů strojového učení. Knihovna poskytující rozhraní pro mnohé metody používané při strojovém zpracování dat je distribuována pod licenci GNU-GPL. Podrobnější informace a dokumentaci lze najít v [35] a [36].

Kapitola 7

Experiment

Měření úspěšnosti čtyř klasifikátorů (k-NN, Rozhodovací stromy, Naivní Bayesův klasifikátor a SVM) bylo provedeno s použitím nástroje RapidMiner Studio. Přesnost se určovala pomocí křížové validace (operátor X-Validation) s následujícím nastavením:

- number of validations = 10
- sampling type = automatic

7.1 Nastavení parametrů pro k-NN

Nejvyšší přesnosti pro předložená data dosáhla metoda k-NN pro tyto parametry:

- k = 6
- measure types = NumericalMeasure
- numerical measure = ManhattanDistance

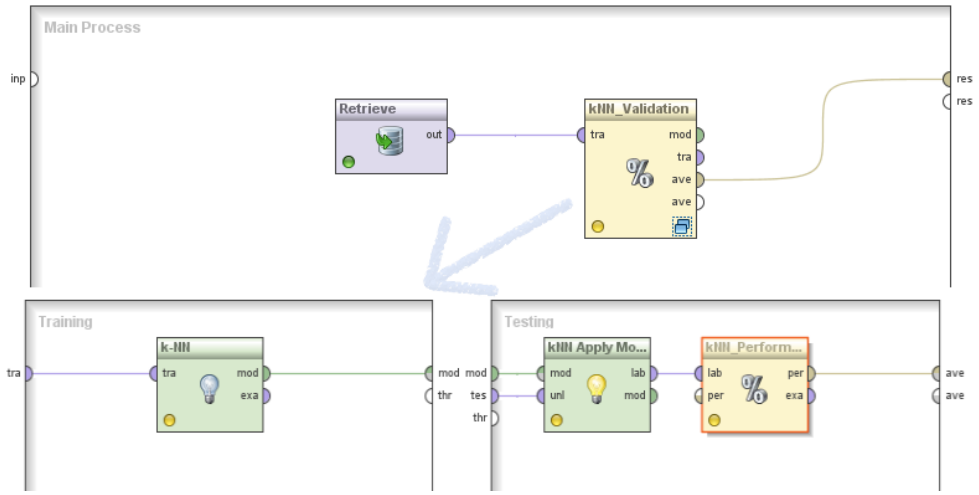
Obrázek 7.1 znázorňuje schéma zapojení operátorů pro k-NN.

7.2 Nastavení parametrů pro Rozhodovací stromy

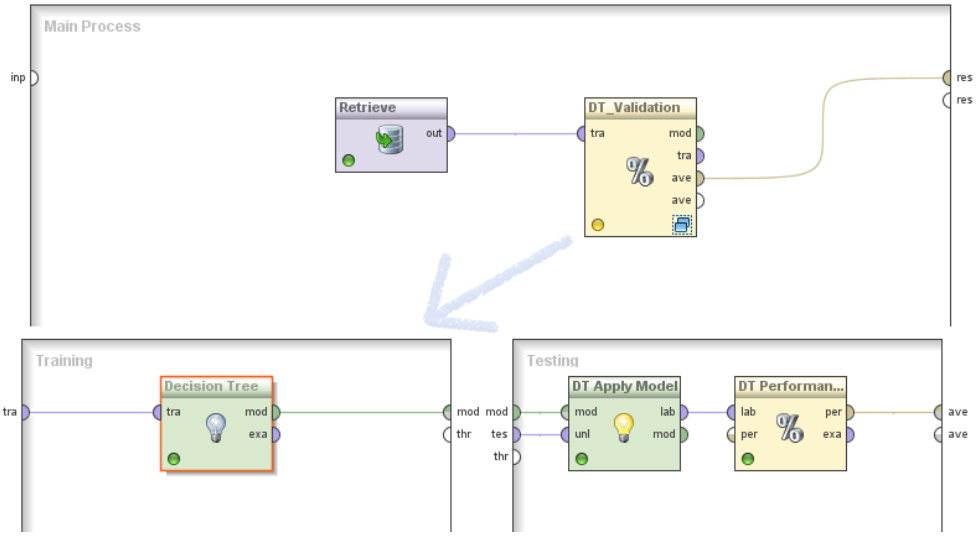
Nejlepší přesnost vykázaly Rozhodovací stromy s tímto nastavením:

- criterion = gini index
- max_depth = 33
- apply pruning = true (default values)
- apply prepruning = true (default values)

Obrázek 7.2 znázorňuje schéma zapojení operátorů pro Rozhodovací stromy.



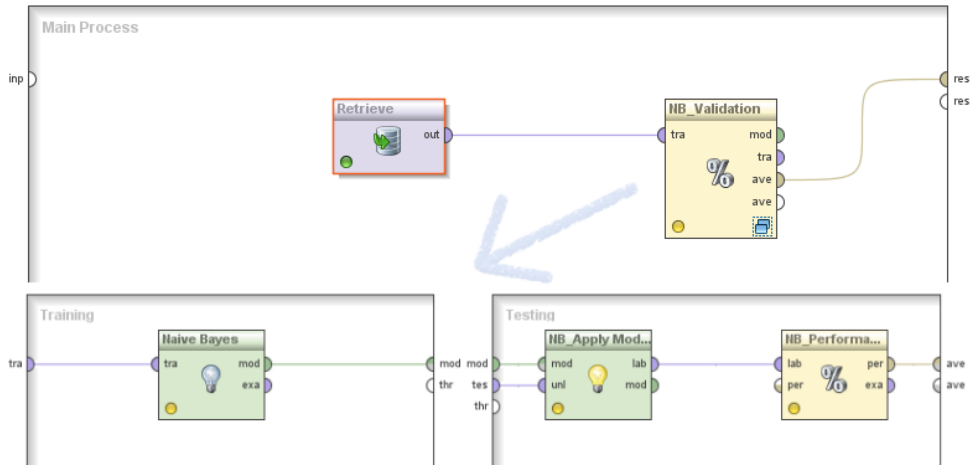
Obrázek 7.1: Schéma zapojení pro k-NN



Obrázek 7.2: Schéma zapojení pro Rozhodovací stromy

7.3 Nastavení parametrů pro Naivní Bayesův klasifikátor

Pro Naivní Bayesův klasifikátor nebyly nastavovány žádné parametry. Obrázek 7.3 znázorňuje schéma zapojení operátorů.

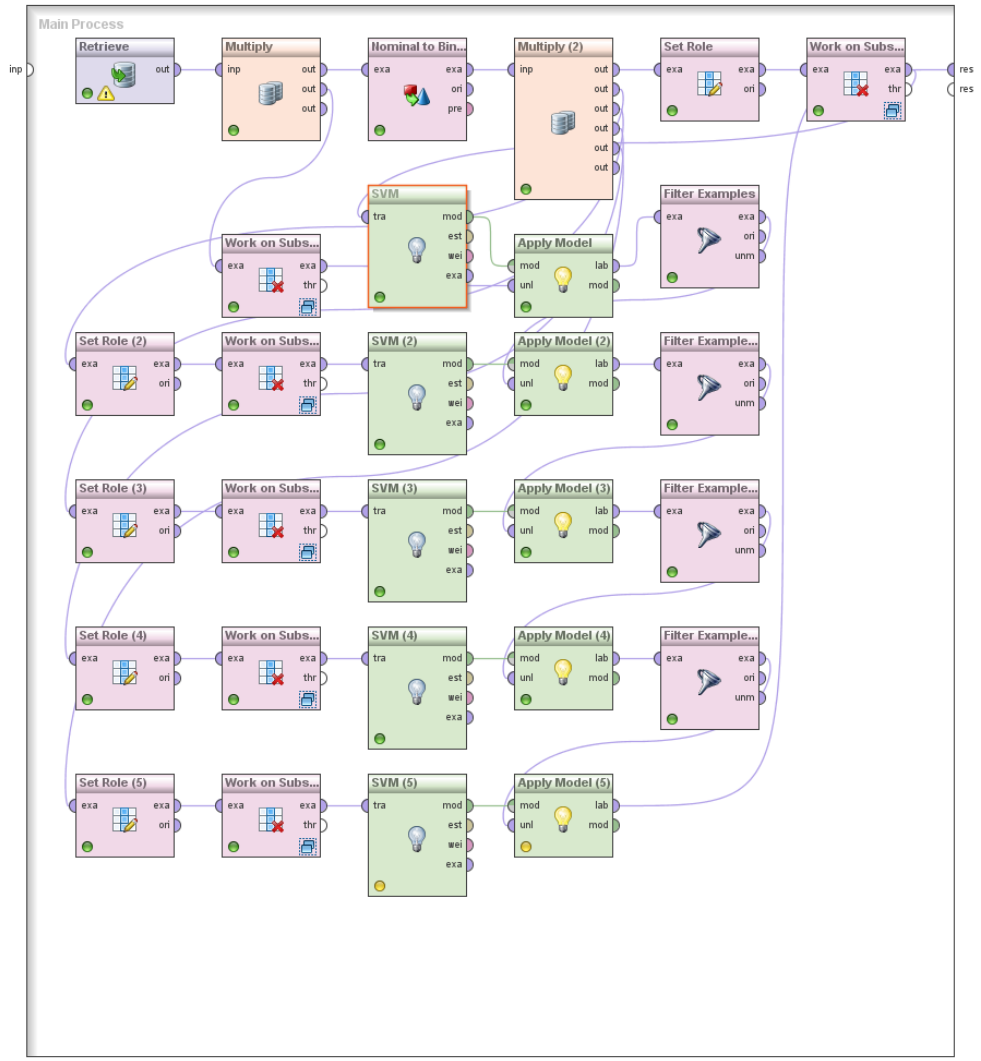


Obrázek 7.3: Schéma zapojení pro Naivní Bayesův klasifikátor

7.4 Nastavení parametrů pro SVM

Jak bylo zmíněno v sekci 2.2.4, metoda SVM je vhodná především pro binární klasifikaci. Pro klasifikaci textů podle šesti funkčních stylů pomocí SVM byly proto naučeny pět klasifikátorů: první klasifikátor rozděloval texty z korpusu do tříd c_1 a $\overline{c}_1$; druhý rozděloval texty označené prvním klasifikátorem jako $\overline{c}_1$ do tříd c_2 a $\overline{c}_2$; třetí klasifikátor pracoval s množinou $\overline{c}_2$ a třídil ji do c_3 a $\overline{c}_3$ atd. Poslední (pátý) klasifikátor dělil zbylé texty do tříd c_5 a $\overline{c}_5 = c_6$ (schéma zapojení je na obrázku 7.4). Parametry všech pěti klasifikátorů byly nastaveny na stejné hodnoty:

- kernel type = dot
- kernel cache = 200
- C = 0.0
- convergence epsilon = 0.001
- max iterations = 100000
- scale = true
- L pos = L neg = 1.0
- ostatní: 0.0



Obrázek 7.4: Schéma zapojení pro SVM

Kapitola 8

Výsledky

Na následujících obrazcích jsou jednotlivé funkční styly očíslovány takto:

1. umělecký funkční styl
2. odborný funkční styl
3. publicistický funkční styl
4. řečnický funkční styl
5. hovorový funkční styl
6. administrativní funkční styl

8.1 k-NN

Nejvyšší dosažená přesnost činila u k-NN 74.98%. Obrázek 8.1 znázorňuje míru senzitivity pro každý funkční styl.

accuracy: 74.98% +/- 4.62% (mikro: 74.98%)

	true 1	true 2	true 3	true 4	true 5	true 6	class precision
pred. 1	201	13	6	9	1	2	86.64%
pred. 2	17	144	17	10	0	21	68.90%
pred. 3	27	23	250	37	26	16	65.96%
pred. 4	13	15	11	131	4	7	72.38%
pred. 5	32	2	12	15	263	10	78.74%
pred. 6	10	11	4	1	6	144	81.82%
class recall	67.00%	69.23%	83.33%	64.53%	87.67%	72.00%	

Obrázek 8.1: Matice záměn k-NN

8.2 Rozhodovací stromy

U Rozhodovacích stromů dosáhla přesnost 79,48%. Na obrázku 8.2 jsou uvedeny hodnoty senzitivity tříd pro tuto metodu.

accuracy: 79.48% +/- 2.94% (mikro: 79.48%)

	true 1	true 2	true 3	true 4	true 5	true 6	class precision
pred. 1	245	6	26	17	13	4	78.78%
pred. 2	1	170	14	10	1	16	80.19%
pred. 3	18	17	218	32	6	7	73.15%
pred. 4	23	8	29	133	12	4	63.64%
pred. 5	13	1	5	9	266	0	90.48%
pred. 6	0	6	8	2	2	169	90.37%
class recall	81.67%	81.73%	72.67%	65.52%	88.67%	84.50%	

Obrázek 8.2: Matice záměn rozhodovacích stromů

8.3 Naivní Bayesův klasifikátor

Naivní Bayesův klasifikátor měl přesnost klasifikace rovnou 81,73%. Míry senzitivity funkčních stylů jsou na obrázku 8.3.

accuracy: 81.73% +/- 1.91% (mikro: 81.73%)

	true 1	true 2	true 3	true 4	true 5	true 6	class precision
pred. 1	201	8	3	3	4	0	91.78%
pred. 2	2	142	2	2	0	7	91.61%
pred. 3	22	21	250	11	6	7	78.86%
pred. 4	61	17	37	179	9	1	58.88%
pred. 5	14	0	5	8	278	0	91.15%
pred. 6	0	20	3	0	3	185	87.68%
class recall	67.00%	68.27%	83.33%	88.18%	92.67%	92.50%	

Obrázek 8.3: Matice záměn Naivního Bayesova klasifikátoru

8.4 SVM

V případě SVM se jednalo o klasifikátor složený z dílčích binárních klasifikátorů (popis v sekci 7.4). Hodnocení takto postaveného klasifikátoru pro všechny třídy by bylo poměrně komplikované, proto jsou na obrázku 8.4 výsledky binární klasifikace pro každou ze šesti tříd zvlášť (*true* je třída c_i, *false* je její doplněk). Toto hodnocení je zajímavé z hlediska porovnání senzitivity funkčních stylů. Nejnižší TPR má umělecký funkční styl, tj. texty tohoto funkčního stylu bývají nejčastěji klasifikovány do jiné třídy. Nízký TPR má také funkční styl odborný. Senzitivita ostatních funkčních stylů se pohybuje nad 90%.

Pokud se podíváme na hodnoty přesnosti (*precision*) tříd, zjistíme, že nejmenší přesnosti dosáhl funkční styl publicistický, druhé nejmenší funkční styl hovorový. To znamená, že do těchto tříd bývají často klasifikovány texty jiných funkčních stylů. Poměrně nízkou přesnost má také funkční styl řečnický, zatímco ostatní funkční styly dosáhly přesnosti 90% a vyšší.

accuracy: 92.59% 1

	true false	true true	class precision
pred. false	1202	103	92.11%
pred. true	9	197	95.63%
class recall	99.26%	65.67%	

accuracy: 94.71% 2

	true false	true true	class precision
pred. false	1287	64	95.26%
pred. true	16	144	90.00%
class recall	98.77%	69.23%	

accuracy: 80.01% 3

	true false	true true	class precision
pred. false	927	18	98.10%
pred. true	284	282	49.82%
class recall	76.55%	94.00%	

accuracy: 91.46% 4

	true false	true true	class precision
pred. false	1192	13	98.92%
pred. true	116	190	62.09%
class recall	91.13%	93.60%	

accuracy: 86.37% 5

	true false	true true	class precision
pred. false	1008	3	99.70%
pred. true	203	297	59.40%
class recall	83.24%	99.00%	

accuracy: 97.82% 6

	true false	true true	class precision
pred. false	1295	17	98.70%
pred. true	16	183	91.96%
class recall	98.78%	91.50%	

Obrázek 8.4: Matice záměn SVM

8.5 ROC křivka a AUC

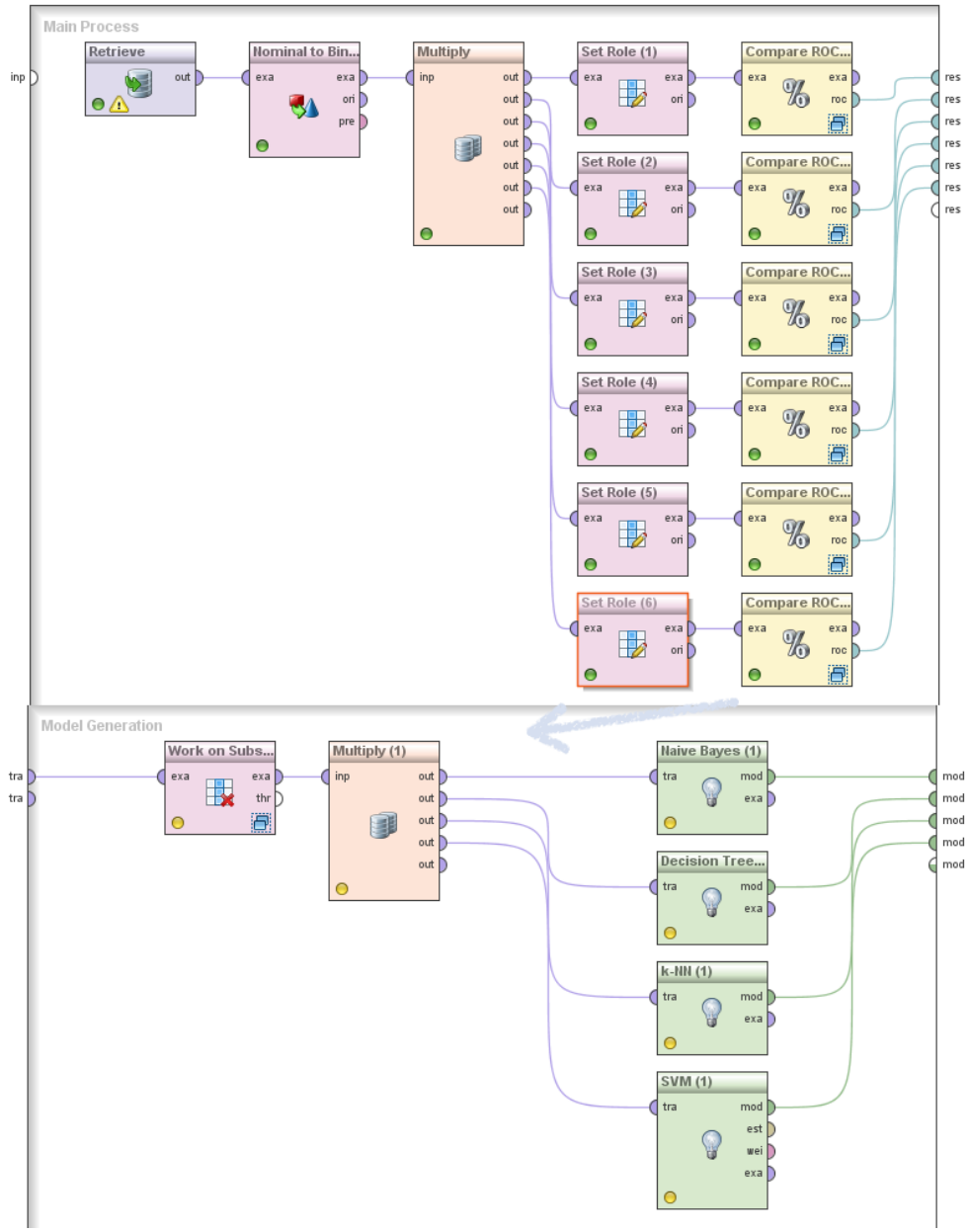
Úspěšnost klasifikátorů byla také porovnána pomocí ROC křivky a AUC. Připomeňme, že AUC je plocha pod ROC křivkou, která nabývá hodnot $\langle 0.5, 1 \rangle$, přičemž větší hodnota znamená větší úspěšnost klasifikátoru. Jelikož jsou tyto ukazatele vhodné pouze pro binární klasifikaci, opět bylo vytvořeno šest klasifikátorů pro každý funkční styl zvlášť.

Obrázky 8.6 až 8.11 porovnávají ROC křivky čtyř klasifikátorů. Schéma zapojení v RapidMiner Studio je na obrázku 8.5. Co se týče hodnot AUC, pohybovaly se ve většině případů nad 0.9. Pouze u metody Rozhodovacích stromů byla AUC výrazně nižší pro dva binární klasifikátory, konkrétně u klasifikace podle funkčních stylů publicistického a řečnického[14].

8.6 Výběr klasifikační metody

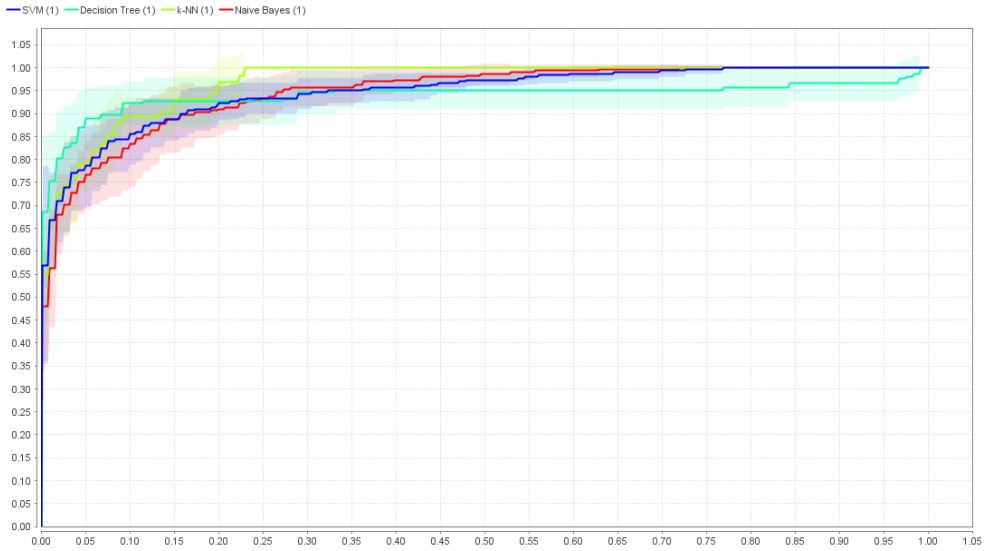
Jelikož implementace metody SVM pro nebinární klasifikační úlohu je obtížnější, byla klasifikační metoda pro realizaci v dané práci navrhovaného algo-

[14]Obrázky ilustrující AUC nenesou velkou informační hodnotu, proto nejsou uvedeny v tištěné verzi práce. Lze je najít na přiloženém CD.

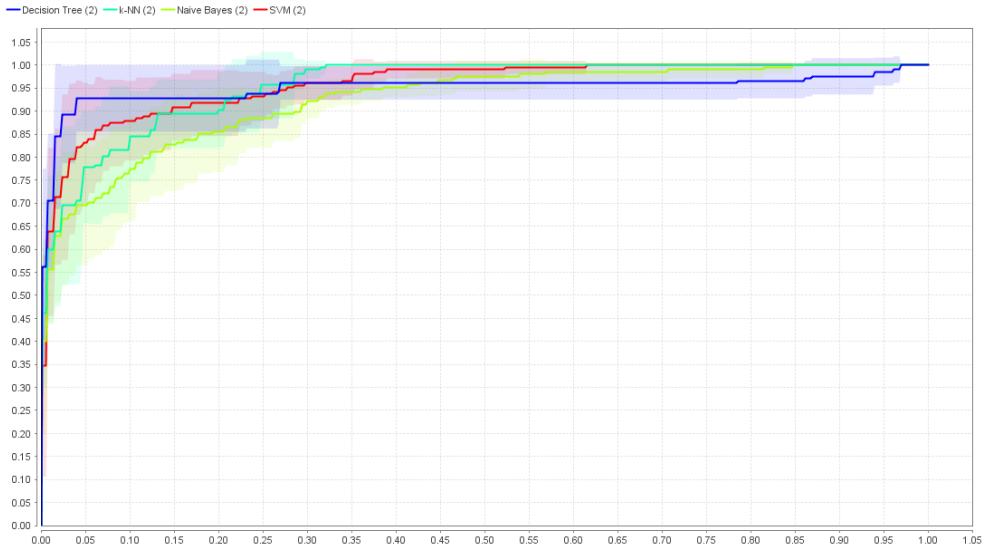


Obrázek 8.5: Schéma zapojení pro ROC křivky

ritmu vybírána mezi zbylými třemi metodami (k-NN, Naivní Bayesův klasifikátor, Rozhodovací stromy). Jak bylo ukázáno v předchozích sekcích, nejvyšší přesnosti dosáhl Naivní Bayesův klasifikátor. Tato metoda má také výhodu v tom, že v porovnání s Rozhodovacími stromy probíhá klasifikace rychle (jedná se o výpočet pravděpodobnosti). Proto byl při implementaci algoritmu použit právě Naivní Bayesův klasifikátor.

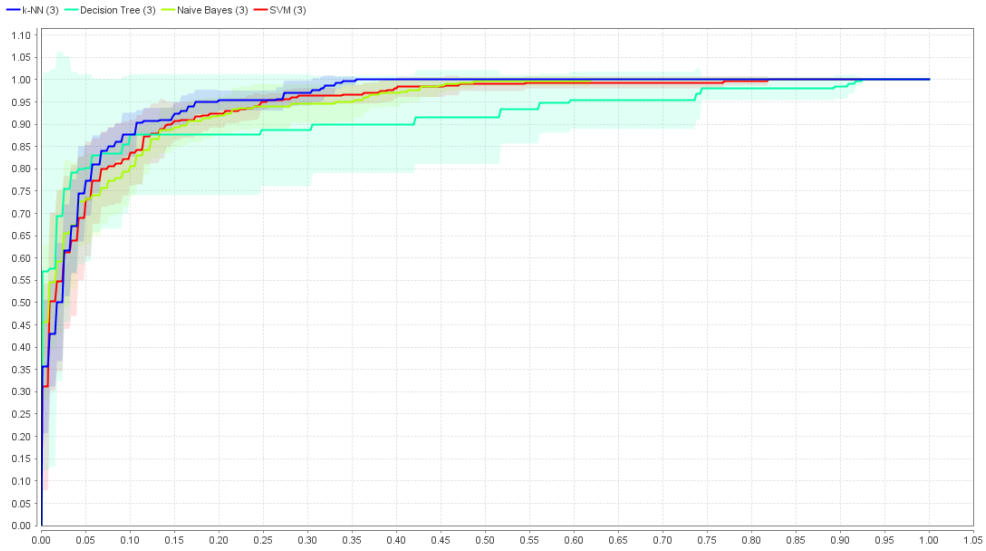


Obrázek 8.6: ROC křivky (umělecký funkční styl)

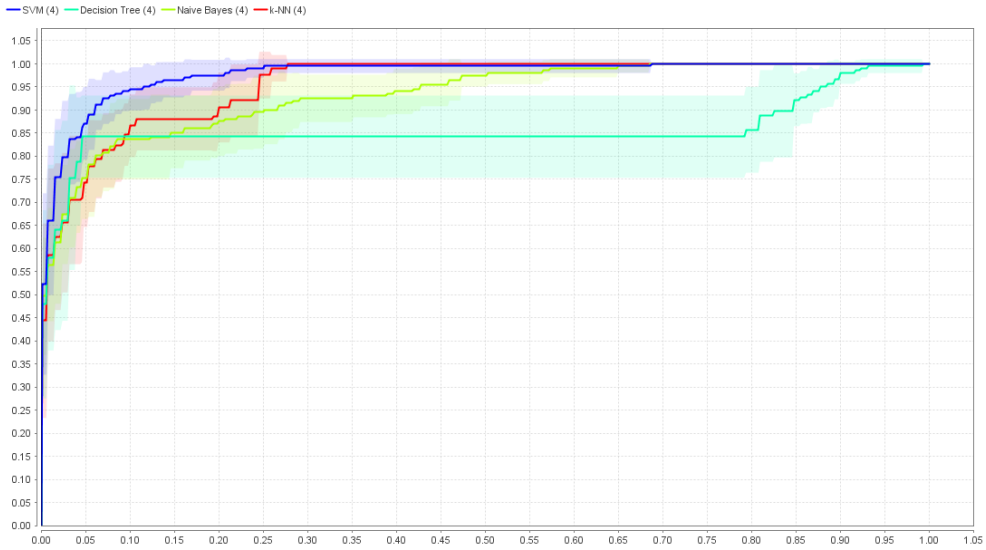


Obrázek 8.7: ROC křivky (odborný funkční styl)

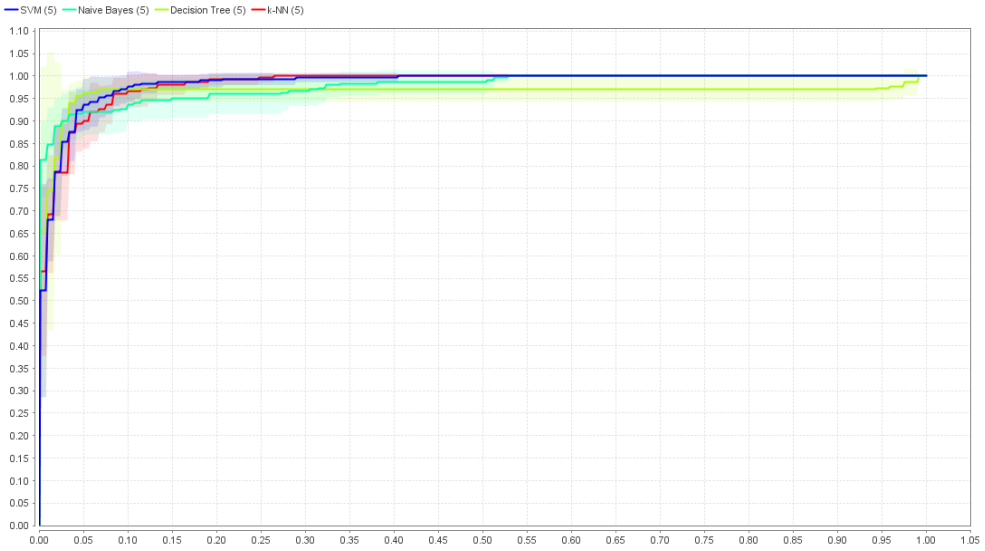
Z obrázku 8.12 je vidět, že Naivní Bayesův klasifikátor má tendenci klasifikovat jako publicistický nebo řečnický funkční styl poměrně hodně textů, které do těchto tříd nepatří. Zvýšit nízkou přesnost u těchto dvou funkčních stylů by se teoreticky dalo přidáním dalších atributů, a to jak strukturálních, tak lexikálních.



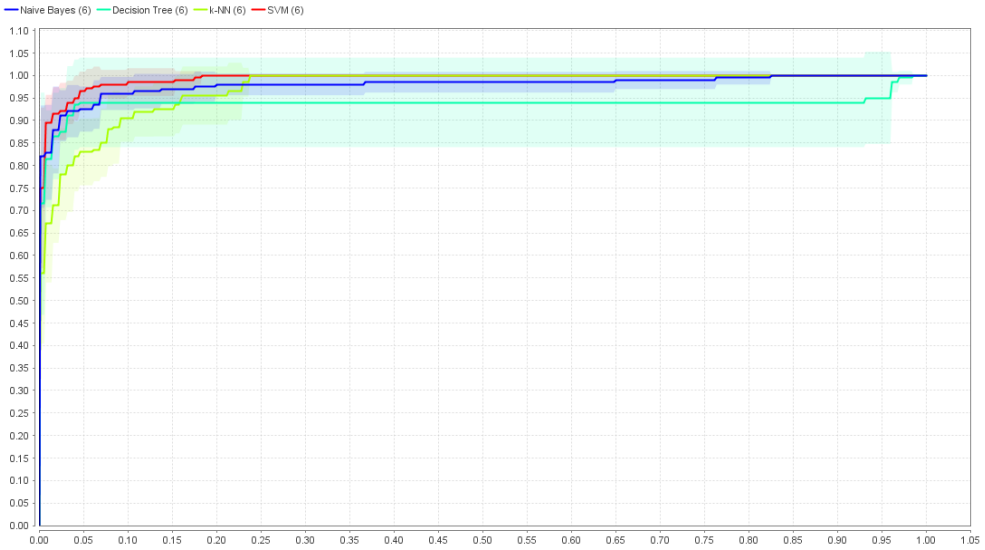
Obrázek 8.8: ROC křivky (publicistický funkční styl)



Obrázek 8.9: ROC křivky (řečnický funkční styl)



Obrázek 8.10: ROC křivky (hovorový funkční styl)



Obrázek 8.11: ROC křivky (administrativní funkční styl)

accuracy: 92.98% 1

	true false	true true	class precision
pred. false	1183	78	93.81%
pred. true	28	222	88.80%
class recall	97.69%	74.00%	

accuracy: 90.40% 2

	true false	true true	class precision
pred. false	1171	13	98.90%
pred. true	132	195	59.63%
class recall	89.87%	93.75%	

accuracy: 77.83% 3

	true false	true true	class precision
pred. false	883	7	99.21%
pred. true	328	293	47.18%
class recall	72.91%	97.67%	

accuracy: 64.59% 4

	true false	true true	class precision
pred. false	778	5	99.36%
pred. true	530	198	27.20%
class recall	59.48%	97.54%	

accuracy: 86.70% 5

	true false	true true	class precision
pred. false	1022	12	98.84%
pred. true	189	288	60.38%
class recall	84.39%	96.00%	

accuracy: 89.08% 6

	true false	true true	class precision
pred. false	1149	3	99.74%
pred. true	162	197	54.87%
class recall	87.64%	98.50%	

Obrázek 8.12: Binární klasifikace pomocí Naivního Bayesova klasifikátoru

Závěr

Cílem práce bylo vytvoření algoritmu pro identifikaci funkčního stylu dokumentů. Základní myšlenkou navrženého algoritmu je využití primárně strukturálních vlastností textů, tj. informací o jejich morfologických a syntaktických rysech a také o jejich horizontálním členění.

Identifikace funkčního stylu textu v navrženém algoritmu je řešena jako klasifikace textu do jedné ze šesti tříd reprezentujících funkční styly. Pro naučení klasifikátoru se používá strojové učení s učitelem.

V rámci práce byly porovnány čtyři metody strojového učení s učitelem: k-NN, Naivní Bayesův klasifikátor, Rozhodovací stromy a SVM. Úspěšnost prvních tří metod se pohybovala mezi 74,98% a 81,73%, přičemž nejnižší přesnosti dosáhla metoda k-NN, nejvyšší přesnosti dosáhl Naivní Bayesův klasifikátor. Hodnocení metody SVM, která je určena primárně pro binární klasifikaci, probíhalo odlišným způsobem, proto nebyla zahrnuta do celkového porovnání.

Kromě úspěšnosti jednotlivých metod byla dále porovnána senzitivita funkčních stylů. Ukázalo se, že nejnižší senzitivitu má funkční styl umělecký. Tímto se potvrdil předpoklad uvedený na začátku práce: umělecký funkční styl je nejproblematičtější z hlediska automatické identifikace, neboť nemá tak typické strukturální rysy, jakými disponují ostatní funkční styly. Analýza přesnosti (*precision*) funkčních stylů dále ukázala, že jako publicistický nebo řečnický funkční styly bývají často klasifikovány texty, které ve skutečnosti patří do jiné třídy.

Úspěšnost klasifikátoru by mohla být zvýšena přidáním dalších strukturálních atributů, popřípadě kombinováním atributů strukturálních a lexikálních. Ohledně lexikálních atributů je nutno zmínit, že za účelem klasifikace textů podle stylu je vhodnější vybírat funkční slova, tj. slova, která se mohou objevovat v textu nezávisle na jeho tématu (například v dané práci byly použity zájmena *jenž*, *týž/tentýž* a *ten*).

Literatura

[1] CHLOUPEK, J.; ČECHOVÁ, M.; KRČMOVÁ, M.; aj.: *Stylistika češtiny*. Praha: SPN, první vydání, 1991, ISBN 80-04-23302-3.

[2] MINÁŘOVÁ, E.: *Stylistika češtiny*. Brno: Masarykova univerzita, první vydání, 2009, ISBN 978-80-210-4973-4.

[3] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H.: *Introduction to Information Retrieval*. Cambridge University Press, 2008 [cit. 30.4.2016], ISBN 0521865719. Dostupné z: `http://www-nlp.stanford.edu/IR-book/`

[4] GROBELNIK, M.; MLADENIC, D.; JERMOL, M.: Exploiting Text Mining in Publishing and Education. In *Proceedings of the ICML workshop on data mining lessons learned*, Sydney, Australia, 2002, s. 34–39.

[5] APTÉ, C.; DAMERAU, F.; WEISS, S. M.: Automated Learning of Decision Rules for Text Categorization. *ACM Trans. Inf. Syst.*, ročník 12, č. 3, 1994 [cit. 30.4.2016]: s. 233–251, ISSN 1046-8188, doi:10.1145/183422.183423. Dostupné z: `http://doi.acm.org/10.1145/183422.183423`

[6] SEBASTIANI, F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, ročník 34, 2002: s. 1–47, ISSN 0360-0300.

[7] MANNING, C. D.; SCHÜTZE, H.: *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press, druhé vydání, 2000, ISBN 0-262-13360-1.

[8] HRALA, M.; KRÁL, P.: Evaluation of the Document Classification Approaches. In *8th International Conference on Computer Recognition Systems (CORES 2013)*, Milkow, Poland: Springer, 27-29 May 2013, ISBN 978-3-319-00968-1, s. 877–885, doi:10.1007/978-3-319-00969-8_86.

[9] KRÁL, P.: Named entities as new features for Czech document classification. In *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014)*, ročník 8404 LNCS, Kathmandu, Nepal, 6-12 April 2014, ISBN 978-3-642-54902-1, s. 417–427, doi:10.1007/978-3-642-54903-8_35.

[10] MITCHELL, T. M.: *Machine Learning*. New York: McGraw Hill, 1997, ISBN 0070428077.

[11] JOACHIMS, T.: Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, London, UK: Springer-Verlag, 1998 [cit. 30.4.2016], ISBN 3-540-64417-2, s. 137–142. Dostupné z: `http://dl.acm.org/citation.cfm?id=645326.649721`

[12] LEE, Y.; LIN, Y.; WAHBA, G.: Multicategory Support Vector Machines. *Journal of the American Statistical Association*, ročník 99, č. 465, 2004: s. 67–81, ISSN 0162-1459, doi:10.1198/016214504000000098.

[13] Ústav Českého národního korpusu: Co je korpus? 2016 [cit. 30.4.2016]. Dostupné z: `https://ucnk.ff.cuni.cz`

[14] FINN, A.; KUSHMERICK, N.: Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, ročník 57, č. 11, 2006: s. 1506—1518, ISSN 1532-2882, doi: 10.1002/asi.20427.

[15] KARLGREN, J.; CUTTING, D.: Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *COLING '94. Proceedings of the 15th conference on Computational Linguistics*, ročník 2, 1994, s. 1071–1075, doi:10.3115/991250.991324.

[16] KESSLER, B.; NUMBERG, G.; SCHÜTZE, H.: Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 1997 [cit. 30.4.2016], s. 32–38, doi:10.3115/976909.979622. Dostupné z: `http://dx.doi.org/10.3115/976909.979622`

[17] ARGAMON, S.; KOPPEL, M.; AVNERI, G.: Routing Documents According to Style. In *Proceedings of First International Workshop on Innovative Information Systems*, 1998, s. 85–92.

[18] PUSTYLNIKOV, O.: Guessing Text Type by Structure. In *Proceedings of the 12th ESSLLI Student Session*, Dublin, Ireland, 2007.

[19] PUSTYLNIKOV, O.; MEHLER, A.: Structural Differentiae of Text Types. A Quantitative Model. In *Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications*, Freiburg, Germany, 2007.

[20] STAMATATOS, E.; FAKOTAKIS, N.; KOKKINAKIS, G.: Text Genre Detection Using Common Word Frequencies. In *Proceedings of the 18th Conference on Computational Linguistics*, ročník 2, Saarbrücken, Germany, 2000 [cit. 30.4.2016], s. 808–814, doi:10.3115/992730.992763. Dostupné z: `http://dx.doi.org/10.3115/992730.992763`

[21] LEE, Y.-B.; MYAENG, S. H.: Text Genre Classification with Genre-revealing and Subject-revealing Features. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002 [cit. 30.4.2016], ISBN 1-58113-561-0, s. 145–150, doi:10.1145/564376.564403. Dostupné z: `http://doi.acm.org/10.1145/564376.564403`

[22] FEJFAR, K.: Metody kategorizace textu. 2007.

[23] TESAŘ, R.; JEŽEK, K.: Klasifikace Suffix Tree frázemi - srovnání s metodou Itemsets. In *Znalosti 2005*, Stará Lesná, Slovakia, 2005, ISBN 80-248-0755-6, s. 144–153.

[24] HRALA, M.; KRÁL, P.: Multi-label Document Classification in Czech. In *16th International conference on Text, Speech and Dialogue (TSD 2013)*, Pilsen, Czech Republic: Springer, 1-5 September 2013, ISBN 978-3-642-40584, s. 343–351, doi:10.1007/978-3-642-40585-3_44.

[25] TOMAN, M.; TESAŘ, R.; JEŽEK, K.: Vliv normalizace slov na klasifikaci textů. In *Znalosti 2007*, Ostrava, Czech Republic, 2007, ISBN 978-80248-1279-3, s. 360–363.

[26] Fakulta informatiky Masarykovy univerzity: An Empirical Study on Two-Class Categorization of Czech Documents. Online [cit. 30.4.2016]. Dostupné z: `http://www.fi.muni.cz/~popel/lectures/ll/tsd04-6.pdf`

[27] KRÁL, P.; CERISARA, C.; KLEČKOVÁ, J.: Lexical Structure for Dialogue Act Recognition. *Journal of Multimedia (JMM)*, ročník 2, June 2007: s. 1–8, ISSN 1796–2048.

[28] PANEVOVÁ, J.: Funkční styly a automatické zpracování jazyka. *Slavia : časopis pro slovanskou filologii*, ročník 67, č. 1-2, 1998: s. 161–167, ISSN 0037-6736.

[29] ČUPÁK, M.: ROC krivka. 2008.

[30] RapidMiner: RapidMiner Studio. 2006–2016. Dostupné z: https://rapidminer.com/products/studio

[31] STRAKOVÁ, J.; STRAKA, M.; HAJIČ, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland: Association for Computational Linguistics, June 2014 [cit. 30.4.2016], s. 13–18. Dostupné z: http://www.aclweb.org/anthology/P/P14/P14-5003.pdf

[32] STRAKA, M.; STRAKOVÁ, J.: Czech Models (MorfFlex CZ 160310 + PDT 3.0) for MorphoDiTa 160310. 2016 [cit. 30.4.2016], LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. Dostupné z: http://hdl.handle.net/11234/1-1674

[33] STRAKOVÁ, J.: MorphoDiTa. 2016 [cit. 30.4.2016]. Dostupné z: http://ufal.mff.cuni.cz/morphodita

[34] HANA, J.; ZEMAN, D.: Manual for Morphological Annotation. Revision for the Prague Dependency Treebank 2.0. 2005 [cit. 30.4.2016]. Dostupné z: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s02s01.html

[35] ABEEL, T.: Java Machine Learning Library (Java-ML). 2006–2012 [cit. 30.4.2016]. Dostupné z: http://java-ml.sourceforge.net

[36] ABEEL, T.; de PEER, Y. V.; SAEYS, Y.: Java-ML: A Machine Learning Library. *Journal of Machine Learning Research*, ročník 10, 2009 [cit. 30.4.2016]: s. 931–934. Dostupné z: http://jmlr.csail.mit.edu/papers/v10/abeel09a.html

Příloha **A**

Seznam použitých zkratek

NLP Natural Language Processing (zpracování přirozeného jazyka)

POS Part of Speech (slovní druh)

SVM Support Vector Machine (metoda strojového učení)

k-NN k-Nearest Neighbor (metoda strojového učení)

ROC Receiver Operating Characteristic (ROC křivka)

AUC Area Under Curve (plocha pod křivkou)

TPR True Positive Rate (úplnost, senzitivita)

FPR False Positive Rate (fall-out)

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresář se spustitelnou formou implementace
- `src`
  - `rm_projects` ..... Rapid-Miner Studio projekty
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `BP_Ekimova_Svetlana_2016.pdf` ..... text práce ve formátu PDF