Určovanie autorstva neznámeho slovenského textu

Analýza a návrh riešenia

Autor: Bc. Radovan Fuska

Vedúci práce: doc. RNDr. Stanislav Krajči, PhD.

# Úvod

Problematike určovania autorstva sa ľudia venujú už dlho. Uplatňuje sa hlavne vo forénznej lingvistike, ale môže sa používať aj na zisťovanie plagiátorstva. Určovaním autorstva sa ľudia zaoberajú už od čias stredoveku, kedy učenci hľadali spôsoby, ako priradiť texty starodávnym autorom. [1] V súčasnosti existuje mnoho spôsobov na analýzu textu za účelom určovania autorstva, ale väčšina z nich je vyvinutá pre svetové jazyky ako napríklad angličtina. Hoci existujú metódy aj pre slovanské jazyky, pre slovenský jazyk ich poznáme veľmi málo. Väčšinu existujúcich metód vhodných pre slovenčinu tvoria všeobecné metódy, ktoré nie sú závislé na jazyku. Cieľom tejto práce je analýza a vyhodnotenie efektívnosti existujúcich metód, návrh novej metódy a jej porovnanie s už existujúcimi metódami.

## Špecifikácia problému

Pri určovaní autorstva sa snažíme textu priradiť jeho autora. Využívame predpoklad, že texty od jedného autora obsahujú nejakú vnútornú podobnosť a zároveň sa líšia od textov od iných autorov. V lingvistike sa tento fenomén nazýva idiolekt – konkrétna podoba jazyka jednej osoby. Teda na určovanie autorstva sa budeme snažiť takéto znaky nájsť, kvantifikovať a pomocou nich nájsť nejakú metriku na odlišovanie textov od rôznych autorov. Z toho vyplýva, že na určovanie autorstva je potrebná vzorka textov od hľadaného autora, aby sa na nej mohli znaky idiolektu autora odmerať a vytvoriť akýsi osobný model podoby jazyka („odtlačok“) pre daného autora.

Pri určovaní autorstva budeme rozlišovať tri varianty tohto problému:

* V prvej variante máme k dispozícii zoznam možných kandidátov a nám stačí len vybrať, ktorý z tohto zoznamu napísal daný text.
* V druhej variante tiež máme zoznam autorov, ale pripúšťame aj variantu, že ani jeden z týchto kandidátov nie je autorom textu, teda že autor textu sa nenachádza v zozname kandidátov.
* V tretej variante poznáme kolekciu existujúcich textov od kandidátneho autora a máme iba určiť, či daný kandidát je alebo nie je autorom daného textu.

Tieto varianty si vyžadujú rozdielne algoritmy na ich výpočet, konkrétne prvá varianta si vyžaduje algoritmus pre klasifikáciu do viacerých tried, druhý variant algoritmus pre klasifikáciu do viacerých tried, ale s možnosťou nulovej klasifikácie zatiaľ čo tretí variant vyžaduje binárnu klasifikáciu. Existujú známe algoritmy na prvý a tretí variant, avšak algoritmy pre druhý variant nie sú až také rozšírené.

Najvšestrannejšie sú algoritmy na binárnu klasifikáciu, pretože sa dajú použiť na všetky tri varianty. Na tretí variant sa dajú použiť priamo. Na prvý a druhý variant sa dajú použiť tak, že sa vytvorí samostatný model pre každého autora a neznámy text sa ohodnotí pomocou modelov každého autora. Ak nastane zhoda u viac ako jedného autora, alebo ak v prvej variante nenastane zhoda u žiadneho autora, ako výsledok sa použije ohodnotenie s najväčšou pravdepodobnosťou resp. zhodou medzi autorom a neznámym textom. Toto sa však dá použiť iba u algoritmov, ktorých výstupom nie je len binárna hodnota, ale reálne číslo, ktoré sa na konci používa na určenie klasifikácie.

Pre druhú variantu sa dá použiť aj kombinácia algoritmu na klasifikáciu do viacerých tried a algoritmus na binárnu klasifikáciu. Algoritmom na klasifikáciu do viacerých tried by sa vybral najlepší kandidát na autora. Algoritmom na binárnu klasifikáciu by sa rozhodlo, či neznámy text je dostatočne podobný resp. zhodný so všetkými autormi resp. s hociktorým z množiny kandidátov. Nevýhodou tohto prístupu je to, že na takéto natrénovanie binárneho klasifikátora sú potrebné aj texty od autorov, ktorí nie sú súčasťou zoznamu kandidátov. Toto by sa však dalo vyriešiť obmenou tohto prístupu, kedy namiesto jedného binárneho klasifikátora by sa použil samostatný klasifikátor pre každého autora, ktorý by len potvrdzoval alebo zamietal klasifikáciu od prvého klasifikátora resp. rozhodoval by medzi výsledkom „konkrétny kandidát“ a „žiaden z kandidátov“.

# História

## Prvé snahy

Pri prvých snahách o určenie autorstva textu dominoval prístup, kedy sa vedci snažili nájsť akýsi autorský znak – jeden znak (väčšinou číslo), ktorého hodnota by vedela jednoznačne určiť autora textu. Výsledkom tohto snaženia však nebolo nájdenie takéhoto znaku, ale skôr nájdenie invariantných vlastností textu (napríklad Zipf pravidlo [1]). Medzi takýchto vedcov patrí napríklad Mendenhall, ktorý sa snažil použiť závislosť medzi dĺžkou slov a ich frekvenciou a Yule, ktorý sa snažil použiť dĺžku viet ako jediný znak na určenie autorstva textu. [1]

## Zovšeobecnený prístup

Po tom, ako sa jednorozmerný prístup na ohodnotenie textu ukázal ako nie veľmi efektívny, sa začalo uplatňovať viacrozmerné ohodnotenie textu. Pri týchto metódach sa využíval nielen jednoduchý skalár, ale aj vektor hodnôt vyskladaných z rôznych znakov a viacrozmerných hodnôt jedného znaku. Príkladmi takýchto prístupov sú Mosteller a Wallace, ktorí využili frekvencie funkcionálnych slov s metódou naivný Bayes. [1]

Príkladom takýchto snáh je aj známa práca od J. F. Burrows [6]:

V tomto článku autor analyzuje frekvencie tridsiatich najčastejšie sa vyskytujúcich slovných typov v anglickom texte.

Autor poukazuje na odlišnosť jednotlivých častí prózy a na jej unikátnosť v tomto ohľade. Autor argumentuje, naproti tvrdeniam ostatných učencov, že individuálny štýl autora textu sa v texte prejavuje nie len v malých rozdieloch v slovách, ale dá sa pozorovať vo väčšej miere pomocou štatistickej analýzy.

Na prisudzovanie autorstva textu, autor používa metódu, ktorá je odlišná od stylometrie a nerozlišuje medzi obsahovými slovami a funkcionálnymi slovami – slovami, ktoré plnia obsahovú funkciu v texte a slovami, ktoré plnia gramatickú funkciu v texte. Autor priznáva možnosť, že zvyky autora textu sa skôr odrazia v slovách, nad ktorými autor textu nemá vedomú kontrolu, avšak argumentuje, že na účely literárneho kriticizmu asistovaného počítačom by sa obsahové slová nemali vynechať.

Autor používa tridsať najčastejšie sa vyskytujúcich slovných typov z textov zo šiestich noviel od Jane Austen na vytvorenie frekvenčného profilu autora textu, ktorý potom porovnáva s profilmi textov noviel od iných autorov ako Henry James, E. M. Forster a Georgette Heyer. Tento profil používa aj na porovnanie naratívu jednotlivých textov od Jane Austen navzájom. Týchto tridsať najčastejšie sa vyskytujúcich slovných typov spolu tvorí asi dve pätiny všetkých slov použitých v textoch. Autor predpokladá, že tieto dve pätiny slov musia mať výpovednú hodnotu o celom texte.

Autor pomocou Pearson-ovho korelačného koeficientu vypočítal korelácie medzi jednotlivými naratívmi v textoch noviel od Jane Austen a ukázalo sa, že jednotlivé naratíva najbližšie korelujú s naratívami rovnakého typu, a teda sú nezávislé na obsahu textu či gramatických vlastností jednotlivých slov.



Tabuľka 1: Korelácie jednotlivých naratív v textoch noviel od Jane Austen.

Autor na týchto koreláciách potom použil dekompozíciu na extrahovanie vlastných vektorov a vlastných čísel z matice korelácií. Ukázalo sa, že najväčší prínos majú prvé dva vektory. Podľa nich autor potom zakreslil do 2D grafu jednotlivé naratíva. Takto sa ukázalo, že naratíva tvoria malé skupinky podľa ich typu. Najrozľahlejšia – najväčšie rozprestretá skupinka je skupinka dialógov postáv. Autor tvrdí, že toto je spôsobené odlišným charakterom jednotlivých postáv novely. Z grafu 1 je taktiež vidno, že najužšiu skupinku tvorí čistá naratíva autora text. Toto naznačuje, že Jane Austen počas jej literárnej kariéry zostala konzistentná vo svojej naratíve.



Graf 1: Porovnanie rôznych naratív v dielach od Jane Austen.

Graf 2: Porovnanie naratív dvoch diel od rovnakého autora.

Autor sa potom zameria iba na čistú naratívu autorov noviel. Taktiež autor rozdelí diela na dvetisíc-slovné časti na účely porovnania jednotlivých častí ako sú úvod a záver diel. Štatistiky týchto častí sa potom v poradí pospájajú do šesťtisíc slovných prekrývajúcich sa blokov. Toto má účel odstránenia štatistických výkyvov z jednotlivých individuálnych častí. Takto vytvorené štatistiky sa potom navzájom rovnakým spôsobom porovnajú.



Graf 3 - 6: Porovnanie čistých naratív diel od rôznych autorov.

Z grafov 3 - 6 je vidieť, že jednotliví autori majú svoje vlastné naratíva a tie sú navzájom odlišné.

Autor taktiež porovnáva aj čistú naratívu dvoch rôznych diel od jedného autora. Na grafe 2 vidieť že tieto naratíva si sú podobné a niektoré časti sa prekrývajú.

V závere autor porovnáva aj jednotlivé časti novely Persuasion samostatne. Z nich je možné vidieť vývoj štýlu autora textu a jeho zmeny v priebehu diela.



Graf 7: Porovnanie jednotlivých častí novely Persuasion.

## Moderné prístupy

V súčasnej dobe sa už zväčša upustilo od ručného hľadania znakov a vzorcov na kvantifikáciu textu a prešlo sa na automatické metódy – metódy využívajúce strojové učenie. [1]

# Súčasný stav

V súčasnosti existuje viacero odskúšaných metód na určovanie autorstva. Veľká časť týchto metód je špecifická pre anglický jazyk. Toto je pravdepodobne zapríčinené tým, že anglický jazyk je vo svete rozšírený, pre jeho analýzu existuje veľké množstvo dát a má relatívne ľahkú štruktúru, vhodnú na počítačové spracovanie bez potreby veľkého predspracovania.

Slovenský jazyk, na rozdiel od anglického, obsahuje zložitejšiu gramatiku, najmä ohýbanie slov, ktoré si na úspešné porozumenie vyžaduje náročné predspracovanie, ako napríklad určenie gramatických kategórií či syntaktický rozbor.

# Znaky textu

Na vytvorenie modelu sa dá použiť viacero rôznych znakov. Znaky sa často rozlišujú podľa ich jazykovednej roviny do ktorej patria resp. podľa toho, z akej roviny pochádzajú ich vstupné údaje. Existujú však aj znaky, ktoré patria do viacerých rovín resp. znaky, ktoré len stavajú na informáciách od iných znakov – tzv. metaznaky. Takéto delenie znakov však na samotný výpočet nie je nevyhnutné (a častokrát je len približné) a slúži hlavne len na lepšiu charakterizáciu znaku. V tejto práci sme vychádzali zo znakov pre anglický jazyk, pretože sú najpoužívanejšie a je ich najväčší počet.

## Anglický jazyk

Pre anglický jazyk sú najčastejšie znaky [2]:

* Lexikálne znaky
	+ Slová

Ide o sledovanie výskytu slov z konkrétneho predom pripraveného zoznamu. Často sa používajú funkcionálne slová.

* + Slovné n-gramy

Podobne ako u Slov, ale ide o krátke postupnosti slov.

* + Funkcionálne slová

Sú to slová, ktoré nenesú obsah textu ale určujú vzťahy medzi slovami vety. Často sú to slovné druhy ako spojky a predložky.

* + Zámená

Rozdelenie použitých zámen. Využíva predpoklad, že rôzny ľudia preferujú rôzne zámená. [3]

* + Modálne slovesá

Slovesá, ktoré vyjadrujú modalitu – pravdepodobnosť, schoponsť, ... (Sú to napríklad slová *can*, *could*, *may*, ...)

* + Slang

Medzi tieto slová patrí Internetový slang, smajlíkov či nadávky. [3]

* + Skratky a stiahnuté tvary

Ide o skratky a stiahnuté tvary anglického jazyka (napríklad *I’m*, *you’re*, *etc.*).

* + Emotikony

Ide o často používaných smajlíkov. [3][4]

* + Chyby hláskovania

Ide o chyby pri hláskovaní ako napríklad opakujúce sa písmeno, chýbajúce zdvojené písmeno, prehodenie písmen, písmeno navyše, chýbajúce písmeno, zamenené písmeno. Tieto chyby sa dajú hľadať pomocou slovníka na všetkých slovách, ale dá sa zamerať aj na špecifické vybrané chyby. [4]

* + Varianty britskej a americkej angličtiny

Napríklad alternatívne slová (*vacation*/*holiday*) a varianty v hláskovaní (*color*/*colour*). [4]

* + Sémantika (mnohoznačnosť, jednoznačnosť)

Ide o porovnávanie významu slov textu, jednotnosť významu slov v texte. Na tento účel je možné použiť špeciálne vytvorené slovníky ako je napríklad WordNet. [4]

* + Typy pomenovaných entít

Ide o relatívnu frekvenciu rôznych typov pomenovaných entít ako napríklad dátum, miesto, peniaze, číslo, radová číslovka, organizácia, percento, osoba a čas. [4]

* Znaky na úrovni písmen
	+ Písmenové n-gramy

Ide o n-gamy po sebe idúcich písmen. Kvôli ich veľkému počtu zvyčajne ide len o bigramy a trigramy. Aj takýto jednoduchý znak vie čiastočne zachytiť znaky textu ako napríklad: prípony, interpunkcia, čas, stiahnuté tvary, smajlíci, funkcionálne slová vo funkcii predložiek, zámená a spojky ako aj alternatívne hláskovania (americká vs. Britská angličtina, napr. -or/-our, -ise/-ize). [4]

* + Prípony

Ide o prípony slov vytvorených pridaním prípony na koreň slova. [4]

* + Interpunkcia

Ide o frekvenciu interpunkčných znamienok (výkričník, otáznik) a ich násobkov (napríklad !! alebo ???). Taktiež ide aj o ich zlé použitie ako napríklad prítomnosť medzery pred dvojbodkou. [4]

* Formátovacie znaky
	+ Dĺžka

Ide o dĺžku textu, viet a slov resp. ich priemernú dĺžku, maximálnu dĺžku a ich distribúciu. [3][4]

* + Formátovanie textu

Ide o počet resp. pomer prázdnych riadkov a celkového počtu riadkov. [3]

* + Ortografia

Ide o pomer slov s veľkými písmenami k celkovému počtu slov v texte a pomer veľkých písmen k celkovému počtu písmen. [3]

* + Úvodné/Záverečné frázy

Ide o časté frázy na začiatku a konci správ, najmä e-mailov. [3]

* Syntaktické znaky
	+ Vetné členy

Ide o unigramy, bigramy a trigramy vetných členov. [5]

* + Syntax

Ide o závislosti resp. vzťahy medzi jednotlivými vetnými členmi [5] a relatívne frekvencie vetných členov ich samotných a ich pomeru k iným znakom. [4]

* Iné
	+ Miery zložitosti

Ide o miery, ktoré sa snažia vypočítať zložitosť textu napríklad vytvorením modelu jazyka a následného ohodnotenia textu podľa neho [5] alebo pomocou analýzy hĺbky či košatosti syntaktického stromu viet. [4]

* + Vzdialenosti stredov zhlukov

Ide o metaznak vytvorený vzdialenosťami vektorov znakov od stredov vytvorených zhlukovacím algoritmom. [5]

Niektoré rozoberieme bližšie:

### Miery zložitosti

Sú to miery, ktoré sa snažia odhadnúť istú zložitosť textu. Tento prístup sa používal najmä pri prvých snahách o určovanie autorstva kvôli jeho jednoduchosti.

Patrí tu napríklad priemerná dĺžka slov, resp. rozdelenie dĺžok slov. Tá sa dá merať rôzne v závislosti od zvolenej najmenšej jednotky. Konkrétnym príkladom je počet slabík alebo počet písmen. Používaný je aj priemerný počet slov vo vete. [1]

Tieto mierky avšak nie sú veľmi efektívne samostatne. Väčšiu efektivitu dosahujú až v kombinácii s inými znakmi.

### Funkcionálne slová

Funkcionálne slová sú slová, ktoré v texte určujú gramatické alebo štrukturálne závislosti jednotlivých slov vety. Sú to väčšinou neplnovýznamové slová. V slovenčine na tento účel slúži hlavne ohýbanie slov. Tieto slová sú nezávislé od obsahu textu. Keďže sú zväčša ovládané gramatikou textu, cielená manipulácia s nimi je nepravdepodobná. [1]

Tento znak meria relatívne rozdelenie týchto slov navzájom. V súčasnosti sa používajú zoznamy obsahujúce stovky slov, ktoré obsahujú slovné druhy ako predložky, spojky, členy, ale taktiež aj plnovýznamové slovné druhy ako sú zámená a modálne slovesá. [1]

### Syntax, vetné členy

Tento znak využíva syntax, resp. vetné členy. Pri predspracovaní textu sú slovám určené ich slovné druhy, na ktorých sa počíta ich relatívna frekvencia, resp. frekvencie na ich krátkych postupnostiach v kombinácii s konkrétnymi slovnými druhmi. [1]

### Funkcionálne lexikálne taxonómie

Tento znak vytvára taxonómie postavené na funkcionálnych slovách, ktoré reprezentujú gramatické a sémantické rozdiely medzi triedami funkcionálnych slov v rôznych úrovniach abstrakcie. Vytvára stromy, ktoré v koreni majú slovných druh a každý potomok je označený podtriedou svojho rodiča (napríklad druhy zámen). V listoch sa nachádzajú konkrétne slová. Tieto stromy sú vytvárané zo slov z uzavretej množiny, takže predspracovanie textu na označovanie slov nie je potrebné. [1]

### Obsahové slová

Na rozdiel od funkcionálnych slov obsahové slová sú slová, ktoré nesú obsah textu. Takéto slová sú teda ovplyvňované preferenciami autora textu (napríklad pri výbere synoným). Jednotlivé preferencie autora sa dajú merať pomocou relatívnych frekvencií synoným. Zvyčajne veľmi zriedkavé slová a slová s rovnomerným rozdelením sa vyraďujú aby na analýzu zostalo niekoľko tisíc slov. Je možné použiť aj postupnosti takýchto slov. Problémom tohto znaku je, že keďže je závislý na obsahu textu, výsledným rozdielom v analýze môže byť len obsahový rozdiel textov. [1]

### Písmenové n-gramy

Písmenové, resp. znakové n-gramy (zvyčajne pre n ≤ 3 kvôli veľkému počtu kombinácií) počítajú relatívne frekvencie, resp. rozloženie n po sebe idúcich znakov. Vedia zachytiť lexikálne, gramatické a ortografické preferencie autorov bez potreby lingvistického vzdelania na interpretáciu výsledkov. Nie sú závislé na jazyku. [1]

Úspešne boli použité v Angličtine, Holandčine, Ruštine, Taliančine, Gréčtine. [1]

## Slovenský jazyk

Na analýzu slovenského textu sa dajú použiť znaky nezávislé na jazyku (zvyčajne sú to znaky pracujúce na úrovni písmen) a znaky špecifické pre slovenčinu prípadne slovanské jazyky. Mohli by to byť tieto znaky:

### Znakové, resp. písmenové

Frekvencie n-gramov písmen. Pri takýchto znakoch si na rozdiel od angličtiny môžeme vybrať, čo budeme považovať za rovnaké písmeno: či písmená líšiace sa len v diakritike budeme považovať za rovnaké alebo odlišné a či budeme brať písmená ako „ch“, „dz“ a „dž“ ako jedno písmeno alebo zoskupenie dvoch písmen. Tiež si môžeme vybrať, či budeme používať aj nepísmenové znaky ako je interpunkcia, medzera alebo číslice.

### Morfologické

Z morfologických znakov by sa dali použiť štatistiky výskytu slov podľa slovných druhov, gramatických kategórií a rôznych delení slov. Pre ohybné slovné druhy sa dajú určiť ich gramatické kategórie. Patria medzi nich aj gramatické kategórie ako rod a pád, ktoré sú na rozdiel od angličtiny špecifické pre slovenčinu. Štatistiky rodu slov by mohli byť obzvlášť zaujímavé keďže rod, v ktorom autor o sebe píše je zvyčajne konštantný vo všetkých jeho prejavoch. Pre všetky slovné druhy sa dajú počítať štatistiky rozdelenia do skupín v rámci ich slovného druhu.

Nevýhodou týchto vlastností je potreba predspracovania textu na ich určenie a taktiež prípadná nejednoznačnosť spôsobená viacerými možnými interpretáciami.

Radíme sem aj štatistiky výskytu interpunkcie.

Tieto znaky vedia v istom zmysle merať rozvitosť viet.

### Syntaktické

Zo syntaktický znakov by sa dali použiť rôzne typy dĺžok, ako sú dĺžky slov, viet, riadkov, odsekov a iných jednotiek. Taktiež sem vieme zaradiť aj štatistiky druhov viet, vetných členov a ich vzťahov. U vetných členov sa dá vytvoriť strom ich vzájomných závislostí, na ktorom sa dajú vypočítať rôzne metriky ako hĺbka, vzdialenosť najvzdialenejších listov či „košatosť“.

Zaradíme sem aj znaky ako počet prázdnych riadkov.

### Lexikálne

Štatistiky vybratých konkrétnych slov a slovných spojení, resp. slovných n-gramov (vrátane unigramov). Tieto slová môžu byť vopred vybraté alebo sa môžu použiť všetky slová textu.

Taktiež si môžeme vybrať, či budeme zohľadňovať ohýbanie slov - či rôzne tvary slova budeme považovať za jedno slovo, alebo odlišné slová. Neohybné slovné druhy toto neovplyvní, ale pre ohybné slovné druhy v prípade, že sa rozhodneme považovať rôzne tvary slova za jedno slovo, to bude vyžadovať predspracovanie textu.

### Chyby

Medzi najosobnejšie znaky môžeme radiť rôzne chyby v texte, obzvlášť ak sú konzistentné pre daného autora. Chyby sa dajú považovať za metaznak o iných znakoch, pretože v podstate ide o prípady, kedy sú porušené pravidlá používania štruktúr textu na rôznych úrovniach, resp. iných znakov. Tieto znaky sa najviac budú vyskytovať v neupravených textoch, ako je emailová komunikácia alebo chatová správa.

Tu vieme rozlišovať medzi chybami na úrovni slov resp. písmen ako napríklad preklepy a chybami na úrovni viet ako napríklad neprávne skloňovanie alebo nezhoda v gramatických kategóriách ako sú rod, číslo, pád u slov tvoriacich vetné členy v určovacom vetnom sklade.

# Metódy

Určovanie autorstva je v svojej podstate klasifikačný problém. Preto je vhodné na neho použiť klasifikačné metódy. Medzi populárne metódy patria [2]:

* Metóda podporných vektorov (SVM)

Ide o metódu, ktorá pracuje s vektormi v n-dimenzionálnom priestore. Táto metóda sa snaží binárne klasifikovať vektory v priestore tým, že vytvorí hranicu podľa ktorej sa určí klasifikácia. Jej idea spočíva v tom, že sa vytvoria tzv. podporné vektory, pomocou ktorých sa nájde lineárna hranica (nadrovina), ktorá rozdelí priestor vektorov na polovicu tak, aby na jednej strane boli vektory jednej klasifikácie a na druhej strane vektory druhej klasifikácie. Táto hranica a určí maximalizovaním vzdialenosti k bodom, ktoré sa nachádzajú najbližšie k tejto hranici.

Keďže metóda rozdeľuje priestor na dve polpriestory nadrovinou, vyžaduje aby trénovacia množina bodov bola lineárne separovateľná. V prípadoch, kedy trénovacia množina nie je lineárne separovateľná je možné použiť transformácie priestoru a metódu použiť na takto transformovanom priestore.

Keďže táto metóda vytvára binárne klasifikácie je najvhodnejšia na tretiu variantu problému určovania autorstva.

* Naivný Bayes

Táto metóda na základe hodnôt vstupného vektora vypočíta pravdepodobnosť klasifikácie daného vektora do každej triedy. Túto pravdepodobnosť vypočíta pomocou apriórnej pravdepodobnosti výskytu každého znaku (prvku vektora) pre danú kategóriu. Pri tomto výpočet nezohľadňuje možnú závislosť jednotlivých znakov navzájom a teda predpokladá, že jednotlivé znaky sú nezávislé.

Vzhľadom na to, že táto metóda počíta pravdepodobnosti pre a teda vytvára klasifikácie pre viacero kategórií je najvhodnejšia pre prvú variantu problému určovania autorstva.

* K-najbližších susedov (k-nearest neighbour)

Táto metóda klasifikuje vstupné vektory na základe ich vzdialenosti vo vektorovom priestore pomocou trénovacej množiny vektorov. Pre daný vektor metóda vypočíta jeho k najbližších susedov a daný vektor klasifikuje podľa väčšinového hlasovania týchto susedov. Teda daný vektor dostane klasifikáciu, ktorú má väčšina jeho susedov v jeho najbližšom okolí.

Táto metóda je vhodná pre prvú variantu problému určovania autorstva.

* Logistická regresia

Táto metóda funguje na rovnakom princípe ako lineárna regresia, ale namiesto lineárnej kombinácie používa logistickú funkciu. Metóda binárne klasifikuje príslušnosť vstupného vektora do danej triedy. Metóda vypočíta lineárnu kombináciu jednotlivých prvkov vstupného vektora z ktorej pomocou ligistickej funkcie vypočíta pravdepodobnosť príslušnosti vektora do danej triedy.

Keďže metóda počíta pravdepodobnosť klasifikácie vektora do jednej kategórie, s vhodným nastavením hranice pre klasifikáciu príslušnosti je vhodná pre všetky tri varianty problému určovania autorstva.

* Rozhodovacie stromy

Ide o metódu, ktorá používa strom resp. cestu stromom na klasifikáciu vstupného vektora do jednej z kategórií. Metóda využíva strom, ktorým prechádza od koreňa po vrcholoch, v ktorých sa rozhoduje ktorým vrcholom bude ďalej pokračovať až k listom, v ktorých sa nachádza výsledná klasifikácia. Vo vrcholoch sa nachádzajú podmienky založené na hodnotách jednotlivých prvkov vstupného vektora. Podľa toho, či daná podmienka je alebo nie je splnená sa rozhodne, ktorým potomkom sa pokračuje z daného vrcholu. Každý vrchol sa môže vetviť na dve alebo viac potomkov.

Počas trénovania sa začína koreňom, ktorý sa postupne rozvetvuje pridávaním resp. vetvením listov na ďalšie vrcholy o úroveň nižšie v strome. Existuje viacero metrík a algoritmov, pomocou ktorých sa dá určiť ďalšie vetvenie a koniec vetvenia stromu, jedným z nich je napríklad pomocou pomerového informačného zisku rozdelenia podľa atribútu.

Existuje aj variant v ktorom sa používa viacero stromov – les stromov. V tomto variante sa vypočíta klasifikácia pomocou všetkých stromov a výsledná klasifikácia sa určí kombináciou týchto výsledkov, zvyčajne pomocou väčšinového hlasovania.

Keďže táto metóda vie klasifikovať do dvoch ale aj viacerých kategórií, je vhodná pre prvú a tretiu variantu problému určovania autorstva.

* Winnowov algoritmus

Ide o metódu, ktorá pracuje na rovnakom princípe ako perceptrón, ale líši sa v jeho učení. Pri predikcii vypočíta lineárnu kombináciu binárnych prvkov vstupného vektora a pomocou naučenej hranie určí binárnu klasifikáciu. Pri učení sa zameriava na aktualizáciu len tých prvkov, ktoré prispeli k nesprávnej klasifikácii (prvkov s hodnotou 0).

Keďže táto metóda poskytuje binárnu klasifikáciu pre konkrétnu triedu, pričom sa zameriava len na znaky, ktoré majú najväčšiu výpovednú hodnotu, je najvhodnejšia pre prvú a tretiu variantu problému určovania autorstva.

Taktiež keďže tento algoritmus sa zameriava len na prvky, ktoré sú pre klasifikáciu významné, jeho naučená klasifikačná funkcia by sa mohla použiť na zistenie užitočnosti použitých znakov.

Častý je aj prípad, kedy sa tieto metódy kombinujú resp. dochádza k výraznému predspracovaniu znakov pred ich použitím v klasifikačnom algoritme.

# Miery

Na analýzu a vyhodnotenie správnosti výsledkov z jednotlivých metód sa dá použiť veľa mier. Keďže ide o problém klasifikácie textov, často sa používajú zaužívané miery z oboru získavania informácií (information retrieval). [2]

Tie sú podľa autora $A$:

* Precision: $P\_{A}=\frac{|spravne priradene\left(A\right)|}{|vsetky priradene\left(A\right)|}$
* Recall: $R\_{A}=\frac{|spravne priradene\left(A\right)|}{|dokumenty od autora\left(A\right)|}$
* Harmonický priemer $F\_{1}=\frac{2P\_{A}R\_{A}}{P\_{A}+R\_{A}}$

Tieto hodnoty potom vypovedajú o kvalite danej metódy na textoch od autora $A$.

Následne ak chceme vedieť kvalitu danej metódy na množine autorov $\{A\_{i}\}$, vieme podľa miery $M$ vypočítať

* $makropriemer\_{M}\left(\left\{A\_{i}\right\}\right)=\frac{1}{n}\sum\_{i}^{}M\_{A\_{i}}$, kde $n$ je počet autorov
* $mikropriemer\_{M}\left(\{A\_{i}\}\right)=\frac{1}{k}\sum\_{i}^{}k\_{i}M\_{A\_{i}}$, kde $k$ je počet dokumentov a $k\_{i}$ je počet dokumentov od $A\_{i}$

Pomocou týchto hodnôt potom vieme navzájom porovnávať jednotlivé metódy.

Je možné však použiť aj klasické metódy založené chybách, resp. odchýlkach od správneho ohodnotenia. Takéto miery sú napríklad

* Mean Absolute Error – priemer absolútnych chýb

$$\sum\_{}^{}\frac{|ε|}{n}$$

* Root Mean Squared Error – odmocnina priemeru druhých mocnín chýb.

$$\sqrt{\sum\_{}^{}\frac{ε^{2}}{n}}$$

# Postup práce

Momentálne je zväčša dokončený prehľad existujúcich riešení a znakov. Naším najbližším krokom je nájdenie datasetu. Na tento účel existuje niekoľko kandidátov a to Slovenský národný korpus, prepisy vystúpení poslancov z Národnej rady a články z novín. Najsľubnejšie vyzerajú články z novín, keďže SNK nie je určený na takéto účely a v prepisoch ide o hovorené slovo a nie písomný prejav autora, teda znaky ako Chyby by sa nedali použiť.

# Literatúra

1. KOPPEL, Moshe; SCHLER, Jonathan; ARGAMON, Shlomo. Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 2009, 60.1: 9-26.
2. ARGAMON, Shlomo; JUOLA, Patrick. Overview of the international authorship identification competition at PAN-2011. In: *CLEF (Notebook Papers/Labs/Workshop)*. 2011.
3. KERN, Roman, et al. Vote/veto meta-classifier for authorship identification. In: *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands*. 2011.
4. TANGUY, Ludovic, et al. A multitude of linguistically-rich features for authorship attribution.
5. SOLORIO, Thamar, et al. Authorship Identification with Modality Specific Meta Features Notebook for PAN at CLEF 2011. 2011.
6. BURROWS, John F. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary & Linguistic Computing*, 1987, 2.2: 61-70.