UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Ines Panker

Avtomatsko določanje avtorstva slovenskih leposlovnih besedil

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Janez Demšar

Ljubljana, 2012

I Z J A V A O A V T O R S T V U

diplomskega dela

Spodaj podpisani/-a Ines Panker,

z vpisno številko 63050070,

sem avtor/-ica diplomskega dela z naslovom:

Avtomatsko določanje avtorstva slovenskih leposlovnih besedil

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal/-a samostojno pod mentorstvom (naziv, ime in priimek)

  doc. dr. Janez Demšar

  in somentorstvom (naziv, ime in priimek)

  -

- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela

- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki »Dela FRI«.

V Ljubljani, dne _____ Podpis avtorja/-ice: _____

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01801/2012
Datum: 03.02.2012

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **INES PANKER**

Naslov: **AVTOMATSKO DOLOČANJE AVTORSTVA SLOVENSKIH LEPOSLOVNIH BESEDIL**

AUTOMATED AUTHORSHIP ATTRIBUTION FOR SLOVENIAN LITERARY TEXTS

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Določanje avtorstva s pomočjo statističnih metod je zanimiv problem, s katerim so se začeli ukvarjati že pred več kot sto leti, v sodobnem času pa je še zanimivejši zaradi široke dostopnosti ogromne količine besedil v digitalni obliki.

V okviru diplomske naloge preglejte obstoječe tehnike določanja avtorstva, tako tiste, ki temeljijo na preprostih statistikah, kot so pogostosti ločil in dolžine besed in stavkov, kot tehnike, ki opazujejo pogostosti besed. Delovanje postopkov preskusite na primerno izbrani zbirki klasičnih slovenskih leposlovnih besedil.

Mentor:

doc. dr. Janez Demšar

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan:

prof. dr. Nikolaj Zimic

Zahvala

Hvala vsem, ki so sodelovali pri izboljševanju kvalitete tega dela. Predvsem pa tistemu, ki se je trudil z motivacijskimi eksperimenti. Hvala.

Kazalo

Uporabljeni simboli in kratice

CA - klasifikacijska točnost

SVM - metoda podpornih vektorjev

kNN - metoda k najbližjih sosedov

TP - resnično pozitivni

TN - resnično negativni

FP - lažno pozitivni

FN - lažno negativni

RVO - Relative Vocabulary Overlap

Povzetek

Avtomatizirano določanje avtorstva je nadpomenka za metode, ki poskušajo na podlagi besedila sklepati na avtorstvo. Pri tem se poslužujejo raznih tehnik podatkovnega rudarjenja. Naša izbrana naloga je bila preizkusiti uspešnost takšnih postopkov na podmnožici slovenskih leposlovnih besedil. Besedila so bila v naših postopkih predstavljena kot vektorji, katerih dimenzije so določali atributi, ki smo se jih na besedilu odločili meriti. Svoje račune smo začeli z merjenjem števila pojavitev ločil in nadaljevali z merjenjem števila pojavitev vseh besed. Pri tem smo se oprli na preproste in najbolj znane klasifikatorje, preizkusili smo SVM, k-najbližjih sosedov, klasifikacijska drevesa in naivni Bayesov klasifikator ter na koncu ugotovili, da nam slednji daje najboljše rezultate. Končni rezultati so bili zelo zadovoljivi, s preprostimi pristopi smo dosegli 78% klasifikacijsko točnost in v povprečju 87% natančnost, pri čemer je bila dve tretjini avtorjev 100% natančnost.

Ključne besede: podatkovno rudarjenje, določanje avtorstva, naivni Bayesov klasifikator.

Abstract

Automatic authorship attribution is an umbrella term for methods trying to derive authorship from text. To achieve this they make use of various data mining techniques. Our chosen task was to test the successfulness of such procedures on a subset of Slovenian literary texts. Each text was represented as a vector with dimensions corresponding to the attributes we decided to measure. We started the calculations by measuring the number of punctuations and continued by measuring the number of word occurrences. We relied on the simple and most known classificators, we tested the SVM, kNN, classification trees and naive Bayes classificator. The last one was found to be giving the best results. Our final results were very satisfactory, with rudimentary approaches we achieved a classification accuracy of 78% and an average precision of 87% with 2 thirds of the authors having precision at 100%.

Keywords: data mining, authorship attribution, naive Bayes classifier.

1 Uvod

Namen naloge je bil raziskati aplikativnost avtomatiziranega določanja avtorstva na slovenskih pisateljih in dokazati izvedljivost koncepta ali pa vsaj zabeležiti vse naše poskuse.

Določanje in analiziranje avtorstva je intrigantno zaradi svoje matematične strani. Njegov rezultat so tehtni argumenti, ki so sestavljeni iz logičnih sklepov in lahko prepričajo skeptike. Avtomatizacija je bila vpeljana z namenom znebiti se pristranskosti.

Določanje avtorstva spada med priljubljene probleme umetne inteligence. Ljudje, ki se poklicno ukvarjajo z določanjem avtorstva, nimajo minuciozno definiranih postopkov. Po določenem času intuitivnega preverjanja teorij, se začenjajo zanašati na izkušnje. Edine računalnikove izkušnje so statistična preteklost, zato smo poskusili s pomočjo računalnikov najti ponavljajoče se statistične vzorce. Besedila so v teh postopkih večinoma prikazana kot vektorji, katerih dimenzije so atributi, ki jih merimo. Atributi so raznovrstni in segajo vse od števila pik in vejic do števila različnih besed napram vsem besedam do števila pojavitev vsake besede ali besedne zveze do rezultatov zapletenih enačb, ki se računajo glede na izbrane lastnosti besedila. Največji problem, ki se bo pojavil, je velika dimenzionalnost omenjenih vektorjev in z njo povezana želja po filtriranju atributov. Problem dobro ponazarja primer ko si kot atribute izberemo vse besede, ki jih najdemo v besedilih. Teh je v večini primerov neprimerljivo več kot je besedil, ob dodajanju novih besedil se njihovo število v primerjavi s številom besedil tudi hitreje veča. Cilja sta dva, doseči čimvečjo klasifikacijsko točnost in natančnost ter s tem ustvariti dovoljšnje zaupanje v rezultate, da postanejo ti uporabni.

Veliko študij je bilo na to temo že napisanih, le redke pa so se ukvarjale s slovenskimi besedili.

2 Stilometrija

Stilometrija je veda, ki preučuje možnosti merjenja lingvističnega stila avtorjev pisane besede. V besedilih išče vzorce, ki enolično definirajo njihovega avtorja. Kot orodje je komplementarna tradicionalnim literarnim teorijam, ki prav tako poskušajo uloviti pogosto izmikajočo se osebnost avtorjevega stila [4,13,12].

V raziskovanju nas vodi instinkt, da bi moral obstajati način, ki bi z uporabno gotovostjo znal podajati sklepe o lastnostih besedila in pisca, tudi tistih, ki v samem tekstu eksplicitno niso zabeležene. Velja predstava, da ima avtorjevo pisanje zavesten in podzavesten vir in da mu je slednji nedostopen. Ta nedostopnost avtorjevi volji daje lastnostim, s katerimi se podzavesten vir izraža, dovoljšnjo zanesljivost, da je njihovo iskanje postalo temelj stilometričnim algoritmom. Zdi se, da so vsi podatki zbrani, le brati se jih moramo naučiti.

Cilj je najti manjšo število značilnosti pisanja, ki se bodo konsistentno pojavljale znotraj besedil istega avtorja in v katerih se bodo avtorji opazljivo razlikovali med sabo.

2.1 Zbirka podatkov

Večina postopkov iskanja vzorcev v dokumentih rabi obsežne zbirke besedil za svoje delo. Govorimo lahko o več 100 ali več 100 tisoč dokumentih. Tu se srečamo s prvo oviro, digitalizacija besedil.

V Sloveniji na srečo že poteka digitalizacija slovenskega leposlovja v javni lasti, tj. tistega, ki so mu potekle avtorske pravice. Prepisovanje se je organiziralo v večji meri pod mentorstvom Mirana Hladnika, profesorja slavistike na Filozofski fakulteti Univerze v Ljubljani, in vsaj deloma s podporo Ministrstva za kulturo. Na žalost ob začetku te diplomske javno in v potrebni obliki ni bilo dostopnih dosti besedil, jih je bilo pa dovolj za izvedbo naših eksperimentov.

2.2 Študije

Stilometrija je stara tehnika. Uporaba računalnikov je le zamenjava orodij.

Najstarejše zasledene akademske publikacije segajo v konec 19. stoletja, ko je T. C. Mendenhall [19] predlagal uporabo dolžine besed kot možen argument ločevanja med avtorji. Idejo je možno osnovati na razmišljanju, da imajo boljši avtorji večje besedišče in da z večjim besediščem pridejo daljše besede v večjo rabo. Kmalu je bila izmed invariantnih značilnosti avtorjevega pisanja izpostavljene tudi dolžina povedi [28]. Razmišljanje se spet ne zdi napačno, posebej v slovenščini, ki ji slovnica daje možnost podredno in zaporedno vezati stavke in jih tlačiti v eno poved. To zahteva specifičen okus in določeno mero spretnosti. Na žalost se je izkazalo [13], da je odklon od povprečne dolžine besede ali povedi prevelik znotraj besedil istega avtorja in premajhen v okviru gledanja korpusa različnih avtorjev. Poleg tega je odvisen od časa in namembnosti dokumenta in marsičesa drugega.

Sledilo je veliko število novih statističnih tehnik in mer, vse so obljubljale zadovoljive uspehe in vse so se izkazale za nezadovoljive v poskusih pod novimi pogoji. Med njimi so bile povprečno število zlogov na besedo [10], porazdelitev besednih vrst [22], razmerje med številom različnih in številom vseh besed [24] (razmerje tip-simbol, kjer tipi predstavljajo število različnih besed, simboli pa število pojavitev vseh besed), pa tudi Simpsonov D index [21] in Yulova karakteristika K [29]. Za vse od njih je obveljalo isto kot za dolžino besed in povedi; njihova moč razlikovati med avtorji je bila premajhna ali premalo natančna. Njihova uporabnost je zato le omejena.

Eno izmed zanimivejših dognanj je bilo Zipfovo opazovanje o številu besed, ki se pojavijo f-krat. Opazil je, da se nekaj besed uporablja zelo pogosto in številne zelo redko. Pogostost pojavitve besede a je torej obratno sorazmerna njenemu mestu (n) v tabeli pogostosti besed. Najpreprostejša oblika Zipfovega zakona je:

$$P_n \sim \frac{1}{n^y}, \qquad (1)$$

kjer je P pogostost pojavitve besede a in y skoraj 1. Kar pomeni, da se druga najpogostejša beseda pojavi pol tolikokrat kot prva in tretja pojavi tretjino tolikokrat kot prva [30]. Predpostavljal je povezavo med parametrom y in starostjo ter inteligenco avtorja [31].

Med pomembnejše in večkrat raziskovane pojme stilometrije spada bogastvo leksike. Njena osnovna mera, velikost besedišča $V(N)$ je direktno odvisna od

števila besed. Ta odvisnost zmanjšuje zanesljivost mere. Z namenom odpraviti omenjeno odvisnost, so bile predlagane alternativne mere. Enostavnejša izmed njih je npr. razmerje tip-simbol, ki je bilo omenjeno prej, zahtevnejše pa Relativno prekrivanje besedišča (Relative Vocabulary Overlap - RVO) [27]. RVO meri stopnjo, s katero dve besedili vlečeta iz istega besedišča, in si s tem obljublja najti več podrobnosti kot takrat konvencionalne metode. Metoda ima dve pomembnejši pomanjkljivosti. Manjša od njiju je, da rabi vsaj dve besedili, ker dela le primerjavo besedil namesto analize. Večjo pomanjkljivost ustvarja tema besedil. Dve besedili, ki tematizirata isti dogodek, se bosta izkazali za bolj podobni kot se lahko dve besedili istega avtorstva in različne tematike.

Spet novi pristopi so predlagali opazovanje sinonimov. Kaj prepričuje nekoga, da raje piše "ampak" namesto "toda" ali "delaven" namesto "marljiv" ali "smrkavec" namesto "pobalin". Pri izbiri dosti manj kot zunanja logika šteje pisateljeva osebna preferenca, kar je ravno to, kar iščemo. Dobili pa smo dva nova problema: pomanjkanje seznama sopomenk in pomanjkanje konsenza o definiciji sopomenk. Če bi drugega rešili, s prvim ne bi imeli več dosti težav. V slovenščini ne premoremo slovarja sopomenk, razlog je še vedno trajajoča debata, kateri pari besed sodijo skupaj.

Drug pogost pristop je analiza na podlagi funkcijskih besed oz. slovničnih besednih vrst. To so besede, ki so neodvisne od tematike, ker same po sebi ne nosijo nobenega pomena. Uporabljajo se kot pripomočki za izražanje misli. Gre za predloge (s, z, h, v, ...), veznike (ki, ko, če, ampak, ...), členke (menda, le, tudi, ...). Njihova največja vrednost bi se naj pokazala pri analizi avtorjev z obširnim besediščem, pri katerih se besede redkeje ponavljajo in je zato bolj smiselno iskati vzorce v pojavitvah slovničnih besednih vrst. Mosteller in Wallace [20] sta testirala uspešnost 30 funkcijskih besed na t.i. "Federalist papers", govora je o 85 časopisnih esejih iz 18. stoletja. Za 12 izmed njih ni neovrgljivega dokaza o avtorstvu, čeprav se današnja znanost skoraj soglasno strinja glede njihovega avtorja. V raziskavi jima je samo na podlagi statističnih analiz in Bayesa uspelo dokazati isto avtorstvo 12ih spornih esejev, kot ga določajo učenjaki. Raziskava je sčasoma postala ena izmed najbolj znanih in največkrat ponovljenih. Burrows [7] je razvil tehniko uporabe skupin iz več kot 50 izmed najbolj pogostih funkcijskih besed. Binongo in Smith [6] pa sta glede na pojavitve 25ih predlogov razlikovala med deli Oscarja Wilda.

Kljub iskanju filtra za slovnične besede, se je izkazalo, da skoraj vse besede nosijo neko informacijo o avtorju. Joachims je [14] razpravljal o 10000 atributih, katerih vpliv na klasifikacijo je raziskoval. Izkazalo se je, da je model, ki je se učil na atributih rangiranih od 201. do 500. mesta, dosegel

skoraj tako dobre rezultate kot model, ki se je učil na atributih mest 1 do 200. Le malo drugačni so bili rezultati, ki jih je dajal tretji model, ki je uporabljal atribute razvrščene od mesta 4001 do 9962.

Tehnik, pristopov, metod ... je velika množica. Vsaka se vsaj zavidljivo dobro obnese v nekih omejenih, specifičnih okoliščinah, navkrižnega primerjanja med njimi pa je dosti premalo. V kateri situaciji je katera metoda najbolj priporočljiva in kateri pristop se dobro znajde v največ situacijah in kakšno sploh je največje število situacij, v kateri bi se ena sama tehnika zadovoljivo dobro odrezala? Velika računska moč računalnikov deluje kot dovolilnica za podleganje skušnjavam po izumljanju novih in novih metod. Vsako idejo se dá preveriti, preden se jo do konca razmisli in na koncu je pregovorno izmed klasifikacijskih metod naivni Bayes tisti, ki daje zavidljivo dobre rezultate.

Vsekakor pa je ta tematika predstavljala dovolj intriganten izziv, da je o njej razmišljala velika množica raziskovalcev tudi, če si vsi niso enotni glede pozitivnih učinkov vsake izmed teh raziskav [16,8,15].

2.3 Narava problema in napak

Stilometrija rešuje računalnikom nedomač problem. Globok prepad je med naravnimi in računalniškimi jeziki. Dvoumnost, metafore, kontekst so nekateri izmed konstruktov, ki za računalnike, kot so danes, spadajo v prostor težko dosegljivega, in hkrati tiste, ki jih poskušamo vključiti v njihov sistem odločanja.

2.3.1 Osnovni delci

V začetku je treba definirati osnovne delce besedila tj. tiste značilnosti teksta, ki nedvoumno nosijo vsaj del informacije, ki jo iščemo. V vsej množici besed, znakov, črk, fraz, tipografskih elementov, postavitve, ... je treba identificirati tisto podmnožico značilnosti, ki določa besedilo, kot simptomi določajo bolezen.

Na nek kvantitativen, predvsem pa zelo osnoven način je treba definirati razlike med besedili. Medtem ko se v literarni teoriji lahko govori o mračnem razpoloženju, ki seva iz neke pesmi, lahko računalnik prebira le pogostost črkovnih kombinacij (tj. besed), ki se nanašajo na mrak/svetlobo. Pomena besed ne more dojeti in posledično tudi sopomenk ne (v slovenščini še ni slovarja sopomenk). Tako se vsaka beseda lahko upošteva kot samostojna značilnost, posledično skupno število možnih značilnosti kmalu preseže meje

obvladljivega. Za število surovih osnovnih delcev, ki so na voljo pred izborom, se tako zdi, da je omejeno le navzdol.

V raziskavah najpogosteje uporabljeni osnovni delci oz. značilnosti besedil spadajo v eno od sledečih 4 skupin [9]: znaki, besede, izrazi ali koncepti. Njihov trenutni zapis lahko gledamo, kot da je hierarhičen. Element vsake naslednje skupine je skupek elementov prejšnje.

Osnovna skupina so znaki. S tem so mišljeni vsi gradniki besedila. Med drugim mednje spadajo črke, številke, ločila, presledki in razni simboli. Na tem nivoju obstajata 2 pristopa. Prvi ustvarja vrečo znakov, z drugimi besedami, gre za zbirko, ki ne vsebuje nobenih podatkov o poziciji. Bolj pogost in uporaben je način shranjevanja vsaj dela informacije o položaju znakov. V praksi to ponavadi pomeni iskanje znakovnih bi- in trigramov. V večini primerov pa so vse takšne metode beleženja posameznih znakov manj uporabne, čeprav so v svojem bistvu najbolj pristen opis obravnavanega dokumenta, ker lahko iz njih v celoti rekonstruirmamo dokument, po katerem so bili povzeti.

S semantičnim bogastvom nekega avtorja govorimo o celih besedah. Vendar je v nekem besedilu lahko več 100 000 unikatnih besed, zato se ponavadi posveča pozornost kriteriju, ki bo izmed njih znal izbrati smiselno in reprezentativno podmnožico. Pogosto se uporabljajo slovarji izrazov, ki nas zanimajo. V tem koraku analiza besed prestopi v analizo izrazov. Vse, kar se ne najde v izbranem slovarju, se zavrže. Edina izjema tega pravila, besede, ki se obdržijo nedotaknjene, so kvečjemu neposredna okolica najdenih izrazov.

Najtežja izmed teh tehnik je iskanje konceptov, kar je bolj cilj kot pot. Iščejo se besede, znaki, besedne zveze, izrazi, stavki, ..., ki bi se jih dalo povezati s konceptnimi identifikatorji, slednji pa naprej označujejo skupine različnih konceptov. Za koncepte same pa ni nujno, da se v besedilu eksplicitno pojavijo. Številne kategorizacije v ta namen uporabljajo križno primerjanje z zunanjimi viri (npr. učna množica dokumentov).

Seveda se te tehnike med sabo ne izključujejo. V začetku projekta je izbira tehnike odvisna od ciljev, med projektom pa od doseganja teh.

2.3.2 Pretirano prilagajanje podatkom (overfitting)

Eno osnovnih pravil statistike je prepoved prenatančnega definiranja.

V "naravnem svetu" so vse stvari variacije. Četudi pripadajo isti vrsti, za katero imamo v slovenskem jeziku isti izraz, kot npr. pomaranča, so vendar

vse le variacije na isto temo, vsaka malo drugačna od vsake druge in vendar vse povsem drugačne od oranžnih žog. S prenatančnim popisom vsega videnega in vključitvijo tega v statističen model, se nam zgodita dve napaki. Prva je, da naključen šum upoštevamo kot veljaven podatek, namesto da bi ga filtrirali. Do druge pa pride šele, ko na podlagi omenjenega statističnega modela zgradimo algoritem za klasifikacijo. Ta algoritem ima t.i. nično hipotezo, H_0, ki je npr. "To je pomaranča", in jo preverja na objektih, ki se mu posredujejo, ti objekti so ponavadi skupki lastnosti. Kmalu lahko pridemo do t.i. napake tipa 1, situacije ko naš algoritem napačno zavrne hipotezo H_0 nad primerkom P, čeprav bi jo bil moral sprejeti. Razlog za zavrnitev je strogost modela, algoritem bo samo že videne pomaranče prepoznal kot take, vse ostale bodo preveč odstopale od modela.

2.3.3 O avtorjih

V razmislek je potrebno vzeti tudi naravo avtorjev in postopke njihovega ustvarjanja. Izpostavili bi radi dejstvo, da se pisatelji razvijajo skozi čas. Ko iščemo vzporednice med deli, ki so nastala v več desetletnih razmahih, nam lahko omenjeno dejstvo kvari statističen model. Zato bi slovnične besedne vrste znale nositi več informacije kot ostale besede.

2.3.4 Prenatančno iskanje

Vzorci, ki se pojavljajo v teh besedilih, so izredno subtilni. Marsikatere sledi, ki je prisotna, nimamo možnosti zaznati, saj se ne upošteva pomena stavkov, metafor, referenc na različne dogodke in osebe, vse stvari, ki jih literarni zgodovinarji lahko upoštevajo.

Iz takih situacij se pojavi vprašanje ali je boljši algoritem, ki mu stavek "Jaz, Ivan Tavčar, sem avtor tega besedila," prinese približno toliko informacije kot katerikoli drug stavek v sestavku ali tisti algoritem, ki si zabeleži, da sta bili beseda 'avtor' in ime 'Ivan Tavčar' uporabljeni v istem stavku. Na prvi pogled se zdi, da bi bil drug algoritem boljši, ker bi po branju izpostavljenega stavka imel hipotezo, ki bi jo lahko z neko utežjo dal med druge kazalce na avtorstvo, ki jih je nabral med analiziranjem besedila. Toda, kaj pa, če bi bil stavek spremenjen v npr. "Kot je bog ustvaril svet, tako se je izpod peresa Ivana Tavčarja rodilo to besedilo."? Kakšna bi bila logika, ki bi v njem uspela najti podatek o avtorstvu? Pomembneje pa, kakšna bi bila pomembnost takšnega odkritja. Veliko je odvisno od konteksta, ki je algoritmom težje dosegljiv. Če bo nek algoritem zvezi pripisoval preveliko težo, bo morda tudi to besedilo, ki ga jaz zdaj pišem, označeno za delo Ivana Tavčarja, ker je dokaz o njegovem avtorstvu evidentno zapisan sredi tega odstavka (in to celo dvakrat). Morda

pa je algoritem pametnejši in razume, da je stavek v narekovajih in zato dvomljivega izvora. Nakar se pojavi vprašanje, ali se lahko odpovem svojemu avtorstvu že s tem, da izbrišem narekovaje dotičnih stavkov. Dilemi je težko priti do dna. Nemogoče vsekakor brez poznavanja konteksta. Zato postane vprašljiva tudi smiselnost iskanja tako imen kot specifičnih besed. Če ne moremo ničesar razumeti, razen tega, da moramo iskati povezavo med vsemi besedami, ki se začnejo z veliko začetnico, in besedo 'lkdocnx', se morda ni uporabno tej povezavi sploh posvečati, ker je, tudi ko jo imamo, ne znamo kategorizirati.

3 Inteligentne metode

Po tistem, ko smo statistično popisali besedila, torej iz njih dobili vse značilnosti, ki so nas zanimale, nastopi naslednji problem. Vreče statističnih podatkov za vsako besedilo so v svoji surovi obliki neprimerljive, zato posežemo po standardiziranih postopkih iz umetne inteligence, ki znajo iskati vzorce v statistikah. V grobem se delijo na nadzorovano in nenadzorovano učenje.

Nadzorovano učenje deli spremenljivke, s katerimi ima opravke, na več opisnih in eno ali več odvisnih. Cilj metod je ustvariti funkcijo, ki bo na podlagi opisnih metod določila vrednost odvisnih in se pri tem čim manjkrat zmotila. Mednje spadajo tudi klasifikatorji.

Nenadzorovano učenje pa obravnava vse spremenljivke enako. Cilj je najti in pokazati vzorce, ki jim spremenljivke sledijo. Možnih namenov pri tem je več, razlog je lahko zmanjšati število spremenljivk na bolj vplivne ali zbrati in razločiti vhodne primere v skupine.

3.1 Klasifikatorji

Klasifikatorji so odločitveni sistemi, ki poskušajo popredalčkati nova opazovanja na podlagi starih. Poljubne kombinacije dogovorjenih značilnosti razvrščajo v pod-populacije.

Statistično gledano govorimo o razvrščanju entitet v razrede, pri čemer so entitete opisi nekih predmetov, razredi pa ciljne skupine, ki jim ti predmeti pripadajo. V namen učenja jim je podana začetna množica entitet imenovana učna množica, katerih razredi so znani.

Formalno gledano je njihova naloga na podlagi podane učne množice $\{(x_1, y_1), \dots (x_n, y_n)\}$ ustvariti pravilo h tako, da je mogoče $h(x)$ izračunati za poljuben x (predvsem tak, ki ni v učni množici) in da je $\hat{y} = h(x)$ čimbolj podobna resničnemu y [2].

Primerjava klasifikatorjevih napovedi z resničnimi rezultati kaže na stopnjo, s katero lahko njegovim rezultatom zaupamo. V tej povezavi se najpogosteje omenja klasifiakcijska točnost in predstavlja procent pravilnih napovedi v

testni množici. Izmed drugih metod, ki ocenjujejo uspešnost kalsifikatorja, sta za nas relevantni posebej še priklic in natančnost.

3.1.1 Algoritmi

3.1.1.1 Naivni Bayes

Naivni Bayes je verjetnostni klasifikator in temelji na Bayesovem teoremu o pogojnih verjetnostih. S teoremom izračunamo verjetnost, da neka entiteta pripada razredu A, če poznamo vrednosti njenih atributov, pri tem pa naivno predpostavimo, da so atributi med sabo neodvisni. Posebej je primeren za visoko dimenzionalne probleme, podobno kot SVM, in je znan po tem, da pogosto daje boljše rezultate kot bolj sofisticirane metode.

3.1.1.2 Metoda podpornih vektorjev (SVM)

Metoda podpornih vektorjev ali z angleško kratico SVM temelji na ideji, da načeloma obstajajo jasne meje med razredi. Če se jih ne da dobiti v n-dimenzionalnem prostoru, jih bo morda moč najti v $(n+1)-$ dimenzionalnem. Vsaka entiteta se predstavlja kot vektor v $n-$ dimenzionalnem prostoru, pri čemer je n število atributov in so koordinate vrednosti atributov. Med razredi poskuša najti optimalno mejno hiperravnino in pri tem sledi pravilu največjega roba (maximum margin). To pravilo išče mejo, ki bi bila enako in najbolj oddaljena od najbližjih entitet vseh razredov.

3.1.1.3 K- najbližjih sosedov (kNN)

KNN je metoda, ki novi entiteti določi razred glede na razrede, ki jim pripada njenih K najbližjih sosedov. Entitete iz učne množice so vektorji v prostoru atributov, faza učenja pa sestoji le iz shranjenja vseh učnih primerov. Testni primer ravno tako zavzame mesto v istem prostoru, nato se mu poišče K najbližjih sosedov, ki glasujejo o razredni pripadnosti testnega primera. K je pri tem vnaprej določen. Za oceno bližine ponavadi uporabimo Evklidovo razdaljo ali kakšno drugo, naravi problema bližjo, metriko. Izbira prave metrike in opcijska utežitev glasovanja so orodja, ki lahko izboljšajo rezultate.

3.1.1.4 Odločitvena drevesa

Drevesa so odločitveni sistemi, ki na vsakem nivoju ponudijo odločitev na podlagi enega atributa in s spustom do najnižjega nivoja privedejo do razreda, ki mu primer pripada. Iz učne množice se drevo zgradi z opazovanjem vpliva atributov na pripadnost razredu. V vsakem vozlišču se učna množica razdeli

na dva ali več delov, ki so po pripadnosti razredom bolj čisti kot je bila vhodna množica. Vsakič se uporabi atribut, ki najbolje razdeli vhodno množico. Njihova najboljša lastnost je argumentiranost rezultatov; za vsako napoved se namreč točno ve, kaj je v koliki meri vplivalo nanjo.

3.1.2 Priklic in natančnost ter specifičnost in občutljivost

V osnovnem klasifikacijskem primeru imamo en razred, množico testnih primerov in željo vsakemu primeru določiti (ne)pripadnost temu razredu oz. potrditi ali ovreči osnovno hipotezo. Glede na resnično stanje naši izračuni razdelijo primere v štiri skupine, ki se imenujejo resnično pozitivni (TP), resnično negativni (TN), lažno pozitivni (FP) in lažno negativni (FN). Tipično (in najlažje) se te kategorije razlagajo na primeru bolezni z osnovno hipotezo $H_0 = "je\ zdrav"$.

		Resnična vrednost	
		H_0 je res	H_0 ni res
Rezultat napovedi	*+ (potrdimo H_0)*	*TP*	*FP*
	- (zavrnemo H_0)	*FN*	*TN*

Tabela 1: Matrika nedoločenosti

Resnični pozitivni in resnično negativni skupini vsebujeta tiste primere, ki smo jih pravilno klasificirali, njihovo zdravstveno stanje torej pravilno določili. Lažno negativni so tisti, za katere so naši testi narobe pokazali, da so bolni, čeprav niso bili. Kot lažno pozitivni pa so označeni tisti, ki v resnici so bolni, naša napoved pa je bila, da niso.

Skupini FP se reče tudi napaka tipa I, skupini FN pa napaka tipa II. Velikokrat se ju prikazuje v t.i. matriki nedoločenosti kot v Tabela 1.

V tem okviru se uvede več pojmov, ki se med sabo prepletajo.

Priklic ali občutljivost je delež potrjeno zdravih ljudi v množici vseh zdravih ljudi:

$$P(TP\ |Z) = \frac{TP}{TP + FN} \qquad (2)$$

Specifičnost je delež potrjeno bolnih ljudi izmed vseh bolnih ljudi.

$$P(TN\ |B) = \frac{TN}{TN + FP} \tag{3}$$

Natančnost je delež potrjeno zdravih ljudi izmed vseh z napovedjo 'zdrav'.

$$P\left(TP \mid {H_0}^+\right) = \frac{TP}{TP + FP} \tag{4}$$

Točnost pa predstavlja delež pravilnih napovedi.

$$P\left({H_0}^+ \; ali \; \overline{{H_0}^-}\right) = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Zanimivi pri teh merah so odnosi med njimi. Občutljivost meri delež pravilnih pozitivnih primerov, specifičnost pa pravilnih negativnih primerov. Natančnost pove koliko izmed pozitivov je bilo pravilnih, priklic pa, koliko izmed resničnih pozitivov je bilo označenih kot pozitiv. Točnost pove, koliko vseh napovedi je bilo pravilnih, natančnost pa, koliko pozitivnih napovedi je bilo pravilnih. Cilj vseh izmed njih je opisati delež napak tipa I in II.

Vse mere uspešnosti klasifikatorja, ki smo jih omenili do zdaj, so ne povsem dorečene v več-razrednih problemih. V splošnem obstajata dva pristopa, eden proti enemu in eden proti vsem. Pomen je razviden iz imen, pri prvem primerjamo vsak par razredov, pri drugem pa vsak razred naspram vsem drugim, ki delujejo kot skupina.

		Napovedani razred		
		A	B	C
Resnični razred	A	T_A	e_{AB}	e_{AC}
	B	e_{BA}	T_B	e_{BC}
	C	e_{CA}	e_{CB}	T_C

Tabela 2: Matrika nedoločenosti za večrazredne probleme

Mi smo se odločili, da bomo kot FN_x vzeli vse primere, ki so iskanega razreda X, pa so bili drugače kvalificirani, kot FP_x pa tiste, ki so bili kot X klasificirani, pa so v resnici nekaj drugega. Za Tabela 2 se za razred A izračunata FP_A in FN_A sledeče:

$$\begin{aligned} FN_A &= e_{AB} + e_{AC}\ , \\ FP_A &= e_{BA} + e_{CA}\ . \end{aligned} \tag{6}$$

Želja snovalcev vseh testov je izprazniti FP in FN skupini, torej narediti test 100% pravilen. Toda izven teorije in poenostavitev takih primerov ni. Prvič, ker to pomeni, da sta podskupini H_0^+ in H_0^- jasno in nedvoumno ločeni, kar se v naravi le redko pojavlja, in drugič, ker za klasifikacijske naloge tako očitnih rezultatov nihče ne zapravlja časa in denarja [23,25].

3.1.3 Brierjeva ocena

Med delom se sklicujemo na še eno mero uspešnosti klasifikatorja, na Brierjevo oceno. Ta je drugačna od ostalih v tem, da ne ocenjuje pravilnost napovedi. Ne pove, ali je bila ali ni bila neka napoved pravilna, ampak poda podatek o tem koliko se napovedane verjetnosti približajo izidom dogodkov, ki jih napovedujejo. Manjši kot je Brier, bolj pravilna je napoved. Najmanjše število, ki ga lahko zavzame, je 0, meja navzgor pa je odvisna od števila razredov [1].

Enačba za testno množico moči N:

$$BO = \frac{1}{N}\sum_{t=0}^{N}\sum_{i=0}^{R}(f_{ti} - o_{ti})^2 \qquad (7)$$

kjer je N število primerov in R število razredov, f_{ti} je napoved t-tega primera za i-ti razred, o_{ti} pa dejanska vrednost t-tega primera za i-ti razred. Velja tudi, da je $o_{ti} = 1$ za dejanski razred in 0 za vse ostale. Notranja vsota oceni točnost napovedi, zunanja pa je golo povprečenje po vseh primerih iz testne množice.

3.1.4 Mere

Za izražanje koristnosti atributov so bile razvite nekatere statistične mere. Njihov namen je na neki lestvici definirati pomembnost atributa v odločitvenem procesu. Gonilna sila za njihovim razvojem pa je želja, da bi se znebili vseh nekoristnih atributov, ker v entiteti ne nosijo nič informacije o njenem razredu. Najpogosteje te mere merijo porazdelitev vrednosti razreda pri znanih vrednostih atributa.

3.1.4.1 Informacijski prispevek

Informacijski prispevek je merilo nečistosti atributa. Pri tem višje "čistosti" dosegajo atributi, ki s svojimi vrednostmi razdelijo primere v enotnejše podskupine, enotnejše z vidika pripadnosti razredom. Njegov temelj je

entropija, mera nedoločenosti. Naključnim spremenljivkam daje entropija visoke vrednosti, če je njihovo vrednost težko napovedati, in nizke, če je njihova vrednost očitna. Razloga za težjo napovedljivost sta veliko število različnih vrednosti v zalogi vrednosti in enakomerno razporejene verjetnosti za pojavitev vsake izmed teh vrednosti.

Če je X naključna spremenljivka z n vrednostmi $\{x_i : i = 1, \dots, n\}$ in je $p(x_i)$ verjetnost, da X zavzame vrednost x_i, je enačba entropije sledeča:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\, log_b\, p(x_i). \qquad (8)$$

V podatkovnem rudarjenju se visoko cenijo atributi, ki glede na svojo zavzeto vrednost razdelijo množico primerov na čistejše podmnožice, pri čemer se čistost ocenjuje kot razredna homogenost podmnožic. Manj razredom kot pripadajo entitete v podmnožicah, večja je homogenost podmnožice.

Če imamo za entitete diskreten razred R, potem se pomembnost atributa A ocenjuje sledeče:

$$\begin{aligned} IP(R, A) &= H(R) - H(R|A) \qquad (9)\\ H(R) &= -\sum_{r \in R} p(r)\, log_b\, p(r)\\ H(R|A) &= \sum_{a \in A}^{n} p(a)\, H(R|A = a)\\ H(R|A = a) &= -\sum_{r \in R} p(r|a)\, log_b\, p(r|a). \end{aligned}$$

Pri tem je $H(R)$ entropija razreda, $H(R|A)$ entropija razreda, če je znan A in p(a) verjetnost, da atribut A zavzame vrednost a.

3.1.4.2 ANOVA

ANOVA (analysis of variance) meri razmerja med povprečji različnih skupin in je generalizacija t −testa, ki je omejen na binarni razred. V srcu ANOVE je ideja, da je varianco možno razdeliti in njene dele pripisati različnim virom. Z identifikacijo virov in deležem variance, ki jo doprinesejo, je mogoče deliti atribute po pomembnosti.

Varianca se računa kot vsota kvadriranih deviacij od povprečja.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \; kjer\; je\; \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (10)$$

Iz populacije, ki jo sestavljajo različni razredi, je mogoče dobiti dve neodvisni oceni. Prvo, ki prikaže stopnjo razlike med razredi, in drugo, ki je občutljiva na napako znotraj razreda. Prva se imenuje ocena variance med razredi, druga pa ocena variance znotraj razredov. Njuna vsota predstavlja skupno varianco sistema.

$$VK_{skupna} = VK_{med\ razredi} + VK_{napake\ znotraj\ razreda} \qquad (11)$$

Razmerje med njima je porazdeljeno po F-distribuciji in se za računanje zato uporablja F-test.

$$F - test = \frac{VK_{med\ razredi}}{VK_{napake\ znotraj\ razreda}} = \frac{\frac{\sum_i n_i(\overline{Y}_\iota - \bar{Y})^2}{K-1}}{\frac{\sum_{ij}\left(Y_{ij} - \overline{Y}_\iota\right)^2}{N-K}} \qquad (12)$$

N je število primerov, K število razredov, $\overline{Y}_\iota$ povprečje i-tega razreda, n_i število entitet v i-tem razredu, $\bar{Y}$ povprečje celotne populacije, Y_{ij} pa vrednost j-tega primera v i-tem razredu.

Signifikantna razlika med obema merama kaže na večjo pomembnost atributa oz. njegovo večjo deviacijo od povprečja med razredi kot znotraj razreda.

3.1.5 Testiranje

Ocena uspešnosti klasifikatorja je v večji meri odvisna od metode testiranja, kot bi si želeli. Kljub uspešni oblikovanju klasifikatorja, je težko napovedati njegovo obnašanje v dejanski uporabi.

Idealno testiranje bi testno množico sestavilo iz klasifikatorju še ne videnih entitet, ki bi skupaj predstavljale reprezentativen vzorec populacije. V nekaterih klasifikacijskih problemih, kot je tudi naš, nimamo dostopa do dovolj velike množice primerov ali pa nam reprezentativnost dela probleme. V teh primerih uporaljamo križno preverjanje.

Bistvo križnega preverjanje je, da v večih iteracijah množico vseh entitet E razdelimo na učno množico L in testno množico T. Klasifikator se uči na učni množici, testira se ga na testni. Njegova najbolj splošna oblika je $k-$kratno križno preverjanje, v katerem se množico E enakomerno in naključno razdeli

na k delov. V vsaki iteraciji se za testiranje ohrani 1 od teh k množic, ostalih $k-1$ se uporabi za učenje. Dobra lastnost metode je, da je vsaka entiteta natančno enkrat uporabljena za testiranje in da se vse uporabljajo tako za testiranje kot za učenje.

Stratificirano $k-$ kratno križno preverjanja pa je uporabljeno takrat, ko je razdelitev na podmnožice takšna, da so vsi razredi iz množice E predstavljeni v podmnožicah L in T v približno istih razmerjih kot v množici E.

Posebna oblika tega preverjanja je še Izpusti enega. V tem primeru je k enak moči E, v vsaki podmnožici je torej je 1 entiteta.

3.2 Diskretizacija

Diskretizacija je postopek kategoriziranja zveznih vrednosti oz. postopek kreiranja kategorij, s katerimi se nadomesti prej zvezne vrednosti. S tem se omeji število vrednosti, ki jih nek atribut lahko zavzame. Potrebna je, ker nekatere uporabljene metode, niso prilagojene na računanje z zveznimi atributi, in koristna, ker zmanjša šum. Njena največja nevarnost je izguba pomembnih podatkov.

V strojnem učenju sta dve glavni delitvi metod. Prva deli eno-spremenljivčne od več-spremnljivčnih, prve naenkrat kategorizirajo le eno spremenljivko, medtem ko druge več spremenljivk simultano. Druge obljubljajo boljšo rešitev, ker upoštevajo korelacije med spremenljivkami. Prve pa so preprostejše (in s tem v praksi izvedljive) in hitrejše.

Druga delitev deli nadzorovane od nenadzorovanih. Medtem ko prve upoštevajo le spremenljivko, ki jo bodo diskretizirale, druge pa upoštevajo tudi razred, ki mu primer pripada.

V primerih nenadzorovanih metod je ponavadi potrebno vnaprej določiti število kategorij, kar zna biti težavno. Z majhnim številom izgubljamo podatke, s prevelikim se izpostavljamo fragmentaciji in povečujemo varianco.

V naših postopkih smo uporabljali diskretizacijo enakih frekvenc. Zelo podobna ji je diskretizacija enakih širin. Metoda enakih frekvenc razdeli vse vrednosti v vnaprej določeno število kategorij tako, da je v vsaki kategoriji enak delež vrednosti. Metoda enakih širin razdeli interval, ki je omejen z največjo in najmanjšo vrednostjo atributa, na dogovorjeno število enako dolgih kategorij, v katere potem razvršča vrednosti.

Metode enakih frekvenc ima posebej probleme z atributi neidealne razporeditve vrednosti. Če nek atribut v polovici primerov zavzame vrednost 0, imamo lahko največ dve kategoriji. V vsaki polovico primerov, v eni vse ničle, v drugi vse ostale vrednosti.

4 Praktično delo

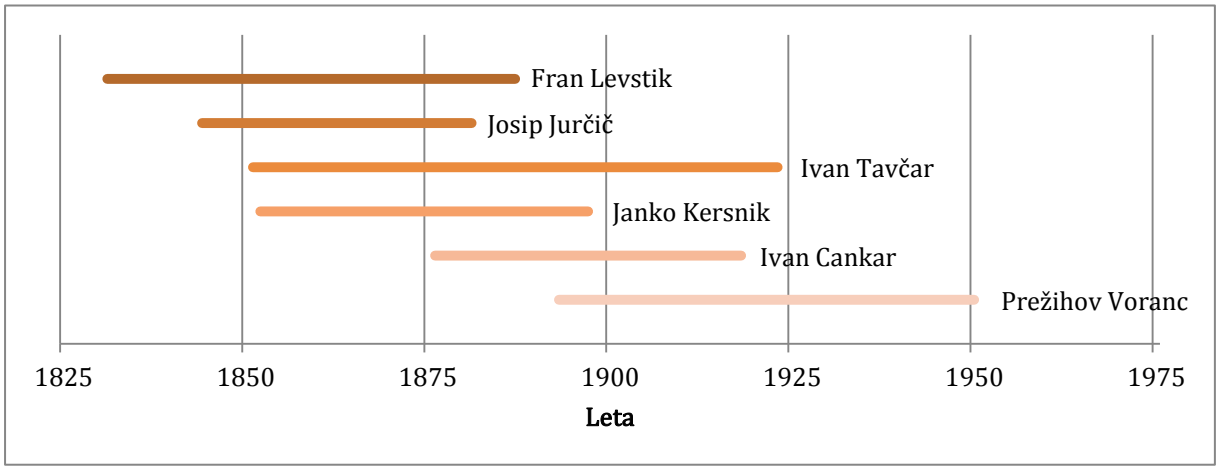
Cilj dela je bil na izbranih literarnih delih različnih avtorjev izvesti statistično analizo stila (t.i. stilometrijo) in se iz pridobljenih podatkov naučiti ločevati med avtorji.

Vsako besedilo je bilo obravnavano kot samostojna entiteta. Vsa statistika je bila narejena na vsakem besedilu posebej, v nasprotju z drugo možnostjo, ki je bila, da vsa besedila nekega avtorja skupaj sestavljajo eno entiteto.

Stilometrija je bila izvedena v dveh fazah, prvi, računsko manj zahtevni, in drugi, računsko bolj zahtevni. V prvi smo merili število pojavitev dvanajstih različnih znakov in drugih parametrov. V drugi pa smo opazovali in primerjali pojavitev besed.

4.1 Besedila

Uporabljenih je bilo 37 besedil šestih avtorjev. Trije avtorji so bili zastopani vsak s sedmimi besedili, eden s šestimi in eden s tremi. Črpana so bila predvsem s spletne strani http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika. Sedem besedil na avtorja se je zdel dovolj velik vzorec, tri besedila s Prežihovim Vorancem kot avtorjem pa so bila posledica slabe izbire pisatelja, vendar nam hkrati služijo kot motnje v testu. Sledi časovni trak življenja avtorjev, ki nam bo morda v pomoč pri razlagi tendenc klasifikacijske netočnosti.



Graf 1: Časovni trak življenja avtorjev

4.1.1 Lematizacija

Slovenska besedila so, zaradi specifičnih lastnosti slovenske slovnice, predvsem sklanjatve in spregatve, računalniškim algoritmom še posebej težko razumljiva. Kritičnega pomena za naše delo je bilo, da prepoznamo sklanjane in spregane besede kot npr predmet, predmeti, predmetov, .. kot različne obraze istega gesla. Na srečo je bil v okviru projekta LemmaGen [17,3] razvit algoritem, ki zna slovensko besedilo poenostaviti na seznam gesel. Pod pogoji licence LGPL prosto dostopen na *http://kt.ijs.si/software/LemmaGen*. Njegovo delovanje je tako, da za vsako podano datoteko ustvari novo datoteko, kamor prepiše vse besede, vendar zamenja vse samostalnike z njihovimi imenovalniki ednine in vse glagole z njihovimi nedoločniki. Njegova pravilnost je 82-97-procentna.

4.2 Stilometrija 1: Ločila

V prvi stopnji stilometrije smo poskušali najti čim več stvari, ki so merljive v besedilu in jih filtrirati na stvari, ki so koristne. Odločili smo se za metodo štetja, v kateri bomo zbirali število pojavitev ločil. V skupino "ločil" smo dodali še podatke o dolžinah besed in stavkov.

Za vsako besedilo so bile izmerjene naslednje vrednosti[1]:

I. povprečno število vejic na stavek
II. povprečno število podpičij na stavek
III. povprečno število klicajev na stavek
IV. povprečno število vprašajev na stavek
V. povprečno število narekovajev na odstavek
VI. povprečno število vezajev/pomišljajev na odstavek
VII. povprečno število tripičij na odstavek
VIII. povprečna dolžina stavka (v besedah)
IX. povprečna dolžina besede (v črkah)
X. aritmerično povprečje dolžine odstavka (v besedah)
XI. povprečno število številk na odstavek
XII. delež različnih besed glede na število vseh besed

Vse značilnosti, ki so se merile, so bile normalizirane, torej povprečja. Vsa povprečja so aritmetična. Iz podobnih raziskav je bilo razvidno, da drugi izmed matematičnih povprečij vedno jemljejo aritmetično, zato smo tudi mi pri tem ostali.

[1] Za orientacijo je bilo uporabljeno [26]

Gledajoč nazaj je treba omeniti eno pomanjkljivost v našem razmišljanju. Pri naši izbiri je prišlo do dveh nivojev povprečij, povprečja na stavek in povprečja na odstavek. Ob času izbiranja mer se je zdelo smiselno, da so znaki kot so tripičja in vezaji dovolj redki, da jih nima smisla meriti na stavek. Toda tako smo dobili vrednosti v različnih skalah in te vrednosti so odvisne ena od druge. Število tripičij npr. se meri na odstavek, ki se meri v besedah in ni standardiziran čez vse avtorje niti čez vsa obravnavana dela enega samega. Plus tega lahko dobi isti avtor povsem drugo število tripičij na odstavek, če se odloči drugače razporediti delo v odstavke.

Probleme bi nam lahko delalo tudi to, da se ista lastnost pisanja lahko izraža preko dveh načinov. V slovenščini se pričakuje, da ima daljši stavek več vejic. Mi merimo obe ti lastnosti, ki pa sta večinoma v odvisniškem odnosu. Morda bomo prišli do napačnih zaključkov, če bomo temu pojavu dali kar dva atributa, medtem ko imajo ostali po enega. Najverjetneje bi nam koristil kakšen filter teh značilnosti že na teoretičnem nivoju. Vendar je težko na matematično nekompatibilnem področju z relativno malo dokazanimi teorijami teoretično zagovarjati kakršnekoli odločitve.

4.2.1 Nenadzorovano učenje

Da bi si pridobili nekaj občutka o prostoru problema, ki ga rešujemo, smo začeli z metodami nenadzorovanega učenja. Vendar nismo bili uspešni. Z nobeno nismo dosegli, da bi pokazala jasne meje med avtorji.

Hierhično razvrščanje je vrnilo težko razumljivo lestvico povezav. Na najglobljem nivoju je resda večinoma pravilno združevalo besedila, višje po drevesu pa so z veliko večino prevladovale napačne povezave. K-means razvrščanje se je odrezalo še slabše. Četudi je bilo število skupin določeno vnaprej in bilo enako št. avtorjev, so izbrane centroide pripadale istim avtorjem, okoli njih pa se je zbrala eklektična zbirka avtorjev. V drugi stopnji stilometrije, ko smo delali analize na vektorjih besed, smo te poskuse ponovili. Vendar smo dobili podobne rešitve, le da je bil postopek njihovega pridobivanja računsko izredno bolj zamuden.

V nadaljevanju smo se zato usmerili v klasifikatorje.

4.2.2 Primerjava klasifikacijskih metod

Z nenadzorovanimi metodami smo bili neuspešni. Odločili smo se narediti primerjavo med klasifikacijskimi metodami, izmed katerih smo izbrali štiri

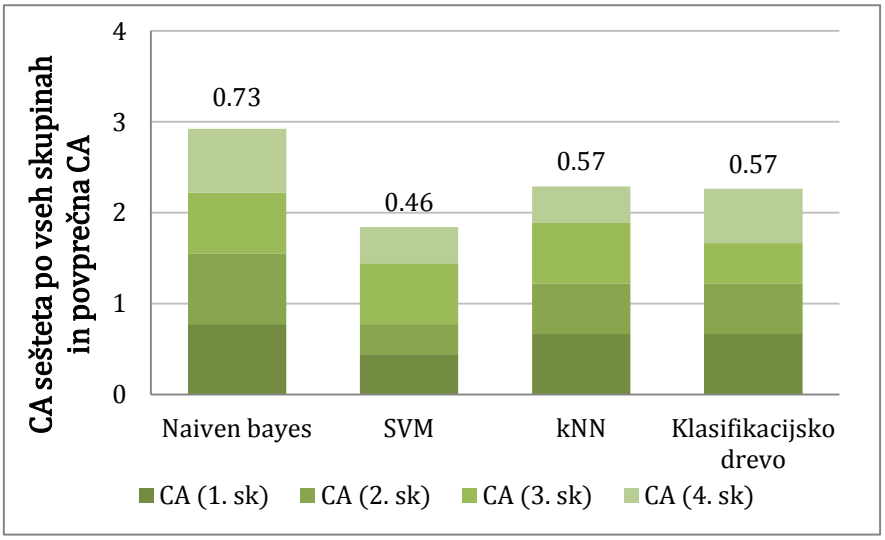
izmed najbolj znanih osnovnih metod: naivni Bayes, odločitvena drevesa, SVM in kNN.

4.2.2.1 Poskus A: Primerjava klasifikacijskih metod

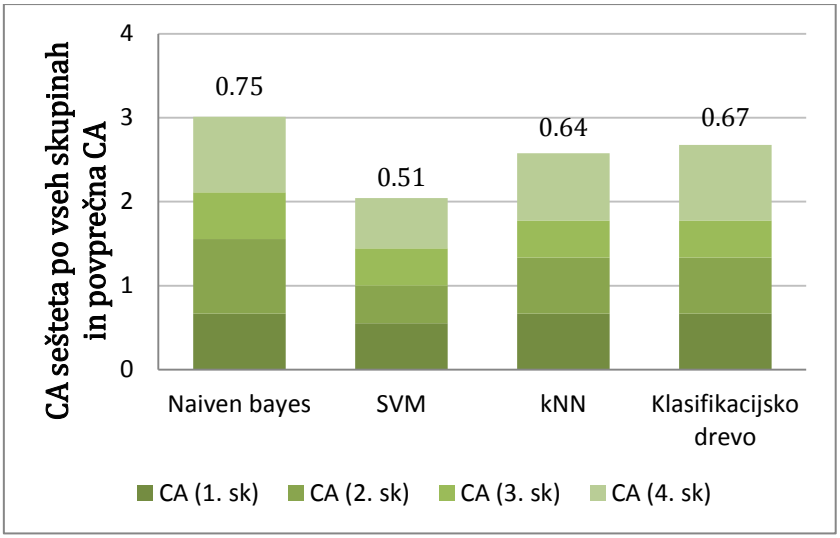
Besedila smo naključno razdelili na 4 skupine. Tri izmed njih smo združili v učno množico, iz četrte pa naredili testno množico. Za vse štiri metode smo nato izmerili klasifikacijsko točnost in Brierjevo oceno. Za diskretizacijo, potrebno za informacijski prispevek, pa je bil odgovoren algoritem enakih frekvenc s štirimi razredi. Ves postopek je bil ponovljen tri-krat za tri različne delitve v 4 skupine.

Pri vseh testih se je najboljše odrezal naivni Bayes.

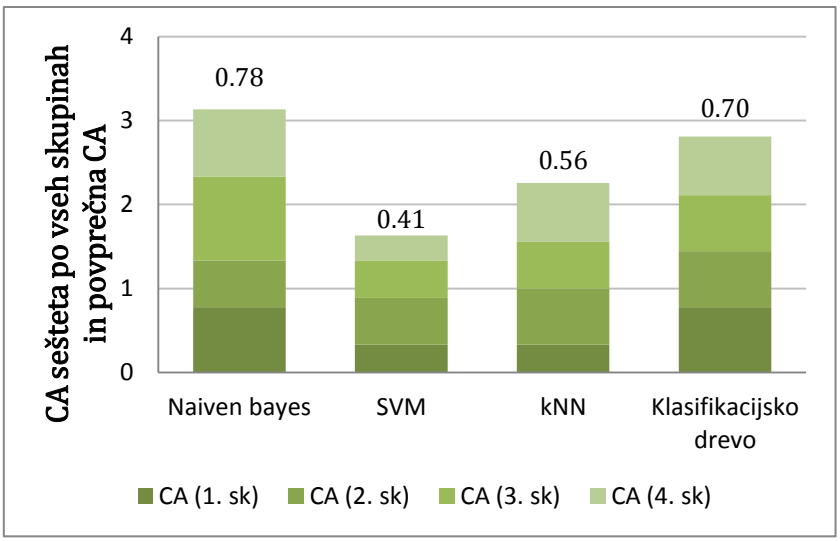
Po klasifikacijski točnosti je v 3 od 4ih primerov (75%) napovedal prav. Medtem se je SVM ves čas ostajal okrog 50%. KNN je v odvisnosti od izbire primerov znal pokazati dobre rezultate, vendar se mu taka nezanesljivost ne šteje nujno med kvalitete. Edina druga metoda, ki se je Bayesu približala in dosledno ohranjala drugo mesto, je bilo klasifikacijsko drevo, vendar je njen uspeh kvarila ista fluktuacija kot pri kNNju. (Grafi 2. – 4.)



Graf 2: Klasifikacijska točnost različnih metod po delitvi 1

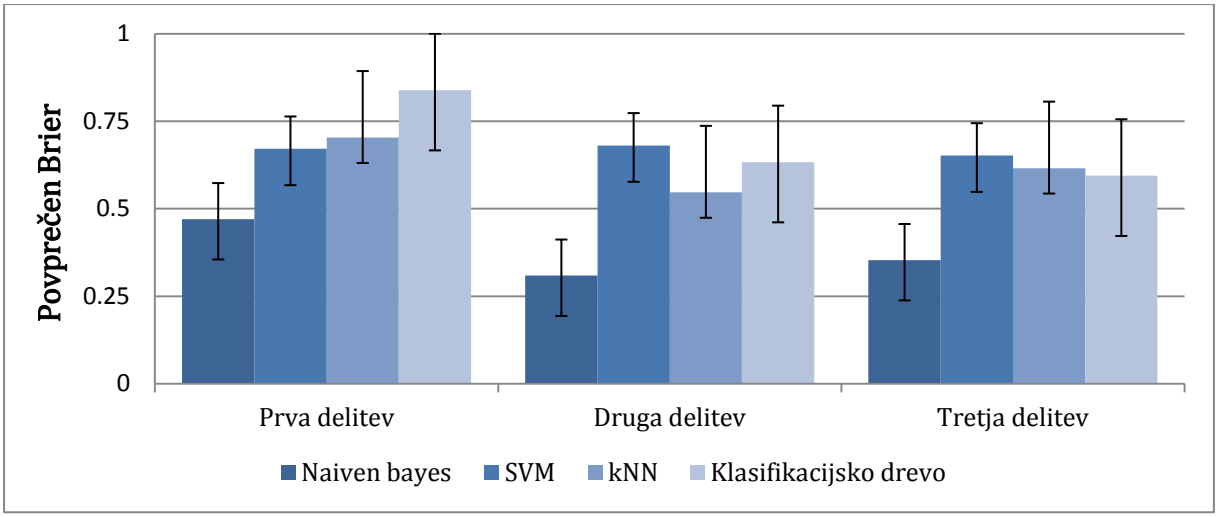


Graf 3: Klasifikacijska točnost različnih metod po delitvi 2



Graf 4: Klasifikacijska točnost različnih metod po delitvi 3

Iz Brierjeve ocene (graf 5.) pa je razvidna previdnost vsake metode pri napovedih. Brier naivnega Bayesa je bil skoraj v vsakem posameznem primeru najnižji od vseh. Tako je možno sklepati, da so verjetnosti, s katerimi napoveduje, uporabni indikatorji zaupanja rezultatom.



Graf 5: Povprečena Brierjeva ocena glede na različne delitve

4.2.2.2 Poskus B: Optimizacija poskusa A

V nadaljevanju smo poskušali pogoje vsaj malo optimizirati. Zaradi prevelikega števila atributov (12) je bil naš vzorec 37 knjig (oz. 28 knjig v učni množici) nereprezentativen. Zanimalo nas je, ali bi lahko dobili boljše rezultate s spreminjanjem razmerja med številom atributov in velikostjo učne množice.

Izbiro atributov smo izvedli z informacijskim prispevkom. Za vsako učno množico smo v prvem koraku izločili najslabših N atributov glede na informacijski prispevek in jo šele nato poslali klasifikatorjem. Za N smo se odločili vzeti 1/3 vseh atributov tj. 4 atribute. Ostali postopki so bili isti kot v prejšnjem poskusu.

Na dno čisto vsake razporeditve atributov sta padla Povprečna dolžina besede in Povprečno število števk. Njun najpogostejši spremljevalec pa je bilo Povprečno število narekovajev.

Rezultati so bili presenetljivi in zelo podobni tistim iz prejšnjega poskusa. Številke klasifikacijske natančnosti so skoraj enake, le da so v vseh primerih za nekaj stotink do dobro desetinko višje. Najboljši rezultat pa ima še vedno Bayes z okrog 75%. Tudi Brierjeva ocena se v svojem povprečju ni omembe vredno spremenila, edina zanimiva razlika pri njej je, da so odkloni od povprečja dosti večji. Odločitveno drevo je v praktično vseh primerih zasedlo tako 0 kot 1 kot poljubne vmesne vrednosti.

4.2.2.3 Poskus C: Klasifikacija

Rezultate Bayesovega klasifikatorja iz Poskusa A smo hoteli pogledati bolj od blizu, predvsem nas je zanimala tista četrtina primerov, ki je bila uvrščena napačno. Upoštevajoč skromne vhodne podatke je četrtina primerov res majhen delež. Kar se zgodi z njimi, kaže matrika zmot.

		Napovedani razred					
		Fran Levstik	*Janko Kersnik*	*Ivan Tavčar*	*Ivan Cankar*	*Prežihov Voranc*	*Josip Jurčič*
Resnični razred	*Fran Levstik*	*14*	*-*	*-*	*-*	*-*	*4*
	Janko Kersnik	*-*	*18*	*-*	*-*	*1*	*2*
	Ivan Tavčar	*-*	*3*	*17*	*-*	*-*	*1*
	Ivan Cankar	*3*	*1*	*-*	*15*	*2*	*-*
	Prežihov Voranc	*-*	*3*	*-*	*-*	*6*	*-*
	Josip Jurčič	*5*	*-*	*3*	*1*	*-*	*12*

Tabela 3: Matrika zmot za poskus C

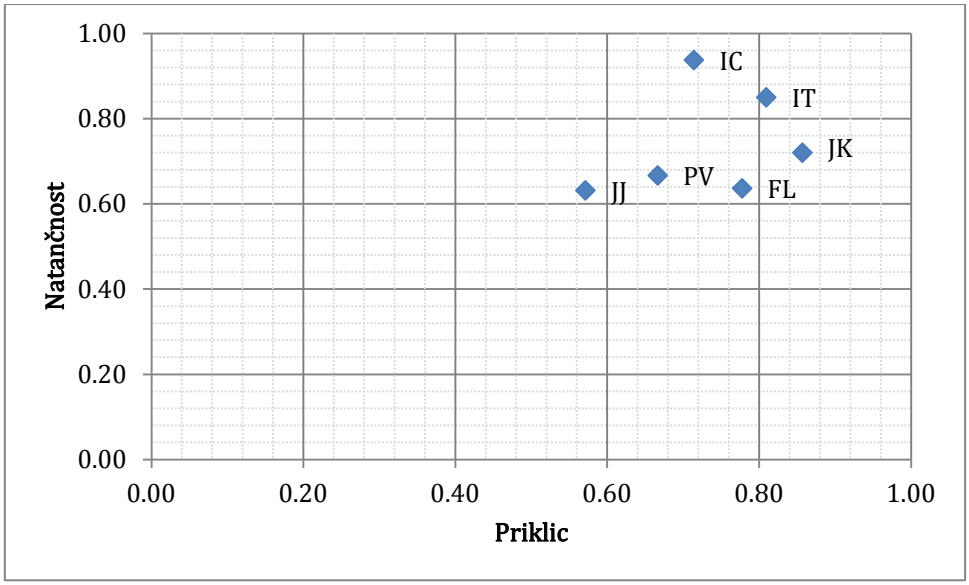
Priklic in natančnost za prikazano matriko sta sledeča.

	Fran Levstik	*Janko Kersnik*	*Ivan Tavčar*	*Ivan Cankar*	*Prežihov Voranc*	*Josip Jurčič*
Priklic	*78*	*86*	*81*	*71*	*67*	*57*
Natančnost	*64*	*72*	*85*	*94*	*67*	*63*

Tabela 4: Priklic in natančnost za poskus C

Priklic definira koliko denimo Cankarjevih knjig klasifikator najde. Natančnost pa koliko izmed klasificiranih Cankarjev je tudi res Cankarjevih.

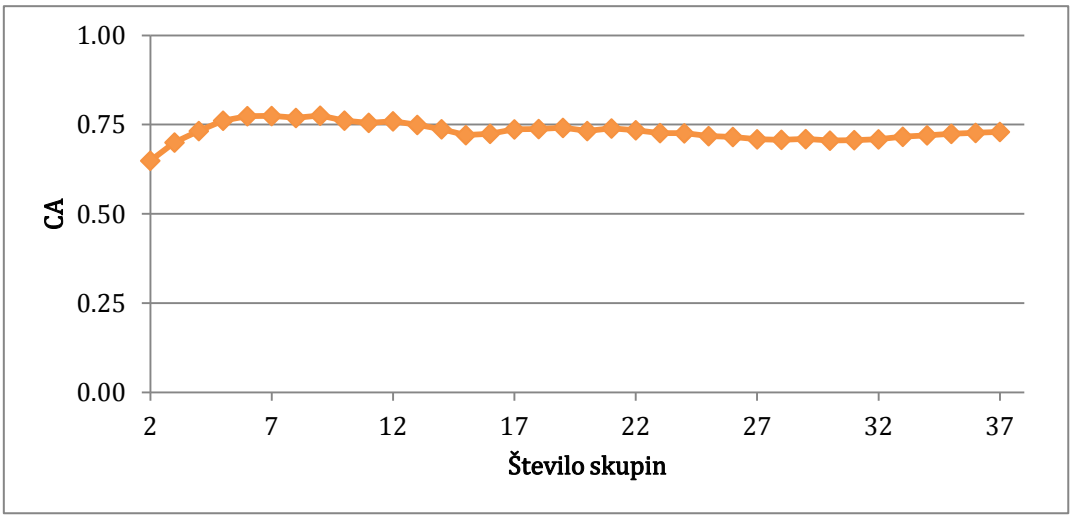
Najslabši so rezultati pri Jurčiču, le dobro polovico njegovih knjig klasifikator prepozna kot take in le malo boljšo polovico od tistih, ki jih označi za Jurčičeva, so tudi v resnici njegove. V pogledu natančnosti je najboljši Cankar, skoraj vse, za kar klasifikator reče, da je njegovo, tudi je njegovo. Najde sicer le 70% njegovih knjig, a skoraj vse od njih so res njegove. Poleg Cankarja je Tavčar edini drug avtor, za katerega smo dosegli večjo natančnost kot priklic, čeprav le komaj. Zanimiv je še Prežihov Voranc, ki je bil predstavljen le s tremi knjigami. Klasifikator ne prepozna le ene njegove knjige, prepozna torej 2/3 njegovih knjig, ravno tako je tudi njegova natančnost na 2/3. Prikaz in natančnost sta grafično prikazana na Grafu 6.



Graf 6: Grafičen prikaz priklica in natančnosti za vsakega avtorja

Ker smo imeli v tem primeru kratke, računsko nezahtevne vektorje entitet, smo se odločili še za test vpliva števila k v k-kratnem križnem preverjanju na klasifikacijsko točnost. Z zelo malo podatki, kot jih imamo v našem primeru, težimo k izbiri "izpusti enega", čeprav velja za računsko najbolj naporno.

Začeli smo z dvema skupinama z po pol entitetami, eno testno in eno učno, in nato število skupin povečevali. Končali smo z 37 skupinami, od katerih je vsaka bila sestavljena iz ene entitete. Prečno preverjanje smo ponovili večkrat, vsakič z drugačno razporeditvijo entitet, da smo dobili statistično boljšo reprezentativnost.



Graf 7: Razmerje med povprečno klasifikacijsko točnostjo in številom k pri večkrat ponovljenem k-kratnem prečnem preverjanju

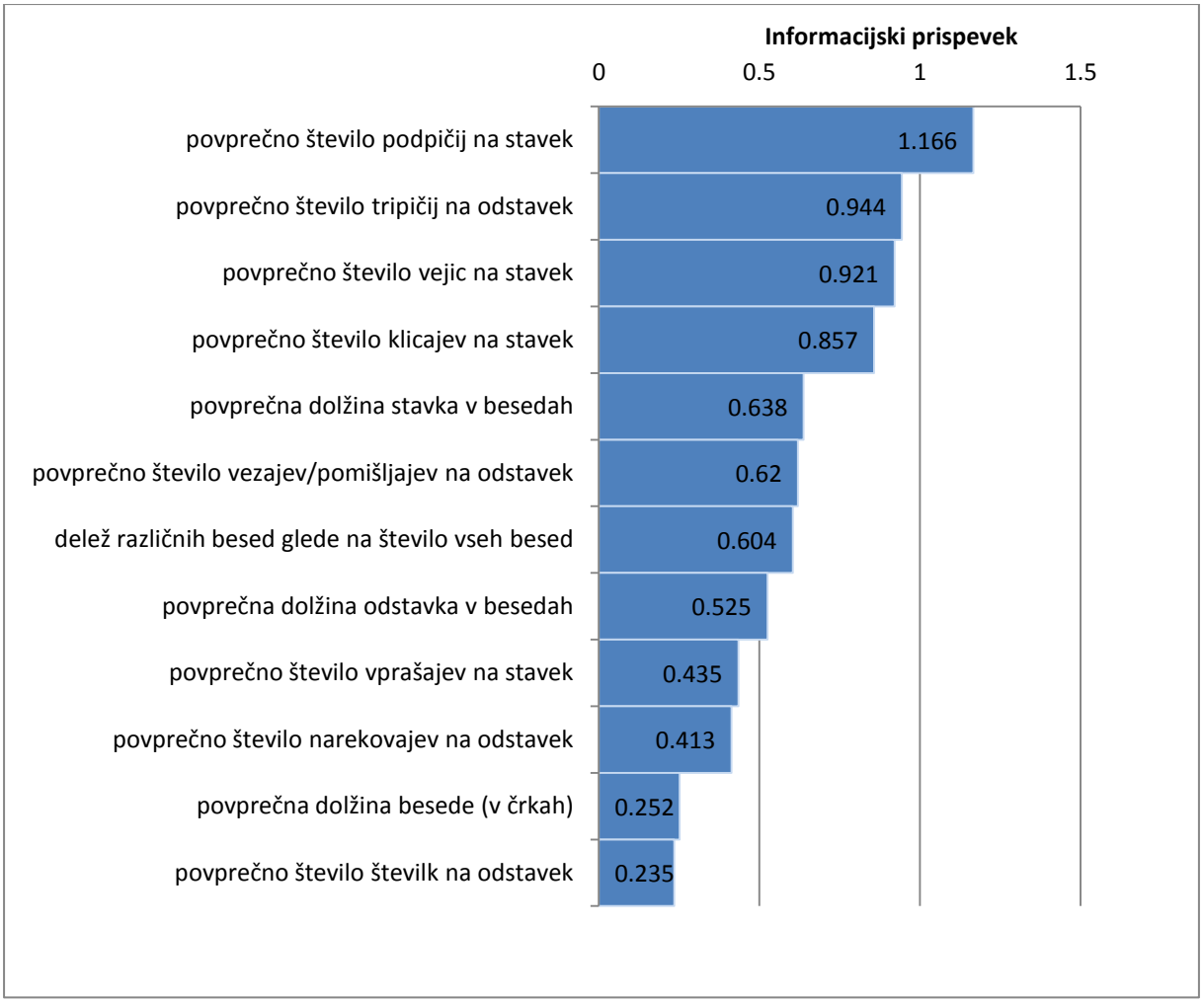
Iz grafa 7 je jasno vidno, da z razpolovitvijo začetne množice največ izgubimo. Učna množica postane premajhna za učinkovito učenje. Točnost pri največjem

številu skupin, je najbolj realna, saj vsako entiteto posebej obravnava in za vsako posebej zabeleži ali je bila prepoznana ali ne. Vmesni rezultati so bili v vsaki iteraciji v veliki meri odvisni od sreče. Njihovo povprečje pa kaže zanimiv maksimum okrog 7-kratnega preverjanja. Z vidika statistike je to področje očitno najbolj primerno za skrivanje napak.

4.2.2.3.1 Uvrstitev atributov

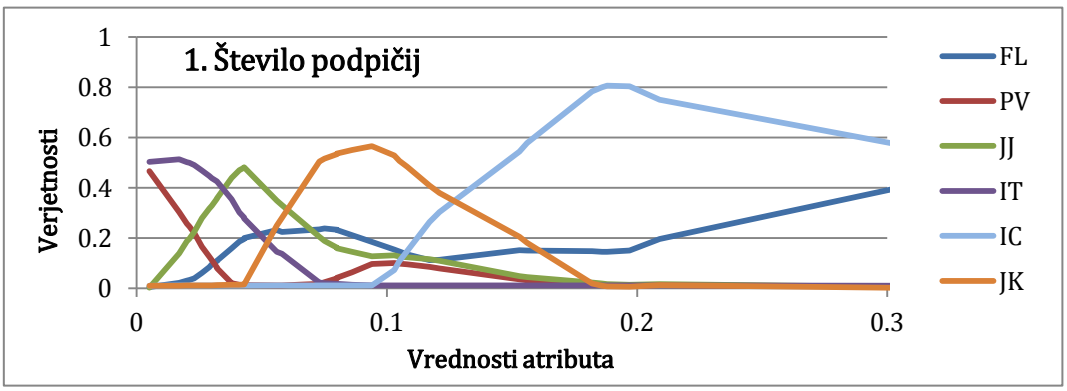
Oglejmo si še stopnje pomembnosti atributov z vidika informacijskega prispevka,ki smo ga v ta namen ves čas uporablajli, in glede na vsa besedila. V spodnjem grafu (Graf 8) je prikazan seznam vseh atr razporejenih po pomembnosti.

Čisto na vrh prideta dve manj pogosti slovnični kategoriji podpičja in tripičja. Na tretje mesto se uvrsti vejica, poleg pike najpogostejše ločilo v slovenščini in po naših novih podatkih očitno eno izmed bolj distinktnih. Število vejic ima svoj vpliv na dolžino stavka in ta se tudi pojavi že na 5. mestu in vendar z dosti nižjo oceno pomembnosti, ki tako dokazuje, da na dolžino stavka vendar vpliva še dosti drugih kriterijev, ki so neodvisni od vejic. Na dno pade povprečno število številk, za kar predvidevamo, da so kriva oštevičena poglavja, ki so vnesla šum v vrednosti. Zanimiva je še vrednost povprečne dolžine besede, ki prav tako pade na konec. Leksika pisateljev je v povprečju dosti večja kot leksika ne-pisateljev, vendar se to ne izraža v povprečni dolžini besed, ki jih uporabljajo. Ne glede na to, od kod črpajo svoje besedišče, koliko tujk in terminov uporabljajo, je povprečna dolžina besede zanemarljiva podrobnost. Isto dokazuje tudi Graf 12, v katerem je pokazana preferenca pri dolžini besed za vsakega avtorja. Za ostale atribute pa prepuščamo bralcu, da si ustvari mnenje o njihovem mestu v hierarhiji. Vse naše trditve so konec koncev le ugibanja.

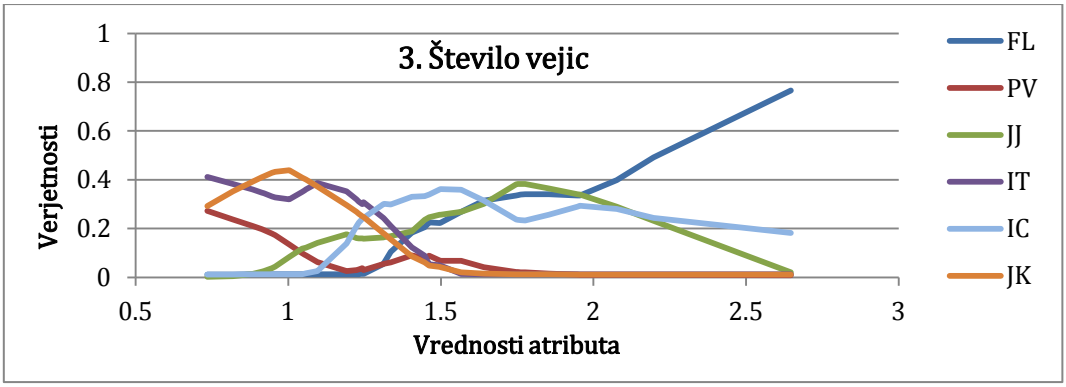


Graf 8: Razporeditev atributov "Ločila" po kriteriju njihovega informacijskega prispevka na podlagi vseh besedil

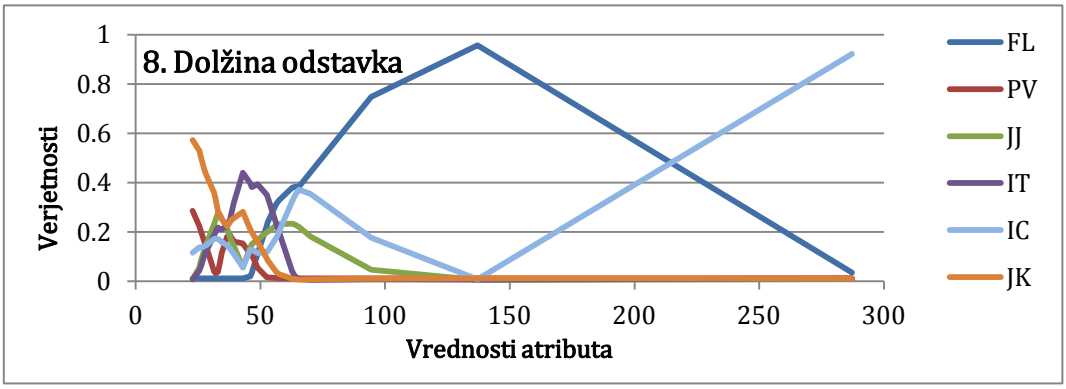
Pokazali bi radi še razlike v verjetnostnih porazdelitvah vrednosti atributov glede na razrede. Spodnji grafi (graf 9. – 12.) prikazujejo verjetnostne napovedi pripadnosti vsakemu razredu v odvisnosti od vrednosti enega atributa. Začnemo z najvišje uvrščenim atributov: številom podpičij.



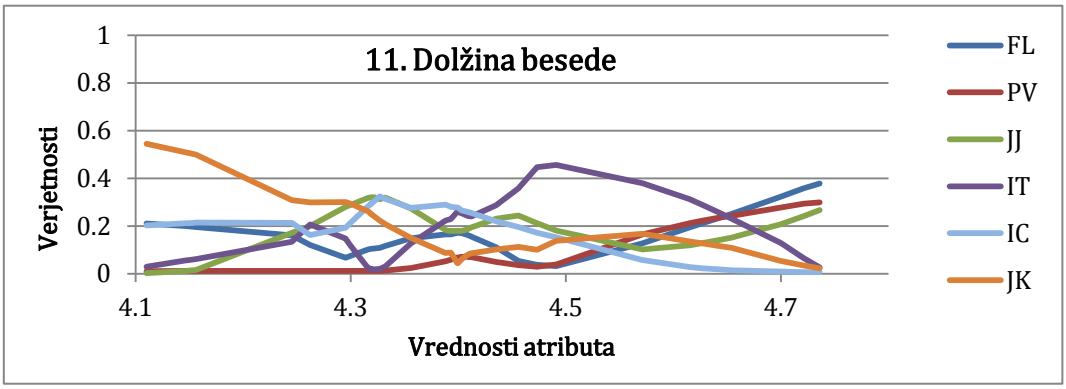
Graf 9: Verjetnostna porazdelitev razredov glede na vrednosti atributa Število podpičij



Graf 10: Verjetnostna porazdelitev razredov glede na vrednosti atributa Število vejic



Graf 11: Verjetnostna porazdelitev razredov glede na vrednosti atributa Dolžina odstavka



Graf 12: Verjetnostna porazdelitev razredov glede na vrednosti atributa Dolžina besede

V prvih dveh grafih je bolj jasno, kateremu razredu najverjetneje pripada primer samo na podlagi tega enega atributa. Če je število podpičij blizu ničle, obravnavamo najverjetneje Prežihovega Voranca ali Ivana Tavčarja. Če je število podpičij približno 0,2, je najverjetneje Ivan Cankar. Pri dolžini besed in dolžini odstavka so takšni sklepi dosti manj ločevalni. Če je dolžina besede približno 4,7, najverjetneje ne obravnavamo Ivana Cankarja, mogoče tudi Janko Kersnika ne, ostali imajo pa približno enake verjetnosti. Dolžina odstavka pa je zanimiva, ker ima pod vrednostjo 100 zelo podobno razdeljene verjetnosti, za vrednosti večje od 100 pa najverjetneje govorimo o Franu

Levstiku ali Ivanu Cankarju. Neomenjeni atributi imajo grafe, ki so, glede na atributovo mesto v hierarhiji, podobni nekaterim od predstavljenih grafov.

4.2.3 Epilog

Bili smo nadvse navdušeni nad rezultati te stopnje. Izredno presenetljivo je, da smo že na tej stopnji dosegli 75% točnost. S samo 12 (oz. 9) merjenimi lastnostmi besedila, smo v 3 od 4 primerov znali določiti avtorja. Natančnost je nekoliko slabša, z izjemo Cankarja lahko govorimo o dobrih dveh tretjinah pravilno klasificiranih primerov. Klasifikacija torej zaenkrat spada med nezaneslije, vendar je avtorski duh v besedilih očitno bolj izrazit, kot smo si upali verjeti, in izslediti ga znajo že osnovne lastnosti kot je dolžina povedi in število vprašajev.

Zanimivo je tudi, da smo lahko s Poskusom B izboljšali rezultate Poskusa A. Dodatni atributi, ki so bili uporabljeni, so očitno pripomogli le k večji nedoločenosti. Čeprav so motili klasificiranje, jih klasifikacijske metode niso znale izločiti. Izjema temu je Bayes. Bayes in SVM sta edini od teh štirih metod, ki sta spretni z veliko dimenzionalnim prostorom problema. KNN in drevesa računata bolj na statistično reprezentativnost učne množice, ki je dosežena z velikim številom učnih primerov. Zanje pa velja, da vlada velika podobnost znotraj razreda in velika raznolikost med razredi. Naša množica je bila zelo majhna, razlika v raznolikosti znotraj in raznolikosti zunaj pa dosti premajhna. Za vse sledeče izračune smo zato uporabljali Bayesa.

4.3 Stilometrija 2: Besede

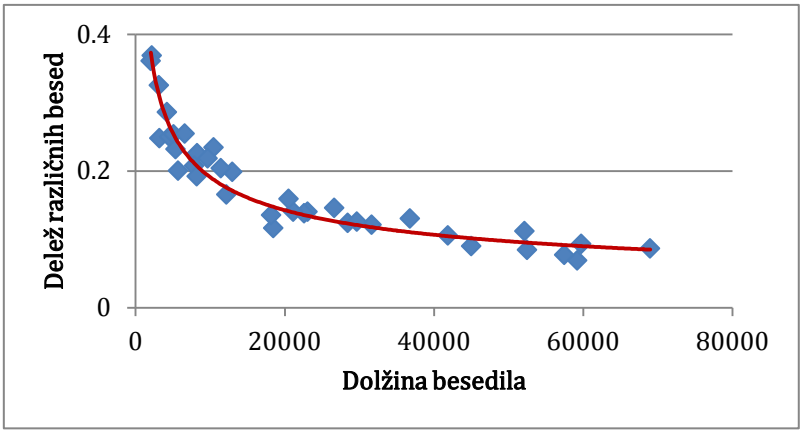
V drugi stopnji smo se posvetili besedam. Za razliko od ločil, katerih mesto večinoma določa slovnica, je izbira besed dosti bolj osebna. Odločiti se je treba med sopomenkami, tujkami, podrednimi vezji in izrazi z za las različnimi pomeni, ki znajo biti vezani na kraj, čas, družbeno mesto, čustva, spol ali pa le zven besede.

4.3.1 Zipfov zakon

Zipfov zakon nas je zanimal, zaradi svoje pogoste citiranosti in zaradi razširitve njegove aplikativnosti na druga, neliterarna, področja [5]. Odločili smo se preveriti, koliko se Zipfova dognanja prilegajo slovenskim besedilom.

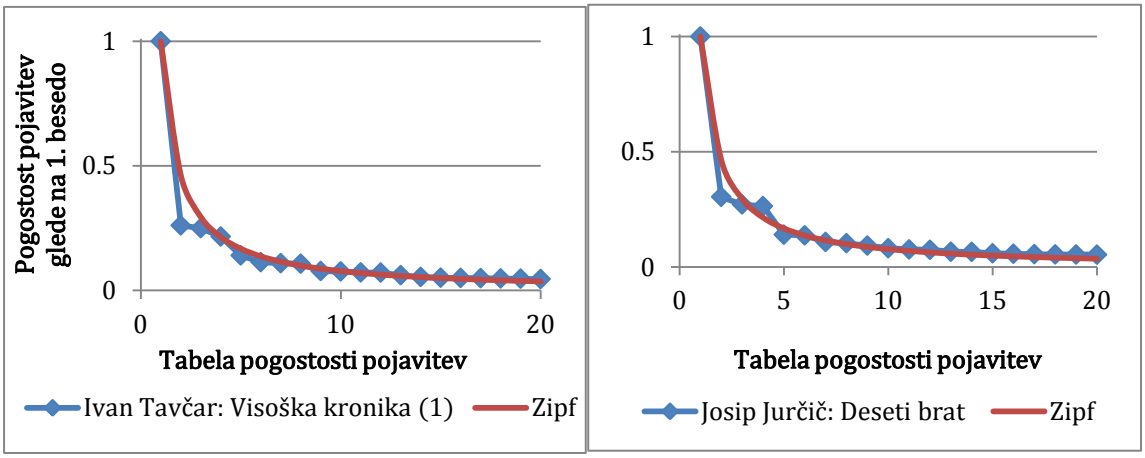
Najprej smo vzeli v obravnavo vsa besedila in pogledali razmerje med dolžino besedila in številom uporabljenih različnih besed. Na graf Zipfove funkcije

smo dali vseh 37 obravnavanih del in dobili Graf 13. Vsako točko, ki predstavlja 1 delo, smo na isti graf postavili glede na število vseh in število različnih besed. Dobili smo dobro ujemanje. V glavnem je potrjeno pričakovanje, da daljše besedilo pomeni več ponavljanja, vendar ta funkcija ni linearna. Zelo kmalu se ljudje začenjajo ponavljati. Če razdelimo besedila v kategorije kratkih, srednje dolgih in dolgih besedil, delež različnih besed najhitreje upada v kategoriji kratkih besedil.



Graf 13: Razmerje med dolžino besedila in številom različnih besed

V naslednjem koraku smo Zipfov zakona testirali na posameznih besedilih. Začeli smo s tremi najdaljšimi besedili treh različnih avtorjev (Josip Jurčič: Deseti brat, Ivan Tavčar: Visoška kronika 1. del in Ivan Cankar: Mimo življenja). Uporabili smo 50 najpogostejših besed, v Grafu 14 pa prikazali le 20, ker se je izkazalo, da že teh 20 sledi zakonu.



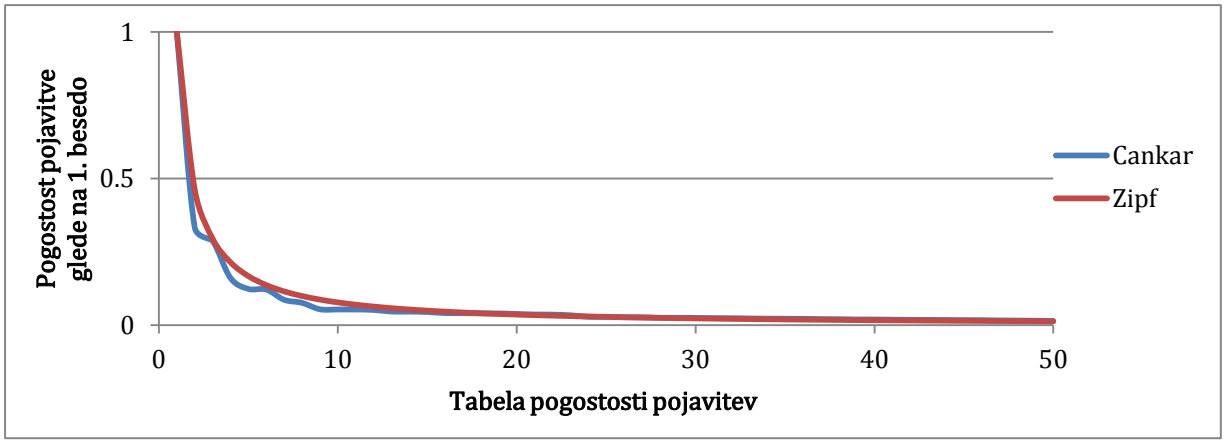
Graf 14: Zipf za Desetega brata in Visoško kroniko

Vsi trije so se obnašali skoraj identično. Približno so sledili enačbi (1), z $y \sim 1.1$. Le v prvih 10ih besedah, ko je statistična natančnost napovedi najmanjša, so nekoliko padli pod njene vrednosti.

Pri testiranju najkrajšega besedila, ki ima vsega skupaj le dobrih 700 besed, je bilo obnašanje enako. Vrednosti so s podobno natančnostjo sledile isti enačbi.

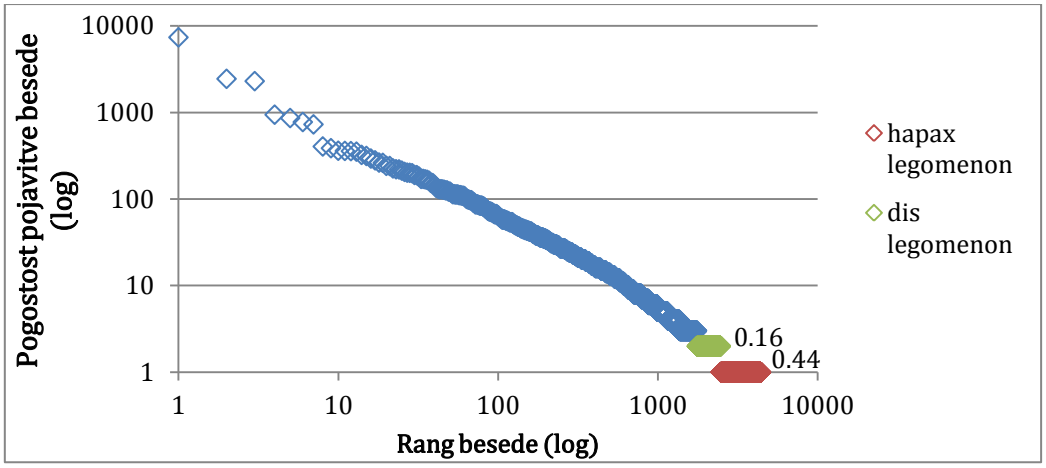
Zadnji test smo izpeljali na vseh sedmih delih Ivana Cankarja, ki smo jih imeli v bazi. Rezultatov nismo znali napovedati. Zipfov zakon implicitno govori o homogenih sklopih. Korpusi nepovezanih besedil nimajo dosti homogenih lastnosti, od njih se pričakuje, da kršijo statistične zakone. Osnovno vprašanje, ki se nam je postavljalo, pa je bilo, ali sedem naključno izbranih Cankarjevih mojstrovin ustvarja homogeno enoto ali ne.

Zipfov zakon se je še vedno skladal z našimi opazovanji, z istim y, vendar z malo daljšim začetnim območjem večjega odstopanja (do 15. besede) (Graf 15).



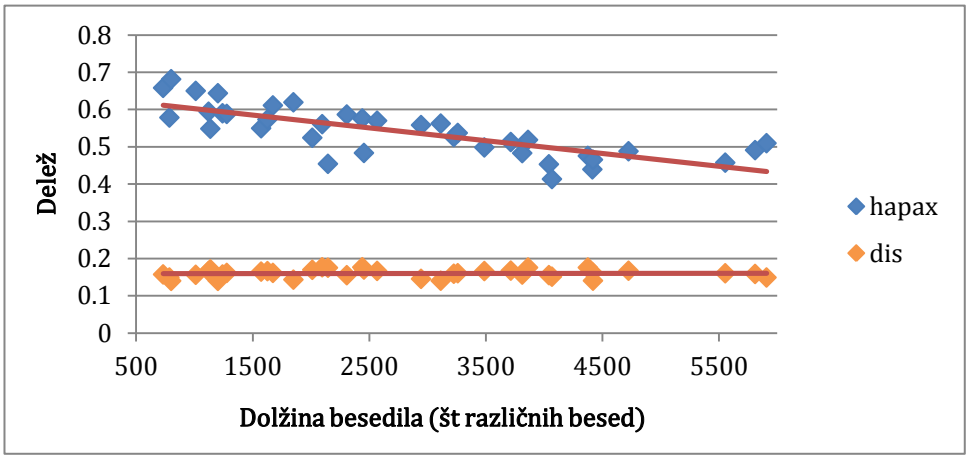
Graf 15: Zipf za Cankarjev korpus

V tem okviru lahko predstavimo t.i. Hapax legomena, to so besede, ki se pojavijo le enkrat v neki omejeni množici, ki jo predstavlja bodisi dokument bodisi zbirka dokumentov bodisi neka druga omejitev. Podobno dis legomena opisuje besede, ki se pojavijo dvakrat. V povprečju se v večjih tekstih pojavi od 40-60% hapax legomena in nadaljnih 10-15% dis legomena (ta delež se nanaša na delež različnih besed) [18]. Raziskave so se tipično delale na korpusih dolgih milijon besed ali več.



Graf 16: Gostota števila pojavitev besed (za Cankar: Mimo življenja)

V naših besedilih je bila zastopanost hapax in dis legomena podobna. Hapax je bilo od 41 do 68 %, ta številka se je zmanjševala z daljšanjem besedila. Dis je bilo med 14 in 18 %, ne glede na dolžino besedila. Grafična prikaza sta Graf 16 in 17.



Graf 17: Delež hapax in dis legomenonov

Pomen hapax je nedorečen. Na prvi pogled bi lahko z njimi ločili dokumente med sabo. Ker se pojavijo le enkrat, bi bila točnost takega modela 100-procentna. V naših besedilih je več kot 10000 besed, ki se pojavijo le enkrat v le enem dokumentu. Te besede bi bile enakovredne naslovom. Ravno to pa bi jih tudi omejevalo. Natančno bi ločili med v naprej dogovorjenimi besedili. Naučili se ne bi ničesar, razen morda kakšnih zanimivih bizarnih povezav. Iz nobenega novega besedila ne bi znali prebrati ničesar uporabnega in ga zatorej ne bi znali klasificirati.

4.3.2 Pomembnost omejene natančnosti računalnikov

V postopkih eksperimentiranja smo naleteli na zanimiv in nevaren problem. Računalniki niso popolni matematični modeli, ovira jih njihova omejena natančnost.

Števila so v računalnikih predstavljene na omejenem številu bitov. V eksponentnem formatu je razpon od najmanjšega do največjega predstavljivega števila zelo velik, vsekakor dovolj velik za naše potrebe. Vendar je število decimalk, ki se jih dá shraniti v tak format, proporcionalno velikosti števila. Fizično nemogoče je v omejenem številu bitov shraniti zelo veliko število z zelo dobro decimalno natančnostjo. Zaokroževanje poskrbi, da se zadnje decimalke porežejo. Težave nastanejo pri matematičnih operacijah. Pri seštevanju zelo velikega in zelo malega števila se lahko zgodi, da celotno malo število pade med porezane decimalke. Rezultat tako postane enak velikemu številu. Podobno se dogaja pri vseh operacijah.

V poskusih smo imeli opravka s po 24 000 atributi zelo majhnih vrednosti in z naivnim Bayesom, ki množi vplive teh atributov. Pri računanju je rezultat od neke točke naprej ostajal nespremenjen. Vse zaokroževanje, ki se je zgodilo, se zdi upravičeno. Če se številka signifikantno ne spremeni, jo lahko pustimo nespremenjeno. Toda vsota vseh neprištetih atributov bi naš dobljen rezultat signifikantno spremenila.

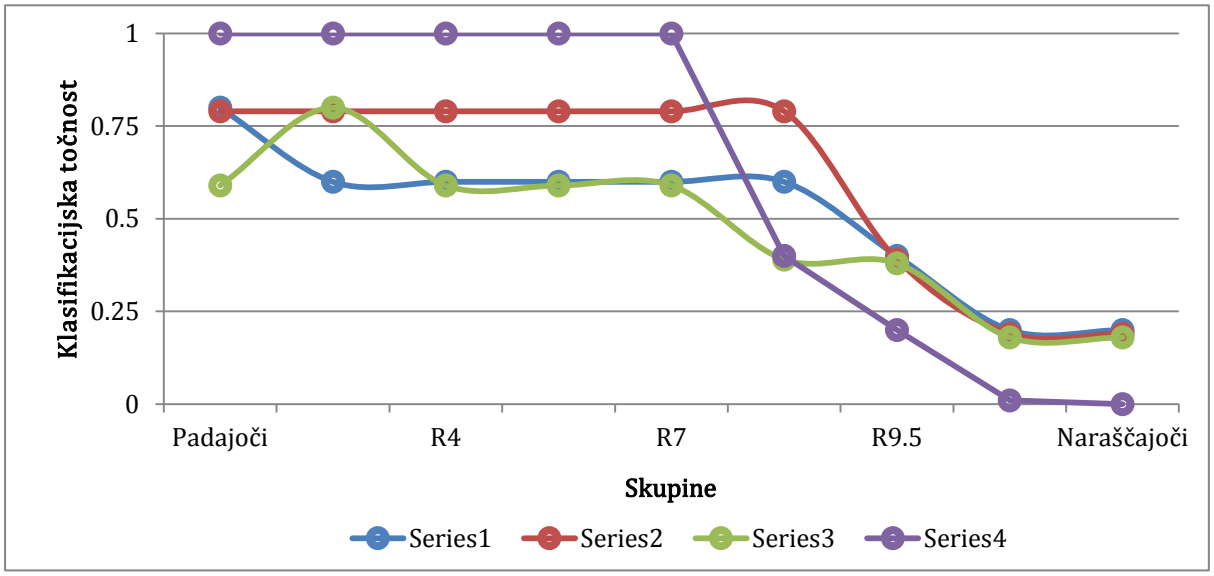
Prvič smo to napako opazili med klasificiranjen z uporabo velikega števila atributov. Napako smo pripisali šumu, ki ga povzročajo nekoristni atributi. Rekli smo, da so nekoristni atributi naključne vrednosti, ki naš izračunan rezultat odmikajo od resnične vrednosti v naključne smeri. V idealnem primeru se med sabo izničijo, v našem primeru pa kot naključno odstopanje od idealnega primera poskrbijo za slabo klasifikacijo. Šele kasneje smo zapazili različne klasifikacijske točnosti ob različnem vrstnem redu atributov. Presenetljiva, ampak nadvse pomembna ugotovitev.

Za ilustracijo smo zgradili eksperiment, ki prikaže odnos, ki ga ima klasifikacijska točnost do vrstega reda atributov.

Iz besedil smo naključno izbrali 4 testne množice, ki so vsaka obsegale 5 entitet 5ih razredov. Te iste 4 testne množice bomo uporabljali skozi vso poglavje Stilometrija 2: Besede.

Vrsti red atributov smo razdelili v 9 skupin: 2 skrajni in 7 vmesnih. 2 skrajni predstavljata situaciji, ko so atributi razporejeni bodisi padajoče bodisi naraščajoče glede na svojo pomembnost. V vmesnih testih smo vrstni red tako priredili, da je najpomembnejših X atributov prišlo na konec seznama

atributov. X je črpal iz urejene množine $\{1000, 4000, 5000, 7000, 9000, 9500, 10000\}$. Iz Grafa 18 je razvidno padanje funkcije klasifikacijske točnosti. Razen majhne deviacije pri R1, se vse do R7 držijo prvotne vrednosti, ki potem nenadoma do R10 pade na novo skrajno vrednost. Razlika med največjo in najmanjšo točnostjo pa je od 40 do 100 procentov, kar sta zaskrbljujoče velika deleža in razloga, da smo previdni pri izbiri zaporedja atributov.



Graf 18: Klasifikacijska točnost glede na vrstni red atributov

4.3.3 Poskus D: manjšanje množice besed

V prejšnjem poglavju smo prikazali glavni razlog za zmanjšanje množice besed. Drug razlog za vlaganje truda v ta namen so prekodimenzionirani vektorji, ki jih dobimo iz izbranih besedil. Posamezna obravnavana dela so obsegala od 700 do 6000 različnih besed, skupno se je našlo slabih 26000 različnih besed, dosti preveč dimenzij za vektorje. Predstavljali pa smo si, da bo težko izmed njih izbrati najboljšo podmnožico.

Poskus, ki smo si ga zamislili, je bil preprost: besede razporediti po njihovi pomembnosti in jih v sistem dodajati postopoma ter se ustaviti, ko bo željeni klasifikacijski točnosti zadoščeno.

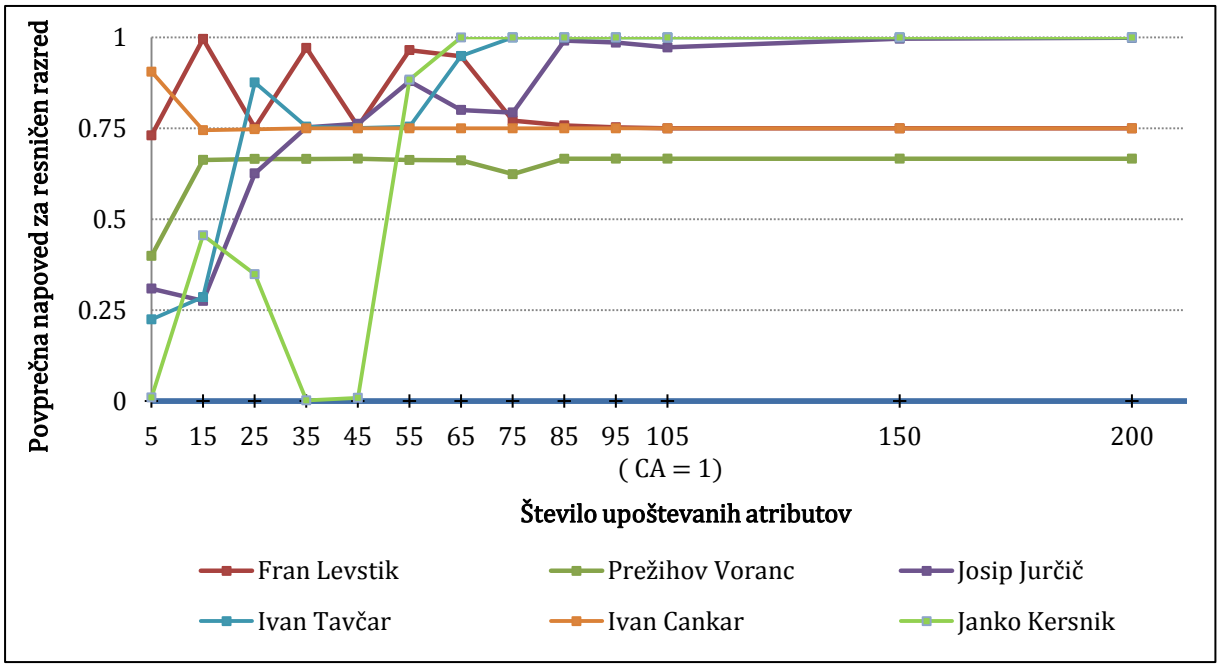
Izmed klasifikacijskih metod smo na podlagi poskusov iz prejšnjega razdelka izbrali naivnega Bayesa, kot merilo pomembnosti smo spet vzeli informacijski prispevek, enega izmed najbolj priljubljenih meril strojnega učenja. Čeprav slednji potrebuje diskretizirane atribute, medtem ko jih prvi ne, je informacijski prispevek dovolj kvalitetna mera z dovolj preprosto računsko zahtevnostjo, da smo se zanjo odločili. Za razporeditev atributov po

pomembnosti smo uporabili diskretizacijo. Bayes je bil pognan na originalnih vrednostih atributov.

Merili smo klasifikacijsko točnost testne množice in procent, s katerim je bil napovedan rezultat za vsako posamezno knjigo. Testirati smo začeli s petimi besedami, ki smo jim potem v skupinah po deset dodajali nove besede.

Poskus smo izpeljali štirikrat, vsakič s testno množico moči pet, ki je vsebovala pet različnih avtorjev. Testne množice se med sabo niso prekrivale.

Rezultati so se za vsak primer zelo razlikovali. Spodnji graf predstavlja funkcije napovedi, kot so se kazale v eksperimetu. Pri meji 105 je v oklepajih dodana še klasifikacijska točnost v tistem trenutku. Graf je omejen na 200 atributov, ker so bile meritve za 500 in 1000 atributov enake meritvam za 200.



Graf 19: Poskus D

V dveh primerih so funkcije zelo oscilirale, preden so se, nekje pri 50 atributih, začele približevati eni vrednosti. Zelo kmalu že, napoved za resničen razred skoči nad 50%. Pri 5 razredih, kolikor jih je bilo v vsaki delitvi prisotnih, je to zelo izrazito glasovanje za enega samega. Že po 60 atributih je vidno, da algoritem večino avtorjev prepozna z več kot 75%-napovedjo. Točnost se v tej točki umiri.

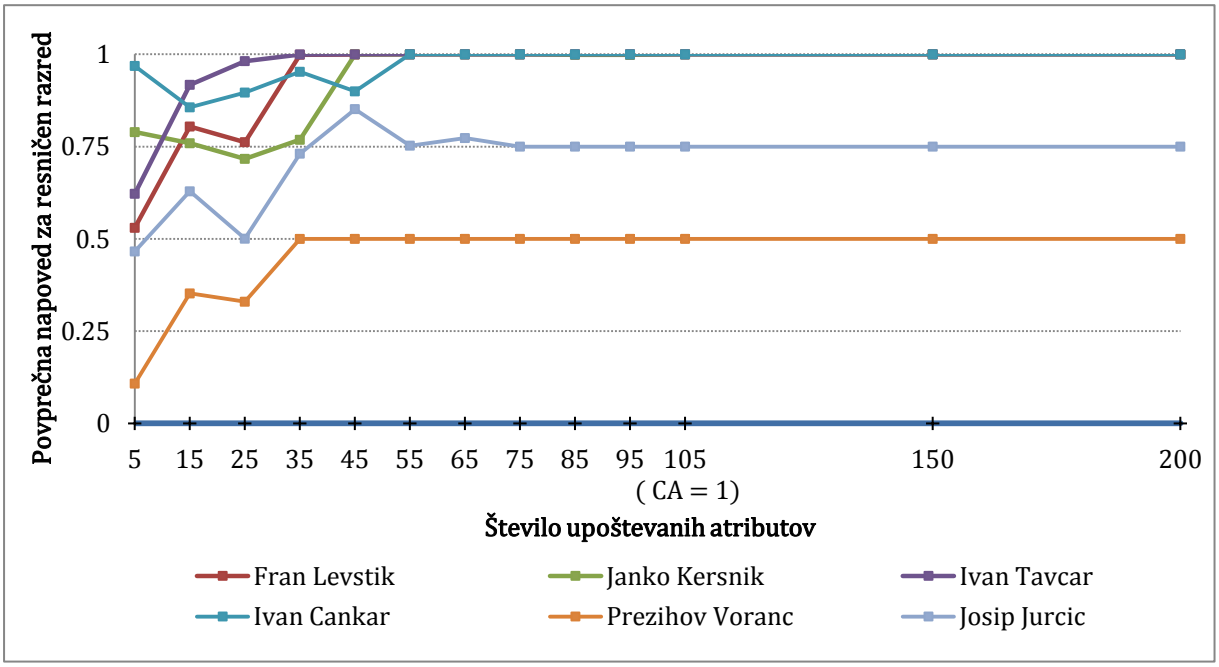
Zdi se nam, da lahko že z 100 najboljšimi atributi zagotovimo zelo dobro klasifikacijo, katere uspešnost z vključevanjem več atributov, po naših podatkih, ne narašča več.

4.3.4 Poskus E: manjšanje množice besed z ANOVA

Kot alternativo uporabi informacijskega prispevka, za katerega je potrebna diskretizacija, smo preizkusili še statistiko ANOVA, ki nima težav z zveznimi atributi. Ves ostali potek poskusa je ostal nespremenjen, vključno s podskupinami besedil, na katerih se je test izvajal.

Rezultati so bili manj uspešni. Pri ANOVI so se visoko uvrstile redkejše besede, vendar ne hapaxe niti v smislu skupnega števila pojavitev niti v smislu števila dokumentov, v katerih se pojavijo.

V primerjavi z informacijskim prispevkom, rabijo funkcije na podlagi ANOVE dlje časa, da nehajo oscilirati. Za najboljšo klasifikacijo smo prej rabili s 100 najboljših atributov, zdaj jih 150. Poleg tega je največja klasifikacijska točnost, ki jo dosežejo, skoraj v vseh primerih pod tisto iz prejšnjega poskusa.



Graf 20: Poskus E

ANOVA se je proti pričakovanjem obnesla zelo slabo. Problem informacijskega prispevka je, da nujno rabi diskretne vrednosti. V našem poskusu noben atribut ni diskreten, diskretizacija pa prinaša izgubo podatkov. Toda diskretizacija je tudi filter šuma. Lahko bi s tem razložili slabše rezultate. Verjetnejši razlog za drugačne rezultate pa je drugačen način nagrajevanja atributov pri obeh metodah. Kot je videti, ANOVA ni primerna za naš spekter problemov.

4.4 Končni preizkus

Po tem, ko smo testirali vse parametre, ki so se nam zdeli relevantni in smo se odločili klasifikacijski model graditi na Bayesu, ki mu informacijski prispevek izbere najboljše atribute, je bil čas za končen preizkus ugotovljenega.

Ker je besedil malo in ker je vsako posamezno zaključena in neodvisna entiteta, smo za preverjanje vzeli skrajno obliko $k-$ kratnega prečnega testiranje, metodo izpusti enega. Za vsako testno skupino oz. testni primer smo vektorje besed in ločil gradili posebej. Vsakič so vektorji obsegali le besede, ki so se pojavile v učni množici, in nikoli besed, ki so se le v testni.

Rezultati niso bili popolni. Dosegli smo 78% klasifikacijsko točnost. Med najboljšimi atributi se nikoli niso pojavila ločila, ki so po pomebmnosti vedno padla v sredino ali pod. Raziskali smo tudi napake, ki jih je naš model, naredil. Zanimalo nas je, če obstaja kakšna vzorec med njimi kot npr. če je do zamenjav prihajalo pri pisateljih, ki so živeli v istih obdobjih ali se zgledovali drug po drugem. Matrika zmot nam na žalost prikazuje drugačno sliko.

		Napovedani razred					
		Fran Levstik	*Janko Kersnik*	*Ivan Tavcar*	*Ivan Cankar*	*Prežihov Voranc*	*Josip Jurcic*
Resnični razred	*Fran Levstik*	*5*	-	-	-	-	*1*
	Janko Kersnik	-	*6*	-	-	-	*1*
	Ivan Tavcar	-	-	*6*	-	-	*1*
	Ivan Cankar	-	-	-	*6*	-	*1*
	Prežihov Voranc	-	-	-	-	*1*	*2*
	Josip Jurcic	-	*2*	-	-	-	*5*

Tabela 5: Matrika zmot

Vsakega pisatelja vsaj enkrat zamenja z Josipom Jurčičem. To dela Jurčiča s stališča našega modela tako splošnega, da mu je vsak stil v neki točki podoben. Zaskrbljujoče je, da Prežihovega Voranca, ki je predstavljen le s tremi knjigami, v dveh tretjinah primerov označi za Jurčiča. Tri knjige so očitno premalo, da bi rezultate jemali resno. Kljub vlogi Jurčiča kot rezervnega razreda, ko klasifikacija ni jasna, pa tudi njegova dela dvakrat zamenjajo lastnišvo. Iz tabele priklica in natančnosti je razvidno, da je razred, ki se ga najtežje klasificira ravno Josip Jurčič.

Zmotno klasificirane knjige so bile sledeče:

- Fran Levstik: Pokljuk
- Prežihov Voranc: Samorastniki

- Prežihov Voranc: Boj na požiralniku
- Ivan Cankar: Jernej in njegova pravica
- Janko Kersnik: Nova železnica
- Josip Jurčič: Črta iz življenja političnega agitatorja
- Josip Jurčič: Doktor Karbonaris
- Ivan Tavčar: Bolna ljubezen.

Zanimivo je, da imajo vse te knjige znotraj obravnavanih knjig vsakega avtorja najmanjše število različnih besed. Le omenjeni knjigi Jurčiča sta dve od treh najkrajših v zbirki. Pri tem je potrebno povdariti, da atribut "število različnih besed" nobenkrat ni bil upoštevan v klasifikaciji. Vendar v tem ne vidimo nobene nujne korelacije.

V primerjavi s podobnim testom iz poskusa C smo zdaj izvrstno napredovali na področju natančnosti. Za 4 razrede je imel klasifikator, ko jih je določil, vedno prav. Tudi na priklicu so vsi dvignili svoje vrednosti. Izjema je le Prežihov Voranc, ki se mu je priklic zmanjšal na polovico.

	Fran Levstik	*Janko Kersnik*	*Ivan Tavčar*	*Ivan Cankar*	*Prežihov Voranc*	*Josip Jurčič*
Priklic	*0,83*	*0,86*	*0,86*	*0,86*	*0,33*	*0,71*
Natančnost	*1*	*0,75*	*1*	*1*	*1*	*0,45*

Tabela 6: Priklic in natančnost za končni test

Zanimivo je, da se nam klasifikacijska točnost ni zares dvignila od časov, ko smo testirali le na ločilih. Z malo pod 75% je prišla na 78%, popravilo pa se je naše zaupanje v klasifikacijo, ki zdaj, razen za Jurčiča, vrača zanesljive rezultate.

4.4.1 IDF

Atribute, ki so glasovali v našem klasifikatorju, si oglejmo še s stališča IDFja. IDF, kratica za inverse document frequency oz. inverzna frekvenca v dokumentih, je v podatkovnem rudarjenju pogosto rabljena mera, ki govori o pogostosti neke besede v dokumentih. Predstavlja jo enačba:

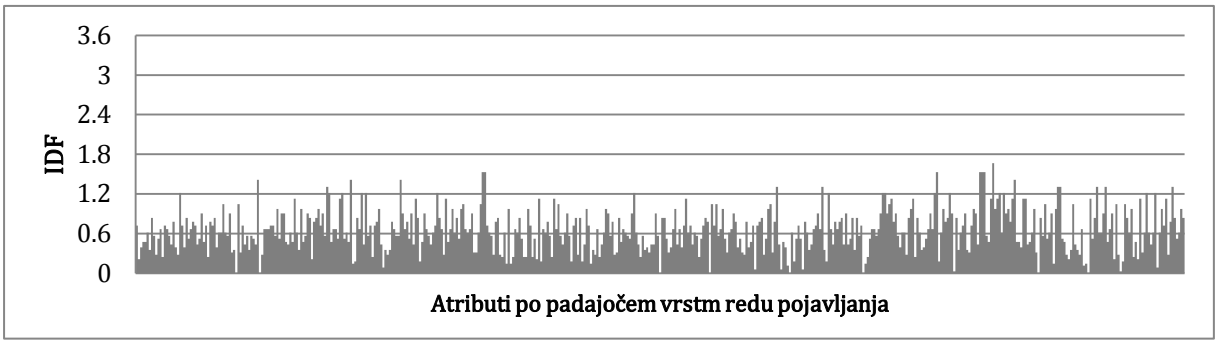
$$idf(a) = log\ \frac{N}{N_a}, \qquad (13)$$

kjer je a beseda, ki ji merimo IDF, N število vseh dokumentov (37 v našem primeru) in N_a število dokumentov, ki vsebujejo besedo a.

Uporablja se v zvezi z mero TF (term frequency), ki prešteje število pojavitev besede znotraj enega dokumenta. Skupaj se ti dve meri med sabo utežita in nagrajujeta le besede, ki so zelo specifične za majhen vzorec besedil, v katerih se tudi dostikrat pojavita.

IDF sama daje višje vrednosti besedam, ki se pojavijo v čim manj dokumentih.

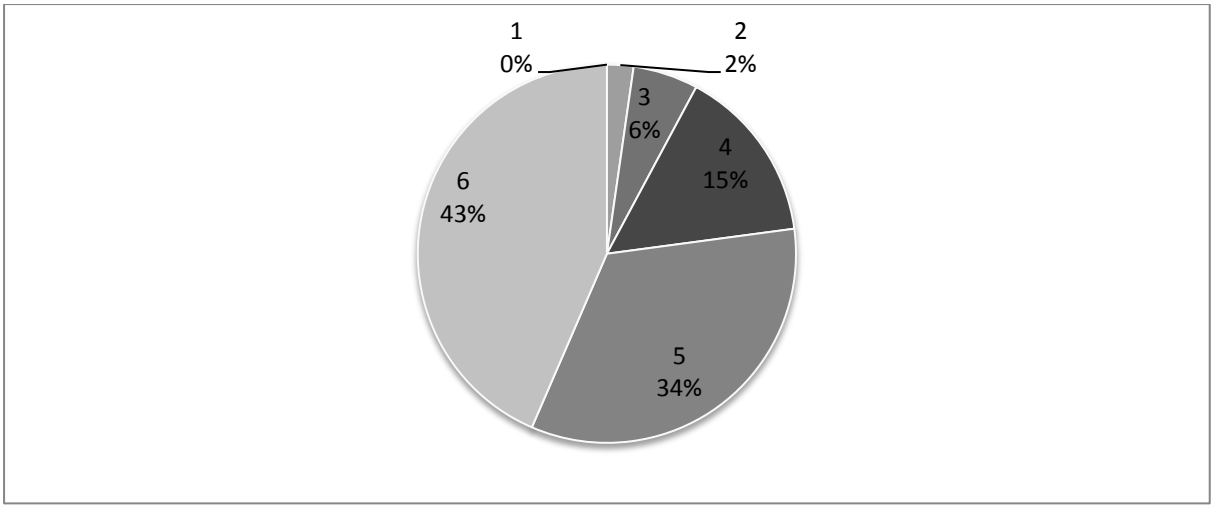
V končnem poskusu smo v vsakem krogu izbrali 200 najpomembnejših atributov, skupaj jih je bilo 485 različnih.



Graf 21: IDF po za atribute iz končnega poskusa za informacijski prispevek

Graf 21 prikazuje IDFje teh besed, ko so razporejene od tistih, ki so v največ krogih glasovale do tistih, ki so v najmanj. Zaloga vrednosti je zaradi $N = 37$ obsegala interval $[0, 3.6]$. Najvišja vrednost, ki je med izbranimi atributi nastopala, je bila le 1,66, kar pomeni, da se je vsaka beseda pojavila v vsaj petini besedil. Dobra polovica se jih je pojavilo v vsaj polovici, saj je mediana pri kar dvajsetih besedilih. Iz tega lahko sklepamo, da hapaxe res niso imele teže. Na besede, ki se pojavijo preredko, se enostavno ne moremo zanesti. Njihove informacije ne moremo preveriti, zato je tudi ne upoštevamo.

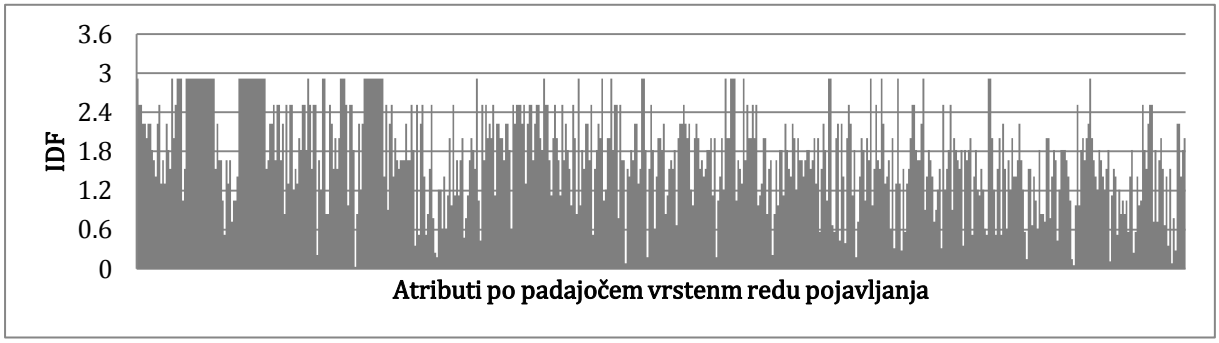
S stališča števila avtorjev, ki so uporabili izbrane besede, se izkaže, da se skoraj polovica besed pojavi pri vseh avtorjih in nadaljna tretjina pri petih avtorjih, kot kaže Graf 22. Informacijski prispevek išče besede, ki se povsod pojavljajo, vendar po različnih pravilih, da lahko po enem atributu določijo verjetnost za vse razrede.



Graf 22: Razdelitev atributov informacijskega prispevka glede na število avtorjev, ki so jih uporabili

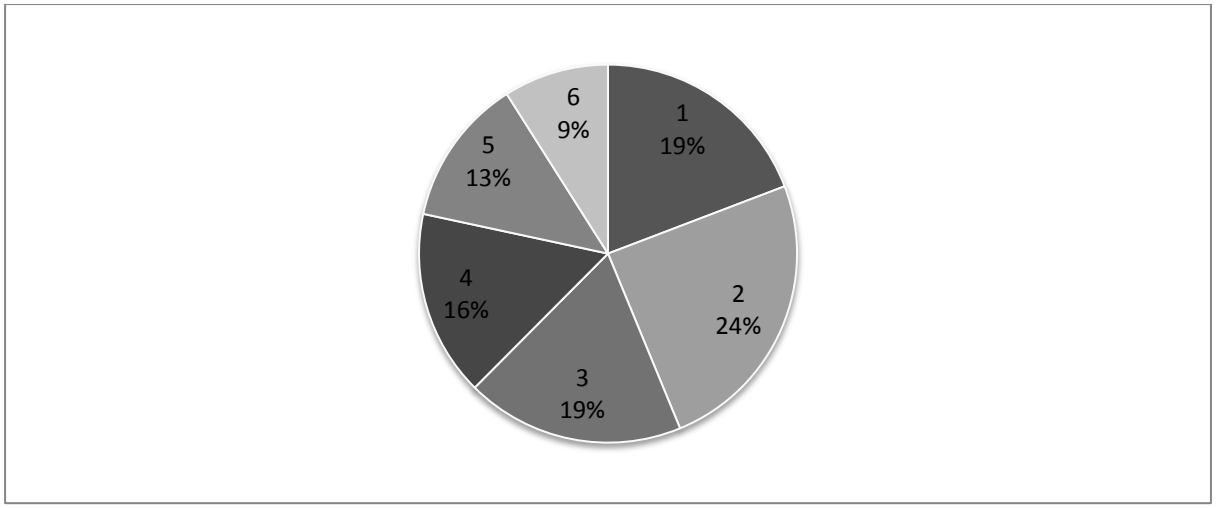
4.4.1.1 IDF z ANOVO

Da bi primerjali ANOVINO izbiro atributov z izbiro informacijskega prispevka še z vidika IDFja, smo isti končni test izvedli še z ANOVO. V skladu s prejšnjimi ugotovitvami so bili rezultati v istih pogojih slabši.



Graf 23: IDF po za atribute iz končnega poskusa za ANOVO

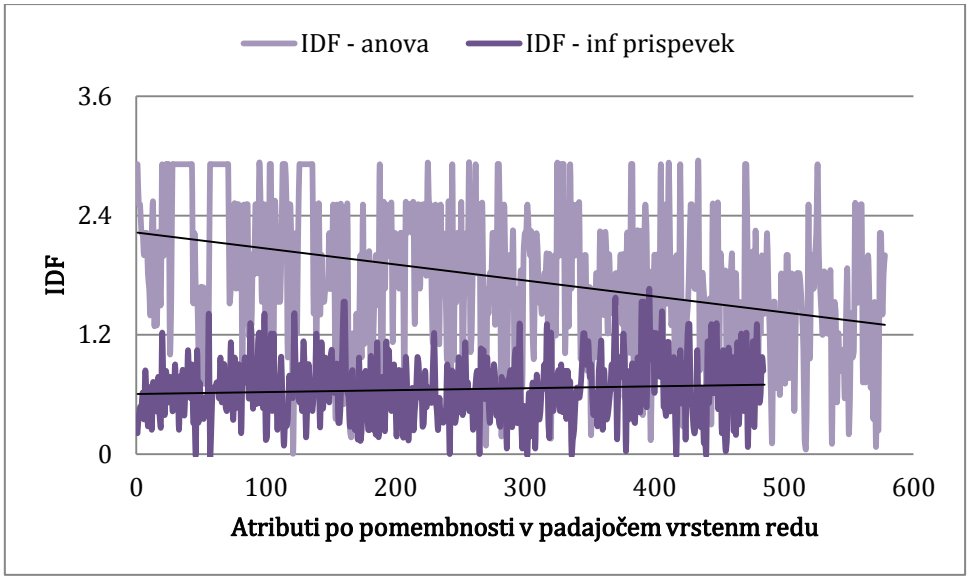
Seznam različnih besed je bil tokrat za petino daljši, 578 besed. Idf-ji pa so bili dosti višji. Mediana pade na šest besedil in minimum na dve besedili, medtem se aritmetično povprečje ustali pri sedmih besedilih. Nič več ni standard pri 20 besedilih, le dobrih 5% izbranih besed se pojavi v 20 besedilih ali več. Presenetjlivo pa je, da je tudi ANOVA, ki očitno preferira visoke idfje, zavrnila hapaxe. Nobena beseda ni bila izbrana, ki bi se pojavila v manj kot 2 besedilih, kar še enkrat potrdi, da je uporabnost hapax v klasifikaciji zanemarljiva.



Graf 24: Razdelitev atributov ANOVE glede na število avtorjev, ki so jih uporabili

Števila avtorjev, pri katerih so se besede, ki jih ANOVA preferira, pojavile, ustvarjajo povprečje nižje kot pri informacijskem prispevku. Slabih dvajset procentov se pojavi le pri enem avtorju, nadaljnih dobrih dvajset pri dveh in še nadaljnih dvajset pri treh. Le vsaka deseta beseda se pojavi pri vseh. ANOVA išče besede, ki bodo ločile avtorje med sabo, ne zato, ker bi se pojavljale v različnih deležih pri vsakem avtorju, marveč, ker se pojavijo le pri manjšem deležu pisateljev.

Graf 25 prikazuje direktno primerjavo v izbiri atributov med ANOVO in informacijskim prispevkom. ANOVA uporabi za klasifikacijo več različnih atributov, njihovi idfji so v začetku dosti višji, vendar tudi dosti hitro padejo. V začetku jemlje besede, ki se pojavijo v 2-3 besedilih, očitno takih uporabnih besed dokaj hitro zmanjka in posegati začne za besedami večjih pojavitev. Informacijski prispevek deluje ravno obratno, začenja z besedami, ki se povsod pojavljajo in nadaljuje z rahlo manj pogostimi. Medtem ko ANOVINA trendna črta pod ostrim kotom pada, se trendna premica informacijskega prispevka komaj zaznavno viša.



Graf 25: Primerjava idf vrednosti izbranih atributov po statistiki ANOVA in informacijskem prispevku

5 Zaključek

V diplomski nalogi smo se spoznali s stilometrijo, da bi njena dognanja preizkusili na manjši množici slovenskih leposlovnih besedil.

V eksperimentih smo preizkusili zelo preproste pristope in vendar dosegli dobre rezultate. S tem smo pokazali, da je stilometrija zmožna najti vzorce. Še vedno nanjo pada senca vseh študij, ki so analizirale različne splošno sprejete in hvaljene metode drugih raziskovalcev in pokazale, da pod novimi pogoji ne blestijo vedno. Zaskbljujoče je pomanjkanje neodvisnih testiranj, ki bi lahko kategorizirala že raziskano in omogočala hitrejši napredek. Ravno tako problematično je pomanjkanje skupne platforme, ki bi nudila pregled in lažjo izmenjavo že raziskanega, kar bi ravno tako omogočilo hitrejši napredek. Mogoče je dejstvo, da je število objavljenih metod tako enormno, razlog, da menim, da je avtomatizirano določevanje avtorstva v nekem deležu jalovo opravilo. Nekdo je nekoč primerjal to nalogo s klasifikatorjem rok in res je, ko imaš 6 rok jih je lahko ločiti med sabo, ko pa zraste to število na 600 000, se razlike med njimi toliko zmanjšajo, da moraš najverjetneje začeti iskati hapaxe, prav tako se zmanjša tudi smisel iskati razlike. Stilometrijo je zato treba jemati z rezervo, povdariti pa je tudi treba, da smo uporabljali res le preproste metode in dobili 78% klasifikacijsko točnost in v povprečju 87% natančnost, ki je bila za dve tretjini avtorjev celo 100%. Ti rezultati so dobri in kažejo na to, da je v vsakem aspektu besedila nek avtorjev podpis.

6 Seznam tabel

Bibliografija

[1] (2011, avg) Brier score. [Elektronski vir]. Dostopno na: http://en.wikipedia.org/wiki/Brier_score

[2] (2011, jun) Classification test. [Elektronski vir]. Dostopno na: http://en.wikipedia.org/wiki/Classification_test

[3] (2011, feb) LemmaGen. [Elektronski vir]. Dostopno na: http://lemmatise.ijs.si/

[4] (2010, jun) Stylometry. [Elektronski vir]. Dostopno na: http://en.wikipedia.org/wiki/Stylometry

[5] (2011, jul) Zipf's law. [Elektronski vir]. Dostopno na: http://en.wikipedia.org/wiki/Zipf's_law

[6] J. Binongo in M. Smith, "A bridge between statistics and literature: The graphs of Oscar Wilde's literary genres," *Journal of applied statistics*, št. 26, str. 781-787, 1999.

[7] J. F. Burrows, "Word-patterns and story-shapes: The statistical analysis of narrative style," *Literary and linguistic computing*, št. 2, str. 61-70, 1987.

[8] Joachim Diederich, Jörg Kindermann, Edda Leopold, in Gerhard Paass, "Authorship Attribution with Support Vector Machines," *Applied Intelligence*, str. 2003, 2000.

[9] Ronen Feldman in James Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press, 2007.

[10] W. Fucks, "On the mathematical analysis of style," *Biometrika*, št. 39, str. 122-129, 1952.

[11] Miran Hladnik. (2011, apr) Slovensko leposlovje na spletu. [Elektronski vir]. Dostopno na: http://slovenskaliteratura.ff.uni-lj.si/sl.html

[12] David I. Holmes. Stylometry: Its Origins, Development and Aspirations. [Elektronski vir]. Dostopno na: http://opim.wharton.upenn.edu/~sok/papers/r/s004.html

[13] D. I. Holmes, "The evolution of stylometry in humanities scholarship," *Literary and Linguistic Computing*, št. 13, str. 111-117, 1998.

[14] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," v *European conference on machine laerning*, 1998.

[15] Matthew L. Jockers in Daniela M. Witten, "A comparative study of machine learning methods for authorship attribution," *Lit Linguist Computing*, str. 215-223, apr 2010.

[16] Patrick Juola, "Authorship Attribution," *Foundations and Trends in Information Retrieval*, št. 1, 2006.

[17] Matjaž Juršič, "Implementacija učinkovitega sistema za gradnjo, uporabo in evaluacijo lemantizatorjev tipa RDR," Fakulteta za računalništvo in informatiko, Ljubljana, Diploma junij 2007.

[18] Andras Kornai, *Mathematical linguistic*. London: Springer, 2008.

[19] T. C. Mendenhall, "The characteristic curves of composition," *Science*, št. 9, str. 237-249, 1887.

[20] F. Mosteller in D. L. Wallace, *Inference and disputed authorship: The federalist*. Reading: Addison-Wesley, 1964.

[21] E. H. Simpson, "Measurement of diversity," *Nature*, št. 163, str. 688, 1949.

[22] H. H. Somers, "Statistical methods in literary analysis," v *The computer and literary style*, J. Leed, Ed. Kent, OH: Kent state university press, 1972.

[23] Inc StatSoft. (2011, avg) Electronic Statistics Textbook. [Elektronski vir]. Dostopno na: http://www.statsoft.com/textbook/

[24] D. R. Tallentire, "Towards an archive of lexical norms - a proposal," v *The computer and literary studies*. Cardiff, University of Wales Press, 1976.

[25] Thomas G. Tape. (2011, apr) Interpreting Diagnostic Tests. [Elektronski vir]. Dostopno na: http://darwin.unmc.edu/dxtests/

[26] Matt Tearle, Kye Taylor, in Howard Demuth, "An algorithm for automated authorship attribution using neural networks," *Lit Linguist Computing*, str. 425-442, okt 2008.

[27] L. Ule, "Recent progress in computer methods of authorship determination," *Association for literary and linguistic computing bulletin*, št. 10, str. 73-89, 1982.

[28] G.U. Yule, "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship," *Biometrika*, št. 30, jan 1939.

[29] G. U. Yule, "The statistical study of literary vocabulary," 1944.

[30] G. K. Zipf, *Human bahaviour and the principle of least effort. An introduction to human ecology*. Boston: Houghton-Mifflin, 1932.

[31] G. K. Zipf, "Observations on the possible effects of mental age upon the frequency-distribution of words from the viewpoint of dynamic philology," *Journal of psychology*, št. 4, 1937.