Inteligentni sistemi, 2. seminarska naloga

Matic Bernik in Robert Tovornik

17. januar 2017

1 Uvod

Glavna tema druge seminarske naloge je tekstovno rudarjenje ter tekstovna klasifikacija. Naloga je bila sestavljena iz več delov. V prvem delu sva izbrala klasifikacijski kriterji, torej lastnost, ki jo bova iz podatkov napovedovala. Izbrala sva napovedovanje avtorja teksta ter spol avtorja. Nato je sledilo dejansko rudarjenje (zbiranje knjig) ter izgradnja korpusa. Sledillo je procesiranje ter obdelava zbranih dokumentov ter tvorjenje znacilk (atributov). Za konec je preostala še klasifikacija dokumentov. Uporabila sva več metod: SVM, naključne gozdove, naivnega Bayesa ter večinski klasfikator.

2 Podatki

Podatki za nalogo, ki so hrajeni v korpusu, so tekstovne knjige zbrane s spletne strani projekta Gutenberg. Vsa dela so bila zbrana v skladu s pravili (uporaba knjižnice R - gutenbergr), preko zrcala. V korpusu so zbrana dela 26 različnih avtorjev, v angleškem jeziku. Hranjena so v direktoriju "books". Ta se nato deli na poddirektorije z imeni avtorjev, ter slednji na dva nova, "txt"ki vsebuje knjige v formatu .txt, ter "header", ki vsebuje datoteke z metapodatki o soimenski knjigi, dostopni s strani gutenberg. V nadalji obdelavi, sva dela razčlenila na "članke", po principu delitve vsakega dela na 20 člankov, kjer vsak vsebuje minimalno 500 besed. Posledično, so nekatera dela, ki zaradi pomanjkanja števila besed niso ustrezala pogojem, izpadla. Na koncu nama je ostalo cca. 7200 člankov.

Glavni problem, na katerega je pri deljenju na članke potrebno biti pozoren, je dejstvo da ima vsaka knjiga na začetku zapisane meta podatke in informacije o samem delu, morda tudi avtorju. Problem sva rešila tako, da pri vsakem delu preskočiva nekaj začetnih odstavkov.

3 Predprocesiranje podatkov

Da poenostavimo procesiranje, ustvarjanje atributov, zmanjšamo število kombinacij, ter da lažje iščemo povezave med besedili, je tekstovne datoteke najprej potrebno predobdelati. Tu pridejo na vrsto klasični postopki, kot so transformacija teksta v male črke, korenizacija, razbijanje na tokene, besede,.. Pri navedenih postopkih sva si pomagala z python knjižnjico nltk.

4 Izbor ter obdelava klasifikacijskih atributov

Ker sva se odločila za napovedovanje avtorjev, teksta, sva morala poiskati atribute, ki nekako definirajo in hkrati ločijo, torej dobro diskiminirajo med lastnostmi posamezniih avtorjev. Zato sva kot osnovo izbrala atribute frekvenc pojavitve nabora "posebnih znakov", to so vejice, pike, narekovaji, dvopicja,.. Kasneje se je izakalo kot zelo učinkovito. Nato sva sestavljala bolj zapletene atribute, ki so zajemala neko povezavo - razmerje znotraj besedila. Npr.: razmerje med številom stavkov ter besed v besedilu, povprečno dolžino stavka,.. Za konec sva sestavila še nabor najpogosteje uporabljenih besed, ozrioma število pojavitev le-teh. Tudi ti atributi so se izkazali kot uspešni.

Slika 1: Skaliran nabor najpogostejših besed v korpusu.

Hitro sva ugotovila, da ker so izbrane knjige različnih velikosti, deliva pa vse po enakem postopku prihaja do velikih razlik pri izračunih. Zato je bilo potrebno vse frekvence tudi normalizirati. Nekatere z dolžino besedila, druge s številom vseh besed, stavkov,..

To je pripeljalo do dodatnih izboljšav v napovedovanju. Približno 4 odstotke.

5 Klasifikacija ter klasifikacijska točnost

Pred pričetkom učenja klasifikacijskih modelov, je bilo potrebno zgraditi dodatno datoteko "dataset.tab", kjer vsaka vrstica predstavlja posamezen članek (delo) ter iz podatkov izpeljane pripadajoče vrednosti značilk.

Pred začetkom klasifikacije, sva uporabila oceno nad atributi (information gain, relieff), da sva potrdila ustreznost izbranih attributov.

	#	Inf. gain	Gain Ratio	Gini	ReliefF
% ','	C	0.447	0.223	0.041	0.048
%"""	C	0.430	0.236	0.036	0.041
avg. sentence length	C	0.401	0.200	0.036	0.022
sentences-to-words ratio	C	0.401	0.200	0.036	0.044
% ':'	C	0.376	0.197	0.060	0.030
%"\n"	C	0.349	0.174	0.042	0.041
% ','	C	0.334	0.167	0.029	0.044
%"wa"	C	0.321	0.161	0.038	0.053
%"d"	C	0.304	0.247	0.048	0.014
%" "	C	0.304	0.152	0.029	0.037
%"mr"	C	0.282	0.205	0.023	0.007
%"["	C	0.280	0.296	0.042	0.014
% ';'	C	0.278	0.139	0.029	0.032
%"is"	C	0.275	0.138	0.024	0.036
%"the"	C	0.268	0.134	0.033	0.048
%"her"	C	0.252	0.131	0.018	0.027
%"had"	C	0.251	0.125	0.026	0.029
% '?'	C	0.241	0.121	0.022	0.020
%"she"	C	0.235	0.128	0.017	0.017
%"""	C	0.234	0.117	0.020	0.039
% '!'	C	0.232	0.120	0.017	0.016
%"my"	C	0.227	0.115	0.026	0.016
%"of"	C	0.223	0.112	0.020	0.040
%"and"	C	0.222	0.111	0.020	0.021
%"t"	C	0.216	0.128	0.015	0.021
%"-"	C	0.214	0.107	0.018	0.009
%"which"	C	0.209	0.105	0.016	0.030
%"said"	C	0.196	0.106	0.022	0.017
%"your"	C	0.193	0.110	0.026	0.012
%"you"	C	0.184	0.092	0.016	0.020
%"be"	C	0.172	0.086	0.015	0.008
%"it"	C	0.170	0.085	0.015	0.012
%"i"	C	0.168	0.084	0.013	0.018
%"will"	C	0.166	0.089	0.020	0.005
%"are"	C	0.160	0.082	0.014	0.004
%"littl"	C	0.156	0.087	0.017	0.024
%"s"	C	0.154	0.077	0.011	0.012
%"me"	C	0.150	0.076	0.015	0.009

(a) Za avtorja

	#	Inf. gain	Gain Ratio	Gini	ReliefF
%"""	C	0.150	0.083	0.079	0.043
%"she"	C	0.136	0.074	0.078	0.016
%"her"	C	0.134	0.070	0.077	0.039
%"mr"	C	0.051	0.037	0.030	0.011
%"wa"	C	0.046	0.023	0.023	0.023
%"had"	C	0.046	0.023	0.024	0.011
%"is"	C	0.038	0.019	0.020	0.025
% ','	C	0.038	0.019	0.019	0.026
%"littl"	C	0.034	0.019	0.019	0.017
avg. sentence length	C	0.032	0.016	0.016	0.015
%"\n"	C	0.031	0.016	0.016	0.021
%"said"	C	0.031	0.017	0.017	0.004
%"thi"	C	0.031	0.015	0.016	0.007
sentences-to-words ratio	C	0.029	0.015	0.015	0.021
%"["	C	0.027	0.028	0.012	0.006
%"d"	C	0.025	0.020	0.011	0.013
%""	C	0.023	0.042	0.009	0.007
%"("	C	0.021	0.016	0.011	-0.002
%"as"	C	0.021	0.011	0.012	0.015
% '('	C	0.021	0.015	0.011	-0.003
% '!'	C	0.019	0.010	0.010	0.003
%"my"	C	0.018	0.009	0.010	0.003
% ';'	C	0.018	0.010	0.009	0.006
%"t"	C	0.018	0.010	0.010	0.009
%" "	C	0.017	0.009	0.010	0.014
%"s"	C	0.017	0.008	0.009	0.011
% '?'	C	0.017	0.008	0.009	0.009
% ','	C	0.016	0.008	0.009	0.017
%"your"	C	0.016	0.009	0.008	0.006
%"to"	C	0.016	0.008	0.009	0.021
%"look"	C	0.016	0.008	0.009	0.003
%"me"	C	0.016	0.008	0.008	-0.002
%"then"	C	0.014	0.007	0.007	0.012
%"like"	C	0.014	0.007	0.008	0.005
%"are"	C	0.014	0.007	0.007	-0.000
%"a"	C	0.013	0.007	0.007	0.000
%"on"	C	0.013	0.006	0.007	0.009
%"could"	C	0.012	0.007	0.007	0.005

(b) Za spol avtorja

Slika 2: Primernost klasifikacijskih atributov

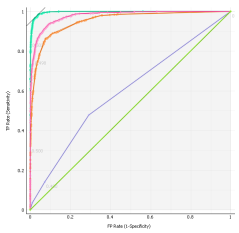
Tukaj lahko opazimo zanimivost pri napovedovanju spola avtorja, kjer je nenadoma prišlo do večjega preskoka v informacijskem pridobitku. Velik del namreč pridobijo besede, ki definirajo spol ('her', 'she'),.

Za učne modele sva izbrala več klasifikatorjev: večinski klasifikator, naivni Bayesov, k najbližjih sosedov, SVM ter naključne gozdove. Kot predvideno po teoriji, se je tudi tukaj izkazalo, da so najprimernješi prav, NB, SVM, ter RF.

Rezultati so sledeči.

Method	AUC	CA	F1	Precision	Recall
SVM	0.914	0.843	0.835	0.841	0.843
Naive Bayes	0.843	0.704	0.704	0.727	0.704
Random Forest	0.818	0.673	0.652	0.662	0.673
kNN	0.564	0.197	0.182	0.185	0.197
Majority	0.500	0.132	0.031	0.017	0.132

Tabela 1: Klasifikacijska točnost pri napovedovanju avtorja

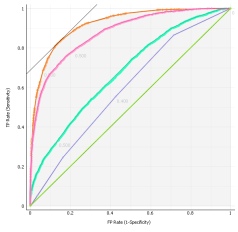


Slika 3: ROC krivulja

Pri napovedovanju spola, pa se je točnost napovedih metod močno spremenila. Izstopali so naključni gozdovi, kot izredno dober klasifikator, medtem ko je prej najboljši SVM močno padel.

Method	AUC	CA	F1	Precision	Recall
Random Forest	0.810	0.866	0.909	0.862	0.866
Naive Bayes	0.783	0.783	0.838	0.813	0.783
Majority	0.500	0.720	0.837	0.518	0.720
kNN	0.541	0.671	0.785	0.636	0.671
SVM	0.614	0.588	0.659	0.689	0.588

Tabela 2: Klasifikacijska točnost pri napovedovanju spola avtorja



Slika 4: ROC krivulja

Pa še nekaj zanimivih vizualizacij. Confusion matrix, scatter plots.

Actual \ Predicted	Alcott	Austen	Carroll	Cather	Chopin	Christie	Defoe	Dickens	Dostoyevsky	Einstein	Eliot	Flaubert	Homer	Ibsen	Joyce	London	Plato	Rand	Shakespeare	Shelley	Twain	Tyler	West	Wharton	Wilde	Woolf	Σ
Alcott	512	1	0	0	0	0	5	13	4	0	4	1	4	0	0	4	0	0	1	0	5	0	0	4	1	1	560
Austen	5	169	1	1	0	0	1	7	2	0	2	0	1	0	0	1	0	0	0	1	3	0	0	5	1	0	200
Carroll	3	0	108	1	0	2	1	11	1	0	4	2	0	0	0	1	0	0	0	0	3	0	0	1	2	0	140
Cather	6	1	0	113	0	0	0	1	2	0	0	0	0	0	0	36	0	0	0	0	6	0	0	11	3	1	180
Chopin	1	1	0	7	11	0	0	2	1	0	3	1	0	0	0	8	0	0	0	0	1	0	0	24	0	0	60
Christie	0	0	0	1	0	14	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	1	0	20
Defoe	6	0	0	0	0	0	396	14	1	0	1	1	1	0	0	7	4	0	5	0	2	0	0	0	2	0	440
Dickens	25	7	3	0	0	0	22	641	14	0	11	6	2	0	0	30	1	0	0	1	24	0	0	4	9	0	800
Dostoyevsky	9	5	0	6	0	0	2	30	131	0	4	2	0	0	0	18	0	0	0	0	8	0	0	3	1	1	220
Einstein	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	20
Eliot	17	8	0	1	1	0	3	13	2	0	128	0	1	0	0	9	5	0	0	1	12	0	0	17	2	0	220
Flaubert	1	1	1	2	0	0	1	19	3	0	0	96	1	0	0	5	0	0	0	0	8	0	0	1	1	0	140
Homer	8	0	1	2	0	0	6	1	3	0	1	2	115	0	0	8	3	0	4	0	3	0	0	0	3	0	160
Ibsen	0	0	0	0	0	0	0	0	0	0	0	0	0	305	0	2	0	0	11	1	1	0	0	0	0	0	320
Joyce	0	0	0	1	1	1	0	1	0	0	0	0	0	1	38	11	0	0	1	0	2	0	0	3	0	0	60
London	7	0	6	15	1	2	1	26	8	0	6	2	0	6	0	714	2	0	0	2	56	0	1	11	14	0	880
Plato	0	0	2	0	0	0	5	4	1	0	0	0	0	0	0	0	347	0	0	0	0	0	0	0	1	0	360
Rand	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	20
Shakespeare	2	0	0	0	0	0	4	3	0	0	1	0	3	1	0	0	3	0	899	0	0	0	0	0	4	0	920
Shelley	2	3	0	0	0	0	2	6	2	0	1	1	2	0	0	2	2	0	2	53	1	0	0	1	0	0	80
Twain	12	0	9	3	0	0	5	32	9	0	16	3	2	6	0	29	1	0	3	0	313	0	0	11	6	0	460
Tyler	2	0	0	4	0	0	0	1	0	0	1	0	0	0	0	11	0	0	0	0	1	0	0	0	0	0	20
West	1	0	0	4	0	0	1	1	2	0	6	0	0	0	0	2	0	0	0	0	5	0	18	19	1	0	60
Wharton	6	2	0	4	2	0	1	12	4	0	9	3	0	1	1	9	0	0	0	2	9	0	0	353	1	1	420
Wilde	2	0	0	5	0	0	6	14	7	1	2	2	2	8	0	17	3	0	3	0	4	0	0	3	241	0	320
Woolf	12	0	0	0	2	0	0	12	0	0	3	0	0	0	0	8	0	0	0	0	3	0	1	13	0	6	60
Σ	639	198	132	170	18	19	463	864	197	20	203	122	134	328	39	935	371	18	929	61	471	0	20	484	295	10	7140

Slika 5: Confusion matrix za avtorja

Actual \ Predicted	F	M	Σ
F	1389	651	2040
M	343	4757	5100
Σ	1732	5408	7140

Slika 6: Confusion matrix za spol

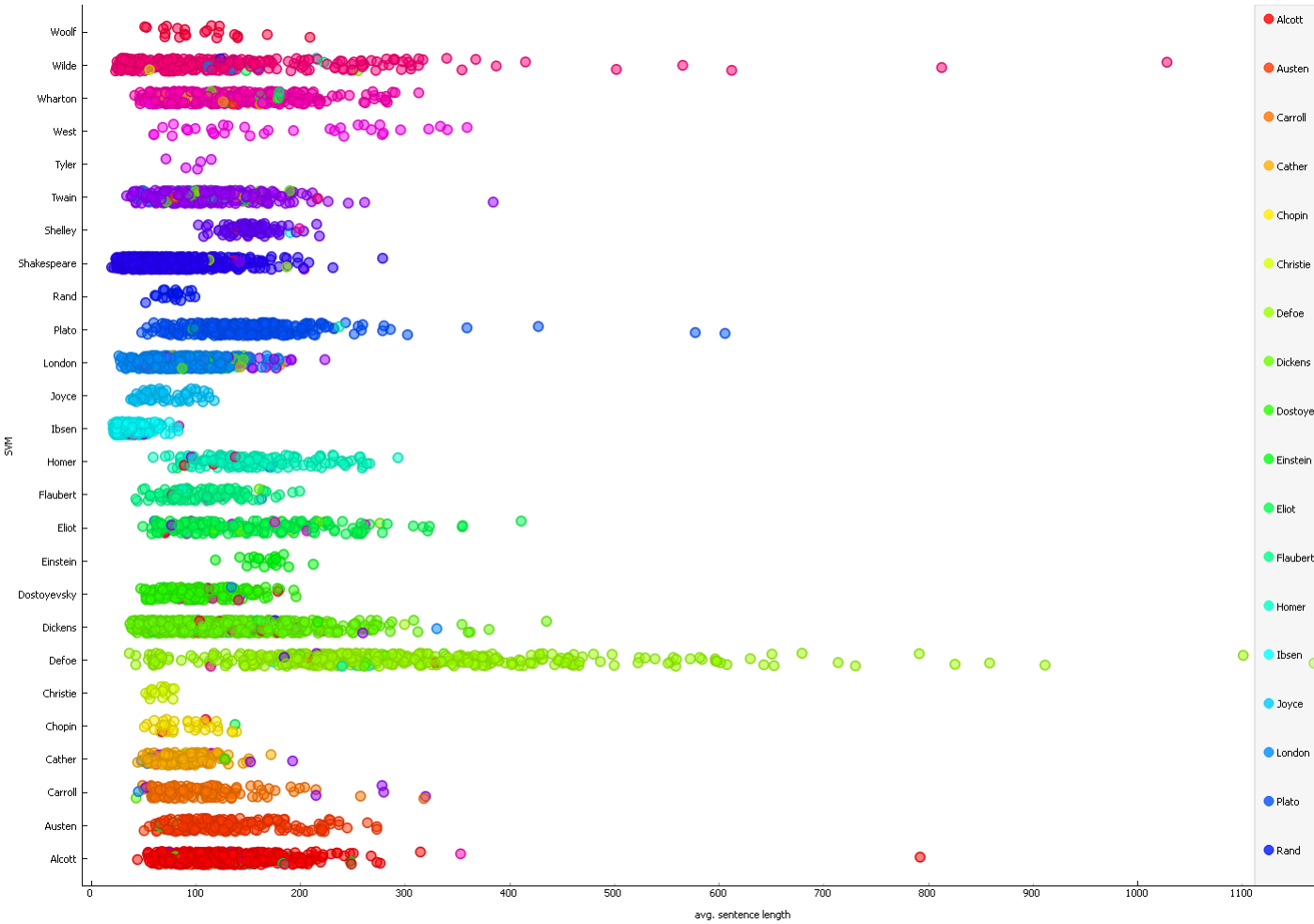
V konfuzni matriki napovedi avtorja, lahko kot zanimivost opazimo povezavo med največ

zgrešenimi napovedmi prav med dvema avtorjema, Mark Twain ter Jack London. Če se malce poglobimo, pridemo do podatka, da sta oba avtorja ustvarjala v istem obdobju (razlika 10ih let), oba sta pisala pustolovske romane, oba sta pisala v 2. osebi,.. Torej lahko rečemo, da obstaja velika podobnost v slogu pisanja.

6 Zanimivi atributi

Med prej naštetimi se je nekaj atributov izkazalo kot takih, ki izjemno dobro določajo specifike avtorjev. Dva izmed njih sva se odločila izpostaviti.

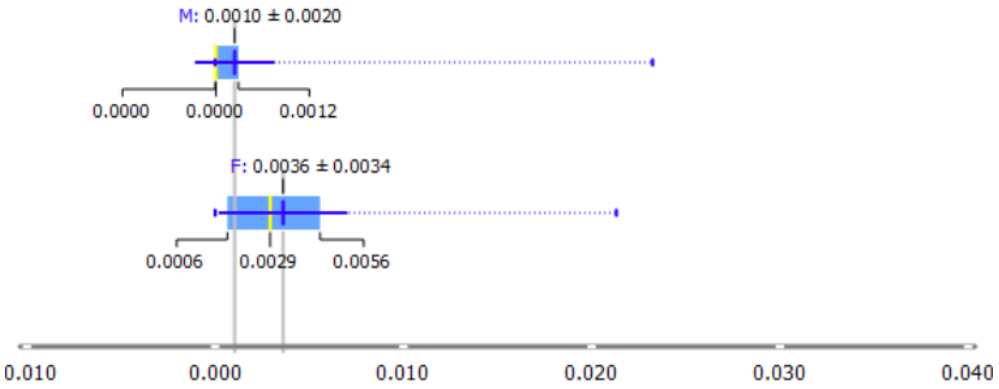
Ker si množice številk težko predstavljamo, sva se odločila da to prikaževa grafično. Tako spodnja slika zelo dobro prikazuje razporejanje del k avtorjem, na podlagi ene izmed slovognih specifik (povprečne dolžine stavka). Zelo dobro je razvidno, da večina barv spada kamor je treba, izstopa jih le nekaj malo.



Slika 7: Graf razpršenosti - povprečna dolžina stavka - SVM

Drug atribut, ki izjemo izstopa, tokrat pri klasifikaciji spola avtorja, pa je že prej ome-

njen atribut 'she', torej atribut ki v tekstu opredeljuje spol. Če pogledamo graf škatle z brki, lahko vidimo da izjemno natančno opredeljuje vrednosti, saj se večina vrednosti nahaja znotraj izjemno majhnega intervala.



Slika 8: Škatla z brki - beseda she

7 Zaključek

Za konec, bi lahko na vse skupaj pogledali še s prakičnega vidika. Namreč če pogledamo atribute, ki zelo dobro določajo avtorje, lahko vidimo nek smisel, ki ga prikazujejo. Na primer.: če opazujemo število uporabljenih ločil, navednic, in pa dolžino povprečnih povedi, doližno odstavkov, vidimo da so to pravzaprav lastnosti, ki določajo pisalni slog avtorja. Po drugi strani če pogledamo najpogostje uporabljene besede v besedilih avtorja, vidimo da to pravzaprav določa razgledanost avtorja in njegov besedni zaklad.

7.1 Stacking

Kot lahko opazimo, znamo spol napovedovati zelo dobro, bolje kot avtorja samega, saj je nabor dosti večji kot pri spolu, kjer sta (vsaj v večini primerov) zgolj dve opciji. Zato lahko pridemo na idejo, da bi napovedni metodi zaporedno združili v upanju izboljšave. Torej najprej napovemo spol, ter ga kot atribut podamo v klasifikator avtorja. Če to naredimo pride do pozitivne spremembe v klasifikacijski točnosti. Nova napovedna točnost avtorja s predhodno najboljšim klasifikatorjem SVM je 87.8 odstona z podporo AUC 0.933. Pripeljali smo torej do izbolšanja za dobre 3 odstotke. Tehnika torej kaže potencial.