Inteligentni sistemi, 2. seminarska naloga

Matic Bernik in Robert Tovornik

17. januar 2017

1 Uvod

Namen seminarske naloge je, delovanje na področju tekstovnega rudarjenja. Naloga je sestavljena iz več delov. Najprej sva izbrala klasifikacijski kriterji, nato je sledilo rudarjenje ter izgradnja korpusa, zatem procesiranje zbranih dokumentov, obdelava dokumentov ter na koncu ob zgeneriranih znacilkah, sama klasifikacija dokumentov. V podanem primeru klasifikacija avtorja ter spola.

2 Podatki

Podatki za nalogo, ki so hrajeni v korpusu, so tekstovne knjige zbrane s spletne strani projekta Gutenberg. Seveda v skladu z navodili (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 premalo 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, zato sva pri vsakem preskocila prvih nekaj odstavkov.

3 Predprocesiranje podatkov

Da poenostavimo procesiranje, ustvarjanje atributov, zmanjšamo število kombinacij, ter da lažje iščemo povezave med besedili, je tekstovne datoteke najprej potrebno obdelati. 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 stevilom stavkov ter besed v besedilu, povprecno dolzina stavka, frekvenca pojavitev globalno najpogostejsih besed v posameznem besedilu,.. Tudi ti atributi so se izkazali kot uspesni. Dodatno sva podala se število pojavitev najpogostejših besed.

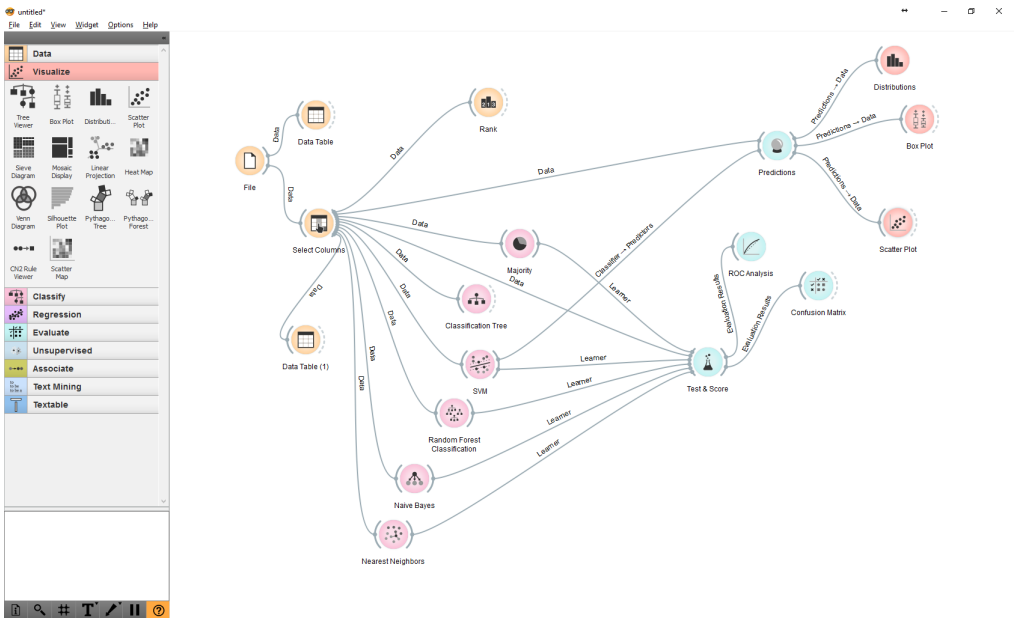
Hitro sva ugotovila, da ker so izbrane knjige različnih velikosti, deliva pa vse po enakem postopku prihaja do velikih razlik pri izračunih. Zato je bilo potrebno vse frekvence tudi normalizirati. Ali z dolžino besedila, ali s številom besed, stavkov,..

To je obrodilo dodatne izboljšave.

5 Klasifikacija ter klasifikacijska točnost

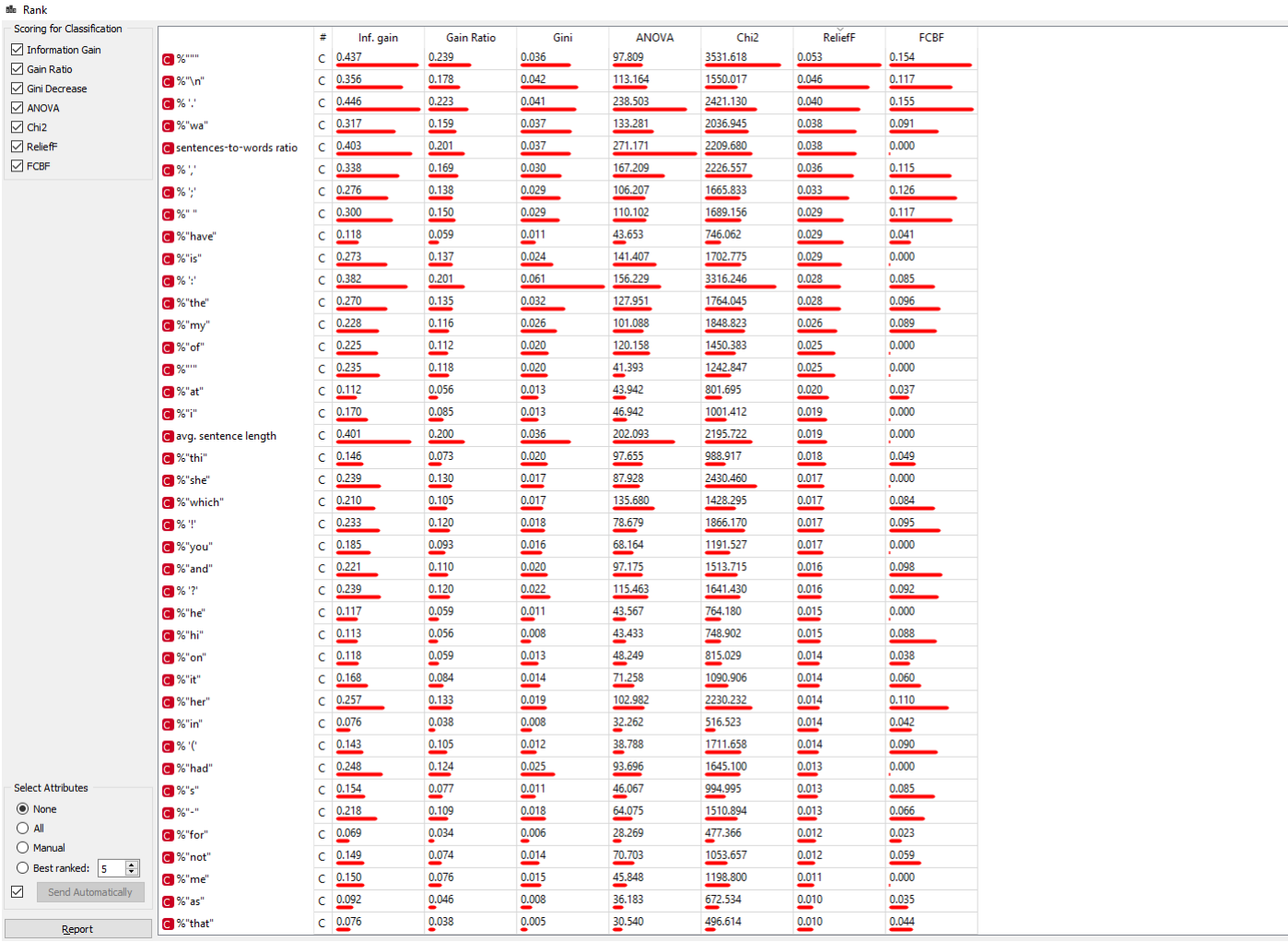
Za klasifikacijo, sva na podlagi v prejšnjem poglavjo navedenih atributov, zgradila datoteko "dataset.tab", kjer vsaka vrstica predstavlja posamezen članek ter pripadajoče atribute.

Datoteko dataset, nato uvoziva v ogrodje orange, s katerim sva računala klasifikacijo.



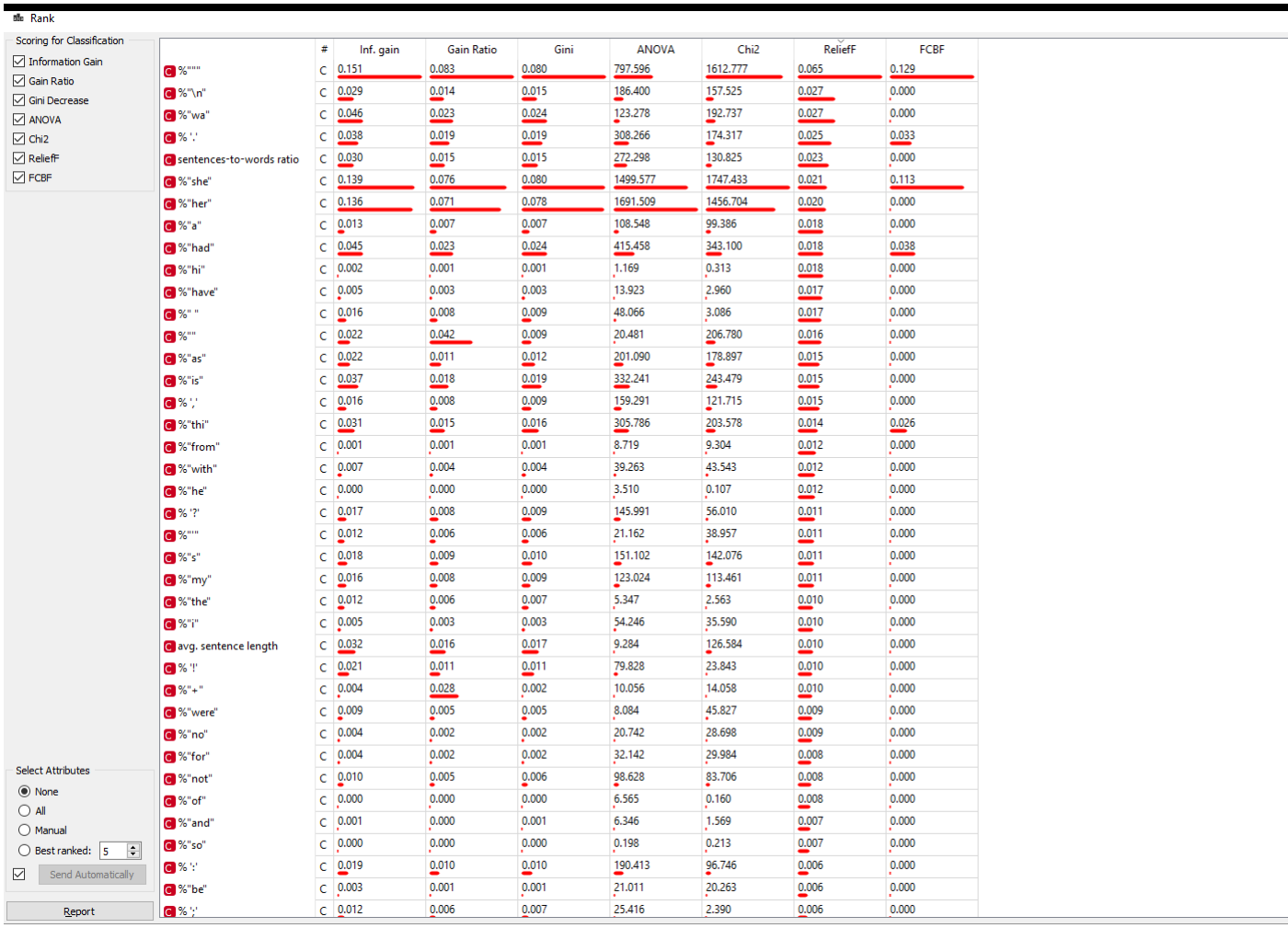
Slika 1: Ogrodje orange, klasifikacijska mreža.

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



	#	Inf. gain	Gain Ratio	Gini	ANOVA	Chi2	ReliefF	FCBF
%"""	C	0.437	0.239	0.036	97.809	3531.618	0.053	0.154
%"\n"	C	0.356	0.178	0.042	113.164	1550.017	0.046	0.117
% '.'	C	0.446	0.223	0.041	238.503	2421.130	0.040	0.155
%"wa"	C	0.317	0.159	0.037	133.281	2036.945	0.038	0.091
sentences-to-words ratio	C	0.403	0.201	0.037	271.171	2209.680	0.038	0.000
% ','	C	0.338	0.169	0.030	167.209	2226.557	0.036	0.115
% ';'	C	0.276	0.138	0.029	106.207	1665.833	0.033	0.126
%" "	C	0.300	0.150	0.029	110.102	1689.156	0.029	0.117
%"have"	C	0.118	0.059	0.011	43.653	746.062	0.029	0.041
%"is"	C	0.273	0.137	0.024	141.407	1702.775	0.029	0.000
% ':'	C	0.382	0.201	0.061	156.229	3316.246	0.028	0.085
%"the"	C	0.270	0.135	0.032	127.951	1764.045	0.028	0.096
%"my"	C	0.228	0.116	0.026	101.088	1848.823	0.026	0.089
%"of"	C	0.225	0.112	0.020	120.158	1450.383	0.025	0.000
%"""	C	0.235	0.118	0.020	41.393	1242.847	0.025	0.000
%"at"	C	0.112	0.056	0.013	43.942	801.695	0.020	0.037
%"i"	C	0.170	0.085	0.013	46.942	1001.412	0.019	0.000
avg. sentence length	C	0.401	0.200	0.036	202.093	2195.722	0.019	0.000
%"thi"	C	0.146	0.073	0.020	97.655	988.917	0.018	0.049
%"she"	C	0.239	0.130	0.017	87.928	2430.460	0.017	0.000
%"which"	C	0.210	0.105	0.017	135.680	1428.295	0.017	0.084
% '!'	C	0.233	0.120	0.018	78.679	1866.170	0.017	0.095
%"you"	C	0.185	0.093	0.016	68.164	1191.527	0.017	0.000
%"and"	C	0.221	0.110	0.020	97.175	1513.715	0.016	0.098
% '?'	C	0.239	0.120	0.022	115.463	1641.430	0.016	0.092
%"he"	C	0.117	0.059	0.011	43.567	764.180	0.015	0.000
%"hi"	C	0.113	0.056	0.008	43.433	748.902	0.015	0.088
%"on"	C	0.118	0.059	0.013	48.249	815.029	0.014	0.038
%"it"	C	0.168	0.084	0.014	71.258	1090.906	0.014	0.060
%"her"	C	0.257	0.133	0.019	102.982	2230.232	0.014	0.110
%"in"	C	0.076	0.038	0.008	32.262	516.523	0.014	0.042
% '('	C	0.143	0.105	0.012	38.788	1711.658	0.014	0.090
%"had"	C	0.248	0.124	0.025	93.696	1645.100	0.013	0.000
%"s"	C	0.154	0.077	0.011	46.067	994.995	0.013	0.085
%"-"	C	0.218	0.109	0.018	64.075	1510.894	0.013	0.066
%"for"	C	0.069	0.034	0.006	28.269	477.366	0.012	0.023
%"not"	C	0.149	0.074	0.014	70.703	1053.657	0.012	0.059
%"me"	C	0.150	0.076	0.015	45.848	1198.800	0.011	0.000
%"as"	C	0.092	0.046	0.008	36.183	672.534	0.010	0.035
%"that"	C	0.076	0.038	0.005	30.540	496.614	0.010	0.044

Slika 2: Za avtorja

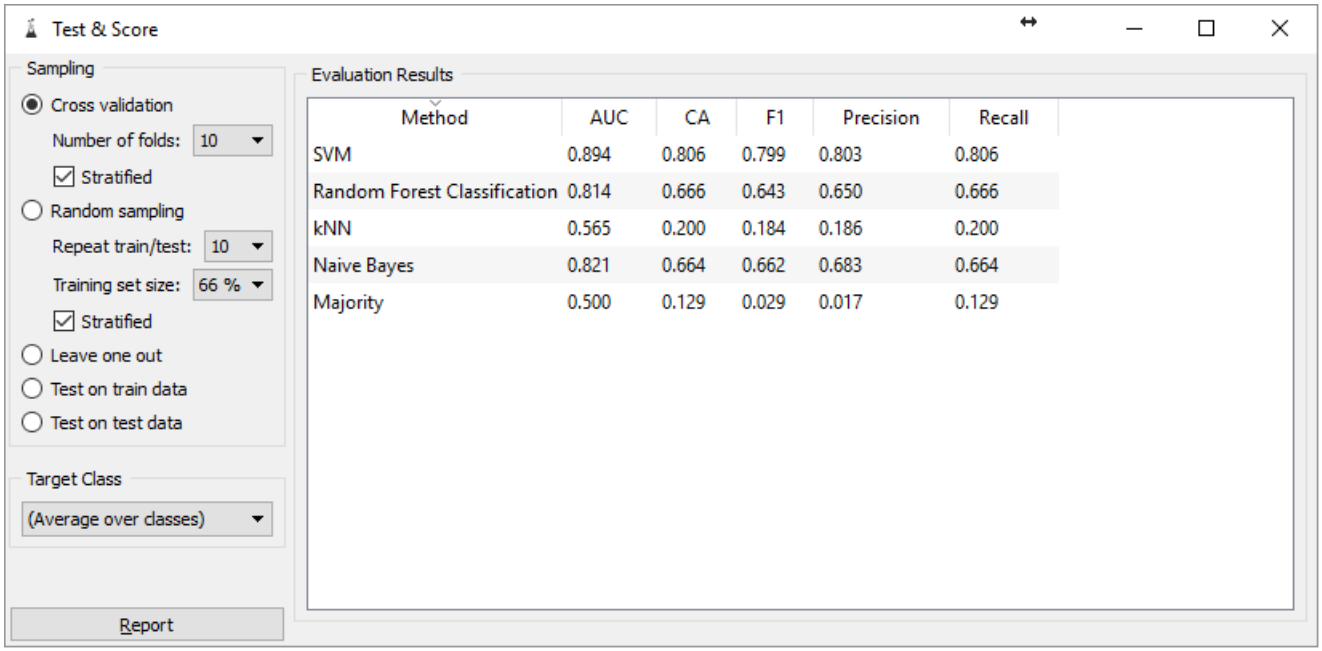


	#	Inf. gain	Gain Ratio	Gini	ANOVA	Chi2	ReliefF	FCBF
%"""	C	0.151	0.083	0.080	797.596	1612.777	0.065	0.129
%"\n"	C	0.029	0.014	0.015	186.400	157.525	0.027	0.000
%"wa"	C	0.046	0.023	0.024	123.278	192.737	0.027	0.000
% '.'	C	0.038	0.019	0.019	308.266	174.317	0.025	0.033
sentences-to-words ratio	C	0.030	0.015	0.015	272.298	130.825	0.023	0.000
%"she"	C	0.139	0.076	0.080	1499.577	1747.433	0.021	0.113
%"her"	C	0.136	0.071	0.078	1691.509	1456.704	0.020	0.000
%"a"	C	0.013	0.007	0.007	108.548	99.386	0.018	0.000
%"had"	C	0.045	0.023	0.024	415.458	343.100	0.018	0.038
%"hi"	C	0.002	0.001	0.001	1.169	0.313	0.018	0.000
%"have"	C	0.005	0.003	0.003	13.923	2.960	0.017	0.000
%" "	C	0.016	0.008	0.009	48.066	3.086	0.017	0.000
%""	C	0.022	0.042	0.009	20.481	206.780	0.016	0.000
%"as"	C	0.022	0.011	0.012	201.090	178.897	0.015	0.000
%"is"	C	0.037	0.018	0.019	332.241	243.479	0.015	0.000
% ','	C	0.016	0.008	0.009	159.291	121.715	0.015	0.000
%"thi"	C	0.031	0.015	0.016	305.786	203.578	0.014	0.026
%"from"	C	0.001	0.001	0.001	8.719	9.304	0.012	0.000
%"with"	C	0.007	0.004	0.004	39.263	43.543	0.012	0.000
%"he"	C	0.000	0.000	0.000	3.510	0.107	0.012	0.000
% '?'	C	0.017	0.008	0.009	145.991	56.010	0.011	0.000
%"""	C	0.012	0.006	0.006	21.162	38.957	0.011	0.000
%"s"	C	0.018	0.009	0.010	151.102	142.076	0.011	0.000
%"my"	C	0.016	0.008	0.009	123.024	113.461	0.011	0.000
%"the"	C	0.012	0.006	0.007	5.347	2.563	0.010	0.000
%"i"	C	0.005	0.003	0.003	54.246	35.590	0.010	0.000
avg. sentence length	C	0.032	0.016	0.017	9.284	126.584	0.010	0.000
% '!'	C	0.021	0.011	0.011	79.828	23.843	0.010	0.000
%"+"	C	0.004	0.028	0.002	10.056	14.058	0.010	0.000
%"were"	C	0.009	0.005	0.005	8.084	45.827	0.009	0.000
%"no"	C	0.004	0.002	0.002	20.742	28.698	0.009	0.000
%"for"	C	0.004	0.002	0.002	32.142	29.984	0.008	0.000
%"not"	C	0.010	0.005	0.006	98.628	83.706	0.008	0.000
%"of"	C	0.000	0.000	0.000	6.565	0.160	0.008	0.000
%"and"	C	0.001	0.000	0.001	6.346	1.569	0.007	0.000
%"so"	C	0.000	0.000	0.000	0.198	0.213	0.007	0.000
% ':'	C	0.019	0.010	0.010	190.413	96.746	0.006	0.000
%"be"	C	0.003	0.001	0.001	21.011	20.263	0.006	0.000
% ';'	C	0.012	0.006	0.007	25.416	2.390	0.006	0.000

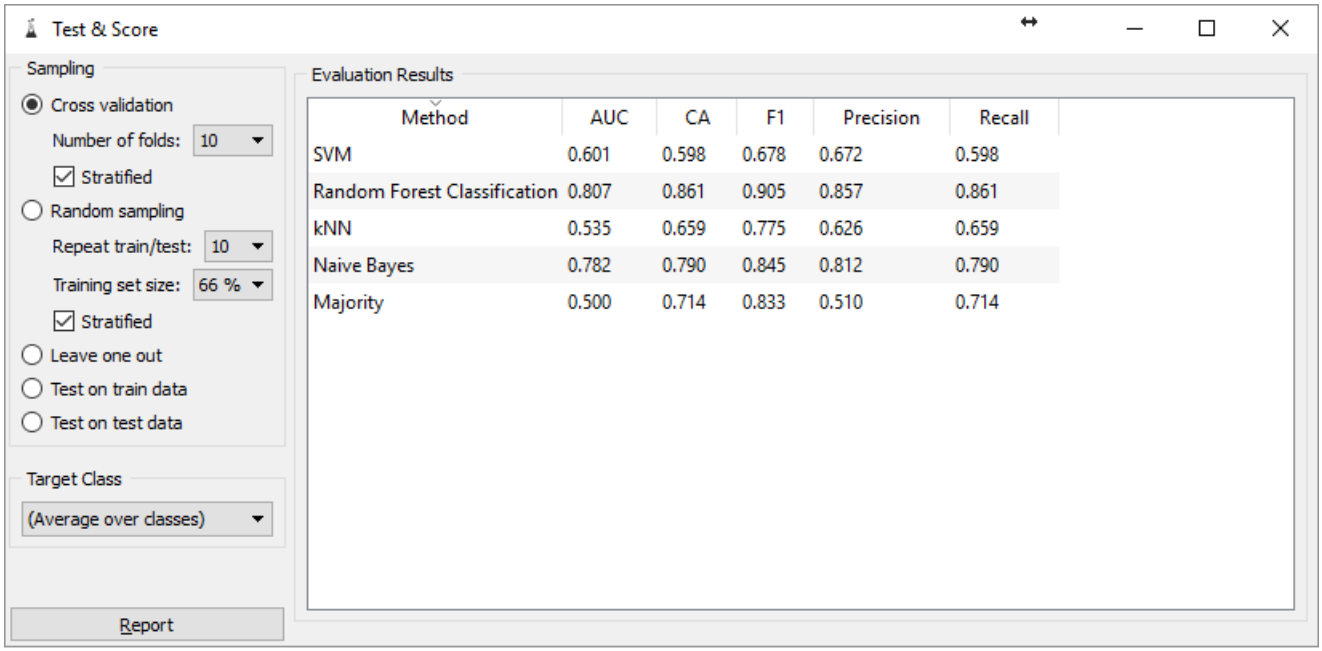
Slika 3: Za spol

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

Rezultati so sledeči.

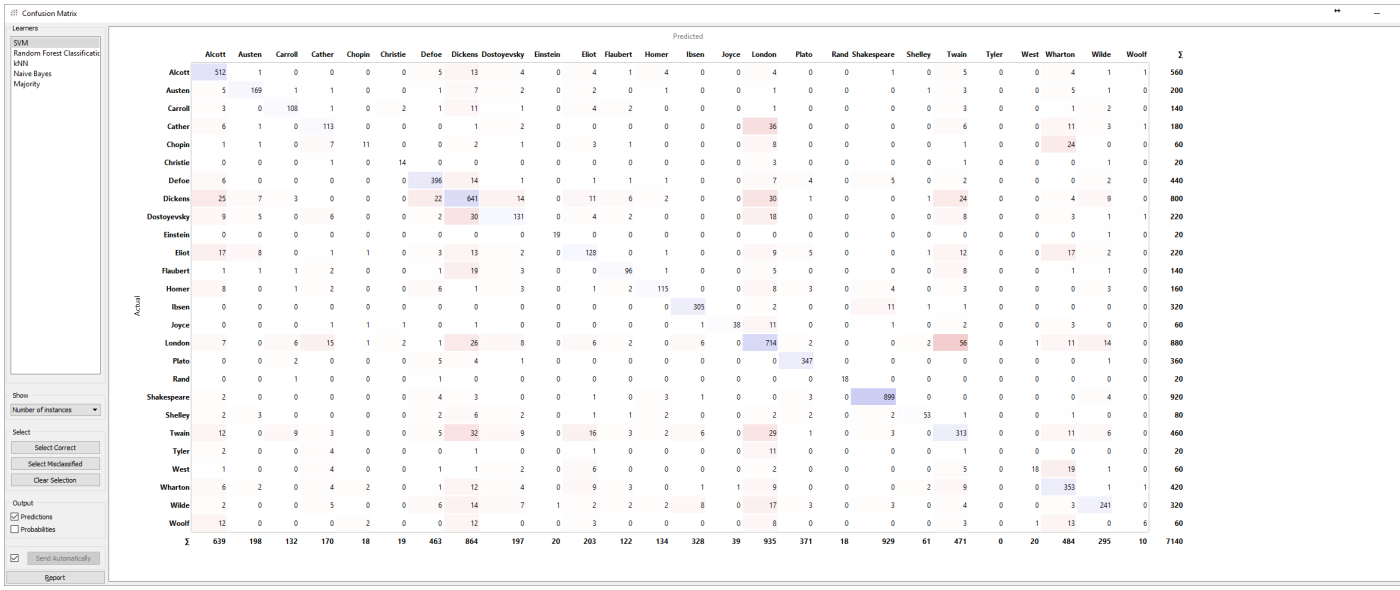


Slika 4: Klasifikacijska točnost pri napovedovanju avtorja

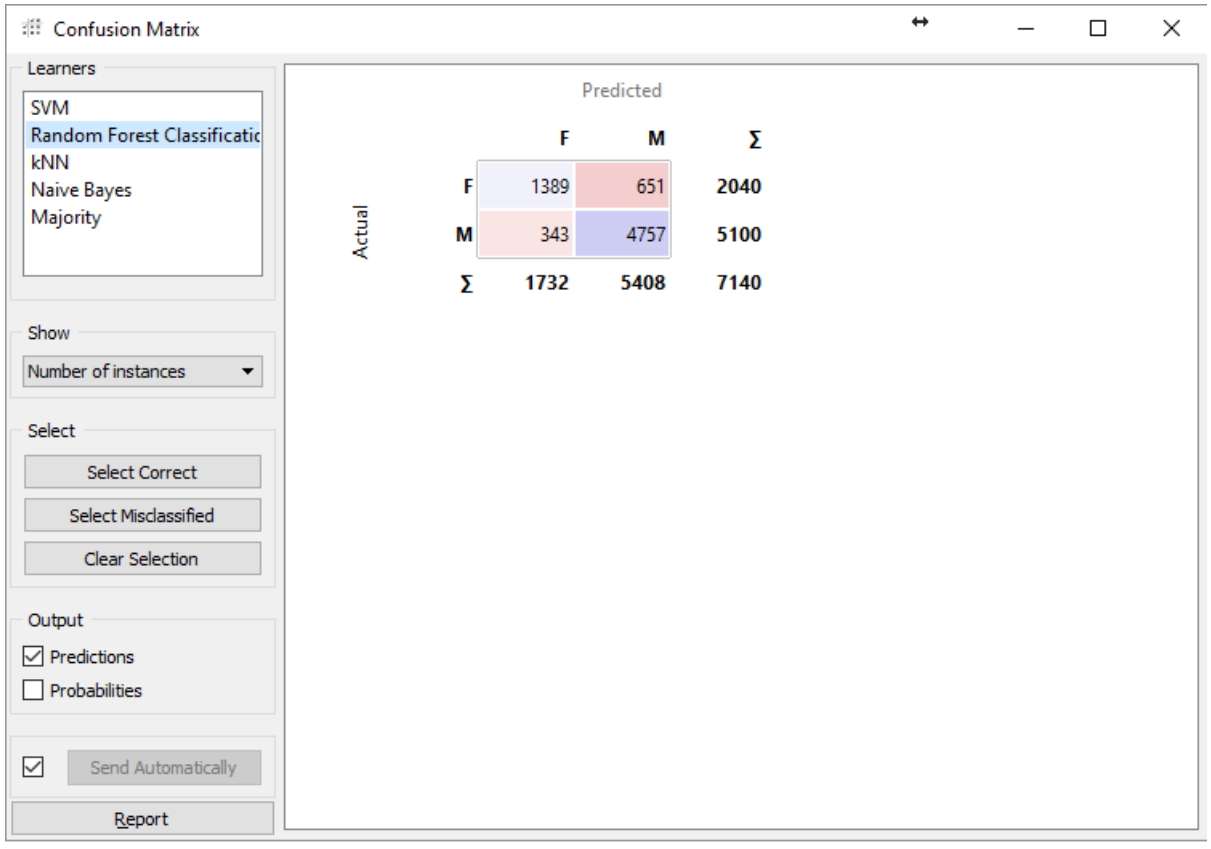


Slika 5: Klasifikacijska točnost pri napovedovanju spola

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

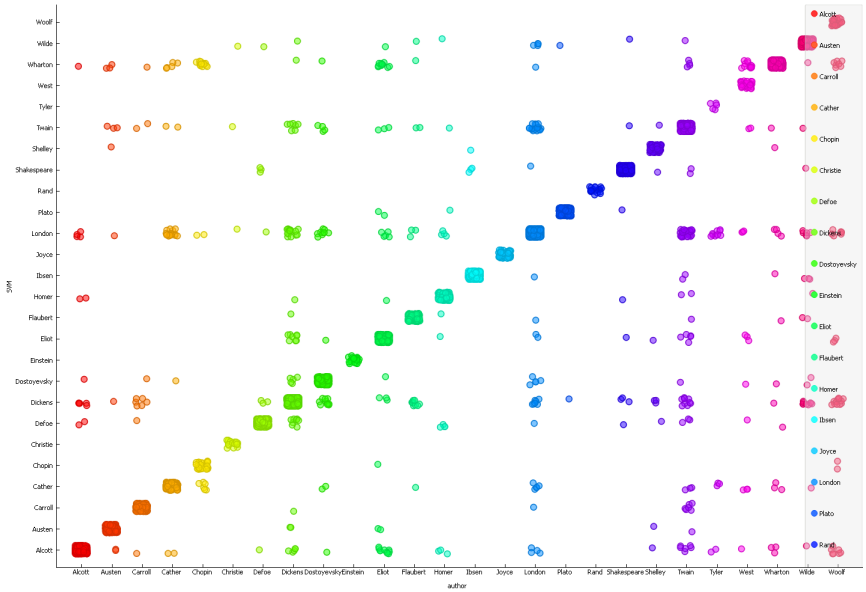


Slika 6: Confusion matrix za avtorja

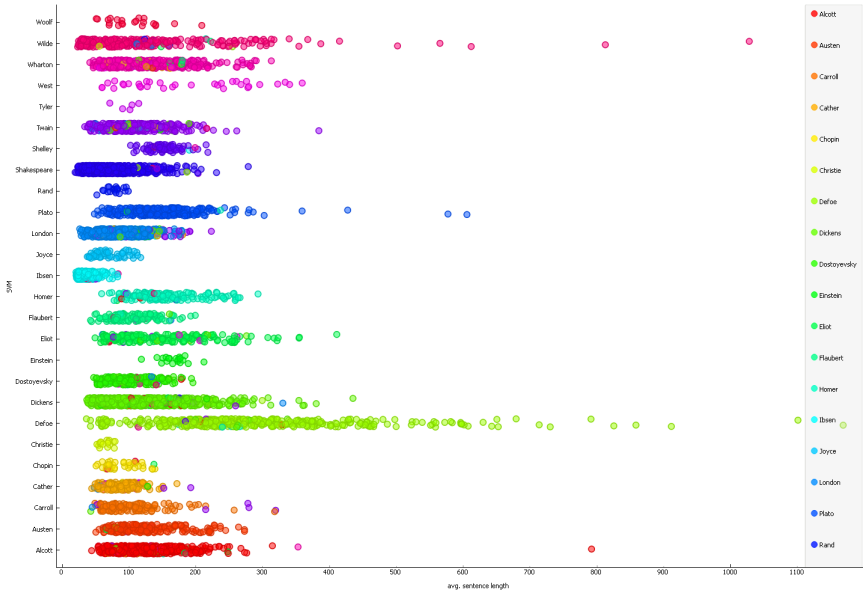


Slika 7: Confusion matrix za spol

Scatter plots

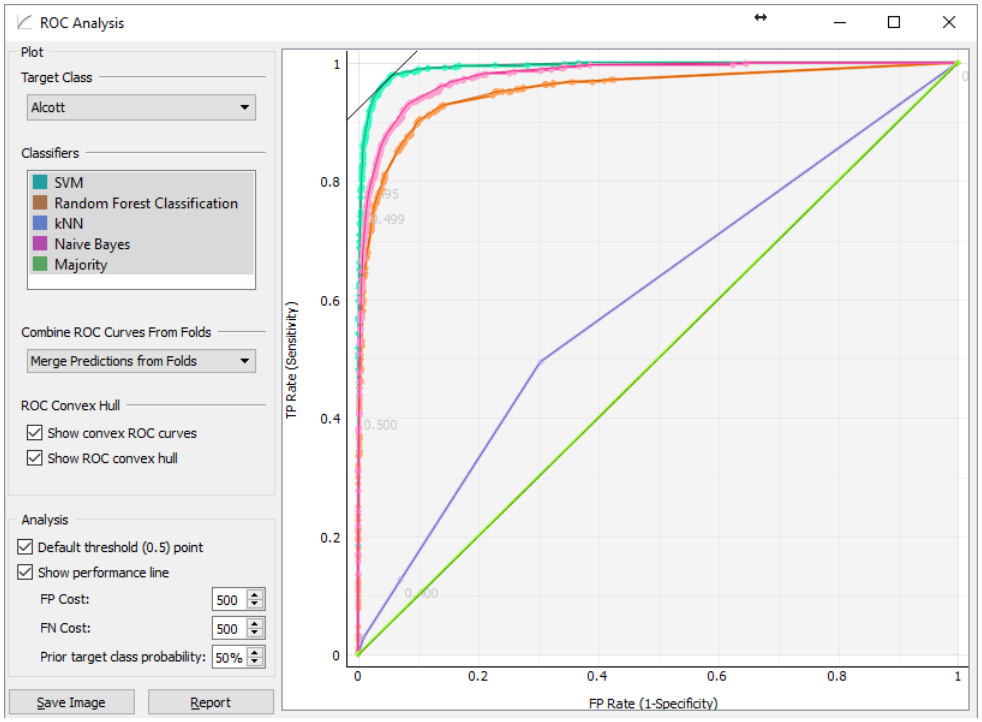


Slika 8: Scatter - author - SVM

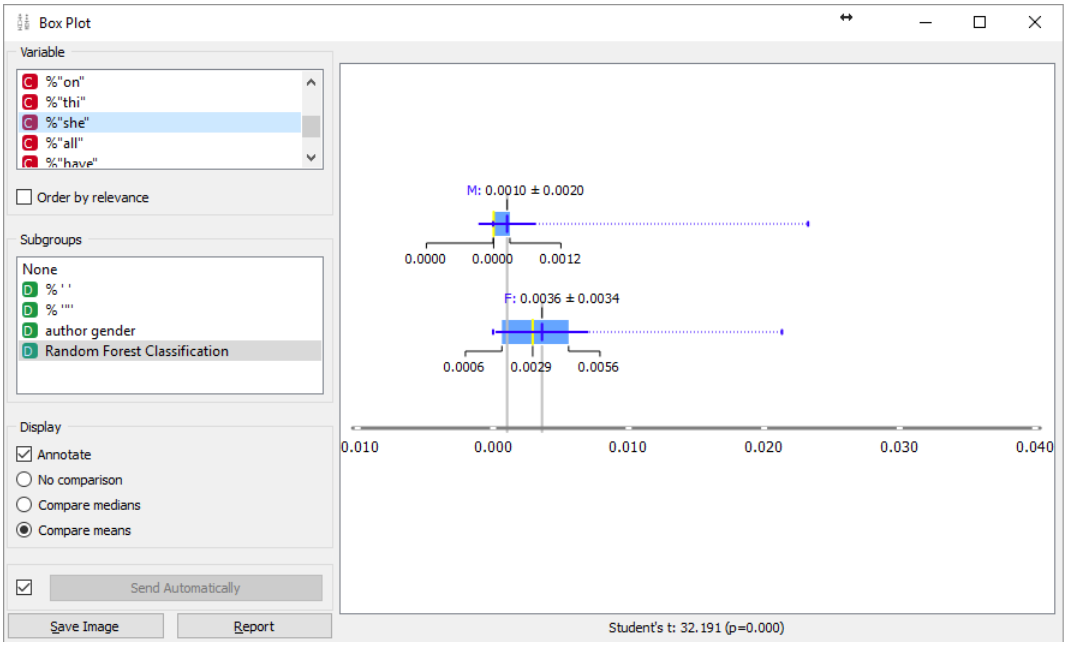


Slika 9: Scatter - dolzina stavka - SVMl

ROC krivulja ter box plot



Slika 10: Author - ROC



Slika 11: Skatla z brki - beseda she