Univerzitet u Sarajevu

Elektrotehnički fakultet

Odsjek za automatiku i elektroniku

Predmet: Data mining

Akademska godina: 2018/2019

Zadaća

Text mining

Student: *Emina Hasanović*

Index:*1297/16907*

Sarajevo, Juni 2019.

Sadržaj

Zadatak 1

Potrebno je implementirati automatsku klasifikaciju tekstualnih datoteka iz BBC dataseta novosti (http://mlg.ucd.ie/files/datasets/bbc-fulltext.zip). Dataset se sastoji iz preko 2000 tekstualnih datoteka, raspoređenih u 5 kategorija. Svaka datoteka sadrži samo tekst novosti nad kojim će se direktno vršiti analiza. Za svaku kategoriju izabrati minimalno 10 datoteka kao trenirajući skup. Izvršiti ekstrakciju teksta iz definiranog skupa datoteka, procesirati tekst iz svake datoteke kroz osnovne korake obrade teksta: tokenizacija, filtriranje i stemming. Zatim formirati reprezentativnu TF-IDF matricu za svaki tekst. Rezultat prvog koraka bit će ukupno 50 procesiranih datoteka i njihovih TF-IDF matrica. Odabrati algoritam klasifikacije i primjenom tekst mininga izvršiti automatsku klasifikaciju novih ulaznih tekstualnih datoteka. Odabrati po 10 datoteka iz svake kategorije i provjeriti da li će algoritam klasifikacije ispravno automatsku kategorizirati datoteku u odgovarajuću kategoriju. Rezultate je potrebno tabelarno prikazati. Također, shodno kvaliteti dobijenih rezultata, izvršiti još jedno testiranje sa povećanim trenirajućim skupom datoteka (minimalno 20 datoteka), te utvrditi da li je došlo do poboljšanja rezultata automatske klasifikacije. Implementaciju izvršiti u programskom jeziku i okruženju po vašem izboru. Kao rješenje zadaće neophodno je priložiti kompletan izvorni kôd rješenja kao i PDF izvještaj u kojem su prikazane rezultantne TF-IDF matrice za trenirajući skup datoteka te tabelarna usporedba rezultata automatske klasifikacije i uspješnosti iste.

Rješenje

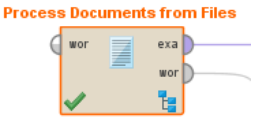
BBC dataset sadrži 2225 dokumenata sa BBC stranice vijesti koji su podijeljeni u 5 grupa: *business (510), entertainment (386), politics (417), sport (511), tech (401).*

Obzirom na to da je bilo potrebno izvršiti odabir seta koji ćemo koristiti kao trenirajući set, odabrano je prvih 10 tekstualnih file-ova iz svake od 5 kategorije, što je činilo ukupno 50 tekstualnih file-ova. Kako što postavka zadaće navodi prvenstveno je bilo potrebno izvršiti pripremu teksutualnih podataka kako bi bilo moguće bilo šta raditi s njima. Tako da je urađena tokenizacija, filtriranje, stemming i lematizacija datih podataka i to radi lakšeg procesiranja. Zajedno sa pripremama cjeloukupno pisanje koda je odrađeno u okviru programskog paketa *RapidMiner,* koji sadrži alate za text mining (sadržane u ekstenziji *Text Processing*). i klasifikatora na bazi *k-NN* algoritma. Razlog za odabir *RapidMiner-a* je jednostavnost implementacije i brzina procesiranja podataka.

Prilikom pisanja koda odabrano je 5 klasa koje su date u BBC setu i to:

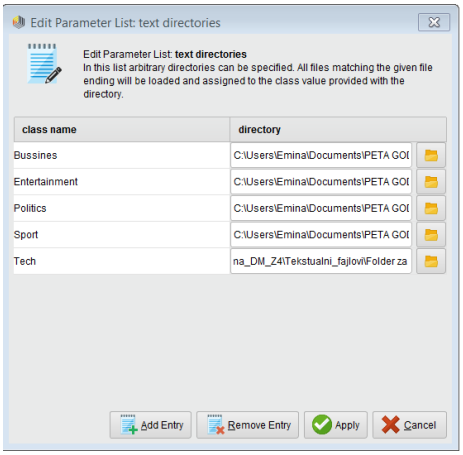
- Business
- Entertainment
- Politics
- Sport
- Tech

Prvi korak je slaganje blokova koji su bili potrebni zaizvršavanje klasifkacije. Prvenstveno je bilo potrebno učitati tekstualne podatke, kako bi se moglo bilo šta raditi. To je urađeno koristeći blok *Process Documents From Files,* koji je prikazan na slici 1.



Slika 1. Blok Process Documents From Files

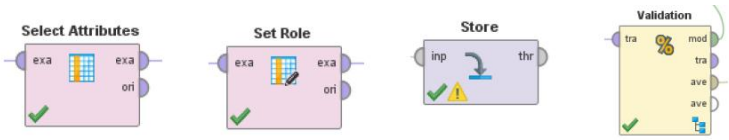
Unutar njega je bilo potrebno podesiti staze do svih klasa koje je potrebno klasificirati,a njih 5 je već navedeno. To je urađeno pronalaskom njihove lokacije na kompjuteru i postavljanja staze do njih, a naknadno im je dato ime koje smo sami odabrali i koje je navedeno na slici 2. Sve ovo navedeno je urađeno unutar *Edit List-e* ovog bloka. Postavljeno je da rezultat **Process Documents from** Files bloka bude *TF-IDF matrica* riječi i na osnovu nje će se vršiti klasifikacija dokumenata.



Slika 2. Postavljanje staze do pojedinih klasa

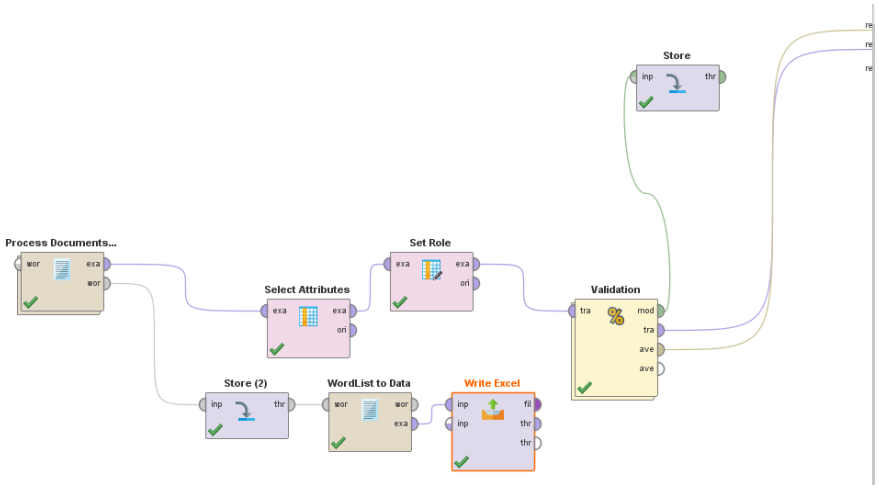
Sljedeći blokovi koje je bilo potrebno koristiti su:

- **Select Atributes** - blok unutar kojeg je bilo potrebno postaviti da ***atribute filter type*** bude *no_missing_values.* (slika 3.)
- **Set Role** - unutar kojeg je potrebno postaviti u odnosu na koji atribut se vrši klasifkacija a to se postiže postavljanjem ***atribute name*** na *label* i ***target role*** na *label*. Također unutar ***Edit Parametar List*** *target role* je potrebno postaviti na regular. (slika 3.)
- **Validation** - blok će naknadno biti objašnjen (slika 3.)
- **Store** - blok koji se koristi za spremanje klasificiranih podataka (slika 3.)



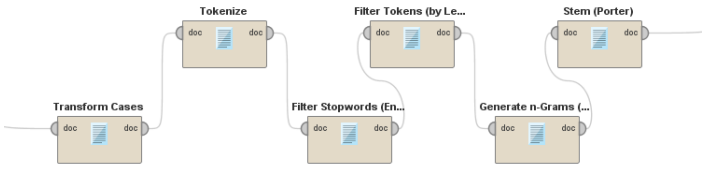
Slika 3. Korišteni blokovi

Cijela blokovska struktura je prikazana na slici 4.



Slika 4. Blokovska struktura

Pored ove osnovne blokovske strukture potrebno je podesiti ostale blokove unutar bloka **Process Documents From Files** i oni su dodavani dvostrukim klikom na ovaj blok. (slika 5.)



Slika 5. Blokovska struktura unutar bloka *Process Documents From Files*

Blokovi koji su sadržani unutar bloka **Process Documents From File** su:

- **Transform Cases** - blok koji se koristi za pretvaranje malih u velika slova.
- **Tokenize** - blok koji se koristi za razdvajanje teksta na riječi.
- **Filter Stopwords (English)** - blok koji se koristi za uklanjanje engleskih riječi tipa veznika npr. and, or,...
- **Filter Tokens (by Length)** - blok koji se koristi za smanjenje riječi na određenu dužinu (4 min length, 25 max length).
- **Generate n-Grams (Characters)** - blok koji se koristi za stvaranje jednog izraza tj. riječi.
- **Stem (Porter)** - blok koji se koristi za pronalaženje korijena riječi.

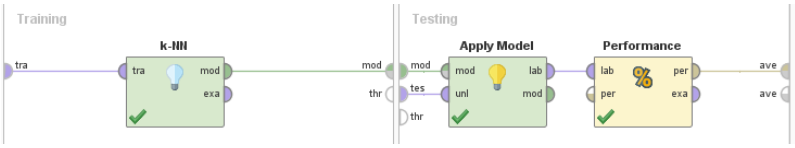
Na slici 6. su prikazani pretprocesirani podaci tj. kako su teksovi raspodijeljeni u tokene, broj njhovih pojavljivanja u cijelom korpusu dokumenata, ukupan broj pojavljivanja riječ, te pojavljivanje riječi u pojedinim kategorijama dokumenata.

Word	Attribut...	Total O...	Docum...	Bussines	Entertai...	Politics	Sport	Tech
abil	abil	4	4	1	0	0	2	1
absenc	absenc	2	2	0	0	0	2	0
abus	abus	3	3	0	0	3	0	0
accept	accept	4	4	0	0	3	1	0
accord	accord	7	6	4	3	0	0	0
account	account	2	2	1	0	0	0	1
accus	accus	2	2	0	1	1	0	0
achiev	achiev	4	3	0	2	0	1	1
act	act	3	3	1	1	1	0	0
action	action	5	5	1	0	1	1	2
activ	activ	7	5	6	0	0	0	1
activist	activist	2	2	0	0	2	0	0
ad	ad	16	15	3	1	6	3	3
adapt	adapt	2	2	0	2	0	0	0
address	address	9	3	0	0	2	0	7
adjust	adjust	3	2	3	0	0	0	0
administr	administr	5	4	3	1	1	0	0
adopt	adopt	2	2	0	0	0	0	2
advis	advis	3	3	1	1	0	0	1
affect	affect	4	4	3	0	1	0	0
agenc	agenc	3	3	3	0	0	0	0
agenda	agenda	3	3	1	0	2	0	0
agre	agre	6	4	0	1	2	3	0
ahead	ahead	5	5	1	1	2	1	0

Slika 6. Pretprocesirani podaci

Pored ovog bloka, još je unutar bloka **Validation** bilo potrebno dodati dodatne blokove i to (slika 7.):

- **k-NN** - blok koji se koristio za definisanje algoritma klasifikacije (uzeta su tri najbliža susjeda) .
- **Apply Model** - blok koji se koristio za testiranje krajnjeg modela.
- **Preformance** - blok koji se koristi za ispisivanje statističkih podataka.



Slika 7. Blokovska struktura unutar bloka *Validation*

Na slici 8. i 9. prikazana je uspješnost modela za treniranje na osnovu mjera: tačnost (accuracy) i greške klasifikacije (classification error). Svih pet kategorije su klasifikovane sa 100%. Dakle, ukupna tačnost klasifikatora je 100%, tj. greška pri klasifikaciji je 0%.

accuracy: 100.00%

	true Bussines	true Entertainment	true Politics	true Sport	true Tech	class precision
pred. Bussines	3	0	0	0	0	100.00%
pred. Entertainment	0	3	0	0	0	100.00%
pred. Politics	0	0	3	0	0	100.00%
pred. Sport	0	0	0	3	0	100.00%
pred. Tech	0	0	0	0	3	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	

Slika 8. Uspješnost seta za treniranje (accuracy)

classification_error: 0.00%

	true Bussines	true Entertainment	true Politics	true Sport	true Tech	class precision
pred. Bussines	3	0	0	0	0	100.00%
pred. Entertainment	0	3	0	0	0	100.00%
pred. Politics	0	0	3	0	0	100.00%
pred. Sport	0	0	0	3	0	100.00%
pred. Tech	0	0	0	0	3	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	

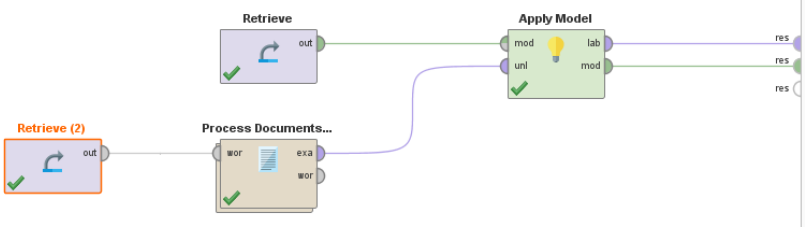
Slika 9. Uspješnost seta za treniranje (classification error)

Na slici 10. je prikazan jedan dio TF-IDF matrica tj. za nekoliko dokumenata od njih 50. Cijela TF-IDF matrica je data kao prilog ovom dokumentu.

Row No.	label	text	metadata_file	metadata_p...	..	abil	absenc	abus	accept	accord	account	accus	achiev	act	action	activ	activist	ad	adapt
1	Bussines	Rank 'set to s...	021.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0.075	0	0.039	0
2	Bussines	Sluggish eco...	022.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Bussines	Mixed signals...	023.txt	C:\Users\Emi...	.	0	0	0	0	0.075	0	0	0	0	0	0.081	0	0	0
4	Bussines	US trade gap ...	024.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Bussines	Yukos loses ...	025.txt	C:\Users\Emi...	.	0.046	0	0	0	0	0.059	0	0	0.052	0	0.127	0	0	0
6	Bussines	Safety alert a...	026.txt	C:\Users\Emi...	.	0	0	0	0	0.087	0	0	0	0	0	0	0	0	0
7	Bussines	Steel firm 'to ...	027.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Bussines	Strong dema...	028.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Bussines	UK firm faces...	029.txt	C:\Users\Emi...	.	0	0	0	0	0.065	0	0	0	0	0.070	0.070	0	0.037	0
10	Bussines	Soaring oil 'hi...	030.txt	C:\Users\Emi...	.	0	0	0	0	0.058	0	0	0	0	0	0	0	0.033	0
11	Bussines	Irish markets ...	031.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Entertain...	UK Directors ...	377.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0.151	0	0	0	0	0	0
13	Entertain...	Halloween wr...	378.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0.130
14	Entertain...	Tarantino 'to ...	379.txt	C:\Users\Emi...	.	0	0	0	0	0.210	0	0	0	0	0	0	0	0	0
15	Entertain...	Boogeyman t...	380.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Entertain...	Lost Doors fr...	381.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0.136	0	0.119	0	0	0	0.051	0
17	Entertain...	Last Star War...	382.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Entertain...	French hono...	383.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Entertain...	Robots marc...	384.txt	C:\Users\Emi...	.	0	0	0	0	0.078	0	0	0	0	0	0	0	0	0
20	Entertain...	Hobbit pictur...	385.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Entertain...	Buffy creator j...	386.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0.090
22	Politics	Howard hits ...	408.txt	C:\Users\Emi...	.	0	0	0	0	0	0	0.057	0	0.050	0	0	0.057	0	0
23	Politics	Tories urge 'c...	409.txt	C:\Users\Emi...	.	0	0	0	0.084	0	0	0	0	0	0	0	0	0	0
24	Politics	Sayeed to sta...	410.txt	C:\Users\Emi...	.	0	0	0	0.092	0	0	0	0	0	0	0	0	0	0

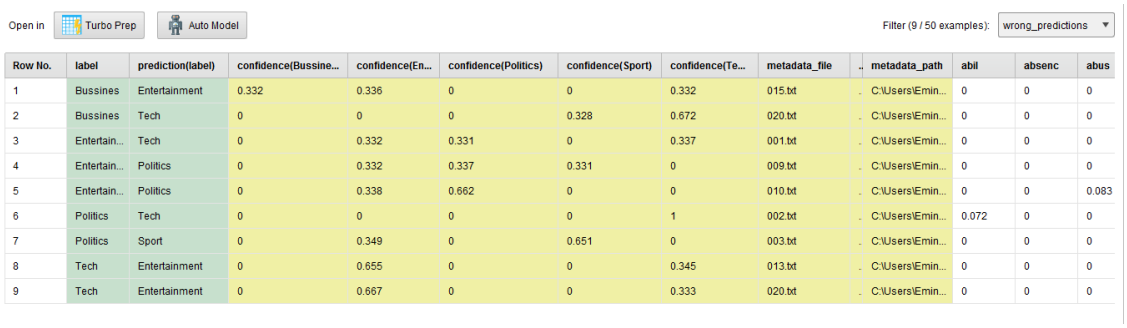
Slika 10. Rezultanta TF-IDF matrica

Na slici 11. prikazana je blokovska struktura za validaciju dobivenog modela. Struktura bloka **Process documents From Files** je ista, ali su podaci za procesiranje iz skupa datoteka za testiranje (prepodešena putanja) i lista spašenih riječi iz koraka treniranja. Blok **Apply Model** koristi spašeni model iz koraka treniranja i preprocesirane podatke iz skupa testiranja.



Slika 11. Blokovska shema za testiranje modela

Međutim ukoliko uzememo drugi set koji nije treniran dobijemo sljedeće podatke (slika 12, 13, 14 i 15). Na osnovu prethodne tabele možemo primijetiti da je pogrešna klasifikacija izvršena nad 9 klasa od 50 što nam daje tačnost od 82% što je i dalje jako dobra tačnost. Slika 12. pokazuje kod kojih se klasa javljaju pogrešne klasifikacije. Slika 13 i 14 pokazuje koje su klasifikacije tačne.



Open in Turbo Prep Auto Model Filter (9 / 50 examples): wrong_predictions

Row No.	label	prediction(label)	confidence(Bussine...	confidence(En...	confidence(Politics)	confidence(Sport)	confidence(Te...	metadata_file	..	metadata_path	abil	absenc	abus
1	Bussines	Entertainment	0.332	0.336	0	0	0.332	015.txt	.	C:\Users\Emin...	0	0	0
2	Bussines	Tech	0	0	0	0.328	0.672	020.txt	.	C:\Users\Emin...	0	0	0
3	Entertain...	Tech	0	0.332	0.331	0	0.337	001.txt	.	C:\Users\Emin...	0	0	0
4	Entertain...	Politics	0	0.332	0.337	0.331	0	009.txt	.	C:\Users\Emin...	0	0	0
5	Entertain...	Politics	0	0.338	0.662	0	0	010.txt	.	C:\Users\Emin...	0	0	0.083
6	Politics	Tech	0	0	0	0	1	002.txt	.	C:\Users\Emin...	0.072	0	0
7	Politics	Sport	0	0.349	0	0.651	0	003.txt	.	C:\Users\Emin...	0	0	0
8	Tech	Entertainment	0	0.655	0	0	0.345	013.txt	.	C:\Users\Emin...	0	0	0
9	Tech	Entertainment	0	0.667	0	0	0.333	020.txt	.	C:\Users\Emin...	0	0	0

Slika 12. Pogrešne klasifikacije

Row No.	label	prediction(label)	confidence(Bussi...	confidence(Entert...	confidence(Politics)	confidence(Sport)	confidence(Tech)	metada...	...	metadata_...
1	Bussines	Bussines	1	0	0	0	0	011.txt	...	C:\Users\Em...
2	Bussines	Bussines	0.668	0	0	0	0.332	012.txt	...	C:\Users\Em...
3	Bussines	Bussines	0.672	0	0	0	0.328	013.txt	...	C:\Users\Em...
4	Bussines	Bussines	1	0	0	0	0	014.txt	...	C:\Users\Em...
5	Bussines	Bussines	1	0	0	0	0	016.txt	...	C:\Users\Em...
6	Bussines	Bussines	0.664	0	0	0	0.336	017.txt	...	C:\Users\Em...
7	Bussines	Bussines	0.646	0.354	0	0	0	018.txt	...	C:\Users\Em...
8	Bussines	Bussines	1	0	0	0	0	019.txt	...	C:\Users\Em...
9	Entertain...	Entertainment	0	1	0	0	0	002.txt	...	C:\Users\Em...
10	Entertain...	Entertainment	0.328	0.342	0.331	0	0	003.txt	...	C:\Users\Em...
11	Entertain...	Entertainment	0	1	0	0	0	004.txt	...	C:\Users\Em...
12	Entertain...	Entertainment	0	1	0	0	0	005.txt	...	C:\Users\Em...
13	Entertain...	Entertainment	0	0.676	0	0.324	0	006.txt	...	C:\Users\Em...
14	Entertain...	Entertainment	0	1	0	0	0	007.txt	...	C:\Users\Em...
15	Entertain...	Entertainment	0	1	0	0	0	008.txt	...	C:\Users\Em...
16	Politics	Politics	0	0	1	0	0	001.txt	...	C:\Users\Em...
17	Politics	Politics	0	0	1	0	0	004.txt	...	C:\Users\Em...
18	Politics	Politics	0	0	1	0	0	005.txt	...	C:\Users\Em...
19	Politics	Politics	0	0	1	0	0	006.txt	...	C:\Users\Em...
20	Politics	Politics	0	0	1	0	0	007.txt	...	C:\Users\Em...
21	Politics	Politics	0	0.339	0.661	0	0	008.txt	...	C:\Users\Em...
22	Politics	Politics	0	0	1	0	0	009.txt	...	C:\Users\Em...
23	Politics	Politics	0	0	1	0	0	010.txt	...	C:\Users\Em...
24	Sport	Sport	0	0	0	1	0	001.txt	...	C:\Users\Em...

Slika 13. Tačne klasifikacije

Row No.	label	prediction(label)	confidence(Bussi...	confidence(Entert...	confidence(Politics)	confidence(Sport)	confidence(Tech)	metada...	...	metadata_...
18	Politics	Politics	0	0	1	0	0	005.txt	...	C:\Users\Em...
19	Politics	Politics	0	0	1	0	0	006.txt	...	C:\Users\Em...
20	Politics	Politics	0	0	1	0	0	007.txt	...	C:\Users\Em...
21	Politics	Politics	0	0.339	0.661	0	0	008.txt	...	C:\Users\Em...
22	Politics	Politics	0	0	1	0	0	009.txt	...	C:\Users\Em...
23	Politics	Politics	0	0	1	0	0	010.txt	...	C:\Users\Em...
24	Sport	Sport	0	0	0	1	0	001.txt	...	C:\Users\Em...
25	Sport	Sport	0	0	0	1	0	002.txt	...	C:\Users\Em...
26	Sport	Sport	0	0	0	1	0	003.txt	...	C:\Users\Em...
27	Sport	Sport	0	0	0	1	0	004.txt	...	C:\Users\Em...
28	Sport	Sport	0	0	0	1	0	005.txt	...	C:\Users\Em...
29	Sport	Sport	0	0	0	1.000	0	006.txt	...	C:\Users\Em...
30	Sport	Sport	0	0.330	0	0.670	0	007.txt	...	C:\Users\Em...
31	Sport	Sport	0	0	0	1	0	008.txt	...	C:\Users\Em...
32	Sport	Sport	0	0	0	1	0	009.txt	...	C:\Users\Em...
33	Sport	Sport	0	0	0	1	0	010.txt	...	C:\Users\Em...
34	Tech	Tech	0	0	0	0	1	011.txt	...	C:\Users\Em...
35	Tech	Tech	0	0	0	0	1	012.txt	...	C:\Users\Em...
36	Tech	Tech	0	0	0	0	1	014.txt	...	C:\Users\Em...
37	Tech	Tech	0	0	0	0	1	015.txt	...	C:\Users\Em...
38	Tech	Tech	0	0	0.330	0	0.670	016.txt	...	C:\Users\Em...
39	Tech	Tech	0.331	0	0	0	0.669	017.txt	...	C:\Users\Em...
40	Tech	Tech	0	0	0	0	1	018.txt	...	C:\Users\Em...
41	Tech	Tech	0	0.336	0	0	0.664	019.txt	...	C:\Users\Em...

Slika 14. Tačne klasifikacije

Index	Nominal value	Absolute count	Fraction
1	Sport	11	0.220
2	Tech	11	0.220
3	Entertainment	10	0.200
4	Politics	10	0.200
5	Bussines	8	0.160

Slika 15. Broj pogrešnih klasifikacija u pojedinačnim kategorijama

Dalje ćemo pokušati da povećamo set za treniranje i da uzmemo 20 file-ova iz svake od 5 klasa za treniranje (100 ukupno), dok je set za testiranje 10 fajlova (50 ukupno). Greška se desi na 4 file-ova od 50 što nam daje tačnost od 92 %, na osnovu čega možemo da zaključimo da se tačnost znatno povećala i da je povećanje seta za treniranje imalo velikog uticaja. Na slici 16. su prikazane pogrešne klasifikacije, a na slici 17 i 18. tačne klasifikacije.

Row No.	label	prediction(label)	confidence(Bussines)	confidence(Entertainment)	confidence(Politics)	confidence(Sport)	confidence(Tech)	metada...	..	metadata_path
1	Bussines	Tech	0	0	0	0	1.000	020.txt	.	C:\Users\Emina\...
2	Entertainment	Politics	0	0	0.678	0.322	0	010.txt	.	C:\Users\Emina\...
3	Tech	Entertainment	0	0.675	0	0	0.325	019.txt	.	C:\Users\Emina\...
4	Tech	Entertainment	0	0.666	0	0	0.334	020.txt	.	C:\Users\Emina\...

Slika 16. Pogrešne klasifikacije u slučaju kada je povećan skup za treniranje

Row No.	label	prediction(label)	confidence(Bussines)	confidence(Entert...	confidence(Politi...	confidence(Sport)	confidence(Tech)	metada...	..	metada...
1	Bussin...	Bussines	0.751	0	0	0	0.249	011.txt	.	C:\Users...
2	Bussin...	Bussines	1	0	0	0	0	012.txt	.	C:\Users...
3	Bussin...	Bussines	0.752	0.248	0	0	0	013.txt	.	C:\Users...
4	Bussin...	Bussines	0.748	0	0	0	0.252	014.txt	.	C:\Users...
5	Bussin...	Bussines	0.500	0	0.250	0	0.250	015.txt	.	C:\Users...
6	Bussin...	Bussines	1	0	0	0	0	016.txt	.	C:\Users...
7	Bussin...	Bussines	1	0	0	0	0	017.txt	.	C:\Users...
8	Bussin...	Bussines	0.644	0.356	0	0	0	018.txt	.	C:\Users...
9	Bussin...	Bussines	1	0	0	0	0	019.txt	.	C:\Users...
10	Entertai...	Entertainment	0	0.752	0	0	0.248	001.txt	.	C:\Users...
11	Entertai...	Entertainment	0.245	0.755	0	0	0	002.txt	.	C:\Users...
12	Entertai...	Entertainment	0	1	0	0	0	003.txt	.	C:\Users...
13	Entertai...	Entertainment	0	1	0	0	0	004.txt	.	C:\Users...
14	Entertai...	Entertainment	0	1	0	0	0	005.txt	.	C:\Users...
15	Entertai...	Entertainment	0	1	0	0	0	006.txt	.	C:\Users...
16	Entertai...	Entertainment	0	1	0	0	0	007.txt	.	C:\Users...
17	Entertai...	Entertainment	0	1	0	0	0	008.txt	.	C:\Users...
18	Entertai...	Entertainment	0	0.662	0.338	0	0	009.txt	.	C:\Users...
19	Politics	Politics	0	0	1	0	0	001.txt	.	C:\Users...
20	Politics	Politics	0.250	0	0.500	0	0.250	002.txt	.	C:\Users...
21	Politics	Politics	0	0.241	0.759	0	0	003.txt	.	C:\Users...
22	Politics	Politics	0	0	1	0	0	004.txt	.	C:\Users...
23	Politics	Politics	0	0	1	0	0	005.txt	.	C:\Users...
24	Politics	Politics	0.332	0	0.668	0	0	006.txt	.	C:\Users...

Slika 17. Tačne klasifikacije u slučaju kada je povećan skup za treniranje

Row No.	label	prediction(label)	confidence(Bussines)	confidence(Entert...	confidence(Politi...	confidence(Sport)	confidence(Tech)	metada...	..	metada...
23	Politics	Politics	0	0	1	0	0	005.txt	.	C:\Users...
24	Politics	Politics	0.332	0	0.668	0	0	006.txt	.	C:\Users...
25	Politics	Politics	0	0	1	0	0	007.txt	.	C:\Users...
26	Politics	Politics	0	0	0.684	0.316	0	008.txt	.	C:\Users...
27	Politics	Politics	0	0	1	0	0	009.txt	.	C:\Users...
28	Politics	Politics	0	0	1	0	0	010.txt	.	C:\Users...
29	Sport	Sport	0	0	0	1	0	001.txt	.	C:\Users...
30	Sport	Sport	0	0	0	1	0	002.txt	.	C:\Users...
31	Sport	Sport	0	0	0	1	0	003.txt	.	C:\Users...
32	Sport	Sport	0	0	0	1	0	004.txt	.	C:\Users...
33	Sport	Sport	0	0	0.210	0.790	0	005.txt	.	C:\Users...
34	Sport	Sport	0	0	0	1	0	006.txt	.	C:\Users...
35	Sport	Sport	0	0	0	1	0	007.txt	.	C:\Users...
36	Sport	Sport	0	0	0	1	0	008.txt	.	C:\Users...
37	Sport	Sport	0	0	0	1	0	009.txt	.	C:\Users...
38	Sport	Sport	0	0	0	1	0	010.txt	.	C:\Users...
39	Tech	Tech	0	0	0	0	1	011.txt	.	C:\Users...
40	Tech	Tech	0	0	0	0	1	012.txt	.	C:\Users...
41	Tech	Tech	0.255	0	0	0	0.745	013.txt	.	C:\Users...
42	Tech	Tech	0	0	0	0	1	014.txt	.	C:\Users...
43	Tech	Tech	0.246	0	0	0	0.754	015.txt	.	C:\Users...
44	Tech	Tech	0	0	0	0	1	016.txt	.	C:\Users...
45	Tech	Tech	0	0	0	0	1	017.txt	.	C:\Users...
46	Tech	Tech	0	0	0	0	1	018.txt	.	C:\Users...

Slika 18. Tačne klasifikacije u slučaju kada je povećan skup za treniranje

Zadatak 2

Za jednu dobijenu grupu klasifikacije odredite koje riječi se najćešće pojavljuju u dokumentima te grupe.

2.1. Za 2 dobijene riječi koja se najčešće koristi u toj kategoriji analizirati twitter postove (možete iskoristiti neku postojeću arhivu) i utvrditi koliko twittova sadrže te riječi.

2.2. Izvršiti sentimentalnu analizu za dokumente iz dobijene grupe klasifikacije.

Rješenje

Ovaj zadatak je također urađen koristeći Rapidminner. Da bi se riječi sortirale po broju pojavljivanja u pojedinoj kategoriji ili po ukupnom broju pojavljivanja riječi korištišten je blok **Sort** (slika 19.) kojem je proslijeđena lista riječi koju je na njegovom izlazu sortirana. Pronađene su riječi koje se najčešće pojavljuju u kategoriji Sport (slika 20.)

Slika 19. Sort blok

Row No.	word	in documents	total	..	..	..	in class (Sport)
1	world	27	65	.	.	.	40
2	athlet	10	29	.	.	.	28
3	indoor	8	19	.	.	.	19
4	championship	11	16	.	.	.	16
5	olymp	8	16	.	.	.	16
6	european	15	32	.	.	.	13
7	record	18	36	.	.	.	12
8	countri	17	31	.	.	.	11
9	cross	6	12	.	.	.	11
10	madrid	4	11	.	.	.	11
11	believ	15	23	.	.	.	10
12	champion	7	10	.	.	.	10
13	london	12	17	.	.	.	10
14	marathon	4	10	.	.	.	10
15	medal	5	10	.	.	.	10
16	athen	4	9	.	.	.	9
17	miss	7	13	.	.	.	9
18	season	10	14	.	.	.	9
19	titl	7	14	.	.	.	9
20	event	9	13	.	.	.	8
21	final	11	14	.	.	.	8
22	ireland	7	17	.	.	.	8
23	radcliff	3	8	.	.	.	8
24	tripl	3	9	.	.	.	8
25	confid	8	10	.	.	.	7

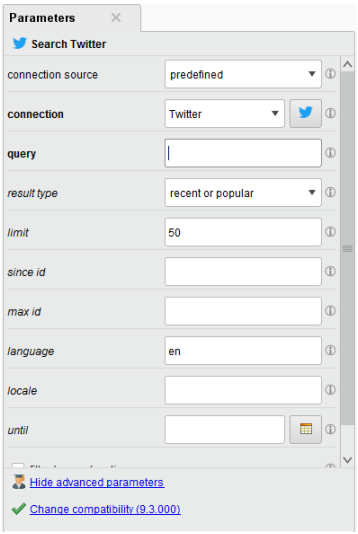
Slika 20. Prvih 25 najčešćih riječi u kategoriji Sport

Analiza najčešćih riječi u Twitter postovima

Analizu ćemo vršiti za najčešće riječi iz kategorije Sport. Twitter postovi se često koriste za zadatke text minnga. Analizu ću raditi na osnovu svog stvarnog profila na twitteru (Emina Hasanović) i njene arhive. Da bi se ovo realizovalo u Rapidminneru potrebno je koristiti operator **Search Twitter** (slika

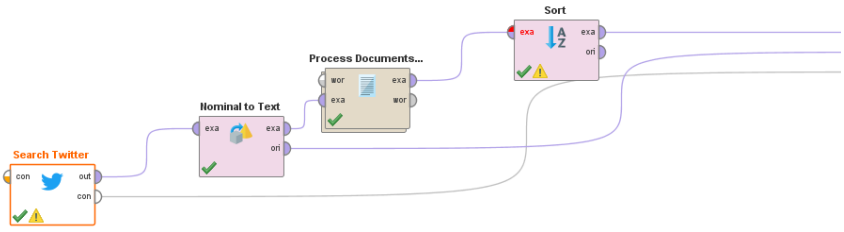
21.) Ovaj operator omogućava konekciju na postojeći profil tako što se podese njegovi odgovarajući parametri (slika 22.). Podesi se tip konekcije, twittovi iz kojih oblasti (parametar query: bit će unesena riječ Sport) će se pretraživati, da li će to biti skoriji ili popularni twittovi, broj twittova, ili da se unese mininalan i maksimalan id twitta pa da se pretražuje na taj način, zatim jezik (ovdje engleski) itd.

Slika 21. Search Twitter operator



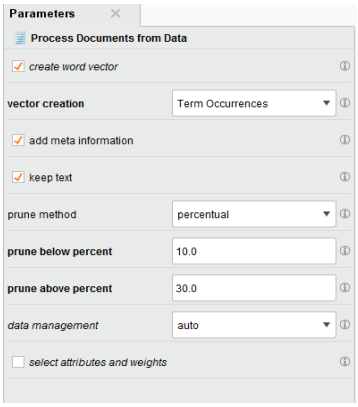
Slika 22. Parametri Search Twitter operatora

Cijela šema za pretraživanje Twittera prikazana je na slici 23.

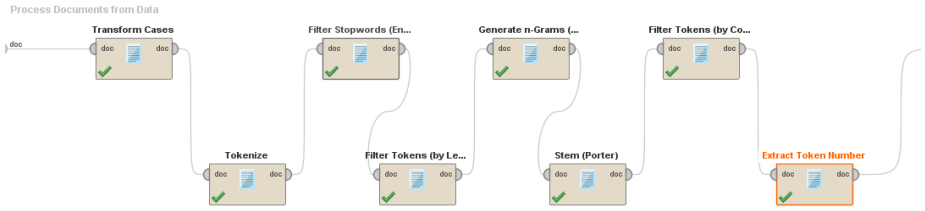


Slika 23. Blok operatora za pretragu Twittera-a u Rapidminner-u

Nominal to Text operator samo konvertuje ulazne podatke u standardnu tabelu podataka, **Process Documents from Data** obavlja slične operacije kao u prethodnim zadacima, samo će još dodatno filtrirati riječi na osnovu sadržaja, **Filter Tokens (by Content),** te je podešeno da izlaz bloka bude vektor broja riječi u dokumentu (term occurrences), parametri ovog bloka su prikazani na slici 24., a njegova unutrašnja struktura na slici 25. **Sort** blok radi sortiranje podataka. Ideja je da se na osnovu operatora Filter Tokens (by Content) filtriraju sve riječi osim jedne/dvije određene koja se pretražuje u twittu te da u njegovom sadržaju ostane samo ta riječ i na kraju samo sortiramo dokumente na osnovu broja tokena, svi koji budu sadržali tražene riječi imat će nenulti broj tokena i doći će na prvi dio tabele sortiranjem, te se također može uočiti koji to tweetowi sadrže više od jednom tražene riječi.



Slika 24. Parametri operatora Process Documents from Data

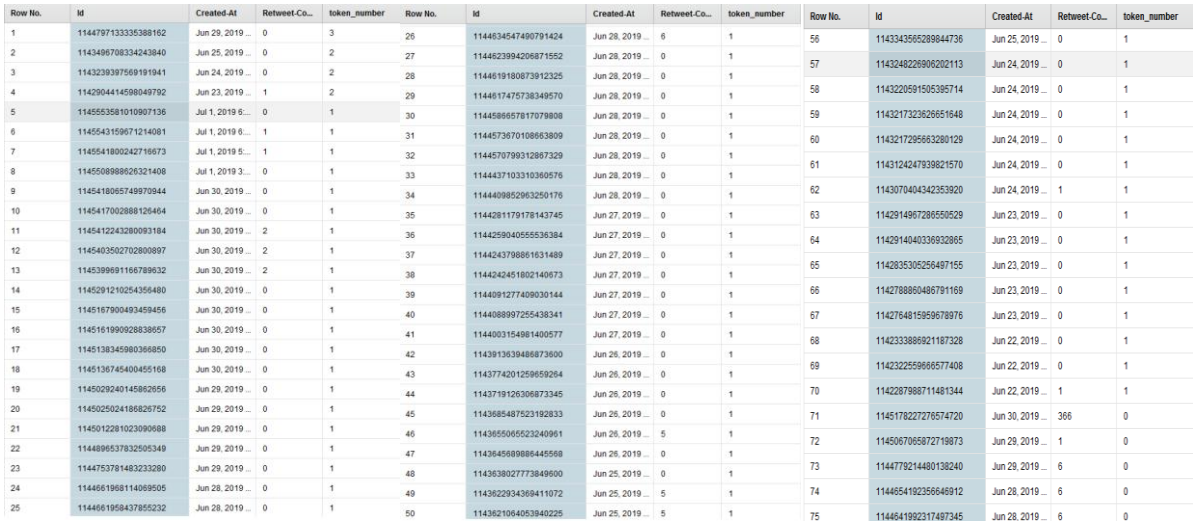


Slika 25. Unutrašnja struktura operatora Process Documents from Data

Dvije najčešće korištene riječi su **world** i **athlet** u prethodnom zadatku. Pretraženo je 1000 twittova, dio te tabele podataka prikazan je na slici 26. koja sadrži sve informacije o twittu, koji je id twitta, kada je objavljen, od kojeg korisnika, kojem korisniku je namijenjen, id korisnika, jezik na kojem je twitt, izvor (link), sam text twitta, geolokacija itd. Nama je najzanimljivije analizirati tekst twitta tj. da li sadrži riječ **athlet** (slika 27).

Row No.	Id	Created-At	From-User	From-User-Id	To-User	To-User...	Lan...	Source	Text	Geo-Loc...	Geo-Loc...	Retweet-...
1	1145553581010907136	Jul 1, 2019 ...	□□mateoooo...	1090975257...	joy_talon	1028867...	en	<a href=...	@joy_talon aw athlet	?	?	0
4	1145508988626321408	Jul 1, 2019 ...	Jay pichardo	38724681	T_Rozz...	1881202...	en	<a href=...	@T_Rozzay3 @SLAMonline @PUMAHoops athlet on a @...	?	?	0
5	1145418065749970944	Jun 30, 201...	Nai	469766933	?	-1	en	<a href=...	"You're an athlet" ebuka abeg □	?	?	0
6	1145417002888126464	Jun 30, 201...	Fad Gadget	2199382054	sgroffy	3242952...	en	<a href=...	@sgroffy Glad to read you and that you are okay, you are a...	?	?	0
10	1145291210254356480	Jun 30, 201...	Mark Cunnin...	365361084	Leavelt...	1644262...	en	<a href=...	@LeaveltYeh Bosch Athlet is decent	?	?	0
12	1145167900493459456	Jun 30, 201...	Mujahid	1518409460	shoaib...	4285547...	en	<a href=...	@shoaib100mph Game is source of get together but your...	?	?	0
13	1145161990928838657	Jun 30, 201...	Son Of CBD Oil	604902486	?	-1	en	<a href=...	Sean Clores Named CCNY Head Men's Basketball Coac...	?	?	0
14	1145138345980366850	Jun 30, 201...	IURYPAIVA	9048697095...	?	-1	en	<a href=...	IAAF releases Consensus Statement on Nutrition for Athl...	?	?	0
15	1145136745400455168	Jun 30, 201...	The Black Le...	9894318538...	?	-1	en	<a href=...	Talib Kweli And Jemele Hill Speak On LeBron James, Do...	?	?	0
17	1145029240145862656	Jun 29, 201...	Sai reddy	2167064587	ysjagan	2965511...	en	<a href=...	@ysjagan Sir, if you could please look in to the link about ...	?	?	0
18	1145025024186826752	Jun 29, 201...	stephanie jon...	565498955	?	-1	en	<a href=...	Sean Clores Named CCNY Head Men's Basketball Coac...	?	?	0
19	1145012281023090688	Jun 29, 201...	Jen Metcalf	2550483518	Leidig2...	1168501...	en	<a href=...	@Leidig24Team Athlet Heilig, as ever, makes infinitely be...	?	?	0
20	1144896537832505349	Jun 29, 201...	rudra □□□□□ r...	155885405	imranir...	3799703...	en	<a href=...	@imranirampal @TheHockeyIndia @Media_SAI @Nike ...	?	?	0
21	1144797133335388162	Jun 29, 201...	DJ MAD MIKE	1047958345...	?	-1	en	<a href=...	When will we wake up?	?	?	0
23	1144753781483233280	Jun 29, 201...	♥□ Charito ♥□	2179046962	CNN	759251	en	<a href=...	@CNN Good athlet but to controversial....	?	?	0
24	1144661968114069505	Jun 28, 201...	Sabrina□IBel...	85036783	olympic...	3163717...	en	<a href=...	@olympicchannel The First Athlet that made ne Fall in Lo...	?	?	0
25	1144661958437855232	Jun 28, 201...	The Betting S...	380539184	undisp...	7659058...	en	<a href=...	@undisputed @RealSkipBayless Jordan is the greatest ...	?	?	0
31	1144623994206871552	Jun 28, 201...	Collective Acti...	1015239163...	?	-1	en	<a href=...	RT SIGAlliance "RT FAMilliat_en: #SIGAWomen - Beginni...	?	?	0
32	1144619180873912325	Jun 28, 201...	designlab m...	21021588	?	-1	en	<a href=...	Sean Clores Named CCNY Head Men's Basketball Coac...	?	?	0
33	1144617475738349570	Jun 28, 201...	Leah □□□	33649176	rege_ry...	3775152...	en	<a href=...	@rege_ryan Everyone's favorite doom and gloom the athl...	?	?	0
34	1144586657817079808	Jun 28, 201...	Isa - France	2999436161	?	-1	en	<a href=...	waouh, this music for an athlet ! yeh ! https://t.co/ZpcEwpY...	?	?	0
35	1144573670108663809	Jun 28, 201...	Cindy Stovall	27267984	?	-1	en	<a href=...	Sean Clores Named CCNY Head Men's Basketball Coac...	?	?	0
36	1144570799312867329	Jun 28, 201...	Faisal Mahm...	2508175433	?	-1	en	<a href=...	30 years ago, Mike Tyson hired Donald Trump to be his p...	?	?	0
37	1144437103310360576	Jun 28, 201...	Manny	952471908	?	-1	en	<a href=...	Yunevich Featured on NFF College Football Hall of Fame ...	?	?	0
38	1144409852963250176	Jun 28, 201...	Pants, Men's ...	474654248	?	-1	en	<a href=...	(eBay Sponsored) ONE PAIR COOPER SOLE CREW SO...	?	?	0

Slika 26. Tabela sa twittovima



Row No.	Id	Created-At	Retweet-Co...	token_number
1	1144797133335388162	Jun 29, 2019 ...	0	3
2	1143496708334243840	Jun 25, 2019 ...	0	2
3	1143239397569191941	Jun 24, 2019 ...	0	2
4	1142904414598049792	Jun 23, 2019 ...	1	2
5	1145553581010907136	Jul 1, 2019 6:...	0	1
6	1145543159671214081	Jul 1, 2019 6:...	1	1
7	1145541800242716673	Jul 1, 2019 5:...	1	1
8	1145508988626321408	Jul 1, 2019 3:...	0	1
9	1145418065749970944	Jun 30, 2019 ...	0	1
10	1145417002888126464	Jun 30, 2019 ...	0	1
11	1145412243280093184	Jun 30, 2019 ...	2	1
12	1145403502702800897	Jun 30, 2019 ...	2	1
13	1145399691166789632	Jun 30, 2019 ...	2	1
14	1145291210254356480	Jun 30, 2019 ...	0	1
15	1145167900493459456	Jun 30, 2019 ...	0	1
16	1145161990928838657	Jun 30, 2019 ...	0	1
17	1145138345980366850	Jun 30, 2019 ...	0	1
18	1145136745400455168	Jun 30, 2019 ...	0	1
19	1145029240145862656	Jun 29, 2019 ...	0	1
20	1145025024186826752	Jun 29, 2019 ...	0	1
21	1145012281023090688	Jun 29, 2019 ...	0	1
22	1144896537832505349	Jun 29, 2019 ...	0	1
23	1144753781483233280	Jun 29, 2019 ...	0	1
24	1144661968114069505	Jun 28, 2019 ...	0	1
25	1144661958437855232	Jun 28, 2019 ...	0	1
26	1144634547490791424	Jun 28, 2019 ...	6	1
27	1144623994206871552	Jun 28, 2019 ...	0	1
28	1144619180873912325	Jun 28, 2019 ...	0	1
29	1144617475738349570	Jun 28, 2019 ...	0	1
30	1144586657817079808	Jun 28, 2019 ...	0	1
31	1144573670108663809	Jun 28, 2019 ...	0	1
32	1144570799312867329	Jun 28, 2019 ...	0	1
33	1144437103310360576	Jun 28, 2019 ...	0	1
34	1144409852963250176	Jun 28, 2019 ...	0	1
35	1144281179178143745	Jun 27, 2019 ...	0	1
36	1144259040555536384	Jun 27, 2019 ...	0	1
37	1144243798861631489	Jun 27, 2019 ...	0	1
38	1144242451802140673	Jun 27, 2019 ...	0	1
39	1144091277409030144	Jun 27, 2019 ...	0	1
40	1144088997255438341	Jun 27, 2019 ...	0	1
41	1144003154981400577	Jun 27, 2019 ...	0	1
42	1143913639486873600	Jun 26, 2019 ...	0	1
43	1143774201259659264	Jun 26, 2019 ...	0	1
44	1143719126306873345	Jun 26, 2019 ...	0	1
45	1143685487523192833	Jun 26, 2019 ...	0	1
46	1143655065523240961	Jun 26, 2019 ...	5	1
47	1143645689886445568	Jun 26, 2019 ...	0	1
48	1143638027773849600	Jun 25, 2019 ...	0	1
49	1143622934369411072	Jun 25, 2019 ...	5	1
50	1143621064053940225	Jun 25, 2019 ...	5	1
56	1143343565289844736	Jun 25, 2019 ...	0	1
57	1143248226906202113	Jun 24, 2019 ...	0	1
58	1143220591505395714	Jun 24, 2019 ...	0	1
59	1143217323626651648	Jun 24, 2019 ...	0	1
60	1143217295663280129	Jun 24, 2019 ...	0	1
61	1143124247939821570	Jun 24, 2019 ...	0	1
62	1143070404342353920	Jun 24, 2019 ...	1	1
63	1142914967286550529	Jun 23, 2019 ...	0	1
64	1142914040336932865	Jun 23, 2019 ...	0	1
65	1142835305256497155	Jun 23, 2019 ...	0	1
66	1142788860486791169	Jun 23, 2019 ...	0	1
67	1142764815959678976	Jun 23, 2019 ...	0	1
68	1142333886921187328	Jun 22, 2019 ...	0	1
69	1142322559666577408	Jun 22, 2019 ...	0	1
70	1142287988711481344	Jun 22, 2019 ...	1	1
71	1145178227276574720	Jun 30, 2019 ...	366	0
72	1145067065872719873	Jun 29, 2019 ...	1	0
73	1144779214480138240	Jun 29, 2019 ...	6	0
74	1144654192356646912	Jun 28, 2019 ...	6	0
75	1144641992317497345	Jun 28, 2019 ...	6	0

Slika 27. Twittovi koji sadrže riječ *athlet*

Sa slike 27 možemo vidjeli da 70 od 1000 twittova sadrži riječ athlet. Tako naprimjer prvi twitt sa id-om: 1144797133335388162 sadrži čak tri riječi, a orginalni tekst twitta je (očitamo iz tabele sa slike 26.):

When will we wake up?
#run #athlete #AthleticClub
#athlet #transpride
#TransIsBeautiful #Trump2020
#woman #realwoman #real #not #a #dream #twisted #reality #WomensWorldCup2019
#WomensWorldCup
https://t.co/LtYQcK22iV

Iz orginalnog teksta twit-a možemo vidjeli da on stvarno sadrži tri riječi s korijenom athlet.

Analogno je urađeno i za riječ world i dobijeno je da 62/1000 sadrže riječ world, dok obje riječi world i athlete zajedno (sadržaj twittova je filtriran po obje riječi) sadrži 14/1000 twittova.

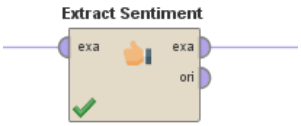
Sentimentalna analiza za dokumente iz jedne grupe klasifikacije

Sentimentalna analiza se može opisati kao analiza mišljenja, osjećanja i subjektivnosti teksta. Postoji određena razlika u razumijevanju mininga mišljenja i sentimentalne analize. Mininig mišljenja izvlači i analizira mišljenja korisnika o nekom entitetu, dok sentimentalna analiza identificira i analizira osjećanja izražena u tekstu. Detekcija osjećanja (eng. emotion detection) je zadatak sentimentalne analize koji podrazumijeva identificiranje različitih osjećanja iz teksta.

U Rapidminneru (kao i u ostalim alatima) se mogu koristiti dvije tehnike sentimentalne analize:

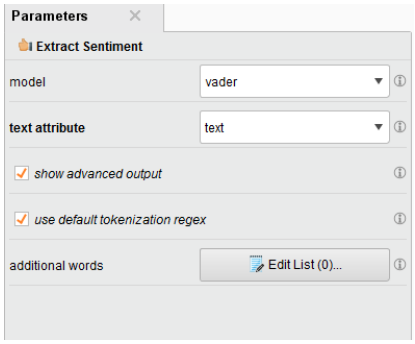
1. Leksikon bazirana
2. Model bazirana, koja koristi machine learning

Ovdje će se koristiti leksikon bazirana sentimentalna analiza. Sentimentalna analiza će biti urađena na kategoriji Sport. Blok koji je korišten u Rapidminner-u je **Extract Sentiment** (slika 28.).



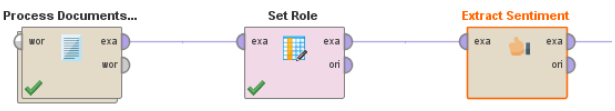
Slika 28. Extract sentiment blok

Ovaj operator kreira sentiment score svake riječi u tekstu (dokumentu) koristeći neki od leksikona. Postavljeni parametri ovog operatora su prikazane na slici 29.



Slika 29. Postavke parametara bloka Extract sentiment

Na osnovu prethodnih postavki možemo uočiti da je korišteni leksikon tipa **vader. VADER** (*Valence Aware Dictionary and sEntiment Reasoner*) je leksikon i pravilo-baziran alat koji je osmišljen prvenstveno za sentimentalnu analizu u tekstovima socijalnih medija. Ovaj operator računa sentiment score svake riječi u tekstu i na kraju sve te score-ove sumira. Ukoliko se koristi advanced output bit će predstavljen i nominalni atribut sa svim riječima koje učestvuju u score-u, sumu pozitivnih komponenti i sumu negativnih komponenti i broj korištenih i nekorištenih tokena. Također, na kraju se specificira text attribute u koji se unosi naziv atributa koji sadrži tekst koji će se obrađivati. Blok dijagram za cijeli proces prikazan je na slici 30. Prva dva bloka Process Documents from Files i Set Role su isti kao i u prethodnim zadacima.



Slika 30. Blok dijagram sentimentalne analize

Rezultati procesa su prikazani na slici 31.

Row No.	label	text	Score	Scoring String	Negativity	Positivity	Uncovered Tokens	Total Tokens
1	Sport	Radcliffe yet ...	3.103	granted (0.26) championships (0.56) upset (-0.41) no (-0.31) huge (0.33) asset (0.38) fantastic ...	1.923	5.026	216	230
2	Sport	Edwards tip...	6.487	championships (0.56) well (0.28) win (0.72) medal (0.54) ability (0.33) best (0.82) medal (0.54) ...	0.462	6.949	248	263
3	Sport	Kenya lift Ch...	1.333	ban (-0.67) apology (0.05) suspended (-0.54) failing (-0.59) ban (-0.67) accepted (0.28) apology...	4.974	6.308	343	371
4	Sport	McIlroy aimin...	11.410	confident (0.56) win (0.72) championships (0.56) great (0.79) wins (0.69) promise (0.33) rewar...	2.821	14.231	529	562
5	Sport	UK Athletics ...	4.462	agrees (0.21) agreed (0.28) great (0.79) championships (0.56) championships (0.56) delighted ...	0	4.462	160	170
6	Sport	Verdict delay...	-1.641	verdict (0.15) delay (-0.33) postponed (-0.21) missing (-0.31) missed (-0.31) won (0.69) won (0....	5.846	4.205	285	308
7	Sport	Call for Kent...	-0.000	cleared (0.10) charges (-0.28) champion (0.74) no (-0.31) failing (-0.59) verdict (0.15) missing (-...	2.564	2.564	279	294
8	Sport	Merritt close ...	6.308	champion (0.74) well (0.28) clear (0.41) missed (-0.31) inferior (-0.44) excellent (0.69) winning (...	1.282	7.590	272	289
9	Sport	London hop...	2.026	hope (0.49) hoping (0.46) banned (-0.51) suspended (-0.54) failing (-0.59) hoping (0.46) satisfa...	1.641	3.667	149	160
10	Sport	Edwards tip...	6.487	championships (0.56) well (0.28) win (0.72) medal (0.54) ability (0.33) best (0.82) medal (0.54) ...	0.462	6.949	248	263

Slika 31. Rezultati sentimentalne analize

Atribut *Scoring String* određuje svaku riječ text-a da li je ona pozitivna, negativna ili neutralna na sljedeći način:

1. Pozitivan: kada je score riječi >= 0.05
2. Neutralan: kada je score riječi > -0.05 i < 0.05
3. Negativan: kada je score riječi <= -0.05

Atribut *Score* određuje ukupan score teksta (što je veća vrijednost tekst je pozitivniji i obrnuto). Atribut *Negativity* predstavlja zbir svih negativnih riječi u tekstu, a atribut *Positivity* zbir svih pozitivnih riječi u tekstu. Atribut *Uncoverd Tokens* pokazuje koliko riječi nije učestvovalo u ovoj analizi dok atribut *Total Tokens* pokazuje ukupan broj riječi teksta/dokumenta.

Iz tabele rezultata možemo učiti da tekst/dokument br. 4 ima najveći score = 11.410 i ukoliko pogledamo koje riječi taj tekst sadrži vidimo da sadrži jako puno pozitivnih riječi kao naprimjer: *confident(0.56), win(0.72), championship(0.56), great(0.79), wins(0.69)* itd. Možemo čak na osnovu ovih nekoliko riječi da ovaj tekst govori o pobjedi nekog tima ili pojedinca na nekom velikom šampionatu, o njihovoj samouvjerenoj pobjedi i sl. Najnegativniji score ima tekst 6, score = -1.641, koji sadrži jako puno negativnih riječi kao naprimjer: *delay(-0.33), postponed(-0.21), missing(-0.31), ban(0.67), guillty(-0.46), suspended(-0.54)* itd. O ovom tekst možemo zaključiti da se na ovom sportskom turniru/takmičenju desilo neko kašnjenje te da je neka osoba suspenzirana, ko se osjećao krivim za te situacije i sl. Slični zaključci mogu biti doneseni i o drugim dokumentima u tabeli.

Zadatak 3

Istražiti način provođenja Topic Modelinga nad dokumentima i primijeniti ga (implementirati) za dokumente iz 1.

Rješenje

Topic modeling

Velike količine podatka se generiraju svakog dana. Kako sve više informacija postaje dostupno, postaje teže naći ono što nam je potrebno, pa su potrebni neki alati, tehnike kako bismo organizirali, pretražili i razumjeli ogromne količine informacija.

Topic modeling omogućava metode za organiziranje, razumijevanje i sumiranje velike količine tekstualnih informacija, i to:

- otkrivanje skrivenih tema u kolekciji tekstova/dokumenata
- označavanje dokumenata na osnovu tih tema
- korištenje tih oznaka da bismo organizirali, pretražili i sumirali tekstove.

Topic modeling može biti opisan kao metoda za pretragu grupa riječi (tema) iz kolekcije dokumenata koji najbolje predstavljaju informacije u kolekciji. Može biti predstavljen i kao forma text mininga – način prikupljanja paterna riječi u tekstovima.

Postoje mnoge različite tehnike koje se koriste za topic modele. Jedna od njih **je LDA (Latent Dirichlet Allocation)** metoda koja se jako puno koristi i **TextRank proces** koji predstavlja algoritam baziran na grafovima kako bi se ekstraktovale relevantne ključne fraze.

Latent Dirichlet Allocation (LDA)

U LDA modelu, svaki dokument je predstavljen kao mješavina tema koje su predstavljene u korpusu. Model pretpostavlja da se svaka riječ u dokumentu pripisuje jednoj od tema dokumenta. Naprimjer, ako razmatramo sljedeći skup dokumenata kao korpus:

Dokument 1: *I had a peanut butter sandwich for breakfast.*

Dokument 2: *I like to eat almonds, peanuts and walnuts.*

Dokument 3: *My neighbor got a little dog yesterday.*

Dokument 4*: Cats and dogs are mortal enemies.*

Dokument 5: *You mustn't feed peanuts to your dog*.

LDA model otkriva koje različite teme dokumenti sadrže i koliko je svaka tema zastupljena u dokumentu. Naprimjer, LDA daje sljedeći rezultat:

Topic 1: *30% peanuts, 15% almonds, 10% breakfast ... (može biti interpretirano da se tema odnosi na hranu).*

Topic 2: *20% dogs, 10% cats, 5% peanuts .... (može biti interpretirano da se ova tema odnosi na ljubimce ili životinje).*

Pa zaključujemo da je:

Dokumenti 1 i 2: *100% Topic 1.*

Dokumenti 3 i 4: *100% Topic 2.*

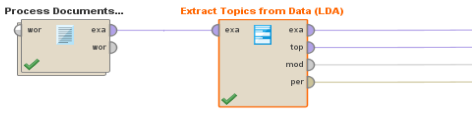
Dokument 5: *10% Topic 1, 30% Topic 2.*

Kako LDA uradi prethodni proces? Collapsed Gibbs sempliranje je jedan od načina kako LDA nauči teme i reprezentacije tih tema u svakom dokumentu. Procedura je sljedeća:

1. Prođi kroz svaki dokument i dodijeli random svaku riječ jednoj od K tema (K je određeno unaprijed).
2. Ovo random dodijeljivanje odredi neku reprezentaciju tema u svim dokumentima i raspodijelu riječi svake teme, iako ne dobru.
3. Potrebno je ovo unaprijediti:
   a. Za svaki dokument d, idi kroz svaku riječ w i izračunaj:
      - p(topic t | dokument d): proporciju riječi u dokumentu d koje su dodijeljenje temi t.
      - p(word w | topic t): proporcije dodijela temi t, kroz sve dokumente d, koje dolaze od riječi w.
4. Dodijeli ponovo riječ w novoj temi t', gdje odaberemo temu t' sa vjerovatnoćom *p(topic t' | document d)*p(word w | topic t')*. Ovaj generativni model predviđa vjerovatnoću da je tema t' generirala rijec word w.

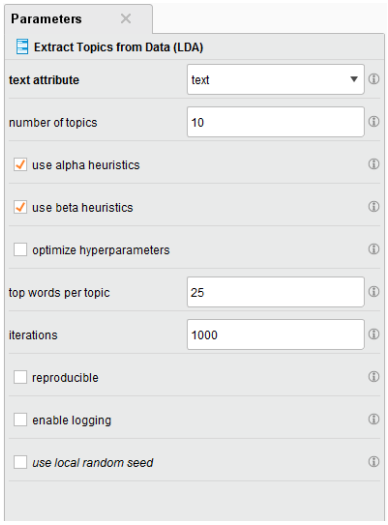
Ponavljajući posljednji korak veliki broj puta dosegne se stacionarno stanje gdje su dodjele temama vrlo dobre. Dodjele su onda korištene da se odredi mix tema svakog dokumenta.

Topic Modeling je urađen na dokumentima iz kategorije Sport da bi se uočilo o kojoj temi tačno svaki dokument govori i da li su dokumenti povezani na neki način (imaju zajedničku temu). Uzeli smo samo 10 dokumenata iz kategorije Sport da bismo lakše uočili sličnosti itd. Shema u Rapidminneru prikazana je na slici 32.



Slika 32. Topic Modelig u Rapidminner-u

Kao što možemo vidjeti na slici 32. u Rapidminneru već postoji gotov operator za LDA tehniku modelinga, njegovi parametri su prikazani na slici 33.



Slika 33. Parametri bloka LDA

Na slici 33. vidimo da je odabrano da se pretražuje 10 tema (jer imamo 10 dokumenata, mogli smo uzeti i različit broj) i da je odabrano da se uzima 25 riječi po temi.

Rezultati su sljedeći (slika 34.). *Prediction(Topic)* predstavlja rezultat tj. gdje je LDA smjestio dati dokument kojoj temi pripada, ostali atributi *confidence(Topic_x)* govore kolika je pouzdanost da dokument pripada svakoj temi od njih 10, zatim *text* atribut koji sadrži orginalni tekst i dodatni metapodaci.

Row No.	documentid	label	prediction(Topic)	confidence(Topic_0)	confidence(Topic_1)	confidence(Topic_2)	confid...	confi...	confi...	co...	confi...	confi...	co...	text	metadata_file	metadata_p...	m
1	0	Sport	Topic_7	0.102	0.041	0.198	0.076	0.091	0.066	0.041	0.289	0.046	0.051	Radcliff...	011.txt	C:\Users\Emi...	M
2	1	Sport	Topic_9	0.057	0.061	0.061	0.026	0.031	0.044	0.083	0.026	0.057	0.555	Edward...	012.txt	C:\Users\Emi...	M
3	2	Sport	Topic_7	0.066	0.083	0.194	0.080	0.048	0.093	0.042	0.294	0.038	0.062	Kenya li...	013.txt	C:\Users\Emi...	M
4	3	Sport	Topic_5	0.109	0.149	0.097	0.084	0.029	0.308	0.088	0.034	0.036	0.066	McIlroy ...	014.txt	C:\Users\Emi...	M
5	4	Sport	Topic_3	0.176	0.147	0.082	0.188	0.088	0.035	0.065	0.053	0.112	0.053	UK Athl...	015.txt	C:\Users\Emi...	M
6	5	Sport	Topic_4	0.116	0.127	0.154	0.049	0.330	0.041	0.067	0.026	0.041	0.049	Verdict ...	016.txt	C:\Users\Emi...	M
7	6	Sport	Topic_4	0.061	0.087	0.183	0.084	0.369	0.034	0.076	0.046	0.030	0.030	Call for ...	017.txt	C:\Users\Emi...	M
8	7	Sport	Topic_8	0.043	0.078	0.130	0.039	0.043	0.052	0.074	0.061	0.398	0.082	Merritt cl...	018.txt	C:\Users\Emi...	M
9	8	Sport	Topic_7	0.106	0.100	0.075	0.119	0.031	0.138	0.050	0.219	0.062	0.100	London ...	019.txt	C:\Users\Emi...	M
10	9	Sport	Topic_9	0.074	0.039	0.057	0.052	0.031	0.026	0.048	0.031	0.066	0.576	Edward...	020.txt	C:\Users\Emi...	M

Iz date tabele možemo uočiti da dokumenti koji imaju id 5 i 6 pripadaju temi 4, analizirat ćemo orginalni tekst da bismo se uvjerili da je to stvarno tako. Tekst dokumenta 6 i 5 je:

DOKUMENT 6:

Call for Kenteris to be cleared Kostas Kenteris' lawyer has called for the doping charges against the Greek sprinter to be dropped. Gregory Ioannidis has submitted new evidence to a Greek athletics tribunal which he claims proves the former Olympic champion has no case to answer. Kenteris and compatriot Katerina Thanou were given provisional suspensions in December for failing to take drugs tests before the Athens Olympics. The Greek tribunal is expected to give its verdict early next week. Kenteris and Thanou withdrew from the Athens Olympics last August after missing drugs tests on the eve of the opening ceremony. They were also alleged to have avoided tests in Tel Aviv and Chicago before the Games. But Ioannidis said: "Everything overwhelmingly shows that the charges should be dropped." Ioannidis also said he has presented evidence that will throw a different light on the events leading up to the pair's

DOKUMENT 5:

Verdict delay for Greek sprinters Greek athletics' governing body has postponed by two weeks the judgement on sprinters Costas Kenteris and Katerina Thanou for missing doping tests. The pair are facing lengthy bans for the missed tests, including one on the eve of last year's Athens Olympics. They were set to learn their fate by the end of February, but late evidence from them has pushed the date back. "A decision is now expected by around mid-March," said one of their lawyers, Michalis Dimitrakopoulos. Kenteris, 31, who won the men's 200m title at the 2000 Sydney Games and Thanou, 30, who won the women's 100m silver medal in Sydney, face a maximum two-year ban if found guilty. The athletes, who spectacularly withdrew from the Athens Olympics, have been suspended by the International Association of Athletics Federations

Vidimo da oba dokumenta govore o grčkim atletičarima, sprinterima, Costas Kenteris i Katerina Thanou koji su propustili doping test i o toj situaciji. Analogno se može uraditi za sve ostale dokumente koje pripadaju istoj temi. Naravno teme se i preklapaju pa svaki dokument na osnovu mjere confident pripada i svim drugim temama u nekom procentu.