Domaći zadatak
Procesiranje prirodnih jezika

Zadatak 2.
Klasifikacija dokumenata

Darko Sarajkić-Markelić
1142

- # Uvod

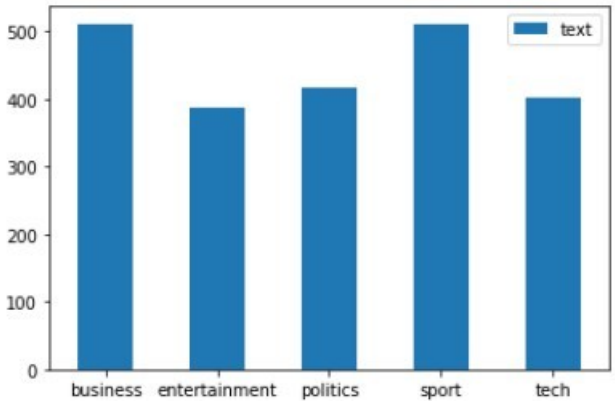
U ovom domaćem zadatku obrađena je tema klasifikacije dokumenata. Cilj ovog zadataka je da se utrenira klasifikator koji na osnovu preprocesiranog teksta može da izvrši klasifikaciju tekstova u neke kategorije. Za predstavljanje podataka korišćen je BagOfWords model.

- # Skup podataka

Skup podataka se sastoji od 2225 novinskih članaka sa portala BBC koji je dostupan na https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification koji su podeljeni u pet kategorija:

1. Business
2. Entertainment
3. Politics
4. Sport
5. Tech

Skup podataka je prilično dobro balansiran (Slika 1.) i broj tekstova po kategoriji je prilično uravnotežen pa nije potrebno vršiti dodatno balansiranje skupa podataka.



Slika 1. Broj tekstova po kategoriji

Pre obrade teksta, prilikom učitavanja teskta u dataframe, svakom tekstu dodeli se labela koja označava klasu kojoj pripada na osnovu file-a u kome se tekst nalazi (Slika 2.).

	text	label
0	Ad sales boost Time Warner profit\n\nQuarterly profits at US media giant TimeWarner jumped 76% t...	business
1	Dollar gains on Greenspan speech\n\nThe dollar has hit its highest level against the euro in alm...	business
2	Yukos unit buyer faces loan claim\n\nThe owners of embattled Russian oil giant Yukos are to ask ...	business
3	High fuel prices hit BA's profits\n\nBritish Airways has blamed high fuel prices for a 40% drop ...	business
4	Pernod takeover talk lifts Domecq\n\nShares in UK drinks and food firm Allied Domecq have risen ...	business
5	Japan narrowly escapes recession\n\nJapan's economy teetered on the brink of a technical recessi...	business

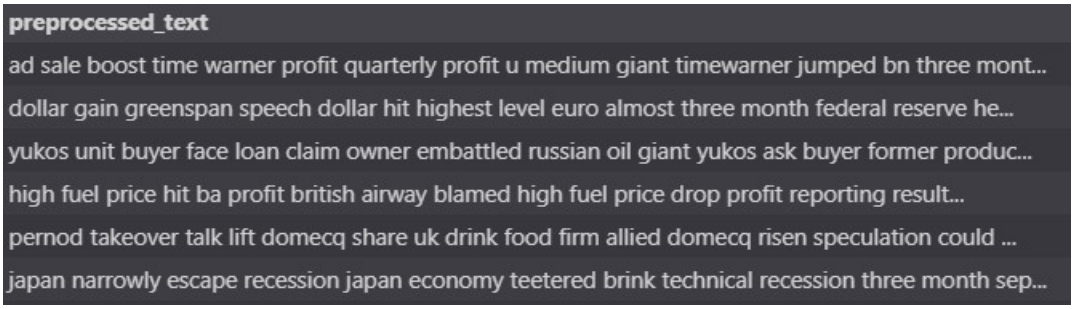
Slika 2. Izgled podataka u dataframe-u

- # Preprocesiranje teksta

Prvi korak koji je potrebno da se učini pre treniranja klasifikatora je preprocesiranje teksta. U ovom zadataku preprocesiranje je odrađeno kroz sledeće korake:

1. Konverzija svih slova u mala slova
2. Uklanjanje svih karaktera koji nisu mala slova (brojevi, interpunkcijski znaci...)
3. Uklanjanje stop reči engelskog jezika iz rečnika koji se nalazi u NLTK biblioteci
4. Lematizacija uz pomoć WordNet lematizatora iz NLTK biblioteke

Nakon preprocesiranja tekstovi su svedeni na oblik (Slika 3.) pogodan za primenu BagOfWords modela koji je potreban da bi se iz njih izvukli atributi koji su pogodni za treniranje klasifikatora.



Slika 3. Tekstovi nakon preprocesiranja

- # Klasifikacija

Najpre je potrebno, pre treniranja klasifikatora, podeliti skup podataka na trening i test skup. Prilikom ove podele korišćena je funkcija *train_test_split()* iz biblioteke *sklearn*, koja vrši podelu. Važno je podesiti uz pomoć parametra *stratify* da se prilikom podele zadrži odnos između svih klasa u trening i test skupu. Ovo se radi da bi se sprečila situacija da se u test ili trening skupu nalazi mnogo veći broj podataka iz jedne klase što bi moglo da da pogrešnu sliku o kvalitetu klasifikatura prilikom testiranja.

Sledeći korak u pripremi za klasifikaciju je izdvajanje atributa iz preprocesiranog teksta. Koristi se BagOfWords model kod koga se kreira matrica gde svaka vrsta predstavlja jedan tekst iz skupa podataka dok je svaka kolona jedena reč iz skupa svih reči koje se javljaju u korpusu. Vrednosi ćelija u matrici zavise od izbora mere koja se koristi. Može se koristiti binarna mera kod koje se sa 0 označava da se u dokumentu koji je određen vrstom ne nalazi reč određena kolonom i obratno za 1. U ovom zadataku su korišćene sledeće mere:

- **TF(Term Frequency) mera**: računa se kao odnos koliko puta se određeni pojam pojavi u tekstu i ukupnog broja pojmova u tekstu.

- **TF-IDF(Term Frequency- Inverse Document Frequency) mera**: slična mera kao i TF ali dodaje se još jedan član prilikom računjana IDF koji određuje važnost određene reči koje se pojavljuje u tekstu u odnosu na ceo skup tekstova. Time se rešava problem da se neke reči čija je vrednost TF-a velika javljaju često u svim tekstovima iz korpusa pa samim tim nemaju veliki značaj prilikom klasifikacije**.**

Za kreiranje BagOfWords modela sa TF i TF-IDF merama korišćena je klasa *TfidfVectorizer* iz biblioteke *sklearn*. Razlika između vektorizatora za dve mere je u postavljanju atributa *use_idf* na *False* u slučaju da želimo da koristimo TF meru. Takođe kreirano je više vektrizatora za svaku meru sa različitim brojem atributa da bi se uporedila razlika u preciznosti klasifikatora sa većim i manjim brojem atributa. Bitno je napomenuti da klasa *TfidfVectorizer* koristi nešto izmenjenu formulu za računanje zato što nakon računjana TF i TF-IDF mera koristi normalizaciju pa se dobijeni vektori razliku od onih koji se dobijaju primenom osnovnih formula za ove dve mere.

Za klasifikaciju se koristi Naivni Bajesov klasifikator implementira u biblioteci *sklearn*. Korišćene su dve verzije Naivnog Bajesovog klasifikatora *MultinomialNB* i *ComplementNB* koje su preporučene za korišćenje prilikom klasifikacije teksta na zvaničnom sajtu.

vectorizer	score
tf_idf_max	96.179775
tf_idf_20000	96.179775
tf_idf_10000	96.179775
tf_idf_5000	96.179775
tf_5000	95.730337
tf_10000	95.505618
tf_idf_1000	95.280899
tf_max	95.05618
tf_20000	95.05618
tf_1000	95.05618
tf_idf_500	94.831461
tf_500	94.831461
tf_idf_100	87.191011
tf_100	85.842697
tf_50	81.123596
tf_idf_50	80.674157

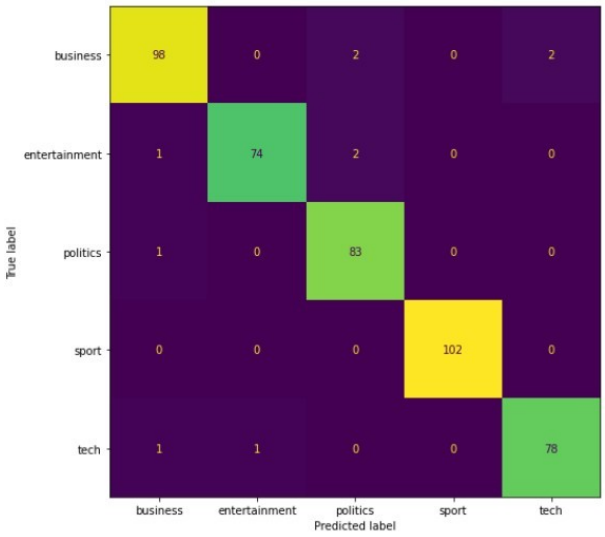
Slika 4. Preciznost klasifikacije za MNB klasifikator sa različitim vektorizatorima

vectorizer	score
tf_idf_max	97.52809
tf_idf_20000	97.52809
tf_idf_10000	97.303371
tf_idf_5000	97.303371
tf_max	96.179775
tf_20000	96.179775
tf_10000	96.179775
tf_5000	96.179775
tf_1000	95.280899
tf_idf_1000	95.05618
tf_500	94.606742
tf_idf_500	93.932584
tf_idf_100	85.168539
tf_100	84.269663
tf_50	75.280899
tf_idf_50	74.606742

Slika 5. Preciznost klasifikacije za CNB klasifikator sa različitim vektorizatorima

Može se uočiti (Slika 4. , Slika 5.) da prilikom klasifikacije bolje rezultate daje *ComplementNB*. Takođe se može primetiti da za oba klasifikatora nešto bolje rezultate daje TD-IDF mera što je i očekivano ako znamo da se njenim korišćenjem dobija vrednost vektora koja uzima u obzir ceo korpus tekstova a ne samo jedan tekst kao kod TF mere.

Na osnovu prethodnog zaključka izabran je *ComplementNB* klasifikator i TF-IDF mera za detaljniju analizu rezultat klasifikacije. Sa kofunzione matrice (Slika 6.) možemo učiti da su sve klase dobro klasifikovane i da ne postoji neki očigledni problemi koji bi trebalo ispraviti



Slika 6. Konfuziona matrica

Upoređivanjem predviđenih klasa i stvarnih klasa (Slika 7.) za test skup podataka dolazimo do zaključka da se najčešće greške dešavaju kada se radi o tekstovima koji su na neki način povezani sa dve oblasti. Na primer ako pogledamo prvi pogrešno klasifikovani tekst vidimo da se javljaju reči “bbc”, “bafta”, “theatre” i klasifikator je ovaj tekst svrstao u zabavu dok zapravo pripada grupi tehnoloških tekstova.

preprocessed_text	true_label	predicted_label
bbc lead interactive bafta win bbc national theatre led field year interactive bafta award natio...	tech	entertainment
making office work mission brighten working life continues time taking long hard look office nex...	business	tech
gallery unveils interactive tree christmas tree receive text message unveiled london tate britai...	entertainment	tech
call save manufacturing job trade union congress tuc calling government stem job loss manufactur...	business	politics
ask jeeves tip online ad revival ask jeeves become third leading online search firm week thank r...	business	tech
china ripe medium explosion asia set drive global medium growth beyond china india filling two t...	tech	business
golden rule boost chancellor chancellor gordon brown given bn boost attempt meet golden economic...	business	politics
tv show unites angolan family angolan family attempting track separated nearly year war succeedi...	entertainment	politics
holmes win top tv moment sprinter kelly holmes olympic victory named top television moment bbc p...	entertainment	sport
baa support ahead court battle uk airport operator baa reiterated support government aviation ex...	politics	business
uk national gallery pink national gallery home uk greatest artwork seen big jump visitor number ...	entertainment	business
muslim group attack tv drama british muslim group criticised new series u drama aired sky one cl...	entertainment	politics

Slika 7. Poređene tekstova iz test skupa koji su pogrešno klasifikovani

Na kraju izvršena je provera klasifikatora na slučajno odabranom tekstu sa novinskog portala da bi se videlo da li klasifikator dobro radi sa tekstovima koji nisu iz originalnog skupa. Odabran je tekst:

"For all the young dogs waiting for a go at Novak Djokovic, it is increasingly hard to imagine any scenario other than them rolling over for a tickled belly when their moment comes.
There has long been a perceived flimsiness around the next generation of men's tennis, who have spent so much time waiting in line that there is a danger of them missing their turn altogether.
As we begin the second week of Wimbledon, it is necessary to again wonder whether the stasis at the top of the game is down to the astonishing endurance and gifts of a few old boys, or the deficiencies of those in pursuit.
Plainly the answer covers bits of both, but with each passing Slam it is getting harder to mount a defence for younger guys who are suddenly not so young any more.

It is an awkward fact that since Stan Wawrinka won the 2016 US Open, there have been 17 Slams and only one has not been claimed by Djokovic, Rafael Nadal or Federer.
That was Dominic Thiem at last year's US Open, to go with three other finals, but he will turn 28 later this year and is missing these championships with an injury.
Those present and remaining in the draw have a chance across the next week, but it is difficult to make a serious case for anyone beyond Djokovic, irrespective of the Serb not yet looking the full ticket.
Other than his second-round win over Kevin Anderson, in which he was excellent, Djokovic had iffy moments against Jack Draper and Denis Kudla, and his difficult relationship with the Wimbledon crowd is a developing situation.
But for those minor considerations, he remains an overwhelming favourite to take his record-equalling 20th Slam on Sunday.

Even at 34, he is a head, shoulders and most of a torso above the chasing pack, with the next modest barrier to a third straight Wimbledon crown being the 17th seed, Cristian Garin of Chile.
They will play on Centre Court on Monday, where part of the intrigue will concern any further blowouts at those in attendance. 'You know that I play 90 per cent of matches against the audience, the field, the opponents and everyone alive,' said Djokovic.
'It's something I'm used to, but I'm a man of flesh and blood. I can't always stay calm. 'When someone provokes me, when it crosses the line of taste, sportsmanship and respect, then I show him where he belongs.'
Garin, a clay-court specialist with no huge weapon, is part of a curious trend at this tournament, whereby no fewer than 13 of the last 16 are on their best Wimbledon run. Only Djokovic, Federer and Roberto Bautista Agut have gone further, and none of that trio is younger than 33."

Nakon primene preprocesiranja i vektorizacije teksta klasifikator je ispravno klasifikovao ovaj tekst u kategoriju *sport*.