Klasifikacija novinskih članaka u kategorije

Đorđe Ivković, SW54-2016
Fakultet tehničkih nauka, Novi Sad

PROBLEM

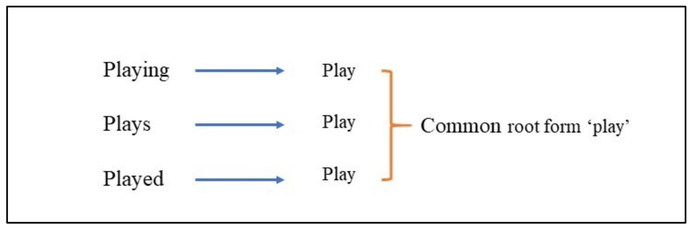
Kreirati model koji ce sa što većom preciznošću klasifikovati novniske članke u predefinisane kategorije. Problem je to što su podaci zadati u tekstualnom formatu, potrebno je na efikasan način modelovati tekstualne dokumente tako da oni odgovaraju ML modelima. Jedno od mogućih rešenja su BoW modeli.

METODOLOGIJA

1. Pre-procesiranje teksta
2. Kreiranje vokabulara
3. S elekcija značajnijih obeležja
4. Obučavanje različitih modela

PRE-PROCESIRANJE TEKSTA

- Uklanjanje svih specijalnih karaktera osim razmaka
- Pretvaranje velikih slova u mala
- Uklanjanje "stopwords" koje ne doprinose predikciji
- Menjanje reči sa njenim korenom upotrebom stemming/lemmatisation.



KREIRANJE VOKABULARA

Koristi se BoW model pomocu koga se kreira vokabular koji koristimo za predstavljanje dokumenata u numeričkom vektorskom obliku.

Za pridruživanje vrednosti rečima koristi se:

- Counting
- Word hashing
- TF IDF

SELEKCIJA NAJVAŽNIJIH OBELEŽJA

Posto je vokabular jako veliki, u ovom projektu preko 20 hiljada reči, potrebno je izvršiti selekciju reci koje najvise uticu na kategoriju novinskog clanka.
Nakon selekcije ostaje oko 1000-2000 najbitnijih reci. U projektu se koristiti:

- Anova-f

OBUČAVANJE MODELA

Pre-procesiranje podatke delimo u razmeri 60:40 na train i test podatke. Nakon toga obučavamo više različitih modela sa različitim kombinacijama pre-procesiranja:

- Support vector machine (SVC)
- K nearest neighbors
- Random Forests

Odabir modela će se vršiti na osnovu njegove tačnosti nad testnim skupom.

REZULTATI

Rezultati dobijeni stemming pre-procesiranjem:

	COUNT	HASH	TF-IDF
SVC	0.96	0.97	0.96
KNN	0.70	0.91	0.78
RF	0.95	0.96	0.96

Rezultati dobijeni lemmatisation pre-procesiranjem:

	COUNT	HASH	TF-IDF
SVC	0.97	0.97	0.97
KNN	0.73	0.90	0.84
RF	0.96	0.96	0.95

ZAKLJUČAK

Na osnovu tabele izabrali SVC ili Random Forest model sa stemming pre-procesiranjem pošto lemmatization ne dobijamo značajno bolje performanse dok je vreme pre-procesiranja mnogo duže.