# OTKRIVANJE MIŠLJENJA I ANALIZA STAVA IZ KOMENTARA O RESTORANIMA

Ovaj projekt nastao je na PMF-u u sklopu kolegija Strojno učenje. Bavi se otkrivanjem mišljenja i analize stava u recenzijama restorana prikupljenih sa internet stranice www.yelp.com.

## Korišteni alati

Programi su pisani u programskoj jeziku python i izvođeni na operacijskom sustavu Ubuntu. Paketi iz pythona korišteni u projektu:

- sys https://docs.python.org/3/library/sys.html
- urllib2 https://docs.python.org/2/library/urllib2.html
- numpy http://www.numpy.org/
- nltk http://www.nltk.org/
- scikit-learn http://scikit-learn.org/stable/
- matplotlib http://matplotlib.org/

## Baza podataka

Baza podataka se sastoji od 25 871 primjera koji su podijeljeni u tri skupine ovisno o stavu kojeg izražavaju : pozitivno(10 768), neutralno(6 555), negativno(8 548). Podaci su prikupljeni radom programa skini_komentare.py koji koristi pakete sys i urllib2 za rad s internet stranicama. Podaci su potom razdvojeni u dvije datoteke: jedna koja sadrži samo komentare te druga koja sadrži pripadni stav. To je zadatak programa razdvoji.py.

## Pretpocesiranje

Pretprocesiranje se sastoji od uklanjanja interpunkcijskih znakova i nebitnih riječi (stop words) te uklanjanja sufiksa koji se pojavljuju korištenjem gramatičkih pravila (word stemming) što se obavlja u programu pretpocesiranje.py. Korišten je paket nltk-natural language toolkit koji sadrži listu nebitnih riječi "stopwords.words('english')" i klasu "PorterStemmer()" koja koristi Porterov algoritam da bi uklonila sufikse.

## Reprezentacija riječi i odabir značajki

Riječi su reprezentirane metodom vreća riječi (bag of words) sa odgovarajućim TD-IDF-om upisanim u vektore. Za reprezentaciju korištena je klasa "TfidfVectorizer()" iz paketa scikit-learn. Funkcija za odabir značajki je također iz paketa scikit-learn. Riječ je o funkciji "SelectKBest()" korištenoj uz statistički $\chi^2$ test. Dva programa provode ovaj dio obrade (svaki za jedan klasifikator): bayes.py i svm.py.

## Parametri za SVM

SVM algoritam se koristi uz parametre koji su pronađeni radom programa search_parameters.py pomoću klasa "GridSearchCV()" i "StratifiedKFold()". Klase su iz paketa scikit-learn.

## Klasifikacija

Korištena su dva algoritma za klasifikaciju : naivni bayesov algoritam i metoda potpornih

vektora(SVM). Klase za oba su uzete iz paketa scikit- learn. Za naivni bayesov algoritam korištena je klasa "MultinomialNB()", dok je za SVM korištena klasa "SVM()". Provođena je deseterostruka cross-validacija klasom "cross_validation()" i mjerom uspješnosti "accuracy". Program koji provodi klasifikaciju naivnim bayesovim algoritmom zove se bayes.py, a program koji provodi SVM svm.py. U oba programa su rezultati prikazani grafom koristeći funkciju "plot()" iz paketa matplotlib.

## Primjer pokretanja

Programi se pokreću bez dodatnih parametara. Jedino moramo biti sigurni da su tekstualne datoteke iz kojih se čitaju podaci dostupne programu.
Primjer:

```
python skini_komentare.py

python razdvoji.py

python pretprocesiranje.py

python search_parameters.py

python bayes.py        python svm.py
```

Izradile : Nikolina Ivezić i Lea Balaško