Odabir značajki

Analiza prirodnog jezika zahtjeva veliki trud prilikom odabiranja značajki. Budući da je skup svih riječi engleskog jezika jako velik, teško je zamisliti koliko različitih rečenica je moguće sastaviti. Odaberemo li krive značajke javlja se velika vjerojatnost da se u nekoj rečenici neće pojaviti niti jedna od tih značajki što rezultira krivom klasifikacijom. Ovakav način razmišljanja vodi nas do zaključka da treba posvetiti veliku pozornost odabiru značajki te da skup značajki mora biti dovoljno velik, reda veličine nekoliko tisuća. Za značajke su korištene riječi i bigrami koji se javljaju u skupu za učenje. Prilikom analize komentara, značajke su promatrane nezavisno, odnosno korištena je tzv "bag of words" metoda. Metoda se sastoji u tome da riječi i bigrame stavimo u "vreću" i analiziranjem njihove frekvencije pojavljivanja i pozitivnosti određujemo pozitivnost dijela komentara ili cijelog komentara. U nastavku navodimo načine odabira značajki i funkcije koje su korištene za njihovu realizaciju. Sve te funkcije nalaze se u **feature _ selection.py** modulu.

TF-IDF

Riječi koje se često javljaju u različitim komentarima manje su korisne za klasikaciju. S druge strane, značajno je ako se neka riječ često pojavljuje unutar neke vrste komentara ili dijela komentara koji se odnosi na određenu kategoriju. Za potrebe klasifkacije teksta svakoj riječi dodjeljuje težina tf-idf (engl. term *frequency-inverse document frequency*):

$$weight(w, d, D) = tf(w, d) \times idf(w, D) \tag{1}$$

gdje je w promatrana riječ (odnosno značajka), d konkretan skup komentara, a D skup svih komentara (primjera za učenje). Frekvencija riječi (engl. *term frequency*) računa se kao relativna učestalost pojavljivanja riječi w unutar skupa d (frekvencija riječi w normalizirana s frekvencijom najfrekventnije riječi u dokumentu). Inverzna frekvencija po dokumentima (engl. *inverse document frequency*) računa se kao logaritam omjera ukupnog broja komentara (primjera) i broja komentara u kojima se pojavljuje promatrana riječ(značajka):

$$idf(w, D) = log \frac{|D|}{|\{d \in D : w \in D\}|} \tag{2}$$

Funkcija **tf_idf** vraća riječi iz skupa komentara sortirane silazno po tf-idf vrijednostima.
Na temelju parsiranja komentara koje je opisano u poglavlju *skupovi za učenje* dobiveni su primjeri rečenica za svaku od kategorija (miris, okus, izgled i osjećaj). Označimo uniju rečenica iz svih kategorija sa K. Funkcija **get_tfidf_values** računa tf-idf vrijednosti svih riječi w unutar skupa komentara koji su iz jedne kategorije(k) $weight(w, k, K)$. Silazno sortirane riječi za svaku kategoriju spremaju u file korištenjem pickle modula. Te su riječi korištene za učenje SVD klasifikatora koji za svaku rečenicu iz komentara određuje koje kategorije ona opisuje.

Za određivanje značajki klasifikatora koji određuju ocjenu komentara po kategorijama i općenito korištene su funkcije **get_tfidf_posneg_category** i **get_tfidf_posneg**. Time dobivamo riječi koje najbolje odvajaju klase (positive, negative i average) unutar svake kategorije posebno te nad svim komentarima skupa za učenje. Konačna ocjena komentara je težinska ocjena pet klasifikatora opisana u poglavlju *ocjenjivanje*.

Najveća frekvencija

Funkcija **most_frequent** vraća riječi koje su najfrekventnije unutar dokumenta. Ova se funkcija koristi prilikom generiranja značajki za SVD klasifikatore. Funkcija **most_frequent_bigrams** vraća riječi koje su najfrekventnije unutar dokumenta. Koristi se za generiranje značajki Naive Bayes klasifikatora.