Izvještaj

Samuel Savanović 1.6.2017

Projekt također dostupan na: https://github.com/samuelSavanovic/sentiment_analasys

Izvori podataka I čiščenje

Svi podaci su preuzeti sa web stranice www.goodreads.com koja sadrzi knjige te njihove recenzije. Specificno korpus se sastoji od 201 komentara (642.4 KiB) na engleskom jeziku preuzetih iz 10 knjiga sa liste najboljih klasika 20-og stoljeća
http://www.goodreads.com/list/show/6.Best_Books_of_the_20th_Centur

sami linkovi knjiga imaju format domain/book/show/id.bookname vise detalja u opisu korpusa corpus_description.txt koji je na engleskom jeziku radi konzistentosti(kako je cijeli projekt I komentari na engleskom).

Sa svakog linka se skida cijela stranica iz koje se parsiraju komentari. Komentari su sadržani u tagu sa ID-om koji odgovara regularnon izrazu **freeText\d** dok su datumi objave komentara sadržani u linku sa klasama “reviewDate createdAt” ručnim testiranjem otrkiveno je da je potrebno maknuti prvi, srednji, I zadni komentar kako bi doslo do podudaranja datuma I komentara (prvi komentar je opis knjige, srednji neka reklama, zadnji je footer stranice). Kako stranica ne prisiljava određeni jezik svi strani jezici su odstranjeni funkcijom Iz english_checker.py

Struktura I rezultati

Projekt je sadržan unutar 5 mapa. Mapa scraping sadrzi prvi dio zadatka seminara I sadrrzi 3 datoteke scrapper.py koji sadrzi glavnu logiku sa parsiranje te poziva pomocne funkcije iz 2 python datoteke english_checker.py koja sadrzi algoritam koji testira jeli tekst na engleskom jeziku te goodreads_books koji parsira I dobavlja knjige. Corpus_builder sadrži corpus_builder.py koji sadrži kod za ručnu klasifikaciju komentara te pravljenje korpousa unutar mape corpus. Mapa features sadrzi feature_extractors.py koja sadrzi funkcije za izlučavanje značajki pomoću leksičkog resursa(reader.py) te feature_set_builder koji izgrađuje prostor značajki za klasifikaciju. Na kraju mapa classifiers sadrži 9 python datoteka (Bayes, DecisionTree, Maxent x (pozitivne, negativne, sve značajke)) te 9 .txt datoteka sa rezultatima izvedbe(točnost te najjinformativnije znacajke) algoritama.

Bayes za sve 3 značajke daje prosječnu točnost od ~63%, DecisionTree 77%, Maxent(100 iteracija) 75%. Problem je kako su skupovi za učenje I testiranje randomizirani pa točnost varira o odabranom skupu za učenje.

Pokretanje

Projekt je napravljen testiran sa python 3.6 iako bi trebao raditi sa bilo kojim python-om 3.0+

Svi klasifikatori se pokreću zasebno unutar classifiers te rade offline na već spremljenom riječniku, pravljenje riječnika se pokreće iz corpus_builder.py koji sve potrebno kako bi se napravio riječnik sličan onom u mapi corpus.