Procesiranje prirodnih jezika

Zadatak 5. Sentiment analysis

Student: Vesna Stojanović 1339

Niš, 2022.

Zadatak

5. Sentiment analysis

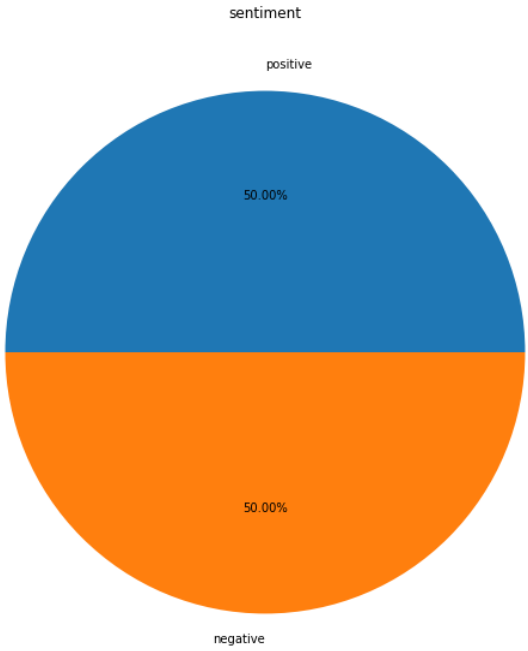
- Naći na web-u odgovarajući dataset koji je pogodan za analizu sentimenta (potrebno je da bude skup filmskih utisaka).
- Prvo je neophodno izvršiti pretprocesiranje dataset-a.
- Iskoristiti gotovo rešenje za Sentiment Analysis iz alata NLTK.
- Nakon toga pokušati poboljšavanje rezultata treniranjem sopstvenog klasifikatora korišćenjem alata NLTK.
- Uporediti dobijene rezultate.
- Uz kod i link do dataset-a, predati i detaljan izveštaj u kome opisujete izradu domaćeg zadatka.

Rešenje

Dataset (Skup podataka) - Izabrani skup podataka nalazi se na sledećem linku:
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.
U ovom zadatku korišćen je dataset koji sadrži listu IMDB review-a (utisaka filmova) u kojima su korisnici komentarisali odgledane filmove. Svaki utisak sadrži komentar i sentiment ('negative' ili 'positive'). Ovo je data set sa 50.000 recenzija filmova za obradu prirodnog jezika ili analitiku teksta.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Slika 1: Prikaz prvih 5 elemenata data set-a



Slika 2: Zastupljenost sentimenata u data set-u

Pretprocesiranje dataset-a

Prvo se importuju biblioteke neophodne za rad kao što su nltk, pandas, matplotlib. Potom se učita IMDB data set koji je prethodno download-ovan i takođe uradimo encoding jer poruke sadrže i emoji-eve i specijalne karaktere i reči koje nisu na engleskom pa je zbog toga potrebno setovati encoding. Prvi korak u rešavanju ovog zadatka je pretprocesiranje. Za rešenje zadatka neophodno je znati tekst utiska i njegov sentiment. Proverom broja sentimenata po svakom tipu sentimenta možemo videti da imamo 25000 pozitivnih i 25000 negativnih utisaka. Promenljivama X i Y dodeljene su kolone koje sadrže utiske i sentimente. Nazivi sentimenata preimenovani su brojnim vrednostima (negative=0, positive=1).

Tokenizacija, filtriranje, eliminacija stop reči, stemovanje

Prvo se vrši tokenizacija reči nad svakim utiskom iz dataset-a. Kreira se skup reči koji čine rečenicu. Iz svake reči odnosno tokena se uklanjaju brojevi i ostali karakteri koji nam nisu potrebni pri analizi i vršimo konverziju u mala slova. Potom ide eliminacija stop reči iz dataset-a. Korišćen je engleski skup stop reči iz NLTK alata jer su utisci na engleskom jeziku. Takođe koristimo speller da ispravimo greške u rečima. Na kraju se vrši stemovanje dobijenih reči i spajanje (join-ovanje) u rečenicu.

Odredjivanje sentimenta utisaka korišćenjem Sentiment Analysis iz NLTK alata

NLTK ima alat za određivanje sentimenta teksta (klasifikaciju) bez prethodne obrade. Ovaj alat je pogodan za tekstove koje nemaju puno rečenica, tako da se ovo može upotrebiti nad data set-om IMDB utisaka filmova. Funkcija polarity_scores se koristi u ovom zadatku. Njoj se prosledi tekst kao argument za koji je potrebno odrediti sentiment, a vraća 4 vrednosti:

- pos: Verovatnoća da je sentiment pozitivan
- neu: Verovatnoća da je sentiment neutralan
- neg: Verovatnoća da je sentiment negativan
- compound: Normalizovana suma svih rejtinga. Ima vrednost izmedju -1 i 1.

Tekst je pozitivan ako je vrednost compound >= 0.05, neutralan ako je compound između -0.05 i 0.05, i negativan ako je compound <= -0.05. Nakon određivanja sentimenta pomoću Sentiment Analysis alata i upoređivanja sa vrednostima iz dataset-a dobijen je rezultat od 69.29% preciznosti.

Ekstrakcija atributa korišćenjem BOW metoda

Za ekstrakciju atributa korišćen je bag of words pristup. Za prebrojavanje pojavljivanja reči korišćen je CountVectorizer iz biblioteke scikit-learn. Ovo je neophodno izvršiti radi kasnijeg izvršenja klasifikacije.

Naivni Bajesov klasifikator iz NLTK

Set podataka deli se na trening i na test u standardnoj razmeri. Za treniranje u NLTK alatu podaci u trening setu treba da budu oblika: (features, label). Iz tog razloga izvršeno je kreiranje dictionary-a funkcijom generate_features, gde je ključ naziv atributa, a value vrednost za taj atribut. Preciznost klasifikatora je 0.83.

Decision Tree klasifikator iz NLTK

Korišćen je i Decision Tree klasifikator koji nam pruža NLTK alat radi poređenja. Korišćen je isti skup podataka koji je korišćen i za Naive Bayes klasifikator. Preciznost klasifikatora u ovom slučaju je 0.69.

Korišćenje TF-IDF mere

Ovo je pouzdaniji pristup. Bag of words pristup sa TF-IDF merom (Term Fruquency - Inverse Document Frequency). Vrednost za TF-IDF dobijamo kao proizvod normalizovane TF i IDF (normalized_term_frequency * idf). Glavni nedostatak je to što odbacuje redosled reči - ignoriše kontekst i značenje reči u dokumentu. Preciznost ove klasifikacije uz korišćenje Naive Bayesovog klasifikatora je 0.65.

Zaključak

Najbolji rezultat definitivno daje Naive Bayes-ov klasifikator 0.83. Naive Bayes-ov klasifikator uz korišćenje TF-IDF mere daje najslabije rezultate 0.65. Decision Tree klasifikator daje slabije rezultate od Naive Bayes-ovog klasifikatora, dobijamo rezultat od 0.69. Ugrađeni Sentiment Analysis alat u okviru NLTK daje preciznost od 0.69. Tokom rada ovog zadatka moglo se primetiti znatno sporije izvršenje Decision Tree klasifikatora i Naive Bayes-ovog klasifikatora uz korišćenje TF-IDF.