Analiza sentimenta komentara

Glavni cilj projekta jeste istrenirati model koji će za datu recenziju ili komentar odrediti da li je pozitivan ili negativan. Na taj način možemo zaključiti da li su mušterije zadovoljne uslugom ili ne. Ovaj model se može primeniti na razne skupove podataka koji obuhvataju komentare korisnika.

Skup podataka na osnovu kojih smo mi vršili treniranje i testiranje je 'RestaurantReviews.txt', datoteka koja sadrži 1000 primera recenzija o restoranima. Sadrži dve kolone: review i sentiment.
Sentiment je vrednost 0 ili 1. 1 ako je komentar pozitivan, 0 ako je negativan.

Datoteka je preuzeta sa:
https://www.kaggle.com/datasets/jurk06/restaurantdataset

Pokretanje programa iz terminala moguće je uz komandu python analiza_sentimenta.py.

Korišćene biblioteke:

```
import pandas as pd
import nltk
import re
import numpy as np
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import
CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
```

Ako biblioteke nisu već instalirane, potrebno ih je instalirati korišćenjem komande pip install ime_biblioteke.

Prvi korak nakon prikupljanja podataka je priprema podataka, koja podrazumeva izbacivanje specijalnih karaktera iz komentara, transformisanje u mala slova i podelu komentara na reči.
Nad rečima se vrši manipulacija korićenjem PorterStemmer() funkcije.

CountVectorizer se koristi za transformaciju podataka (kolona review) u matricu dimenzija (1000, 1500) . Za svaki od 1000 komentara se za svaku od 1500 reči koje se najčešće pojavljuju, postavlja 1 ili 0 u zavisnosti da li se reč nalazi u komentaru ili ne. Y predstavlja kolonu sentiment (0 ili 1).

Sledeći korak je treniranje modela i klasifikacija. 20% podataka čini test skup a 80% podataka se koristi za treniranje. Osim Multinomial Naive Bayes metode razmatrane su i metode logistička regresija i Bernoulli Naive Bayes. Multinomial Naive Bayes metoda je odabrana zbog najboljih rezultata. Podešavanjem parametara ove metode pronađen je parametar sa najvećom preciznošću.
Korišćenjem sklearn.metrics računa se tačnost koja iznosi 80%.

Izbor parametra alpha vrši se proverom koja daje najveću tačnost. Dobijena vrednost za alpha je 0.1, za koju se dobija tačnost 80.5%, pa je ta vrednost i korišćena.

Funkcijom sentiment koja kao parametar uzima test primer, vrši se provera. Funkcija vraća 0 ako je u pitanju negativan i 1 ako je u pitanju pozitivan komentar.

```
def sentiment(test_komentar):
     test_komentar = re.sub(pattern='[^a-zA-Z]',repl=' ', string =
test_komentar)   test_komentar = test_komentar.lower()   test_komentar_reci =
test_komentar.split()
      ps =
PorterStemmer()
  test = [ps.stem(rec) for rec in test_komentar_reci]
test = ' '.join(test)


  tmp = cv.transform([test]).toarray()
return classifier.predict(tmp)
```

Zaključak je da model sa tačnošću od 80% uspešno detektuje da li je komentar pozitivan ili negativan, tako da je glavni cilj projekta ostvaren i model se može uspešno koristiti za klasifikaciju recenzija.

Ana Cvijović, 646-2019
Kristina Mojsić, 626-2019