SBNZ: Domaći 3

SV81-2021

Mitrović Andrej

tim_6

Pristup problemu:

Prvi korak u klasifikaciji pjesama je analiza i pretprocesiranje podataka iz zadatog skupa. Pozivanjem info funckije nad Data Frame-om vidimo da imamo 4800 kolona bez praznih podataka. Analizom individualnih strofa zaključujemo par koraka potrebnih za pretprocesiranje teksta:

- pretvorimo sva slova u mala
- uklonimo sve sto nije slovo
- uklonimo dodatne razmake nastale u prošlom koraku
- uklonimo riječi kraće od tri slova
- definišemo skup „stopwords“, nepounoznačnih riječi koje ce biti ignorisane tokom vektorizacije

Nakon pretprocesiranja analiziramo algoritme koje cemo koristiti za vektorizaciju i klasifikaciju.

Za metriku koristimo F1 score: $F1 = 2 * (P * R) / (P + R)$

Gdje su P i R: $P = TP / (TP + FP)$ $R = TP / (TP + FN)$

Isprobani algoritmi:

Vektorizacija

Pokušaćemo uraditi vektorizaciju pomoću Bag of Words i TD-IDF (Term Frequency-Inverse Document Frequency). Mana BoW se javlja kod velike dimenzionalnosti jer broji ponavljanje svake posebne riječi. TD-IDF pored računanja pojave riječi u dokumentu (strofi), računa i količnik ukupnog broja dokumenata i broja dokumenata sa određenom riječju. Samim tim je i zahtjevniji od BoW ali se rješava problema manje značajnih riječi koje se često pojavljuju.

Klasifikacija

Klasifikaciju ćemo testirati na sledećim algoritmima: Logistička Regresija, Perceptron, SVM

LR je efikasna kod binarne klasifikacije, a u našem slučaju imamo tri muzička žanra. Slično tome i Perceptron je efikasan kod binarne klasifikacije i oba algoritma ne rade dobro sa kompleksnom raspodjelom podataka.

SVM je efikasan kod visoke dimenzionalnosti i kod ne-linearne raspodjele, ali može biti spor kod u obuci na velikom skušu podataka.

Ostvareni rezultati:

Vektorizacija (SVM klasifikacija)

BoW F1 score: 0.525

TDIDF F1 score: **0.567**

Klasifikacija (TD-IDF vektorizacija)

Logička Regresija F1 score: 0.520

Perceptron F1 score: 0.495

SVM F1 score: **0.607**

Konačno rješenje:

Za krajnje rješenje uzeta je kombinacija TD-IDF vektorizacije da bi na taj način izbjegli riječi koje se često ponavljaju, a nemaju veliki značaj i klasifikacije korišćenjem Support Vector Machine.

Dalje je rješenje popravljeno unapređenjem funkcije koja se koristi za pretprocesiranje teksta. Uklanjanjem riječi iz liste „stopwords“, F1 score je dostigao vrijednosti veće od 0.7