Извештај за задатак 3 – класификација

Даница Газдић
SV 12/2020

Милош Обрадовић
SV 55/2020

1. Проблем

У овом извештају се решава проблем класификације над проблемом идентификовања жанра песме на основу строфа. Задатак је решен употребом *SVM* класификатора, а за векторизацију је коришћен *TF-IDF* векторизатор.

2. Решење

Прво су сва слова пребачена у мала, затим су избачени знаци интерпункције и *stop words.* Затим је искоришћен *TF-IDF* векторизатор како би се текст претворио у вектор нумеричких вредности које представљају меру оригиналности. На крају је уз помоћ *SVM* класификатора извршена класификација на жанрове. Грешка је мерена *f1 micro* мером. У првом покушају нису избачене стоп речи и добили смо решење $F1 = 0.7067$, а приликом избацивања стоп речи добили смо нешто боље решење $F1 = 0.7083$. Пошто ниједна библиотека не садржи стоп речи за српско говорно подручје, са неколико *github* репозиторијума смо покупили стоп речи за српски, хрватски и босански језик.