Višestruka Regresija

Tim:

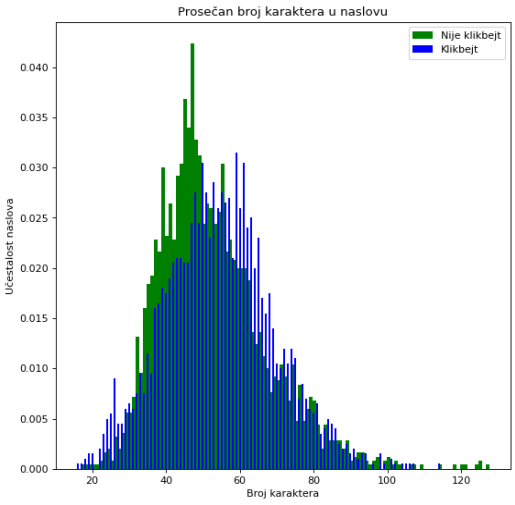
- Marko Vukotić SW-71/2018
- Miloš Popović SW-24/2018

Zadatak:

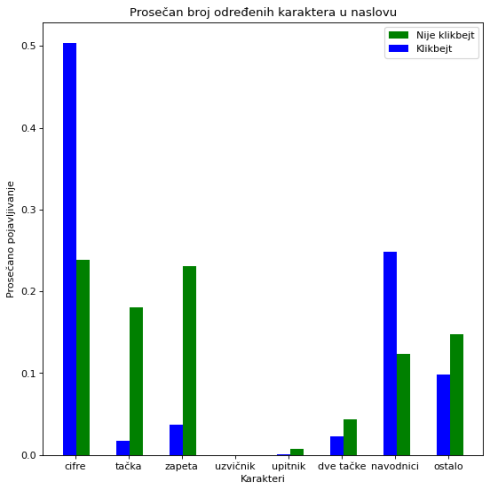
Klasifikovati naslove onlajn medijskih članaka na engleskom jeziku (text) u dve klase, naslov nije klikbejt ili naslov je klikbejt

Analiza podataka:

1. Broj karaktera
   Klikbekt naslovi imaju u proseku više karaktera od regularnih naslova, sa grafika vidimo da je prosek karaktera pomeren blago ka desno
2. Količina specijalnih karaktera u naslovima
   Cifre i navodnici se češće pojavljuju u klikbejt naslovima, dok se tačka i zapeta znatno ređe pojavljuju
3. Upotreba velikih i malih slova
   Klikbejt naslovi u proseku imaju više velikih slova

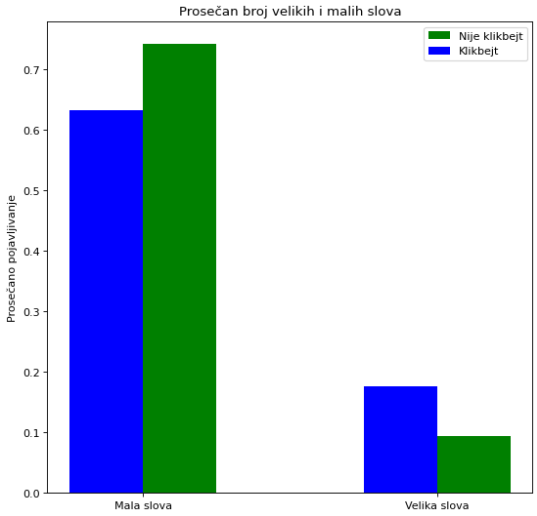


1. Broj karaktera



2. Učestalost specijalnih karaktera

Kod za generisanje grafika samo ostavili na ovom colabu
https://colab.research.google.com/drive/1TW2GpJvV32QVB3LsU9_GHUW_zJdbg61P?usp=sharing



3. Upotreba velikih i malih slova

Primetili smo da je za ovaj problem neophodno posmatrati i znake interpunkcije i velika i mala slova.

Pretprocesiranje podataka:

Što gore to bolje!

1. Tekst **nije** konvertovan u mala slova
2. Nisu izbačeni stop wordovi pošto su uticali negativno na rezultat
3. Svi brojevi su zamenjeni sa 999
   Kako klikbjet naslovi imaju više brojeva, svaki broj gledamo kao isti da bi povećali sličnost klikbejt naslova
4. Tokenizacija reči je rađena na osnovu sledećeg regex izraza
   `([A-Za-z]+|[^a-zA-Z\d\s]+|[0-9])`

   Tokenizovane su reči, brojevi (999) i pojedinačni karakteri
5. Vektorizovano korišćenjem `Tf idf Vectorizer`

Algoritam

Korišćen je SVM algoritam iz scikit-learn biblioteke sa linearnim kernelom i regularizacionim parametrom C=1.5

`sklearn.svm.SVC`

Evaluacija

Evaluacija je rađena podelom u Train i Test skup. Train 80% Test 20%