Završni izvještaj

Podatci o članovima tima:

Anela Rački, Bastijan Kobler, Petar Hrsto, Tena Čop i Nikola Klobučar studenti diplomskog studija istraživačke informatike na Filozofskom Fakultetu u Zagrebu. Pohađaju kolegij Obradu prirodnog jezika u sklopu kojeg je pokrenut projekt strojne analize mnijenja.

Podatci o projektu:

Općenito: U ovom projektu smo radili na treniranju programa u prepoznavanju sentimenta u hrvatskom jeziku. To smo napravili pronalaženjem lingvističkih uzoraka u tekstu koji otkrivaju nečiji stav o nekom svojstvu nekog objekta.

Podatci: Prikupljeni sa stranice za recenziju knjiga (<https://www.najboljeknjige.com/recenzije>) koristeći beautiful soup 4(bs4). Čistili smo podatke tako što smo ih segmentirali na rečenice (razdijelnici su ., ! i ?) programom te ručno provjerili radi točnosti.

Anotiranje: Imali smo 4 anotatora. Svaki anotator je anotirao od 2 250-3 000 rečenica. Koristili smo Cohhen's kappa kako bi provjerili konzistentnost između anotatora rezultat je bio K=0.6089.

Statistika: Sveukupan broj rečenica je bio 3 157. Broj tokena je bio 89 264. Imali smo tri klase u koje smo podijelili rečenice: pozitivno, neutralno i negativno. Koristili smo 2 209 rečenica radi treninga našeg programa, a broj rečenica korištenih radi testiranja je bio 948.

Rezultati projekta:

Model XGBoost – Točnost: 85,86 %

Model XGBoost – F1: 41,02 %

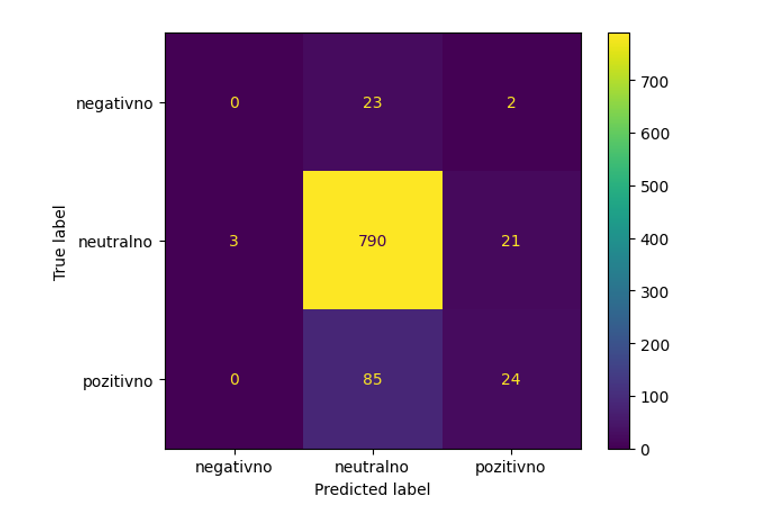
Model SVM – Točnost: 86,18 %

Model SVM – F1: 33,19 %

Model KNN – Točnost: 85,86 %

Model KNN – F1: 30,80 %

Confusion matrix



Aktivnosti projekta:

Anotator i programer

Aktivnosti članova:

Anela Rački – Anotator

Bastijan Kobler – Anotator

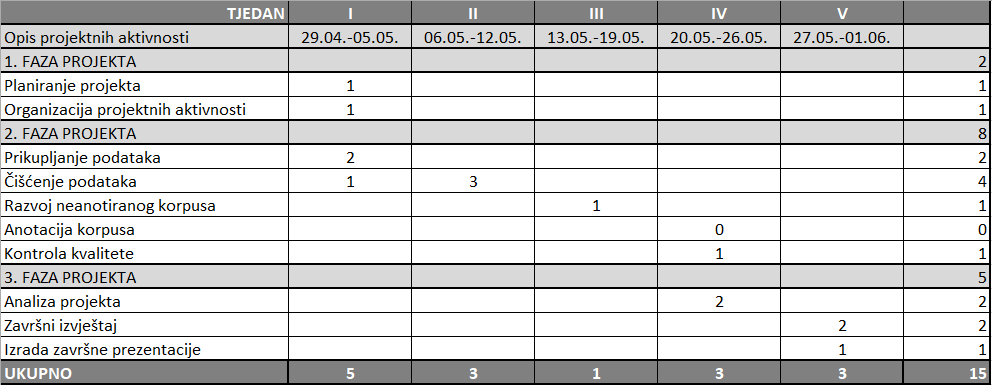
Petar Hrsto – Anotator

Tena Čop – Anotator

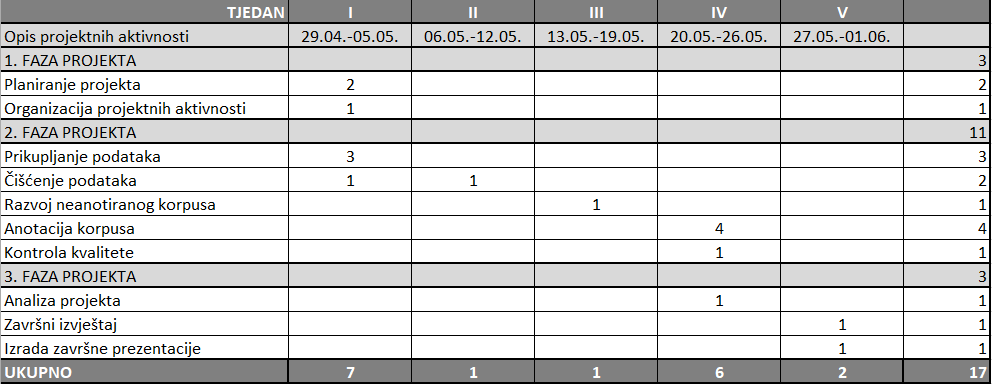
Nikola Klobučar – glavni programer

Evidencija radnih sati:

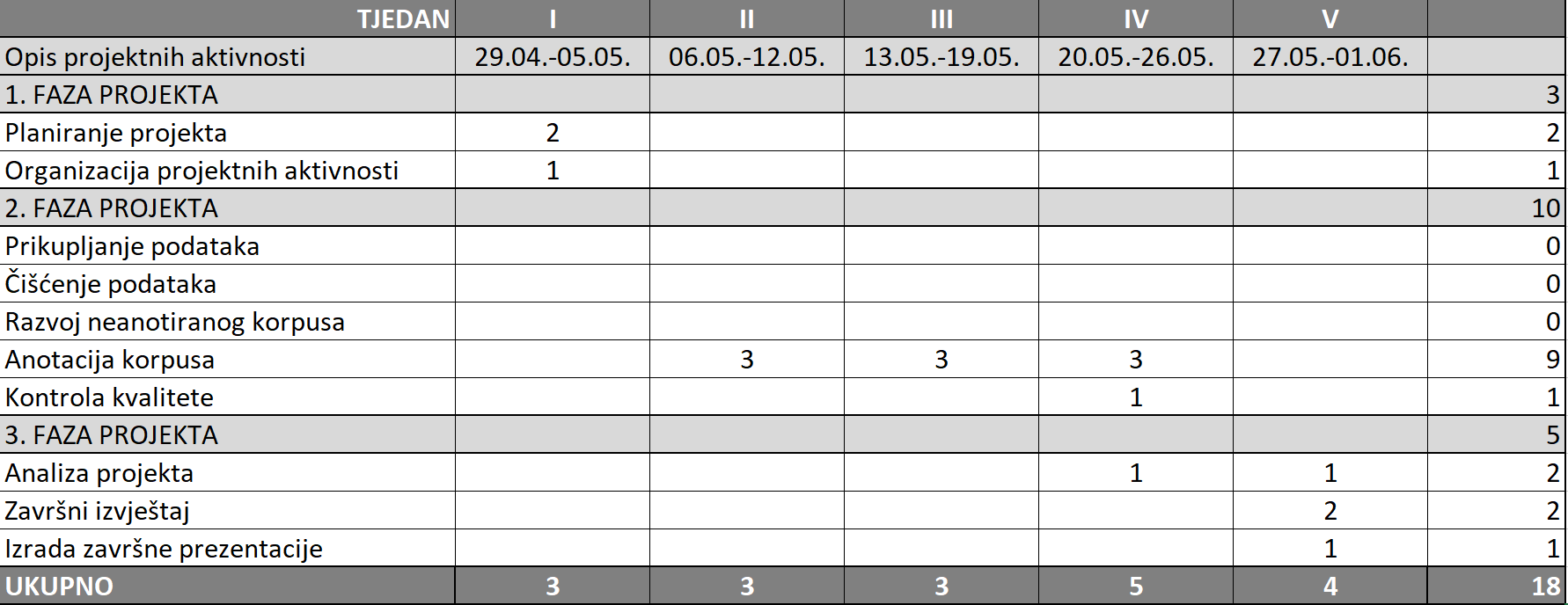
Nikola Klobučar



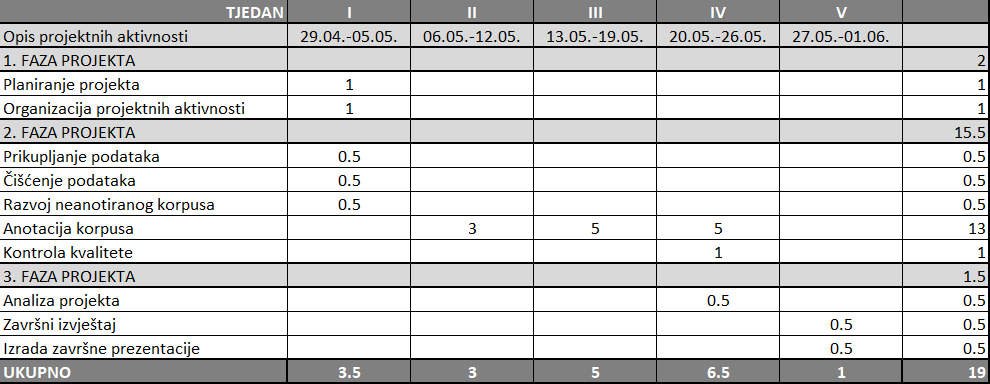
Tena Čop



Anela Rački



Bastijan Kobler



Petar Hrsto

Završna evaluacija:

Iako naš program pokazuje relativno visoki stupanj točnosti u modelima(<85%) zbog stanja naših podataka(data bias) gdje imamo većinski neutralne rečenice(2 000+) za razliku od pozitivnih(300) i negativnih(manje od 100) stupanj F1 je u nižim pragovima. To nam pokazuje da pri prikupljanju podataka nismo odabrali najbolji izvor mnijenja već dobar objektivni osvrt na temu podataka. Što nije optimalno za analizu mnijenja.