**Obrada prirodnih jezika**

Projekat – dokumentacija

Vladimir Milijić 3366/20

Dušan Janić 3317/19

Lena Jokanović 3299/20

Projekat je podeljen na tri glavne faze:

Faza 1: prikupljanje podataka

Faza 2: anotacija podataka

Faza 3: modeli

Svaka od ovih faza je detaljnije opisana u nastavku ovog dokumenta.

FAZA 1 – prikupljanje podataka

U prvoj fazi je bilo potrebno naći parove funkcija-komentar. Zbog toga što je u našem timu bilo tri osobe, bilo je potrebno prikupiti 750 parova funkcija-komentar.

Ovi parovi funkcija-komentar su se uglavnom dohvatali sa projekata sa predmeta “Principi softverskog inženjerstva“ sa osnovnih studija na smeru za softversko inženjerstvo.

Skupili smo 4 projekata iz ovog predmeta na kome je rađen isključivo jezik php i koji je morao biti izkomentarisan odgovarajućim doc blokovima. Pored toga iskorišćen je i privatan repozitorijum u kome se nalazio programski kod isključivo za vežbanje jezika PHP.

Zbog toga što kolege nisu želele da njihovi projekti budu okačeni javno, smatramo ih kao da su privatni repozitorijumi.

Svaki od članova iz tima je imao zadatak da skupi 225 parova funkcija-komentar i da prilagodi već postojeće komentare na nekin način sa datim upitima.

Svaki par funkcija-komentar je smešten u jedinstveni fajl sa unikatnim imenom fajla po obrazcu: redni broj upita \_ ime funkcije.

Naravno, postoji tabela sa nazivom faza\_1 u kojoj su izlistani svi ovi parovi.

Pored gore navedenih parova, postoji jos njih 75 koje smo skupili zajedno da bi se zajednički anotirali, ali o tome više u drugoj fazi projekta. To je ukupno 10% svih parova funkcija-komentar.

FAZA 2 – anotacija podataka

Nakon faze prikupljanja podataka, došli smo do trenutka kada je bilo potrebno anotirati prikupljene parove funkcija-kometar.

Podaci su bili anotirani sa četri klase:

0 - Nema sličnosti

1- Slaba sličnost

2- Jaka sličnost

3- Identičnost

Svaki član grupe je samostalno anotirao po 225 podataka onako kako je on mislio da je ispravno.

Ovi podaci su smešteni u tabeli sa nazivom faza\_2.

Preostalih 10% parova funkcija-komentar su anotirani zajednički.

Svrha anotiranja 75 podataka zajednički je bila da se mi kao osobe koje anotiraju podatake na neki način usaglasimo. Svako od članova tima je anotirao podatake kako je on mislio da je ispravno i zatim se vodila diskusija koja od anotacija je bila najviše ispravna. Uglavnom su se mišljenja poklapala izuzev na malom skupu parova funkcija-komentar.

U takvim slučajevima smo birali za anotaciju klasu sa najvećim brojem glasova članova tima.

Ovih 10% od ukupno prikupljenih podataka su izdvojeni u posebnu tabelu sa nazivom “faza2\_zasebneOcene”.

Ocene tj. procenat slaganja anotatora su smeštene u fajlu sa imenom similarityscore, ali su takođe prikazani u nastavku:

Vlada-Dusan: 69/75 (91.77%)

Dusan-Lena: 66/75(87.78%)

Vlada-Lena: 68/75 (90.44%)

Grupni prosek: 89.99%

Svi prikupljeni podaci su u skladu sa specifikacijama izloženim u postavci projekta.

FAZA 3 – modeli

Pošto su se u ovom trenutku bili skupili svi potrebni podaci, mogli smo da počnemo sa pravljenjem i treniranjem naših modela.

Kao što je i zahtevano u samoj postavci zadataka, korišćeni modeli su bili logistička regresija, naivni bajesov klasifikator i metoda potpornih vektora.

Pre samog treniranja modela, tekst koji je predstavljao deo ulaza za treniranje modela je na neki način morao da bude obrađen, tj pretprocesuiran.

Što se tiča tehnika obrade teksta, proces stemovanja i obaranja velikih slova na mala smo uvek koristili kao baseline tehnike. Razlog tome je što u našem jeziku postoji mnogo oblika za istu reč, postoje padeži itd...Smatrali smo je veoma potrebno to uraditi jer bi se gotovo sigurno tačnost modela znatno uvećala. Takođe, pretvaranje velikih slova u mala je imalo smisla jer je to ista reč, a pri treniranju modela bi se tretirala kao različita.

**1) KLASIFIKACIJA**

Pre nego što su se ulazni podaci takvi kakvi jesu pustili kao ulazni set za treniranje u model, morali smo da odradimo jos neke tehnike obrade teksta.

Primenjena je bag of words(BoW) metoda. Svaka reč koja se javila u tekstu komentara je predstavljena sada preko vektora. Svaka od ovih reči je bila zasebna kolona u skupu za treniranje. BoW je značajno doprineo tačnosti modela, oko ~10%.

Pored bag of words metoda , primenjena je i tdif tehnika, ali je ona imala gotovo iste rezultate kao o bag of words, tako da smo se na kraju odlučili da zadržimo bag of words kao jednu od tehnika.

Nakon gore navedenih metoda, na kraju smo probali da tekst komentara gledamo kao bigrami i trigrame.

I bigrami i trigrami su poboljšali tačnost za pribižno istu vrednost od par procenata(2% - 3%). Zbog veoma male sličnosti, odlučili smo se sa bigrame.

Sada je ulazni skup podataka x\_train bio spreman da se pusti kao ulaz u sva tri modela.

Što se tiče L1 i L2 parametara, razlika je bila zanemarljiva, tako da smo izabrali L2 koji je zapravo bio default vrednost.

Za samo pronalaženje najboljih hiper parametara kao i za 10-slojnu cross validaciju smo koristili GridSearchCV.

Iz objekta ove klasa smo dobili najbolje optimizovani model korišćenjem metode .best\_estimator\_.s

Sa algoritmom za linearnu regresiju su dobijene sledeće informacije:

Cross validation score: 0.62771

Na skupovima za testiranje: 64.667

Sa algoritmom potpornih vektora su dobijene sledeće informacije:

Cross validation score: 0.62771

Na skupovima za testiranje: 64.667

Sa naivnim bajesovim klasifikatorom su dobijene sledeće informacije:

Cross validation score: 0.44329

Na skupovima za testiranje: 76.66

**2) Regresija**

Zbog toga što regresija radi sa kontinualnim vrednostima, naše već postojeće klase(0,1,2,3) smo morali da pretvorimo u kontinualne vrednosti na sledeći način:

Klasa 0: 0 – 0.25

Klasa 1: 0.25 – 0.5

Klasa 2: 0.5 – 0.75

Klasa 3: 0.75 – 1

Od algoritama za mašinsko učenje smo koristili linearnu regresiju za predviđanje kontinualnih vrednosti.

Rezultati dobijeni preko regresije su znatno lošiji u odnosu na rešavanje putem klasifikacije.