Obrada prirodnih jezika

Projekat za školsku 2019/2020. godinu

Tema projekta

Tema predmetnog projekta za školsku 2019/2020 godinu se tiče problema klasifikacije kratkih tekstova koji su u vidu komentara pisani u programskom kodu. Izrada projekta podrazumeva izgradnju i anotaciju odgovarajućeg skupa takvih komentara. Kreirani skup je zatim potrebno iskoristiti za obučavanje i evaluaciju nekoliko klasifikacionih algoritama zasnovanih na mašinskom učenju. Projekat se može implementirati u programskom jeziku i paketu po izboru. U nastavku će biti detaljnije opisana svaka od faza u izradi projekta.

Projekti će se izrađivati grupno. Proces prijavljivanja grupe je opisan u odeljku o propozicijama izrade projekta. Neophodno je učešće svih članova grupe u svim fazama izrade projekta, tj. nije dozvoljena podela posla između članova grupe po fazama.

Faza 1 - Prikupljanje podataka

Komentari se u programskom kodu javljaju napisani na različitim prirodnim jezicima, ali će u okviru ovog projekta od interesa biti samo jedan od sledeća dva – engleski i srpski. Pored toga, pri izradi projekta biće od interesa komentari za jedan od sledećih programskih jezika: Java, JavaScript, Python, PHP, C, C++, C#, SQL. Iako je osnovna tema projekta ista za sve studente, svaka projektna grupa će razmatrati samo jednu kombinaciju programskog i prirodnog jezika (npr. jedna grupa će se baviti klasifikacijom komentara na srpskom pisanih u okviru Java koda, druga klasifikacijom komentara na engleskom iz PHP koda, itd.).

Proces prikupljanja podataka podrazumeva formiranje dovoljno velikog seta komentara u okviru programskog koda napisanog na određenom programskom jeziku, pri čemu komentari u prikupljenom kodu treba da budu napisani na istom prirodnom jeziku (srpskom ili engleskom). Kao izvor podataka za formiranje ovakvog seta mogu poslužiti lični repozitorijumi studenata (prethodni projektni zadaci za druge predmete, diplomski radovi, itd.), javno dostupne baze repozitorijuma poput GitHub-a, itd. Formirani set za posmatrani par prirodni/programski jezik treba da sadrži barem 3000 komentara ako se radi o komentarima na engleskom, odnosno barem 2000 komentara ako se radi o komentarima na srpskom. Kao pozitivan dodatan faktor u ocenjivanju će se uzimati u obzir prikupljanje većeg broja komentara od navedenog minimuma. Dozvoljeno je da se u znatno manjoj meri u skupu podataka jave i komentari na drugom od posmatrana dva prirodna jezika ili dvojezični komentari (npr. za kod iskomentarisan predominantno na srpskom, a manjim delom na engleskom ili dvojezično, i na srpskom i na engleskom, dozvoljeno je ubaciti u skup i te dvojezične komentare i komentare na engleskom).

Prikupljene podatke treba sačuvati u sledećem formatu:

1. Prikupljeni komentari treba da budu sačuvani u vidu tab-separated UTF-8 TXT fajla sa sledećim kolonama za svaki komentar:

1. *NaturalLanguageID* – dvoslovna ISO oznaka prirodnog jezika komentara: *EN* za engleski, *SR* za srpski, *EN/SR* za dvojezične komentare. Ovo polje će u okviru jedne grupe imati predominantno onu vrednost određenu prirodnim jezikom koji grupa razmatra.
2. *ProgrammingLanguageName* – ime programskog jezika na kome je kod napisan (Java, PHP, C,…). Ovo polje će u okviru jedne grupe uvek imati istu vrednost.
3. *RepoID* – jedinstveni identifikator repozitorijuma/projekta iz koga su podaci dobijeni. Ovaj identifikator može da bude u formatu definisanom od strane same grupe.
4. *SourceID* – jedinstveni identifikator izvora tj. fajla sa programskim kodom iz koga su podaci dobijeni. Ovaj identifikator može da bude u formatu definisanom od strane same grupe.
5. *CommentID* – jedinstveni globalni identifikator za posmatrani komentar. Ovaj identifikator treba da bude u formatu *RepoID/SourceID/LineNumber*, gde je *LineNumber* red fajla sa programskim kodom u kome se posmatrani komentar nalazi ili počinje.
6. Tekst samog komentara. U ovom polju treba čuvati samo koristan tekst komentara, bez znakova koji označavaju početak (ili kraj) komentara u posmatranom programskom jeziku, ali specifične tagove koji ukazuju na određenu vrstu komentara treba sačuvati (npr. *@author*). Kod komentara koji se protežu na više linija koda (ali je jasno da se radi o jednom komentaru) znak za novi red treba kodovati sa *n*.

2. Izvori podataka treba da budu popisani u posebnom tab-separated UTF-8 TXT fajlu u čijoj prvoj koloni se nalazi *RepoID* repozitorijuma/projekta, u drugoj *SourceID* izvora/fajla, u trećoj broj linija koda u fajlu, a u četvrtoj URL do fajla sa programskim kodom ili opis repozitorijuma/projekta i izvora/fajla ako se ne radi o javno dostupnom kodu.

Faza 2 - Anotacija podataka

U fazi anotacije je potrebno za svaki od prikupljenih komentara iz prethodne faze ručno označiti kojoj od sledećih klasa pripada:

- *Functional* – komentari koji opisuju funkcionalnost postojećeg koda. Ove komentare je potrebno bliže svrstati u sledeće potkategorije:
  - *Functional-Inline* – inline komentari koji opisuju funkcionalnost određenog koda
  - *Functional-Method* – multiline/dokumentacioni komentari koji opisuju funkcionalnost određene funkcije/metode
  - *Functional-Module* – multiline/dokumentacioni komentari koji opisuju funkcionalnost određenog modula/klase
- *ToDo* – komentari koji opisuju predviđenu buduću funkcionalnost koja još uvek nije implementirana u kodu, ili bag/nedostatak u postojećem kodu koji bi trebalo ispraviti.
- *Notice* – komentari koji predstavljaju upozorenja ili napomene namenjene drugim programerima ili korisnicima koda. Oni mogu da se odnose na zastarelost delova koda, da daju primere korišćenja/pozivanja nekih delova koda, da upozoravaju na određene izuzetke, i sl.
- *General* – komentari koji sadrže opšte informacije, poput podataka o autorstvu, licencnih i copyright napomena, napomena o korišćenim bibliotekama u okviru koda, o verziji konkretnog modula/klase itd.
- *Code* – komentari koji predstavljaju iskomentarisane/izbačene linije programskog koda, što se obično radi kod debugovanja/testiranja ili kao način čuvanja starih varijanti koda.
- *IDE* – šablonski komentari koji su ili automatski generisani od strane nekog razvojnog okruženja (IDE) ili namenjeni razvojnom okruženju (npr. direktive za kompajler).

Anotaciju je potrebno obaviti tako što će se u tab-separated TXT fajlu sa komentarima iz prethodne faze dodati nova, sedma kolona, u koju će se za svaki komentar ubeležiti naziv njegove klase.

U slučaju da jedan deo posmatranog komentara jasno pripada jednoj od kategorija, a drugi nekoj drugoj, treba razdvojiti taj komentar na dva manja i onda njih svrstati u odgovarajuće klase.

Komentare bi prilikom anotacije trebalo ravnomerno rasporediti između svih članova grupe, tako da svako anotira približno istu količinu podataka. Pritom, očekuje se da na početku i periodično u toku anotacije članovi grupe razmotre karakteristične problematične situacije u anotaciji i da ih reše na sistemski ujednačen način u okviru grupe.

Iako je za potrebe izgradnje skupa podataka dovoljno sprovesti jednostruku anotaciju (tj. anotaciju u kojoj samo jedna osoba anotira određeni komentar), radi merenja stepena saglasnosti anotatora potrebno je na kraju procesa anotacije nasumično izdvojiti 10% od ukupnog broja komentara. Taj podskup podataka treba da anotiraju svi članovi tima zasebno i bez međusobnih konsultacija. Izdvojeni podskup podataka, zajedno sa dobijenim individualnim anotacijama, treba sačuvati u formi dodatnog tab-separated TXT fajla – prva kolona takvog fajla treba da sadrži *CommentID* vrednosti komentara iz podskupa, a preostale kolone anotirane klase od strane svakog od članova grupe. Na osnovu ovog fajla tj. ovog podskupa podatka potrebno je izračunati procentualan stepen podudarnosti anotacija između svaka dva člana grupe, kao i grupni prosek binarnih stepena podudarnosti.

Faza 3 - Obučavanje i evaluacija klasifikacionih modela

Obučavanje i evaluaciju modela je potrebno sprovesti korišćenjem 10-slojne stratifikovane unakrsne validacije, korišćenjem odgovarajuće metrike za merenje performansi. Algoritmi koje treba razmotriti su sledeći:

- Multinomijalni naivni bajesovski klasifikator
- Multivarijacioni Bernulijev naivni bajesovski klasifikator
- Logistička regresija
- Metoda potpornih vektora (linearna, bez kernela)

Kod logističke regresije i metode potpornih vektora je potrebno sprovesti optimizaciju hiperparametra C / lambda koji određuje jačinu regularizacije, korišćenjem ugnježdene unakrsne validacije. Inicijalnim ispitivanjem, korišćenjem default vrednosti za ostala podešavanja, treba utvrditi koja od varijanti funkcije regularizacije (L1 / L2) i funkcije gubitka kod metode potpornih vektora (L1 / L2) daje bolje rezultate – ako nema primetnih razlika, preporučljivo je koristiti L2 oblike funkcija.

Klasifikaciju je potrebno razmotriti u dve varijante – jednu u kojoj svi komentari *Functional* tipa pripadaju jednoj klasi, i drugu u kojoj postoje zasebne klase za svaki podtip *Functional* komentara.

Za sve navedene algoritme i za obe varijante klasifikacije treba ispitati efekte različitih tehnika pretprocesiranja teksta. Počevši od osnovnih podešavanja gde se kao odlike koriste svi unigrami (*bag-of-words* model), sistematski razmotriti sledeće:

- Normalizaciju svih tekstova na mala slova (lowercasing)
- TF, IDF i TFIDF ponderovanje
- Filtriranje stop-reči i/ili stemovanje reči (po izboru)
- Frekvencijsko filtriranje reči

- Korišćenje bigramskih i trigramskih odlika

Kao listu stop-reči moguće je koristiti neku od javno dostupnih lista ili formirati sopstvenu. Za stemovanje reči na engleskom koristiti Porterov stemer. Za stemovanje reči na srpskom koristiti neki od stemera iz paketa *SCStemmers*[1] (dozvoljeno je i korišćenje alternativnih implementacija ovih stemera).

Propozicije izrade projekta

Optimalna veličina grupe je četvoro studenata, ali će za određene kombinacije prirodni/programski jezik biti dozvoljene i grupe od troje, odnosno petoro studenata (za kombinacije za koje objektivno postoji malo/puno dostupnih podataka). Grupe od troje članova će biti dozvoljene samo za rad sa komentarima na srpskom, dok će grupe od petoro članova biti dozvoljene samo za rad sa komentarima na engleskom. Tročlane grupe će moći da iz evaluacije izostave jednu od varijanti naivnog bajesovskog klasifikatora, po izboru. Takođe, minimalan broj komentara (na srpskom) koje takve grupe treba da prikupe i anotiraju će biti 1500 (umesto 2000). Od petočlanih grupa će se očekivati prikupljanje i anotacija barem 3500 komentara (na engleskom, umesto 3000), ali zahtevi u pogledu obučavanja i evaluacije modela klasifikacije ostaju isti kao i kod standardnih, četvoročlanih grupa.

Pre otpočinjanja rada na projektu, neophodno je formirati i zvanično prijaviti grupu putem mejla. Prilikom prijave grupe, neophodno je navesti spisak članova grupe, kao i spisak od bar 3 željene kombinacije prirodni/programski jezik, po redosledu interesovanja. Grupi će zatim u najkraćem roku biti zvanično dodeljena ona kombinacija koja već nije zauzeta od strane neke ranije prijavljene grupe.

Za slučaj nemogućnosti samoorganizovanja u grupe, studenti mogu i da se individualno prijave za izradu projekta. U tom slučaju, biće od strane predavača organizovani u grupe sa ostalim studentima koji su se individualno prijavili ili će, u slučaju nedovoljnog broja tako prijavljenih studenata, biti pridruženi nekoj od već formiranih grupa. U oba slučaja, individualno prijavljeni studenti neće imati mogućnost izbora kombinacije prirodni/programski jezik koju razmatraju.

Ova postavka predmetnog projekta će važiti do prolećnog semestra naredne školske godine. Grupe je potrebno formirati i prijaviti najkasnije do 01.08.2020, bez obzira na konkretan ispitni rok u kome se planira odbrana. Individualne prijave treba dostaviti najkasnije do 01.06.2020. Ni grupne ni individualne prijave nakon tih datuma neće biti uzimane u obzir.

Grupe koje žele da brane projekat u određenom ispitnom roku treba da pošalju urađeno rešenje i projektnu dokumentaciju do početka istog ispitnog roka. U projektnoj dokumentaciji treba detaljno opisati svaku od faza izrade projekta. Ukoliko su projektna rešenja i dokumentacija adekvatni, u dogovoru sa studentima biće određen termin odbrane projekta u toku ispitnog roka. Odbrane će principijelno biti moguće u svim ispitnim rokovima previđenim za predmete iz letnjeg semestra, uzimajući u obzir epidemiološku situaciju u konkretnom roku.

[1] https://vukbatanovic.github.io/SCStemmers/