Obrada prirodnih jezika

Projekat za školsku 2024/2025. godinu

Tema projekta

U školskoj 2024/2025 godini tema predmetnog projekta je izrada *baseline* pristupa i evaluacija naprednijih otvorenih rešenja za zadatak prepoznavanja imenovanih entiteta (*Named Entity Recognition – NER*) u obradi tekstova na srpskom jeziku. Projekat se može izrađivati individualno ili grupno, pri čemu je maksimalna veličina grupe 5 članova. Ukoliko se projekat izrađuje grupno, neophodno je učešće svih članova grupe u svim fazama izrade projekta, tj. nije dozvoljena podela posla između članova grupe po fazama. Projekat je potrebno implementirati u programskom jeziku Python.

Izrada projekta podrazumeva prikupljanje odgovarajućeg skupa tekstova na srpskom jeziku iz nekoliko različitih tematskih domena, kao i ručnu anotaciju prikupljenih podataka u pogledu prisustva osnovna tri tipa imenovanih entiteta (osobe – PER, lokacije – LOC, organizacije – ORG) po IOB2 sistemu. Tematski domeni koji se mogu razmatrati su sledeći:

- Novinski
- Književni
- Pravno-administrativni
- Tviter/X poruke
- Neki drugi domen po izboru, u dogovoru sa predavačem (npr. medicinski tekstovi, finansijski tekstovi, itd.)

Kreirani skup podataka je zatim potrebno iskoristiti za evaluaciju nekoliko različitih otvorenih statističkih modela. Modeli koji se mogu razmatrati su sledeći:

- **CLASSLA** paket (https://github.com/clarinsi/classla), sa odvojenim modelima za
  - Standardni (novinski) jezik
  - Nestandardni jezik (Tviter)
- **BERTić-NER** (https://huggingface.co/classla/bcms-bertic-ner)
- **COMtext.SR NER**, sa paralelnim modelima za ekavicu (https://huggingface.co/ICEF-NLP/bcms-bertic-comtext-sr-legal-ner-ekavica) i ijekavicu (https://huggingface.co/ICEF-NLP/bcms-bertic-comtext-sr-legal-ner-ijekavica)
- **SrpCNNER** (https://live.european-language-grid.eu/catalogue/ld/9484)

U nastavku će biti detaljnije opisana svaka od faza.

Faza 1 - Prikupljanje podataka

Proces prikupljanja podataka podrazumeva formiranje dovoljnog skupa tekstova za odabrani tematski domen, tako da se takav skup može nakon anotacije iskoristiti za evaluaciju statističkih modela. Minimalna veličina skupa tekstova i broj domena koje treba razmotriti zavise od toga da li se projekat izrađuje individualno ili u grupi, odnosno od broja članova grupe:

- Individualna izrada - bar dva tematska domena po izboru, minimalne dužine teksta od bar 5000 tokena (sa interpunkcijom) po domenu.
- Grupna izrada, 2 člana grupe – bar tri tematska domena po izboru, minimalne dužine teksta od bar 5000 tokena (sa interpunkcijom) po domenu
- Grupna izrada, 3 člana grupe – bar četiri tematska domena po izboru, minimalne dužine teksta od bar 5000 tokena (sa interpunkcijom) po domenu
- Grupna izrada, 4 člana grupe – bar pet tematskih domena po izboru, minimalne dužine teksta od bar 5000 tokena (sa interpunkcijom) po domenu
- Grupna izrada, 5 članova grupe – bar šest tematskih domena po izboru, minimalne dužine teksta od bar 5000 tokena (sa interpunkcijom) po domenu

Za tokenizaciju tekstova treba koristiti tokenizator za srpski jezik iz paketa CLASSLA tj. ReLDI tokenizator (https://pypi.org/project/reldi-tokeniser/). Kao izvor podataka za formiranje skupa tekstova mogu poslužiti bilo koji javno dostupni veb sajtovi sa sadržajem na srpskom jeziku. Pri tome, da bi se izbegla mogućnost preterane prilagođenosti modela podacima za evaluaciju, neophodno je obezbediti da se nijedan od prikupljenih tekstova ne nalazi među podacima koji su korišćeni za obučavanje razmatranih NER modela. Izvori podataka koji su korišćeni za obučavanje navedenih NER modela, uz prateću dokumentaciju, su sledeći:

- CLASSLA – 4 korpusa na srpskom i hrvatskom jeziku:
  - SETimes.SR (https://www.clarin.si/repository/xmlui/handle/11356/1843)
  - hr500k (https://www.clarin.si/repository/xmlui/handle/11356/1792)
  - ReLDI-NormTag-NER-sr (https://www.clarin.si/repository/xmlui/handle/11356/1794)
  - ReLDI-NormTag-NER-hr (https://www.clarin.si/repository/xmlui/handle/11356/1793)
- BERTić-NER – isti korpusi kao CLASSLA
- COMtext.SR NER – COMtext.SR.legal korpus pravnih tekstova, u varijantama na ekavici i na ijekavici (https://github.com/ICEF-NLP/COMtext.SR)
- SrpCNNER – Skup književnih tekstova ELTeC-srp (https://github.com/COST-ELTeC/ELTeC-srp)

Finalni skup prikupljenih tekstova treba pročistiti od dupliranih unosa istog teksta. Prikupljeni tekstovi treba da budu sačuvani u vidu UTF-8 enkodovanih TXT fajlova. Očekuje se da se u metapodacima za prikupljene tekstove zabeleži i URL ili jedinstveni identifikator izvora iz koga je svaki tekst dobijen.

Faza 2 – Anotacija podataka

U ovoj fazi potrebno je svaki od prikupljenih tekstova ručno obeležiti u pogledu prisustva osnovnih imenovanih entiteta (osobe – PER, lokacije – LOC, organizacije – ORG) po IOB2 sistemu. Za sprovođenje anotacije dozvoljeno je, ali ne i neophodno, koristiti bilo koji program za anotaciju, bilo neki postojeći, bilo neki namenski razvijen za potrebe projekta. Pre sprovođenja pune anotacije potrebno je formulisati kratka uputstva za anotaciju, koja bi trebalo da sadrže jasne instrukcije za sistematsko postupanje u karakterističnim problematičnim situacijama.

Ukoliko se projekat izrađuje grupno, u sprovođenju anotacije treba pratiti i ostale korake u standardnoj metodologiji označavanja podataka, što podrazumeva:

1. Kalibraciju – proveru upotrebljivosti uputstava za anotaciju uz pomoć malog podskupa primera tekstova (oko 10% od ukupnog broja) koje svi članovi grupe treba da paralelno anotiraju, zasebno i bez međusobnih konsultacija. Ako se u ovom koraku uoče nedostaci u uputstvima (npr. preko niske saglasnosti u anotaciji), treba ih doraditi i ponoviti kalibraciju.
2. Sprovođenje anotacije – podatke bi trebalo ravnomerno rasporediti između svih članova grupe, tako da svako anotira približno istu količinu podataka iz svakog od tematskih domena. Očekuje se da anotacija tekstova bude jednostruka, ali nije zabranjeno višestruko paralelno označavanje, ako članovi grupe procene da se time znatno podiže konzistentnost generisanih oznaka.
3. Analizu kvaliteta anotacije – određivanje saglasnosti anotatora na osnovu kalibracionog skupa (procentualan stepen saglasnosti između svaka dva člana grupe, kao i grupni prosek binarnih stepena saglasnosti).

Na kraju anotacije treba sprovesti statističku analizu zastupljenosti oznaka u finalnim podacima. Anotirani tekstovi treba da budu sačuvani u vidu UTF-8 enkodovanih i vertikalizovanih CONLLU fajlova sa anotacijama imenovanih entiteta. Očekuje se da se u metapodacima za prikupljene tekstove zabeleži i URL ili jedinstveni identifikator izvora iz koga je svaki tekst dobijen. Podatke iz skupa za kalibraciju (u slučaju grupnog rada) treba sačuvati zasebno, sa onoliko anotacija/kolona koliko ima članova grupe, pri čemu za sve anotacije treba koristiti isti format zapisa.

Faza 3 - Evaluacija statističkih modela

U ovoj fazi potrebno je razmotriti nekoliko NER modela koji su navedeni na početku projektnog zadatka. Za rad sa modelima BERTić-NER i COMtext.SR NER preporučuje se korišćenje interfejsa *Simple Transformers* (https://simpletransformers.ai/).

Pri tome, CLASSLA modele za standardni i nestandardni jezik treba tretirati kao odvojene modele, dok u primeni COMtext.SR modela treba primenjivati odgovarajuću varijantu modela shodno izgovoru (ekavica/ijekavica) na kome je konkretan tekst napisan. Broj modela koje treba razmotriti zavisi od toga da li se projekat izrađuje individualno ili u grupi, odnosno od veličine grupe:

- Individualna izrada – bar dva modela po izboru
- Grupna izrada, 2 člana grupe – bar tri modela po izboru
- Grupna izrada, 3 člana grupe – bar četiri modela po izboru
- Grupna izrada, 4 ili 5 članova grupe – svih pet navedenih modela

Pri evaluaciji navedenih NER modela neophodno je izvršiti dodatno mapiranje između njihovih izlaza i posmatrana tri tipa osnovnih imenovanih entiteta (osobe – PER, lokacije – LOC, organizacije – ORG):

- Kod CLASSLA i BERTić-NER modela, oznaka DERIV-PER se mapira na oznaku PER, dok se oznaka MISC mapira na oznaku O.
- Kod COMtext.SR NER modela, oznake ADR i TOP se mapiraju na oznaku LOC, a oznake COURT, INST, COM i OTHORG se mapiraju na oznaku ORG. Sve ostale oznake osim PER se mapiraju na oznaku O.
- Kod SrpCNNER modela, oznaka PERS se mapira na oznaku PER, a sve ostale oznake osim LOC i ORG se mapiraju na oznaku O.

Pored evaluacije navedenih modela, neophodno je obučiti i evaluirati i *baseline* pristup zasnovan na individualnoj klasifikaciji svakog tokena zasebno pomoću multinomijalnog naivnog Bajesovog

klasifikatora. Kao minimalne odlike za takav model potrebno je razmotriti: *bag-of-words* tj. *one-hot* vektorsku reprezentaciju trenutno posmatranog tokena, isti tip reprezentacije za prethodnih nekoliko tokena (od dva do pet, po izboru), kapitalizaciju trenutno posmatranog tokena, kapitalizaciju prethodnih nekoliko tokena (od dva do pet, po izboru), redni broj posmatranog tokena u rečenici. Dozvoljeno je i preporučeno i uključivanje drugih odlika u navedeni *baseline* model.

Obučavanje i evaluaciju *baseline* modela je potrebno sprovesti putem 10-slojne unakrsne validacije, korišćenjem odgovarajućih metrika za merenje performansi. Nije potrebno sprovoditi optimizaciju bilo kog hiperparametra, tj. dozvoljeno je korišćenje njihovih *default* vrednosti. Evaluaciju odabranih naprednijih NER modela treba sprovesti nad celokupnim izrađenim i anotiranim skupom podataka. Evaluaciju svih modela treba sprovesti u dve varijante – jednoj u kojoj se svi B- i I- tagovi tretiraju kao odvojene klase (B-PER, I-PER, B-LOC, I-LOC, itd.), i drugoj gde se kao klase gledaju samo tipovi entiteta (PER, LOC, ORG) a B- i I- prefiksi se ignorišu.

Propozicije izrade projekta

Studenti se mogu sami organizovati u grupe, a mogu i da se individualno prijave za izradu projekta. U slučaju grupnog rada, neophodno je formirati i zvanično prijaviti grupu putem mejla, na adresu: vuk.batanovic@etf.bg.ac.rs. Prilikom prijave grupe, neophodno je navesti spisak članova grupe i izbor modela i tematskih domena tekstova koje bi grupa želela da razmatra.

Ova postavka predmetnog projekta će važiti do prolećnog semestra naredne školske godine. Studenti koji žele da brane projekat u određenom ispitnom roku treba da pošalju urađeno rešenje i projektnu dokumentaciju pre odbrane na adresu: vuk.batanovic@etf.bg.ac.rs.

U projektnoj dokumentaciji treba opisati svaku od faza izrade projekta. Ovo podrazumeva opisivanje procesa prikupljanja podataka, izvora podataka, navođenje kriterijuma koji su u tom procesu korišćeni i opisivanje kako je proces obavljen sa tehničkog aspekta. Pored toga, dokumentacija mora sadržati opis anotacije podataka, uključujući uputstva za anotaciju, kao i opis tehničke strane označavanja podataka. Takođe se očekuje da izveštaj sadrži deskriptivni statistički prikaz prikupljenih i anotiranih podataka. Za fazu evaluacije statističkih modela podrazumeva se da izveštaj sadrži pregledni tabelarni prikaz rezultata različitih modela, kao i analizu i diskusiju dobijenih rezultata. Dokumentacija ne treba da sadrži iskopirana detaljna objašnjenja iz nastavnih materijala za korišćene tehnike i algoritme.

Ukoliko su projektna rešenja i dokumentacija adekvatni, u dogovoru sa studentima biće određeni termini odbrane projekta u toku ispitnog roka. Odbrane će biti moguće u svim ispitnim rokovima predviđenim za predmete iz letnjeg semestra.

Ocene će se dobijati na osnovu broja prikupljenih bodova na skali 0-100, prema sledećoj raspodeli:

- 30 poena – faza prikupljanja podataka
- 30 poena – faza anotacije podataka
- 30 poena – faza evaluacije statističkih modela
- 15 poena – kvalitet i potpunost priložene projektne dokumentacije

Za svaku od prve tri stavke neophodno je da grupa ostvari barem polovinu od minimalnog broja poena. Drugim rečima, nije moguće odbraniti projekat bez sprovođenja i opisivanja sve tri faze izrade.