NASTAVNO-NAUČNOM VEĆU ELEKTROTEHNIČKOG FAKULTETA U BEOGRADU

Komisija za studije II stepena Elektrotehničkog fakulteta u Beogradu imenovalo nas je u Komisiju za pregled i ocenu master rada Marka Mitića, sa brojem indeksa 3085/11, pod naslovom „Kategorizacija teksta napisanog na srpskom jeziku". Komisija je pregledala priloženi rad i dostavlja sledeći

I Z V E Š T A J

1. Biografski podaci

Marko Mitić je rođen 18.02.1986. u Pirotu, gde je završio osnovnu i srednju školu. Elektrotehnički fakultet u Beogradu je upisao 2005. godine. Osnovne studije je završio na Odseku za računarsku tehniku i informatiku sa prosečnom ocenom 8.13, diplomirao 16.06.2011. godine sa ocenom 10. Nakon diplomiranja je na istom fakultetu upisao master studije i položio sve ispite sa prosečnom ocenom 9.40. Počevši od 2007. godine radio je u firmama Omega, PSTech, Vlatacom, HTEC i NVIDIA.

2. Predmet, cilj i metode master rada

Predmet ovog master rada je određivanje oblasti odnosno kategorizacija teksta na srpskom jeziku tehnikom mašinskog učenja, na osnovu postojećih klasifikovanih tekstova oglasa. Klasifikacija kod mašinskog učenja ima za cilj da odredi klasu odnosno kategoriju kojoj neka instanca pripada. Ove metode se koriste za filtriranje email spama, prepoznavanje oblika, lica, znakova (OCR), prepoznavanje govora, određivanje tema tekstova, u biometriji kod prepoznavanja otisaka prstiju itd.

Sva implementacija izvedena je u programskom jeziku Python i korišćenjem njegovih biblioteka. Prikupljanje podataka odnosno oglasa urađeno je pomoću Scrapy framework-a. Za perzistenciju podataka izabrana je MongoDB NoSQL baza prvenstveno zbog dinamičkih šema, lake prenosivosti, jednostavnosti i brzini upita.

Suština rada, vektorizacija tekstova, klasifikacija i evaluacija je urađena korišćenjem "scikit-learn" Python biblioteke. To je jedna od najkorišćenijih biblioteka otvorenog koda koja obuhvata širok spektar implementiranih algoritama mašinskog učenja.

Metode rada se ogledaju u testiranju različitih klasifikatora nad istim trening i test podacima. Korišćeni su različiti pristupi/algortimi za klasifikaciju:

- Logistička regresija
- Naïve Bayes (Multinominalna i Bernuli)
- kNN i NearestCentroid (metode najbližeg suseda i najbližih centroida)
- SVM – Metoda podržavajućih vektora
- Kombinacije klasifikatora za različitim tehnikama učenja

Rad obuhvata prikupljanje podataka za treniranje modela preslikavanja u oblasti na primeru oglasa na srpskom jeziku, realizovanje jednostavnog stemera za srpski jezik, treniranje i testiranje klasifikatora i evaluacija rezultata i performansi merenjem različitih parametara.

Cilj je da se prikažu mogućnosti tekst klasifikacije za srpski jezik, uticaj steming procesa i prikaz i analiza performansi i rezultata kao i upoznavanje sa načinom rada algoritama korišćenih u procesu.

3. Sadržaj i rezultati

Master rad je podeljen u 6 poglavlja sa slikama, tabelama, graficima i segmentima programskog koda. U radu je citirano 15 referenci.

U prvom poglavlju, uvodu, predstavljen je predmet rada i dat je kratak uvod u veštačku inteligenciju i mašinsko učenje sa posebnim osvrtom na klasifikaciju teksta.

Drugo poglavlje predstavlja opis upotrebljenih tehnologija i analiza postojećih rešenja. Ovde je dat detaljniji uvod u klasifikaciju teksta i pregled najzastupljenijih klasifikatora (Naïve Bayes, kNN i SVM). Takođe, dat je opis korišćenih tehnologija u samom radu, konkretno, scikit-learn biblioteka, MongoDB i osvrt na web crawling i opis Scrapy biblioteke.

U trećem poglavlju je opisana implementacije softverskog sistema. Ovde je objašnjen i proces prireme teksta, steming proces, na koji način su podaci prikupljeni, kako se čuvaju i kako je vršen trening i testiranje klasifikatora. U ovom poglavlju su dati i segmenti progamskog koda iz softverskog sistema.

Četvrtko poglavlje, evaluacija dobijenih rezultata, daje prikaz rezultata testiranja, odnosno rezultate klasifikacije. Prvo je dat opis metrika kojima se meri uspešnost klasifikacije a zatim nekoliko test primera i sami rezultati.

Peto poglavlje predstavlja zaključak gde je i dat osvrt na rezultate klasifikacije tekstova na engleskom jeziku sa istim klasifikatorima.

Šesto poglavlje sadrži spisak korišćenih referenci.

4. Zaključak

Iako je srpski jezik komplikovaniji za klasifikaciju zbog svog morfološkog bogatstva rezultati pokazuju da je moguće vršiti klasifikaciju jako uspešno i da su rezultati približno isti kao i kada se radi o tekstovima na engleskom jeziku, a koristeći iste klasifikacione algoritme.

Na osnovu svega izloženog članovi Komisije prelažu Nastavno-naučnom veću Elektrotehničkog fakulteta u Beogradu da prihvati master rad kandidata Marka Mitića, diplomiranog inženjera elektrotehnike, pod naslovom "Kategorizacija teksta napisanog na srpskom jeziku" i odobri njegovu javnu odbranu.

U Beogradu, 22.09.2014.

Članovi Komisije

Docent dr Miloš Cvetanović

Prof. dr Boško Nikolić