Crna Gora
UNIVERZITET CRNE GORE
PRIRODNO-MATEMATIČKI FAKULTET
Broj 2195
Podgorica, 18 07 2018 god.

Vijeću Prirodno-matematičkog fakulteta

Predmet: Izvještaj Komisije za ocjenu podobnosti teme i kandidata za izradu magistarskog rada *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata"* kandidata Koste Pavlovića, specijaliste računarskih nauka.

Na osnovu člana 60 Statuta Univerziteta Crne Gore, a u vezi sa članom 24 Pravila studiranja na postdiplomskim studijama, na sjednici Vijeća PMF-a od 23. maja 2018. godine imenovani smo za članove komisije za ocjenu podobnosti teme i k3andidata za izradu magistarskog rada *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata"* kandidata Koste Pavlovića, specijaliste računarskih nauka.

1) Podaci o kandidatu

Kosta Pavlović je rođen 8. maja 1994. godine u Beranama. Osnovnu školu i gimnaziju završio je u Kolašinu i bio je dva puta izabran za đaka generacije. Nosilac je diplome „Luča". Tokom srednje škole, na državnim takmičenjima iz programiranja imao je zapažene rezultate osvajajući uvijek jedno od prva dva mjesta. Bio je i član reprezentacija koje su predstavljale Crnu Goru na međunarodnim informatičkim olimpijadama.

Prirodno-matematički fakultet u Podgorici, studijski program Računarske nauke, upisao je 2012. godine. Specijalistički rad „Deep learning tehnike za klasifikaciju rukom pisanih cifara" odbranio je 2016. godine. Od 3. novembra 2016. godine saradnik je u nastavi na Prirodno-matematičkom fakultetu. Sve ispite na magistarskim studijama položio je sa najvišom ocjenom.

Kosta Pavlović je angažovan i kao saradnika u nastavi na ljetnjim školama programiranja za učenike osnovnih i srednjih škola tokom 2014., 2015., 2016. i 2017 godine, a i na predavanjima koja su tokom ovih godina održavana na PMF-u. Aktivno učestvuje u pripremama učenika za državna i međunarodna takmičenja iz programiranja.

Dobitnik je godišnje studentske nagrade koju dodjeljuje Univerzitet Crne Gore za 2014. godinu, nagrade Opštine Kolašin za najboljeg studenta za 2015. godinu, nagrade 19. decembar za studente 2015. godine koju dodjeljuje Glavni grad, kao i Plakete Univerziteta Crne Gore za 2016. godinu.

2) Obrazloženje teme

a) Naučna oblast

Predložena tema pripada naučnoj oblasti ***Mašinsko učenje – računarske nauke.***

b) Predmet rada

Rad digitalnih biblioteka i automatizovanih kancelarija zasniva se na aplikacijama za analizu i obradu slika dokumenata. Ulazni dokumenti su veoma različiti po svojoj strukturi i uključuju pisma, račune, ugovore, faksove i formulare. Ručna obrada ovih podataka je veoma nepraktična i s aspekta organizacije i s aspekta finansija. U cilju prevazilaženja

navedenih nedostataka razvijeni su sistemi za automatizovanu obradu velikog broja adminstrativnih dokumenata u što je moguće kraćem vremenu i uz minimalno uplitanje zaposlenih. U zavisnost od tipa dokumenta, definišu se različiti elementi obrade: klasifikacija dokumenata, obrada višestraničnih dokumenata, razumijevanje sadržaja, usmjeravanje (routing) i izvlačenje informacija.

Posebna pažnja je posvećena procesu izvlačenja informacija koji se može opisati kao pronalaženje relevantnih strukturiranih informacija iz djelimično strukturiranih slika dokumenata. Na taj način se iz npr. slike računa za komunalne usluge mogu izvući i sačuvati u bazi podataka polja kao što su datum, iznos i broj računa.

Tri opšta principa izvlačenja informacija su: primjena klasifikatora, prepoznavanje strukturalnih šablona i kodiranje pravila za ekstrakciju. Kreiranje klasifikatora (npr. Support Vector Machines, rekurentne neuronske mreže, drveta odlučivanja...) vrši se na osnovu velikog broja dokumenata koji služe kao skup za obučavanje. Klasifikator najčešće vraća vjerovatnoću da neka riječ ili rečenica pripada nekom polju. Međutim, ovi klasifikatori nisu pogodni ako se struktura dokumenata promijeni npr. dodavanjem novog polja.

Mnogi komercijalni sistemi ovaj proces zasnivaju na fiksiranim prostornim šablonima koji preslikavaju lokaciju koja je dobijena iz programa za optičko prepoznavanje karaktera u odgovarajuće polje. Ovakav pristup ne daje dobre rezultate kada se raspored polja na dokumentu mijenja. Treći pristup je da se kodiraju pravila za esktrakciju. Na primjer, ukupna vrijednost računa se sastoji od cifara, decimalne tačke i oznake valute i nalazi se lijevo od labele koja označava *Total*. Ovaj proces se obavlja ručno prilikom implementacije, gdje se svako pravilo opisuje nizom fiksiranih parametara. Svaka promjena na postojećem dokumentu ili uvođenje novog tipa dokumenta zahtijeva ponovnu implementaciju pravila.

c) Naučni cilj rada

Cilj rada je da razvije inovativan metod učenja pravila za ekstrakciju informacija. U radu se predlaže metod baziran na kodiranju parametrizovanih pravila za ekstrakciju informacija i učenju tih parametara primjenom genetskog algoritma. Metod treba po preformansama biti uporediv sa drugim rješenjima a u određenim aspektima i biti superioran u odnosu na njih. Predloženi metod ne samo da smanjuje potrebu za ponovnom implementacijom pravila već i omogućava inkrementalno učenje na novim dokumentima iz iste klase. Kao demonstracija uspješnosti pristupa, planira se implementacija robustnog sistema koji može izvlačiti informacije iz dokumenata koji pripadaju širokom spektru klasa.

d) Naučne metode

U realizaciji postavljenog cilja koristiće se sljedeće metode:

- Postavljanje i testiranje hipoteze
- Metoda eksperimenta

Očekuje se da predloženi algoritam ima preciznost približno jednaku relevantnim alternativnim metodima iz literature, pri čemu se očekuje unapređenje u pogledu veličine skupa dokumenata za obučavanje.

Svi algoritmi biće implementirani u programskom jeziku Python. Genetski algoritam biće testiran sa različitim podešavanjima početnih parametara. Sprovešće se više serija eksperimenata kako bi se došlo do sistema sa najboljim mogućim rezultatima.

e) **Aktuelnost problematike**

Ručna obrada podataka u digitalnim bibliotekama i automatizovanim kancelarijama nije praktična i zahtijeva mnogo resursa. Poseban problem predstavlja heterogenost dokumenata, pa je proces izvlačenja informacija tj. pronalaženja relevantnih strukturiranih informacija iz djelimično strukturiranih slika dokumenata, veoma složen. Veliki broj naučnih radova iz ove oblasti, kao i neprekidno razvijanje specijalizovanih aplikacija za izvlačenje informacija, pokazuju da je ova tema izuzetno aktuelna i sa naučnog i sa komercijalnog aspekta.

3) Zaključak

Uvidom u podnesenu dokumentaciju, Komisija je utvrdila da predložena tema kandidata Koste Pavlovića ima jasno definisane ciljeve i metode istraživanja i očekivane rezultate.

Predlažemo Vijeću Prirodno-matematičkog fakulteta da odobri izradu magistarskog rada kandidata Koste Pavlovića pd nazivom *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata"*

U Podgorici, 20. juna 2018. godine

Komisija

Prof. dr Predrag Stanišić, član

Prof. dr Savo Tomović, član

Doc. dr Goran Šuković, mentor