Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.42:004.738.12
DOI: https://doi.org/10.24867/15BE13Dragutinovic

PREPOZNAVANJE PODATAKA O LIČNOSTI U TEKSTUALNIM DOKUMENTIMA

RECOGNITION OF PERSONAL DATA IN TEXTUAL DOCUMENTS

Đorđe Dragutinović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu prikazana je aplikacija za automatsko prepoznavanje podataka o ličnosti u tekstualnim dokumentima – specificirani su zahtevi i dizajn aplikacije, opisani su bitni elementi njene implementacije i demonstrirano je korišćenje aplikacije. Aplikacija je implementirana koristeći tehnike prepoznavanja imenovanih entiteta.*

Ključne reči: *imenovani entiteti; podaci o ličnosti; NER; SpaCy; Classla.*

Abstract – *This paper presents an application for automatic recognition of personal data in textual documents. The requirements and design of specified application are specified, the essential elements of its implementation are described and usage of the application is demonstrated. The application is implemented using named entity recognition techniques.*

Keywords: *named entities; personal data; NER; SpaCy; Classla.*

1. UVOD

Prema Zakonu o zaštiti podataka o ličnosti i Opštoj uredbi o zaštiti podataka (eng. *General Data Protection Regulation*, skr. GDPR), "podatak o ličnosti" je svaki podatak koji se odnosi na fizičko lice čiji je identitet određen ili odrediv, neposredno ili posredno, na osnovu oznake identiteta ili jednog, odnosno više obeležja njegovog identiteta. Neki od ličnih podataka koji se najčešće pominju i koriste su ime i prezime, adresa, jedinstveni matični broj građanina, broj telefona, ali u podatke o ličnosti spadaju i fotografija, otisak prsta, podaci o zdravstvenom i imovinskom stanju, verskim i političkim opredeljenjima i drugi [1, 2].

Pojava ovih podataka sve je češća u elektronskim dokumentima koji se čuvaju na računaru, a korisnici računara često nisu ni svesni da veliki broj dokumenata sadrži podatke o ličnosti.

Iz tog razloga, potrebno je rešiti problem prepoznavanja podataka o ličnosti u dokumentima, odnosno omogućiti analizu elektronskih dokumenata koji su sačuvani u memoriji računara i prepoznavanje podataka u ličnosti u tim dokumentima. Ovaj rad bavi se rešavanjem navedenog problema.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Stevan Gostojić, vanr. prof.

Navedeni problem najlakše je rešiti metodom koja se naziva prepoznavanje imenovanih entiteta *(eng Named Entity Recognition*, skr. NER). Ova metoda predstavlja jednu od tehnika obrade prirodnih jezika i spada u oblast veštačke inteligencije, a njen cilj je da u nestruktuiranom tekstu prepozna imenovane entitete i izvrši njihovu klasifikaciju, odnosno definiše kojoj kategoriji pripadaju. Termin „imenovani entitet“ odnosi se na sve što se može jednoznačno identifikovati, odnosno na sve što se može razlikovati od svih ostalih entiteta sa sličnim atributima.

Ostatak rada organizovan je na sledeći način: u drugom poglavlju analizirani su radovi koji su rešavali slične probleme. U trećem poglavlju specificirana je aplikacija koja je implementirana u okviru ovog rada. Četvrto poglavlje opisuje modele koji su iskorišćeni u implementaciji aplikacije. U poglavlju 5 prikazano je kako se opisano rešenje može iskoristiti za prepoznavanje imenovanih entiteta. U šestom poglavlju dat je zaključak o radu i navedene su njegove prednosti i mane, pri čemu je dat predlog kako se navedene mane mogu ispraviti.

2. STANJE U OBLASTI

Rešavanjem problema prepoznavanja imenovanih entiteta i podataka o ličnosti u tekstualnim dokumentima bavi se veliki broj postojećih radova. U radu [3], autori su prikazali kako se postojeći, javno dostupni modeli za prepoznavanje imenovanih entiteta mogu iskoristiti u analizi biografija. Skup podataka sastojao se od 249 ručno anotiranih biografija, preuzetih sa internet enciklopedije „Vikipedija“. Za prepoznavanje imenovanih entiteta iskorišćena su 4 postojeća, javno dostupna modela. Evaluacija korišćenih modela izvršena je prikazom preciznosti (eng. *precision*) i opoziva (eng. *recall*), ali su autori definisali i novi način evaluacije uspešnosti, kako bi se uzele u obzir i situacije kada je imenovani entitet fraza koja se sastoji iz više reči, a model uspešno prepozna samo jedan deo te fraze. Rezultati su pokazali da je preciznost prepoznavanja imenovanih entiteta varirala između 73% i 96%, a opoziv je varirao između 64% i 97%, u zavisnosti od korišćenog modela, a svi modeli su imali problem u prepoznavanju imenovanih entiteta koji se sastoje iz više reči.

Rad [4] bavi se poređenjem uspešnosti postojećih modela u prepoznavanju imenovanih entiteta u novinskim člancima napisanim na engleskom i ruskom jeziku. Korišćeno je ukupno 7 NER modela, a za njihovo testiranje iskorišćena su četiri javno dostupna skupa anotiranih novinskih članaka. Evaluacija korišćenih modela izvršena je prikazom F1 vrednosti (eng. *F1-score*), koja predstavlja harmoničku srednju vrednost preciznosti i opoziva. Rezultati su pokazali da su modeli mnogo preciznije prepoznavali

imenovane entitete u člancima napisanim na engleskom jeziku, a razlog za to leži u načinu obučavanja korišćenih modela.

Nekoliko radova bavilo se i problemom izdvajanja informacija iz dokumenata. Izdvajanje informacija podrazumeva da se mogu ručno definisati tipovi entiteta koje je potrebno prepoznati u dokumentima. U radu [5] prikazano je kako se može vršiti izdvajanje entiteta iz uputstava za upotrebu. Korišćeni skup podataka dobijen je ručnim anotiranjem 453 stranice različitih uputstava za upotrebu, napisanih na engleskom jeziku. Rezultati su pokazali da je preciznost prepoznavanja ručno definisanih kategorija entiteta varirala između 85% i 98%, što dovodi do zaključka da se NER modeli mogu koristiti i za izdvajanje informacija iz teksta.

3. SPECIFIKACIJA

Zadatak ovog rada je napraviti aplikaciju sa grafičkim korisničkim interfejsom, koja će korisnicima omogućiti da izvrše prepoznavanje imenovanih entiteta u tekstualnim dokumentima sačuvanim u memoriji računara. Poseban fokus potrebno je staviti na prepoznavanje podataka o ličnosti (pre svega, imena i prezimena osoba), kako bi korisnici imali uvid u to u kojim se dokumentima i u kom kontekstu pojavljuju ti podaci. Tipovi dokumenata koje je potrebno podržati su:

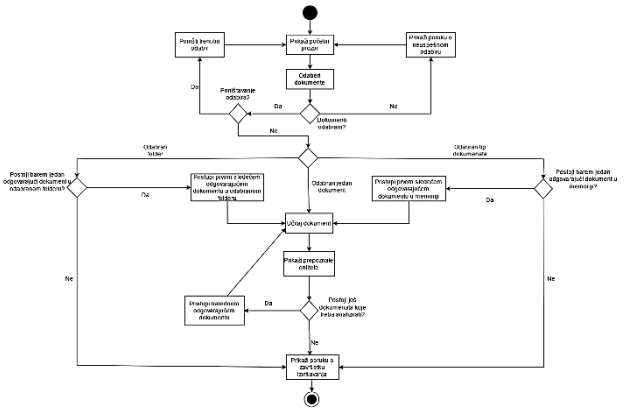
- *PDF*
- *txt*
- *doc(x)*

Samim tim, potrebno je korisnicima dati mogućnost da odaberu formate (jedan ili više) koje žele da uključe u pretragu. Takođe, osim mogućnosti odabira formata, korisnici bi trebalo da imaju i mogućnost da odaberu pojedinačni fajl ili folder u okviru kog će se vršiti analiza dokumenata. Prilikom odabira pojedinačnog fajla, korisnika treba ograničiti da može da bira samo dokumente odgovarajućeg formata. Ukoliko korisnik odabere folder u okviru kog će se vršiti analiza, izvršiće se rekurzivan pristup svim fajlovima i podfolderima unutar izabranog foldera.

Ukoliko korisnik pokuša da započne analizu dokumenata, a da prethodno nije odabrao format dokumenata koje želi da analizira ili pojedinačni dokument ili folder, korisniku je potrebno prikazati odgovarajuću poruku i vratiti ga na početak, kako bi mogao da pokuša ponovo. Ukoliko je korisnik odabrao pojedinačni fajl ili folder, potrebno mu je ponuditi mogućnost da poništi svoj odabir, ukoliko je pogrešio, te ga vratiti na početni ekran i omogućiti mu ponovno biranje ukoliko tu mogućnost iskoristi.

Prikazivanje prepoznatih imenovanih entiteta trebalo bi izvršiti tako što će se na glavnom prozoru aplikacije prikazati lokacija i kompletan sadržaj dokumenta koji se analizira u tom trenutku, pri čemu će imenovani entiteti biti prikazani drugačijom bojom u odnosu na ostatak teksta, a korisniku treba prikazati legendu, kako bi znao koji tipovi entiteta su obojeni i kojom bojom. Potrebno je obezbediti podršku za rad aplikacije sa dokumentima koji su napisani na srpskom ili engleskom jeziku, pri čemu je potrebno prilikom učitavanja dokumenta prepoznati na kom jeziku je dokument napisan, te u zavisnosti od toga iskoristiti odgovarajući NER model, takav da za prepoznati jezik daje najbolje moguće rezultate.

Dijagram aktivnosti specificirane aplikacije prikazan je na slici 1.



Slika 1. *Dijagram aktivnosti specificirane aplikacije*

4. IMPLEMENTACIJA

U implementaciji aplikacije za prepoznavanje imenovanih entiteta u tekstualnim dokumentima korišćeni su sledeći programski jezici, okruženja i biblioteke:

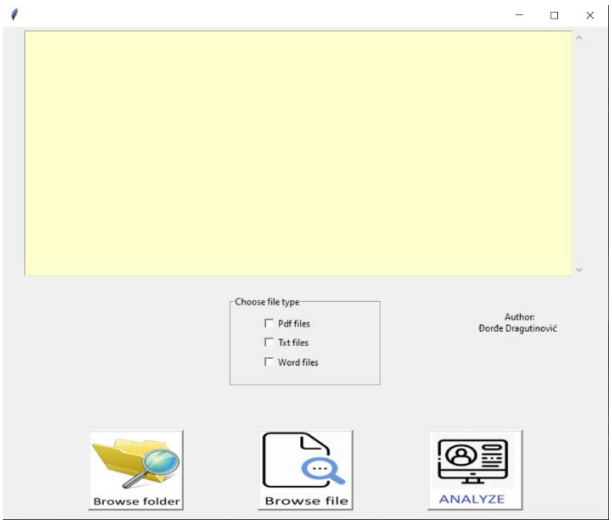
- *Python* – programski jezik visokog nivoa, opšte namene
- *PyCharm* – integrisano razvojno okruženje za programiranje u *Python*-u
- *Tkinter* – alat za kreiranje aplikacija sa grafičkim interfejsom u *Python*-u

Za prepoznavanje imenovanih entiteta u dokumentima iskorišćena su dva besplatna, javno dostupna modela:

- *SpaCy* – biblioteka otvorenog koda koja se koristi za naprednu obradu prirodnih jezika. Razvijena je od strane nemačke softverske kompanije „Explosion" i napisana je u programskom jeziku *Python*. Moguće je vršiti *alpha* tokenizaciju, *part-of-speech* tagovanje, kategorizaciju teksta, lematizaciju, segmentaciju rečenica, dodatno obučavanje modela i mnoge druge funkcije. U osnovi rada biblioteke nalaze se konvolutivne neuronske mreže i *Thinc* biblioteka za mašinsko učenje [6]. U implementiranoj aplikaciji, ovaj model iskorišćen je za prepoznavanje imenovanih entiteta u dokumentima napisanim na engleskom jeziku.
- Classla (*CLARIN Knowledge Centre for South Slavic languages*) – javno dostupni model za obradu i analizu tekstova napisanih na slovenskim jezicima (srpski, hrvatski, slovenački, makedonski, bugarski). Razvijen je od strane slovenačkog nacionalnog instituta CLARIN, a kao baza za kreiranje ovog modela iskorišćen je postojeći *Stanford NER* model, te se prepoznavanje imenovanih entiteta vrši kombinacijom HMM (eng. *Hidden Markov Model*) i MEMM (eng. *Maximum Entropy Markov Model*) algoritama. Korišćenjem ovog modela moguće je vršiti tokenizaciju teksta, *part-of-speech* tagovanje, lematizaciju, segmentaciju rečenica, parsiranje zavisnosti i prepoznavanje imenovanih entiteta [7]. U implementiranoj aplikaciji, ovaj model iskorišćen je za prepoznavanje imenovanih entiteta u dokumentima napisanim na engleskom jeziku.

5. DEMONSTRACIJA

Početni prozor implementirane aplikacije sastoji se iz 3 glavne celine. Prvu celinu predstavlja veliko polje (*textBox*) u okviru kog će se ispisivati rezultati operacija, kao i prepoznati entiteti. Ispod tog polja nalaze se 3 *checkButton* polja, pomoću kojih korisnik može odabrati koje tipove dokumenata želi da analizira (PDF, txt ili doc(x) dokumenti). Na dnu prozora nalaze se 3 dugmeta: "*browse folder*", koje nudi mogućnost odabira diska/foldera koji će se pretražiti, "*browse file*", koje nudi mogućnost odabira fajla koji će biti analiziran i "*analyze*", koje omogućava pokretanje procesa prepoznavanja entiteta. Početni prozor implementirane aplikacije prikazan je na slici 2.



Slika 2. *Početni prozor implementirane aplikacije*

Ukoliko je korisnik uspešno odabrao dokument ili folder koji želi da analizira, klikom na dugme "*analyze*" vrši se učitavanje sadržaja i prepoznavanje imenovanih entiteta. Na početku analize za svaki fajl će biti prikazana njegova lokacija, a taj tekst obojen je ljubičastom bojom.
Tekst analiziranog fajla biće prikazan u *textBox*-u u gornjoj polovini prozora, pri čemu će sve reči koje ne predstavljaju imenovane entitete biti obojene crnom bojom, a imenovani entiteti biće obojeni crvenom ili zelenom bojom, u zavisnosti od klase u koju su smešteni.
Entiteti koji predstavljaju imena i prezimena osoba, što su podaci o ličnosti koji se najčešće pojavljuju u dokumentima, biće obojeni crvenom bojom, dok će svi ostali imenovani entiteti (lokacije, organizacije, vremenske odrednice i drugi) biti obojeni zelenom bojom, a u donjem levom delu prozora pojavljuje se „oblačić" koji predstavlja legendu i korisniku pojašnjava način na koji su obojeni entiteti. Izgled prozora nakon prepoznavanja imenovanih entiteta u odabranom fajlu prikazan je na slici 3.

6. ZAKLJUČAK

U prethodnim poglavljima opisana je aplikacija koja korisnicima omogućava da izvrše prepoznavanje podataka o ličnosti u tekstualnim dokumentima sačuvanim u memoriji računara, napisanim na srpskom ili engleskom jeziku. Za implementaciju aplikacije iskorišćena je metoda koja se naziva prepoznavanje imenovanih entiteta (skr. NER).

Slika 3. *Prepoznavanje imenovanih entiteta u odabranom fajlu*

Aplikacija je kreirana korišćenjem programskog jezika *Python*, razvojnog okruženja *PyCharm* i biblioteke *Tkinter*.

Za prepoznavanje i klasifikaciju imenovanih entiteta iskorišćena su dva gotova modela: *SpaCy* i *Classla*. Najveće prednosti specificiranog i implementiranog rešenja jesu postojanje grafičkog interfejsa i mogućnost lakog odabira lokacije koju želimo da pretražujemo. Takođe, prednost implementiranog rešenja je i to što se korisnicima nudi mogućnost da odaberu tipove dokumenata koje žele da analiziraju.

Osim toga, dobra strana rešenja jeste način prikaza prepoznatih podataka o ličnosti u dokumentu. Umesto da bude naznačena samo pozicija podatka o ličnosti unutar teksta, prikazuje se kompletan tekst dokumenta, pri čemu su podaci o ličnosti posebno obojeni.

Najveća mana specificiranog i implementiranog rešenja jeste za nijansu slabija preciznost prepoznavanja imenovanih entiteta (izražena preko *precision* vrednosti) u odnosu na slična rešenja, analizirana u poglavlju 2. Još jedna mana implementiranog rešenja jeste nedostatak mogućnosti da se korisniku, ukoliko je odabrao folder ili disk u okviru kog će se vršiti analiza dokumenata, prikaže koliko se ukupno dokumenata (koje je moguće analizirati) nalazi u tom folderu/disku.

Ipak, navedene mane mogu se ispraviti kako bi aplikacija bila još učinkovitija, a modeli još precizniji u prepoznavanju imenovanih entiteta. Problem slabijih rezultata prepoznavanja entiteta može se rešiti tako što će se ručno izvršiti anotiranje velikog broja tekstova, dokumenata i novinskih članaka (posebno onih napisanih na srpskom jeziku), što će biti iskorišćeno za dodatno obučavanje modela.
Treba uzeti u obzir da bi trebalo anotirati tekstove iz što više različitih domena i oblasti, kako bi model „naučio" što više različitih reči, što bi omogućilo da ta znanja primeni za prepoznavanje entiteta u dokumentima koje korisnik odabere.
Mana koja se odnosi na prikaz broja dokumenata koji se nalaze u odabranom folderu mogla bi se rešiti tako što bi se, nakon odabira foldera, a pre pokretanja analize,

prebrojali svi dokumenti koji su sačuvani u odabranom folderu.

Taj broj bio bi prikazan korisniku, koji bi u tom slučaju mogao da proceni približno vreme izvršavanja analize i prepoznavanja imenovanih entiteta u dokumentima iz odabranog foldera.

7. LITERATURA

[1] Zakon o zaštiti podataka o ličnosti („Sl. glasnik RS“, broj 87/2018). [Online] Dostupno na: https://www.paragraf.rs/propisi/zakon_o_zastiti_podataka_o_licnosti.html. [Datum pristupa: 7. jul 2021.]

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Online] Dostupno na: https://gdpr-info.eu. [Datum pristupa: 7. jul 2021.]

[3] S. Atdag and V. Labatut, “A comparison of named entity recognition tools applied to biographical texts“, in Proc. of the 2nd International Conference on Systems and Computer Science, Villeneuve d'Ascq (FR), 2013 [Online]. Dostupno na: https://arxiv.org/abs/1308.0661. [Datum pristupa: 7. jul 2021.]

[4] S. Vychegzhanin and E. Kotelnikov, “Comparison of named entity recognition tools applied to news article”, in Proc. of the 2019 Ivannikov Ispras Open Conference (ISPRAS) [Online]. Dostupno na: https://ieeexplore.ieee.org/document/8991165. [Datum pristupa: 8. jul 2021.]

[5] C. M. Correia da Costa, G. Veiga, A. Jorge Sousa and S. Nunes, “Evaluation of Stanford NER for extraction of assembly information from instruction manual”, in Proc. of the 17th IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), April 2017 [Online]. Dostupno na: ieeexplore.ieee.org/document/7964092. [Datum pristupa: 9. jul 2021.]

[6] D. Altinok, *Mastering SpaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*, Packt Publishing LTD., Birmingham, UK, 2021.

[7] V. Batanović, N. Ljubešić, T. Samardžić and M. Miličević Petrović, “Otvoreni resursi I tehnologije za obradu srpskog jezika”, in Proc. of the Primena slobodnog softvera i otvorenog hardvera 2020 (PSSOH 2020), Beograd, Srbija [Online]. Dostupno na: www.researchgate.net/publication/349304650. [Datum pristupa: 13. jul 2021.]

Kratka biografija:

Đorđe Dragutinović rođen je 6.5.1997. god. u Zrenjaninu. Diplomski rad na Fakultetu tehničkih nauka odbranio je 2020. godine, a iste godine upisuje master studije na smeru Računarstvo i automatika – Inteligentni sistemi.

Kontakt: djordje.dragutinovic@uns.ac.rs