Klasifikacija pravnih dokumenata na osnovu više oznaka

Mihajlo Perendija

Petar Bašić

Uvod

Klasifikacija bilo kakvih dokumenata je neophodan i zahtevan posao koji mora biti odrađen u svakom poslovnom procesu, a naročito u domenu prava. U svrhe olakšavanja pretrage, pronalaska, upoređivanja i analize zgodno je posedovati dokumente klasifikovane na osnovu oznaka koje su im dodeljene. Domen prava kao takav zahteva postojanje velikog broja ovakvih oznaka, a proces njihovog dodeljivanja pojedinačnom dokumentu podrazumeva čitanje celokupnog dokumenta od strane domenski stručnog lica.

Napredak tehnologija iz domena obrade prirodnog jezika, u trenutnom stepenu razvoja, omogućava razvijanje softvera koji bi služili kao ispomoć u procesu označavanja dokumenta davanjem predloga za konkretne oznake. Uloga čoveka bi se svela na kontrolu i odabir najboljih oznaka čime bi se celokupan posao velikim delom automatizovao i ubrzao.

Oblast kompjuterske obrade pravnih dokumenata je još uvek mlada i kao takva ima mnoštvo problema koje je potrebno prevazići. Jedan od njih je i nedostatak adekvatnih skupova podataka usled velikog početnog napora koji je potrebno uložiti od strane stručnjaka. Napredak ipak postoji. U ovom radu će biti korišćen skup podataka koji je rezultat rada grčkih naučnika sa Atinskog univerziteta, a koji se sastoji od približno 24 hiljada pravnih akata iz zakonodavstva EU, kojima su pridružene oznake. Oznake su preuzete iz EUROVOC-a (Evropski rečnih pravnih pojmova), a njihov broj je oko 3600.

Glavni ciljevi ovog rada su istraživanje, vršenje eksperimenata i identifikacija najboljih metoda za predstavu pravnih dokumenata i za određivanje njima pripadajućih oznaka. Prilikom vršenja eksperimenata uzeta je u obzir i specifična struktura pravnih dokumenata koji su često izdeljeni na nekoliko sekcija.

Naredna poglavlja bave se korišćenim skupom podataka i metodama reprezentacije i klasifikacije pravnih tekstova koje su isprobane. Naposletku su prikazani i upoređeni rezultati, opisani problemi sa kojima su se autori susreli i donet zaključak na osnovu celokupnog istraživanja.

Skup podataka

Korišćeni skup podataka potiče iz EURULEX57k i predstavlja njegov podskup od približno 24 hiljade označenih dokumenata. Sami dokumenti su zakonodavnog tipa, a tiču se Zakona u okviru Evropske unije. Prosečno su dužine 727 reči, a njihova sadržina je izdeljena u 5 sekcija i to: naslov, zaglavlje, recitale (reference), glavni deo i priloge. Ovako izdeljenim dokumentima je pridružen skup oznaka koje se na njih odnose, a zatim su organizovani u JSON datoteke.

U svrhe obučavanja modela mašinskog učenja skup podataka je izdeljen na podatke za treniranje modela (train), za validaciju modela (validation) i podatke za testiranje (test).

Predprocesiranje skupa podataka

Nakon procesa ručnog dodeljivanja oznaka podacima, koje je izvršila Evropska kancelarija za publikacije, dobijen je skup podataka u kome je skoro polovina oznaka dodeljena u manje od 10 dokumenata. Ovakva distribucija oznaka može biti problematična prilikom obučavanja modela zbog njihove nejednake zastupljenosti, te je prvi korak obrade skupa podataka obuhvatio izbacivanje svih oznaka koje se javljaju u malom broju dokumenata. Od početnog broja od oko 3600 oznaka, ovim postupkom došlo se do skupa od 1289 najzastupljenijih oznaka, dok je izbacivanjem dokumenata koji su bili označeni samo izbačenim oznakama, eliminisano njih oko trideset.

Originalni dokumenti sadrže mnoštvo karaktera i/ili skupa karaktera koji nisu od velikog značaja za njihovu semantiku, kao na primer zagrade, crtice, apostrofi, i navodnici, odnosno svi karakteri koji nisu alfanumerici. Takvi karakteri nisu pogodni ni za računarsku reprezentaciju teksta, stoga je izvršeno njihovo uklanjanje. Dodatno, neki od dokumenata su sadržali karaktere enkodirane na pogrešan način (non-UTF8) te su i oni uklonjeni.

Priroda jezika je takva da se vrlo slične reči predstavljaju drugačije u zavisnosti od konteksta u kome se koriste, kao na primer nastavci reči za množinu ili pripadnost. Ukoliko bi svaki oblik reči bio uzet u obzir prilikom mašinske obrade, a ukoliko se razmatraju metode koje se zasnivaju na statističkoj obradi dokumenata, bila bi potrebna znatno veća količina resursa i vremena, dok bi dobijeni rezultati bili neznatno bolji, a potencijalno i lošiji. Iz ovih razloga je za potrebe nekih od reprezentacija nad tekstom izvršena tzv. lematizacija ili stemovanje svake reči. Oba od navedenih pristupa u osnovi svode neku reč na njen korenski oblik.

Podaci su nadalje obrađeni uklanjanjem tzv. „stop“ reči (stop words), odnosno reči koje ne nose značenje same po sebi, a javljaju se u velikom broju u svakom dokumentu. Ove reči su najčešće prilozi, predlozi, članovi i veznici.

Naposletku, isprobana je i metoda kojom se iz teksta izdvajaju samo imenice i glagoli, kao reči koje nose najveću količinu informacija.

U zavisnosti od izbora metode za vektorizaciju (predstavu teksta u numeričkom obliku) i metode mašinskog učenja, izvršene su i isprobane razne kombinacije prethodno navedenih metoda predprocesiranja.

Metode vektorizacije

Tekstualni oblik predstave dokumenta nije pogodan za mašinsku obradu, te ga je potrebno izraziti numerički. Numerička predstava nekog teksta je najčešće vektor koji sadrži njemu specifične vrednosti.

TF-IDF

TF-IDF je statistička mera koja procenjuje koliko je neka reč relevantna za neki dokumenti u kolekciji dokumenata. Objedinjuje dve metrike:

TF metriku koja određuje koliko se često neki pojam pojavljuje u dokumentu (term frequency) i

IDF metriku koja pokazuje specifičnost pojma za dokument predstavljenu kao inverznu funkciju broja dokumenata u kojima se pojam pojavljuje. Primer na skupu podataka koji se ovde razmatra može biti reč (skraćenica) „EU“, koja se pojavljuje vrlo često u svim dokumentima iz korpusa. Ukoliko se ta reč u nekoj dokumentu pojavi više puta nego što je prosek, računarski model, ukoliko koristi isključivo predstavu reči zasnovanu na TF metrici, može doneti zaključak da se taj dokument po nečemu izdvaja, što nije slučaj. IDF mera u takvim slučajevima smanjuje uticaj koji takvi pojmovi mogu imati na konačno donošenje odluka.

Drugim rečima, TF-IDF dodeljuje pojmu *i* težinu u dokumentu *j* koja je:

* Najveća kada se pojam *i* pojavljuje puno puta u malom broju dokumenata (dajući tim dokumentima moć razlikovanja od ostalih)
* Manja kada se pojam *i* pojavljuje manje puta u dokumentima ili se pojavljuje u velikom broju dokumenata (što daje manju relevantnost pojma)
* Najmanja kada se pojam *i* pojavljuje u gotovo svim dokumentima

Da bismo dokument predstavili u vektorskom obliku, potrebno je odrediti TF-IDF meru za svaku reč u tom dokumentu, a u odnosu na celokupni korpus. Vektor kojim se predstavljaju dokumenti je dugačak onoliko koliko ima jedinstvenih reču u celom skupu podataka. Svaka dimenzija vektora predstavlja jednu od reči. Vektorska reprezentacija jednog dokumenta je jedan takav vektor koji poseduje TF-IDF vrednosti na pozicijama koje odgovaraju rečima koje se u dokumentu i pojavljuju.

Word2vec

Ovaj pristup, na osnovu celog korpusa reči, generiše vektorski prostor od tipično nekoliko stotina dimenzija. Word2vec model, korišćenjem neuronske mreže, rekonstruiše lingvistički kontekst reči, odnosno uzima u obzir u kom kontekstu se neka reč pojavljuje u tekstu. Svakoj reči u korpusu se zatim dodeljuje odgovarajući vektor iz takvog prostora. Sličnost reči je tako predstavljena udaljenošću njihovih vektora u prostoru. Krajnja vektorska reprezentacija jednog dokumenta se dobija agregacijom vektora svake reči koja se u njemu pojavljuje. Dodatno, prilikom agregacije, vektoru svake reči može biti dodat ili oduzet značaj. Ovaj značaj može biti predstavljen kao jednostavna frekvencija pojavljivanja reči u dokumentu ili pak njena TF-IDF vrednost.

BERT