UNIVERZITET U BEOGRADU

FAKULTET ORGANIZACIONIH NAUKA

**ZAVRŠNI RAD**

**TEMA:**

Otkrivanje struktura veština za posao pomoću mašinskog učenja i sistema za preporuku

Informacini sistemi i tehnologije

Modul: Poslovna inteligencija

Mentor: Ime i prezime studenta:

Doc.dr.Miloš Jovanovic Boris Fidler

2017/3124

Beograd 2018.

Komisija koja je pregledala rad kadndidata

Boris Fidler

pod naslovom

Otkrivanje struktura veština za posao

pomoću mašinskog učenja

i sistema za preporuku

i odobrila odbranu:

dr. Miloš Jovanović, docent, mentor

page3image5878864

dr. Milan Vukićević, docent, član komisije

page3image5878864

dr. Sladjan Babarogić, vanredni profesor, član komisije

page3image5878864

# Apstrakt

Ljudski resursi predstavljaju jedan od najbitnijih faktora u jednoj organizaciji. Da bi omogućili da bude zaposlen kvalitetan kadar, neophodno je da svako radno mesto i njegove potrebe budu detaljno analizirani. U ranijim periodima to je bio posao koji se “ručno” radio time što bi zaposleni u odeljenju bili u komunikaciji sa menadžerima odeljenja koja otvaraju pozicije kako bi definisali potrebe novootvorenog radnog mesta. Zatim bi na osnovu primljenih radnih biografija radili selekciju i dalje korake kako bi se došlo do idealnog kandidata. Dolaskom ere kompijuterskih tehnologija i interneta, količina podataka potrebnih za obradu je prevazišla ono što bi se smatralo mogućim za obradu od strane čoveka, neophodno je bilo uključiti i neku vrstu softvera koja bi pomogla u tome. Odeljenje za ljudske resurse se godinama oslanjalo i bilo ograničeno na obradu ponuda od kandidata preko osnovnih menadžerskih aplikacija međutim njih su zamenili ili bolje rečeno njima je pomoglo uključivanje sistema za preporuku.

U ovom radu će biti prikazana analiza sistema za preporuku uz pomoć kojih će proces pravljenja oglasa i regrutacija potencionanih kandidata bili umnogome olakšani.

Uvodni deo rada će opisati proces regrutacije kakav je bio pre Internet ekspanzije i nakon.

Drugi deo rada će biti posvećen opisu trenutnih najvećih sistema koji se bave spajanjem poslodavaca i kandidata. Rad će se takođe baviti pregledom oblasti iz relevantne literature, prikazom prethodnih i srodnih istraživanja i njihovih zaključaka. Takođe će biće definisani kriterijumi potrebni za uspešno određivanje veština potrebnih za određeni posao. Nakon određivanja algoritama i razumevanja podataka na osnovu zadatih kriterijuma biće prezentovan rezultat kao i zaključak sprovedenog istraživanja. Cilj rada jeste da primenom sistema za preporuku omogući otkrivanje strukture sličnosti između veštine, grupe veština, opisa poslova (u formi slobodnog unosa), i veze između tih veština kako bi se poboljšao proces zapošljavanja kao i proces razvoja karijere za pojedinca. Sistem će dati uvid na sve relevatne veštine za određene poslove i time pomoći poslodavcu da pronađe pravog zaposlenog a onome ko traži posao omogućiti da stekne prave veštine koje bi mu obezbedile posao.

U poslednjem delu ovog rada prikazani su rezultati i benefiti dobijeni projektovanjem ovog modela, kao i zaključci sprovedene analize.

# Abstract

Human resources are one of the most important factors in an organization. In order to enable a high-quality grade of people to be employed, it is necessary that each workplace and its needs be thoroughly analyzed. In earlier periods, it was done as a "manual" or better said “by hand” job, that would have employees in the department communicate with department managers that open positions, to define the needs of a newly created job opening. Then, on the basis of the received working biographies, the selection and further steps would be made to arrive at the ideal candidate. With the advent of era of computing technologies and the Internet, the amount of data required to process has outperformed what would be considered possible for human processing, it was necessary to include some kind of software to help it. The HR Department has relied on, for years, limited processing of bids from candidates through basic managerial applications, but they have been replaced or, better to say, helped to include a recommendation system.

In this paper will be presented an analysis of the recommendation system by means of which the process of making advertisements and recruitment of potential candidates will be greatly facilitated. The introductory part of the paper will describe the recruitment process as it was before the Internet expansion and after. The second part of the paper will be devoted to the description of the current largest systems dealing with the merging of employers and candidates.

The work will also deal with the review of areas from the relevant literature, the presentation of previous and related research and their conclusions. It will also define the criteria needed to successfully determine the skills required for a specific job. After determining the algorithms and understanding the data based on the given criteria, the result will be presented as well as the conclusion of the conducted research. The aim of the paper is to enable the recommendation system to reveal the structure of similarities between skills, skills group, job descriptions (in the form of free entry), and the links between these skills in order to improve the employment process as well as the process of career development for the individual. The system will give insight into all the relevant skills for specific jobs and thus help the employer find the right employee and allow the job seeker to acquire the right skills to secure the job.

The last part of this paper presents the results and benefits obtained by designing this model, as well as the conclusions of the conducted analysis.

BIOGRAFIJA

e-mail: [borisfiddler@gmail.com](mailto:borisfiddler@gmail.com)

Lični podaci:

* Ime: Boris
* Prezime: Fidler
* Datum i mesto rođenja: 28.05.1987., Beograd, Zvezdara

Obrazovanje:

* Elektrotehnička škola “Rade Končar”, Beograd
* Osnovne akademske studije, Fakultet organizacionih nauka, studijski program: Informacioni sistemi i tehnologije
* Master studije, Fakultet organizacionih nauka, studijski program: Informacioni sistemi i tehnologije, modul: Poslovna inteligencija

Radno iskustvo:

* 2008-2012 , IT Tehnical Support Officer u Rkeeper, Srbija
* 2016-2017 , Data Governance Officer u Unicredit Srbija, Srbija
* 2017-2018, Software Developer u Unicredit Srbija, Srbija
* 2018-traje, RPA Developer u Nielsen, Srbija

Dodatne kvalifikacije:

* 2017, SAS Programing 2 – Data manipulation techniques , SAS institute, Srbija
* 2017, SAS Data integration Studio 2 – Additional topics, SAS institute, Srbija
* 2018, Advanced PL/SQL Programing, Oracle, Srbija

Lista slika, dijagrama i tabela

**Sadržaj**

[Apstrakt 4](#_Toc525401233)

[Abstract 5](#_Toc525401234)

[Prvo poglavlje 9](#_Toc525401235)

[1.1. Uvod 9](#_Toc525401236)

[1.2. Dosadašnji najvažniji rezultati u zadovoljavanju potreba u predmetnoj oblasti 10](#_Toc525401237)

[1.3. Ciljevi i formulacija problema 11](#_Toc525401238)

[1.4. Ciljna grupa i lična motivacija 12](#_Toc525401239)

[Drugo poglavlje 13](#_Toc525401240)

[2.1. Pregled stanja u predmetnoj oblasti 13](#_Toc525401241)

[2.2. Upravljanje ljudskim resursima 14](#_Toc525401242)

[2.3. Proces regrutacije i selekcija kandidata 15](#_Toc525401243)

[2.4. Uticaj razvoja interneta u načinu oglašavanja poslova 15](#_Toc525401244)

[2.5. Veštine kao deskriptor poslova u IT industriji 16](#_Toc525401245)

[Treće poglavlje 17](#_Toc525401246)

[3.1. Problem istraživanja i metodologija 17](#_Toc525401247)

[3.2. Mašinska obrada prirodnog jezika 18](#_Toc525401248)

[3.3. n-gram 20](#_Toc525401249)

[3.4 Sistemi za preporuku 22](#_Toc525401250)

[3.4.1. Zasnovani na sadržaju 23](#_Toc525401251)

[3.4.2. Zasnovani na saradnji 24](#_Toc525401252)

[3.4.3. Hibridni sistemi 24](#_Toc525401253)

[3.5. TF-IDF 25](#_Toc525401254)

[3.6. Kosinusna slicnost između vektora (Coasine similarity) 27](#_Toc525401255)

[3.7. Razvoj modela 28](#_Toc525401256)

[Četvrto poglavlje 29](#_Toc525401257)

[4.1. Rezultati istraživanja i diskusija 29](#_Toc525401258)

[Peto poglavlje 30](#_Toc525401259)

[5.1. Zakljucak 30](#_Toc525401260)

[Reference 31](#_Toc525401261)

# Prvo poglavlje

## Uvod

Proces regrutacije zaposlenih jeste centralna funkcija odeljenja ljudskih resursa pošto upravo ti novi zaposleni postaju faktor u proizvodnji nove vrednosti. Ukratko cilj jeste da proces regrutacije da na svom izlazu novozaposlenog radnika koji će za to preduzeće doneti najveću vrednost.

Sa aspekta poslodavca to podrazumeva pravljenje pozicije koja je okarakterisana nizom zahteva u vidu obrazovanja, kurseva i verovatno najbitnije veština za uspešno obavljanje posla. Od velikog je značaja osigurati potreban broj zaposlenih sa odgovarajućim sposobnostima i kvalifikacijama, kako bi se ostvarili postavljeni ciljevi , zadaci i misije organizacije.

Popunjavanje radnih mesta počinje procesom privlačenja, odnosno regrutovanja ljudskih resursa, a nastavak procesa je odabir, odnosno selekcija kandidata. Da bi se to uradilo pre svega je neophodno otvoriti nove pozicije i u njima definisati potrebne veštine koje kandidat treba da ima kako bi uspešno obavljao posao.

Često imamo situaciju u kojoj naziv pozicije ima dvosmislen naziv ili gore čak da pogresno opisuje potrebe tog posla i veština koje su potrebne za njega. Jedan od načina da se to prevaziđe bi bio da se grupe poslova sortiraju na osnovu naziva. Primenom mašinskog učenja i algoritama takav proces se može unaprediti i ubrzati time što ne bi radio samo analizu naziva pozicija već i veština koje su navedene u okviru njih. Ponavljanje termina među veštinama moze ukazati na one koji najbolje opisuju taj posao ali takođe i da izdvoji opšte od naročitih. Tako odrađena analiza bi u perspektivi mogla da bude aplicirana i na većem skupu podataka sa podjednako dobrim rezultatima.

## Dosadašnji najvažniji rezultati u zadovoljavanju potreba u predmetnoj oblasti

Potraga za poslom zapocinje procesom u kome pojedinci traze zaposlenje u sferi koja najvise odgovara njihovim kriterijumima i sposobnistima.

Sa druge strane, pronalazenje odgovarajuceg kandidata je kljucni zadatak za odeljenje ljudskih resursa. Oni na osnovu potrebe kompanije otvaraju pozicije, vrse selekciju i na kraju obezbedjuju kvalitetan kadar. Oba problema, trazenja posla i trazenja zaposlenog, se svode na isti u tom smislu da se radi o kontinuiranom procesu i komunikaciji izmedju njih koja kao rezultat moze obezbediti odgovarajuceg novozaposlenog za kompaniju i posao za zaposlenog. (Domeniconi, Moro, Pagliarani, & Pasolini, 2016)

Informacione tehnologije su u proteklim godinama promenile nacin na koji nacin ljudi dolaze do zaposlenja kao i nacin na koji rade uopste. Sprovedeno istrazivanje top 1000 kompanija u Nemackoj pokazalo je da je pojava interneta zamenila stampane medijume oglasavanja kao glavni kanal za regrutaciju. Sa oko 78% upraznjenih mesta koja su objavljena na sajtovima kompanija i 49% novootovrenih pozicija koje su oglasene putim internet portala za zaposaljavanje dolazimo do zakljucka da kanali putem interneta su daleko nadmasili starije nacine oglasavanja stampanim putem.

Takodje istrazivanje kaze da se procenat zaposlenja putem interneta podigao za 58% 2004. godine. (Malinowski, Keim, Wendt, & Weitzel, 2006)

Sve veci broj onih koji traze posao to danas cine upravo tako sto dele svoja akademska dostignuca i profesionalne informacije putem interenta. U isto vreme kompanije sve brze prihvataju svet u kome se proces regrutovanja desava “online”. Na osnovu istrazivanja “Jobvite”, 68% online ljudi koji traze posao su diplomci ili post-diplomci, dok 94% poslodavaca koriste ili planiraju da krenu da koriste drustvene mreze za regrutovanje novozaposlenih.

Ako se uzme samo pretraga na “Google.com” 30% sto je oko 300 miliona pretraga po mesecu je upravo u vezi sa zaposaljavanjem. (Patel, Kakuste, & Eirinaki, 2017)

## Ciljevi i formulacija problema

Platforme za zapošaljavanje putem interneta polako postaju primarni kanal za oglašavanje i pronalaženje kandidata za većinu kompanija. Dok je takav pristup omogućio da se vreme pronalaženja mogućih kandidata smanji kao i da se značajno umanje troškovi u smislu mesta za oglašavanje, problem tradicionalnih načina je nastavio da postoji.

Pristup logičkog operatora pretrage je jedan od njih gde se sistemom samo direktnog poklapanja dolazilo do određenih kandidata što je prouzrokovalo da mnogi propuste svoju šansu da budu izabrani. Sistemi preporuke imaju za cilj da upravo taj nedostatak nadomeste i da omoguće poslodavcima da dođu do idealnih kandidata. Da bi se to ostvarilo postoje različiti načini kako se uz pomoć sistema za preporuku upravo moze doći do željenih rezultata.

Pronalazenje pravog kandidata “ručno” je dugotrajan i mukotrpan proces, stoga su izmišljene metode koje pomažu u ovom procesu pod nazivom sistema za preporuku za pronalaženje poslova što se tiče mogućih kandidata tj. sistema za regrutovanje sa aspekta poslodavaca.

Odeljenje ljudskih resursa ima zadatak za napravi selekciju kandidata sa odgovarajućim veštinama koje tom preduzeću zaista i trebaju. Koncept veštine je od izuzetnog značaja zato što u mnogim slučajevima može mnogo bolje da oslika potrebu poslodavca ili onoga što moguci kandidat moze da ponudi, nekada i bolje od recimo diplome ili završenog kursa.

Danas, ako izuzmemo posebno pravljene programe unutar preduzeća iliti “in-house”, društvene mreže poput LinkedIn, Facebook, Twitter, Dice.com i sl. igraju jako bitnu ulogu u procesu regrutovanja zbog informacija koje su dostupne na tim mrežama. (Domeniconi, Moro, Pagliarani, & Pasolini, 2016)

## 1.4. Ciljna grupa i lična motivacija

U novijem dobu gde je ekspanzijom digitalnih podataka i pojavljivanjem e-platformi za zaposljavanje bila je neohodna reorganizacija nacina na koji kompanije obavaljaju odredjene aktivnosti u odredjenim sferama. Jedna od tih je bila sfera regrutacije.

Postavljanja poslovnih ponuda na internet stranicama kompanija se uglavnom vrsilo na delu stranice pod nazivom Karijera (“Carrier”). Zainteresovani bi odlazili na te stranice i aplicirali putem neke online prijave. Medjutim vremenom su se razvile i platforme specijalno dizjanirane za proces regrutacije tj. zaposaljvanja. Na tim mestima su zainteresovani mogli da naprave svoj profil koji bi zatim popunili sa kljucnim informacijima u vezi sa njihovim skolovanje, prethodnim iskustvom i vestinama koje poseduju. Nakon toga bi ukoliko bi se otvorila nova pozicija imali mogucnost da apliciraju na istu i time steknu sansu u buducem zaposlenju. Nazalost ovakav sistem je bio odrziv samo u ranijim fazama interneta. Trenutna situacija je bila takava da bi za odredjenu poslovnu ponudu stizale hiljade odgovora tj apliciranja. Iz ovoga se stvorila potreba za sistemima za preporuku.

Sa druge strane cak i trazenje posla je nekada mogao biti iscrpljujuca aktivnost. Najcesci pristup bi bio da se uz pomoc par kljucnih reci izvrsi pretraga na stranici potencionalnog poslodavca. Rezultat pretrage bi vratio listu poslova koji sadrze neku od tih reci. Nazalost takav rezultat nije garantovao da bi kandidatu odgovarao posao na osnovu njegovih sklonosti i znanja.

Ovaj rad može poslužiti obema stranama u smislu da za poslodavca znači da može da postavi oglas sa najrelevantnijim veštinama i zahtevima za taj posao, dok kandidat može u svakom trenutku da bude upoznat sa nivoom kompleksnosti posla na traženim pozicijama. To znači da bi obe strane profitirale zato što bi sistem za preporuku alalizom strukture traženih veština pomogao potencionalnim kandidatima da steknu prave veštine koje bi im obezbedile posao.

# Drugo poglavlje

## 2.1. Pregled stanja u predmetnoj oblasti

Odeljenje za ljudske resurse se godinama oslanjalo i bilo ograniceno na obradu ponuda od kandidata preko osnovnih menadzerskih aplikacija. Ovakvi sistemi je pokretalo interna komunikacija izmedju LJR i drugih odeljenja.

Ali sa ekspanzijom kolicine podataka na internetu i uopste podataka u digitalnom formatu i razvojem e-biznisa zahtevalo je odredjenu reformu nacina na koji su kompanije do tog trenutka obavljale proces regrutacije. Platforma za regrutovanje putem interneta jeste jedna od najuspesnijih otkrica putem koje se poslodavci i kandidati otkrivaju. Ovakve platforme su dozivele pravu ekspanziju zbog sve tezeg procesa regrutovanja novih potencijala. Za svaki postavljen oglas za posao, hiljade biografija se svakodnevno moze poslati za otvorene pozicije. Analogno tome takodje postoji ogroman broj radnih biografija koje polako postaju dostupne online.

Takava ogromna kolicina dosupnih informacija u vidu poslovnih ponuda, njihovih zahteva kao i radnih biografija je postala odlicno mesto za unapredjivanje kvaliteta u smislu kolicine poklapanja zahteva i ponude. Naravno taj kvalitet je ostao na niskom nivou zbog pristupa logickog operatora (true-false) i time velike kolicine podataka i sansi za dobro poklapanje ostaju neiskorisceni. Iz toga se stvorila potreba za koriscenjem sistema za preporuku kako bi poslodovaci uspesno mogli da obrade ogromne kolicine podataka brzo i efikasno. Svakako resenje nije moguce ostvariti u kratkom roku i samsi sistemi za preporuku predstavljaju izazov u polju istrzivanja koje se i dalje razvija.

Kako bi se ta ideja dalje razvijala u ovom radu cemo se baviti nekim problemima i potencionalnim resenjima za unapredjenje sistem za prepruku kod otkrivanja vestina potrebnih za sam posao.

Peronalizovani sistemi kao sto su sitemi za porepruku su u proteklim godinama privukli paznju velikog broja istrazivaca. Od kada se termin “Sistemi za preporuku” prvi put pojavio u izdanju magazina “Communications of the ACM“ vreme i istrazivanje je ulozeno kako bi se takvi sistemi poboljsali i ucinili pouzdanijim. Jedan deo istrazivaca se fokusirao na spajanje sistema za preporuku na osnovu sadrzaja sa sistemom za preporuku na osnovu saradnje kako bi se prevazisao problem oskudnosti podataka dok su se drugi okrenuli ka dimenzionalnoj redukciji korisnik-podatak matrice koja je prisutna u sistemima za preporuku zasnovanim na saradnji. (Resnick & Varian, 1997)

## 2.2. Upravljanje ljudskim resursima

Svaka organizacija ima potrebu za ljudima, ali i ljudi imaju porebu da budu deo oragnizacije, kako bi zajedničkim delovanjem ostvarili zadate ciljeve poslovanja. Svaka strategija poslovanja, neminovno polazi od ljudskog kapitala, koji ima najvažniju ulogu u procesu rada. Kako tehničko-tehnološke, ali i društvene promene, zahtevaju različite profile kadrova, nivo zahtevanog obrazovanja zaposlenih se povećava, a s druge strane, sve veća podela rada uslovljava stručne specijalizacije, što dovodi do ograničene mobilnosti ljudskih resursa. Planiranje rasta i razvoja organiazcije mora da bude prožeto nastojanjima da se obezbedi potrebna strukutura zaposlenih, dok su ljudski resursi upravo i inicijator rasta i ostalih promena u poslovanju organizacije.

Kako su kadrovi osnovi stub svakog poslovanja, njihovo planiranje mora da bude deo strategije rasta organizacije, a kako bi se ostavrila veza između planiranja razvoja kadrova i poslovne strategije, moraju da postoje osnovni nivoi planiranja. (Petković, 2014)

Aktivnosti sektora za ljudske resurse su: planiranje ponude i tražnje za ljudskim resurisma, analiza posla, regrutovanje potencijalnih kandidata, selekcija kandidata i njihova socijalizacija po prijemu na rad, ocenijivanje performansi zaposlenih, njihova obuka i razvoj karijere, nagrađivanje, radni odnosi i kolektivno pregovaranje, upravljanje procesom napuštanja organizacije, zdravlje i bezbednost zaposlenih na radu i ostalo.

## 2.3. Proces regrutacije i selekcija kandidata

Kada se spomene proces regrutacije misli se na privlacenje potencionalnih kandidata u odredjenom vremenskom roku i u broju koji je predvidjen sa odgovarajucim kvalifikacijama da se prijave za posao unutar organizacije

Ekonomsko stanje zemlje odredjeuje uslove i znacajno moze da utice na mogucnost organizacije da privuce kvalitetan kadar. Ukoliko se radi o stabilnoj ekonomskoj situaciji, sa niskim nivoom nezaposlenisti, tada postoji verovatnoca da ce organizacija morati da se takmici sa drugim slicnim organizacijama kako bi privukla one vredne zaposaljavanja. Ako se ipak radi o slabijoj ekonomiji dolazi se do situacije, gde ukoliko postoji visoka nezaposlenost, veliki broj prijava moze stici na otvorenu poziciju sa samo par onih koji zaista ispunjavaju kriterijume poslovnih zahteva novootvorenog radnog mesta.

Nacin na koji ce kompanije oglasiti svoje pozicije varira od angazovanja ekternih firmi koje traze zaposlene za odredjeni procenat do postavljanja oglasa na sajtu kompanije ili novina. Takodje postoje varijante staziranja kako bi se procenio kvalitet kadra.

Medjutim najznacanije promene u procesu regrutacije su one koje su nastale pojavom i upotrebom online sistema regrutacije.

## 2.4. Uticaj razvoja interneta u načinu oglašavanja poslova

U danasnje vreme mnoge organizacije postavljaju otvorene pozicije na posebnim lokacijama na interenetu poput “Carrier Builder”,”Monster.com”,”Linedin.com”,”dice.com” i sl. time kompletno prelazeci na elektronski nacin regrutovanja. To naravno nije bez razloga uradjeno, vec je posledica prednosti koje takav sistem sa sobom donosi, poput nizih cena oglasavanja, lakoce i brzine postavljanja kao i sireg skupa ljudi nego sto bi to bilo moguce sa starijim sistemima. (Gusdorf, 2008)

Onima koji traze posao ovo takodje olaksava zato sto u jednom danu mogu poslati svoje radne biografije time aplicirajuci na vise mesta. Ovo nazalost stvara probleme u odeljenju ljudskih resursa zato sto zbog povecanog obima prijava moraju napraviti selekciju onih koji ne zadovoljavaju kriterijumom potrebe otvorene pozicije. Medjutim kako se vestine i kvalifikacije danas sve vise razvijaju i postoji citav dijapazon istih postoji uvek sansa da neki kandidat nije uzet u obizr iako je odgovarao. Da bi se ovakve situacije izbegle i unapredio sistem je upravo cilj i svrha ovog rada.

## 2.5. Veštine kao deskriptor poslova u IT industriji

# Treće poglavlje

## 3.1. Problem istraživanja i metodologija

U ovom radu ce biti izvseno istrazivanje struktura različitih veština koje se traže, otkrivati faktori i hijerarhija veština, sve sa ciljem boljeg razumevanja skupa potrebnih veština za zapošljavanje, i preporučivanje neophodnih veština.

Uzorak podataka predstava deo od 4.6 miliona oglasa za posao koje su izvučene od jednog od najvećih američkih sajtova za pronalaženje poslova u svetu informacionih tehnologija ”dice.com”. (PromptCloud, 2017)

Dodatno, analiziraće se i tekstovi oglasa, i izvlačiti liste veština koje najbolje opisuju ponude. U poslednjem koraku se, uz pomoć sistema za preporuku zasnovanih na sadržaju, predlažu veštine za posao, po principu da slični poslovi imaju slične zahteve tj. zahtevane veštine.

## 3.2. Mašinska obrada prirodnog jezika

Svakodevno, ljudi razmenjuju hiljade reci koje drugi ljudi mogu da interpretiraju na razlicite nacine. Jednostavnije receno radi se o komunikaciji, ali naravno uvek treba imati na umu da reci mogu imati mnogo dublje znacenje u zavisnosti od konteksta. Masinska obrada prirodnog jezika (Natural Language Processing - NLP) se fokusira na konteksutalni patern pre nego na vokalni nacin na koje su reci izgovorene (Mills, 2018)

Masinska obrada prirodnog jezika je u svojoj osnovi forma vestacke inteligencije koja analizira ljudski jezik. Postoje varijateti medjutim sve imaju zajednicku osobinu da predstavljaju tehnologiju koja pomaze masinama da razumeju nas jezik, pa cak i da komunicijaju sa nama koristeci nas jezik. Kao izvor za svoja saznanja ona koristi razlicite dicipline ukljucujuci racunarske nauke i računarsku lingvistiku u svom zadatku koji je smanjenje jaza izmedju ljudske komunikacije i razumevanja od strane racunara. (SAS, 2018)

Kao ljudi mi pricamo jezikom koji je nama poznat medjutim jezik racunara je masinski jezik koji je nama nerazumljiv. Na svom najnizem nivou masine komuniciraju na osnovu jedinica i nula u vidu signala. Od vremena kada su se podaci spustali na magnetnu traku do danasnjeg vremena gde vecina modernih uredjaja poseduje neku vrstu personalnog asistenta koji je takodje naravno digitalni.

Upravo je masinska obrada prirodnog jezika omogucila racunarima da citaju tekstove, razumeju govor, interpretiraju ih i iz toga izvuku delove koji su od znacaja

Svoju istoriju je zapocela 1950 godine kao mesavina vestacke inteligencije i ligvistike. Njegovi koreni vuku iz tehnologije za izvlacenje teksta (Informational retreval - IR), koja primenjuje statisticki zasnovane tehnike da indeksira i pretrazi velike kolicine teksta efikasno.

Odlike prirodnog jezika kao sto su velicina, nestruktuirana priroda kao i cinjenica da je podlozno interpretaciji dovelo je do problema kada su u pitanju standardni pristupi parsiranju.

Masinska obrada prirodnog jezika mora da omoguci izvlacenje sematnike iz teksta uzimajuci u obzir delove govora tj. teksta kao sto su imenice, glagoli, pridevi i sintaksa recenice. (Prakash, Lucila, & Wendy, 2011)

I dok su nadgledano i nenadgledano masinsko ucenje siroko rasprostranjeni nacin za modeliranje ljudskog jezika takodje postoji potreba za sintaksnom i semantickom razumevanju koji u vecem delu masinskog ucenja. Masinska obrada prirodnog jezika pomaze u otkljanjanju problema viseznacnosti u jeziku i postavljanju numericke strukture koju kao izvor koriste programi za prepoznavanje govora ili analizu teksta. Ona ukljucuje razlicite tehnike intepretiranje ljudskog jezika koja se krece od statistickih metoda i masinskog ucenja do metoda zanovanih na pravilima i algoritmima.

Masinska obrada prirodnog jezika podrzumeva tokenizaciju i parsiranje, kao i pravljenje recnica na osnovu recnika i morfoloske analize i vadjenja korena reci (“lemmatization and stemming”) (Manning, Prabhakar, & Schütze, 2008). Ovakvim pristupom inverzno iz jezika dolazimo do teksta i na kraju iz teksta do skupa reci.(Slika 1.)

Recnik se sastoji od Tekst cini niz reci Jezik je konstruiran

skupa reci iz recnika iz skupa svih mogucih tekstova



(Neves, 2016)(Slika 1.) Skup reci u nizu predstanja recnik, dok skup svih recnika cini jezik

## 3.3. n-gram

Kategorizacija teksta predstavlja kljucni zadatak u obradi dokumenata, cime se stice mogucnost obrade ogromne kolicine podataka koji se nalaze u elektronskoj formi. (Trenkle, 2001) Ono sto predstavlja realnu poteskocu u obradi teksta u elektronskim dokumentima jesu tekstualne greske bile ona gramatickog ili sintaksnog tipa. Da bi sistem teksutalne kategorizacije bio proglasen pouzdanim sva razlicite vrste izvoda podataka neophodno je da ipak ima neku dozu tolerancije na oredjene greske kao i nacin da ih prevazidje.

Dokumenta u elektronskom formatu poticnu iz raznih izvora. Neki su generisani kao deo programa koji je zaduzen za pisanje teksta i kao takvi podredjeni su poslovnoj logici ili tzv. Speel-check programa dok su drugi slobodnog formata unosa poput email-a. Druga pomenuta grupa je uglavnom ona koja se stvara na licu mesta bez neke preprovere ili provere nakon stvaranja, kao sto su skenirani dokumenti uz pomoc posebnih softvera za prebacivanje iz papirne u elektronsku formu.

Prilikom skeniranja ili bolje receno prepoznavanja teksta sa papira neminovno je da ce nastati neka greska i upravo je to tip situacije koji bi zahtevao dalju proveru koja je skupa i komplikovana i gde bi neki sistem za proveru teksta bio dobrodosao.

U ljudskom jeziku neke reci se pojavljuje cesce od drugih. Jedan od osnovnih nacina da se iskaze takva ideja je danas poznata kao Zip-ov zakon koja je prikazana ispod na grafikonu (Grafikon 1.) (Kingsley, 1950). On kaze da je pojavljivanje n-te najcesce reci u ljudskom jeziku u tekstu inverzno proporcionalno n. Ovo znaci da najcesca rec u nekom jeziku se pojavljuje dva puta cesce od druge reci i tri puta cesce od trece itd, implicirajuci da uvek postoje reci koje su na neki nacin dominatne u tom jeziku. Ovakav zakljucak se odnosi i na reci generalno ali i na odredjene teme.

|  |  |  |
| --- | --- | --- |
| RANG | UČESTALOST | REČ |
| 0 | 44610 | the |
| 1 | 24997 | to |
| 2 | 19904 | of |
| 3 | 18037 | and |
| 4 | 17251 | a |
| 5 | 16694 | in |
| 6 | 8911 | s |
| 7 | 8728 | for |
| 8 | 8546 | is |
| 9 | 8022 | The |
| 10 | 7934 | that |

(Grafikon 1.) Zip-ov zakon distribucije reci u prirodnom jeziku

n-gram predstavlja kontinuiranu sekvencu od n reci iz neke recenice ili teksta na osnovu govora. To znaci da to moze biti veznik, slovo, rece, slozenica ali ono sto ih sve odlikuje jeste da se posmatraju kao atomska vrednost. N-gram od jedne “stavke” tj. reci se naziva “unigram”,od dve “bigram”,tri “trigram” itd. (Hong, Nduyen, Duong, & Snasel, 2016)

Ovakva podela ce biti od izuzetne koristi u ovom radu posto se u pretprocesiranju radi obrada teksutalnih polja za vestine kao i za sam opis posla.

## 3.4 Sistemi za preporuku

Sistemi za preporuku sve brze postaju defakto nacin za preporuku u razlictim aplikacijama za preporuku prozivoda,usluga i sveukupno informacija korisnicima istih. Mnoge internet aplikacije su se pridruzile trendu koriscenja sistema za preporuku da bi obezbedile veci profit time sto bi korisnicima smanjile vreme pretrage i dolazenja do idealnog proizvoda ili usluge.

Neke od takvih kompanija su “Amazon”, “Microsoft”,”Aliexpress”,”Netflix” i slicno.

Sve ove kompanije su uspesno integrisale sisteme preporuke za komercijalnu upotrebu i time uvecale prodaju a samim tim i prihode prilikom prodaje preko interneta i jos bitnije obezbedile lojalnost kupaca. Druge kompanije su takodje razvijale inhouse genericke sisteme preporuke a neki od njih su “Net Perceptions,Epiphany”, “Art Technology”, “Broad Vision”.

Kompanije koje koriste sisteme za preporuku fokusiraju se na povecanje prihoda na osnovu bolje personalizove ponude i poboljsanja zadovoljstva korisnika. Po pravilu oni ubrzavaju vreme pretrage i olaksavaju da korisnici dodju do sadrzaja koji je relevantan za njih. Sa tako personalizovanim sardzajem komapnije dobijaju prednost na trzistu i smanjuje se mogucnost gubitka korisnika od strane konkurencije. (Rodríguez, 2018)

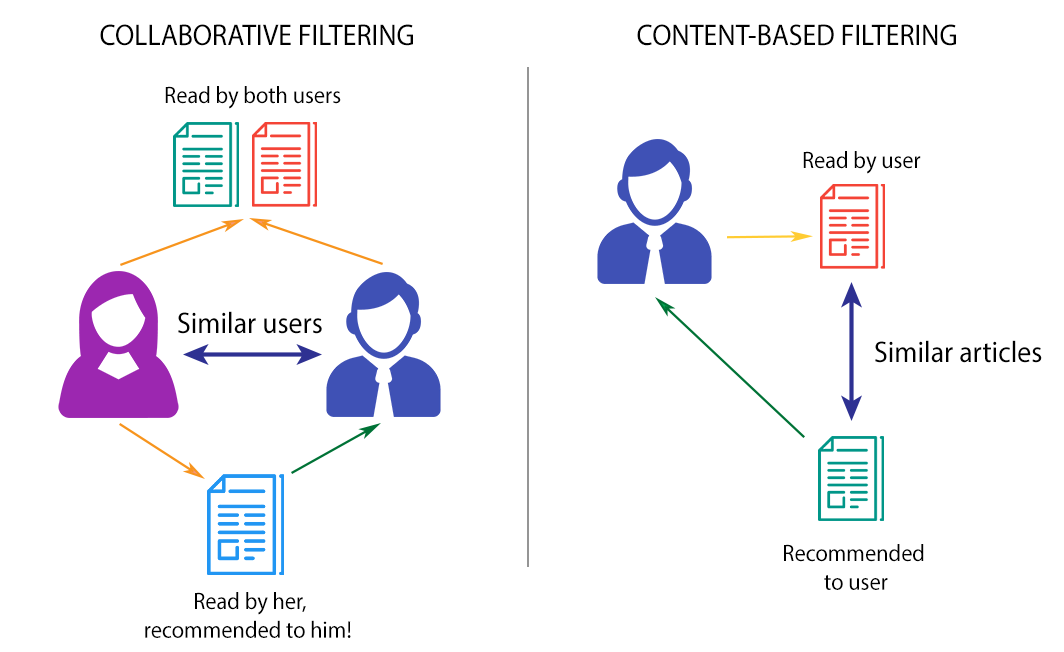
Na osnovu ovoga mozemo napraviti klasifikaciju sistema za preporuku na:

* Zasnovani na sadrzaju - koji koriste karakteristine informacije
* Zasnovani na saradnji - koriste korisnik-element interakcije
* Hibridni sistemi za preporuku – kobinuju gore navedene sisteme za preporuku sa ciljem da se izbegnu nedostaci oba

Oba sistema imaju, naravno, svoje prednosti i mane. Sistemi za preporuku zasnovani na sadrzaju su ograniceni u njihovim mogucnostima u smislu da ce preporucene stvari bili vise slicne onim stvarima na osnovu koji je preporuka i napravljena. Sa druge strane sistemi za preporku zasnovani na saradnji daleko su bolji od onih zasnovanih na sadrzaju u otkrivanju skrivenih paterna. Takav sistem preporuke se vise koncetrise na korisnika i njegove preference pre nego na sadrzaj onoga sa cime je korisnik imao dodira. (Hopmans, 2015)

Nedostatak sistema za preporuku zasnovanih na saranji jeste to sto su mu za dobre preporuke potrebne vece kolicine podataka u vidu istorije pregleda korisnika da bi otkrio dobre paterne. Tako nesto nije problem kada se radi sistemima zasnovanim na sadrzaju koji mogu da daju rezultate sa malo a nekada i bez ikakve informacije o istoriji podataka i stoga su laksi za implementaciju.

Pojednostavljena reprezentacije razlike ova dva sistema je prikazana na Slici 2. Gde se jasno vidi da u prvom slucaju ukoliko su oba korisnika imala interakciju sa istim ili slicnim sadrzajem onda sistem sa odredjenom sigurnoscu moze da preporuci odobi B nesto sa cime je osoba A imala interakciju. U dugrom slucaju se vidi da ukoliko je korisnik imao interakciju sa nekim sadrzajem sistem moze da mu preporuci drugi sadrzaj na osnovu slicnosti izmedju njih.



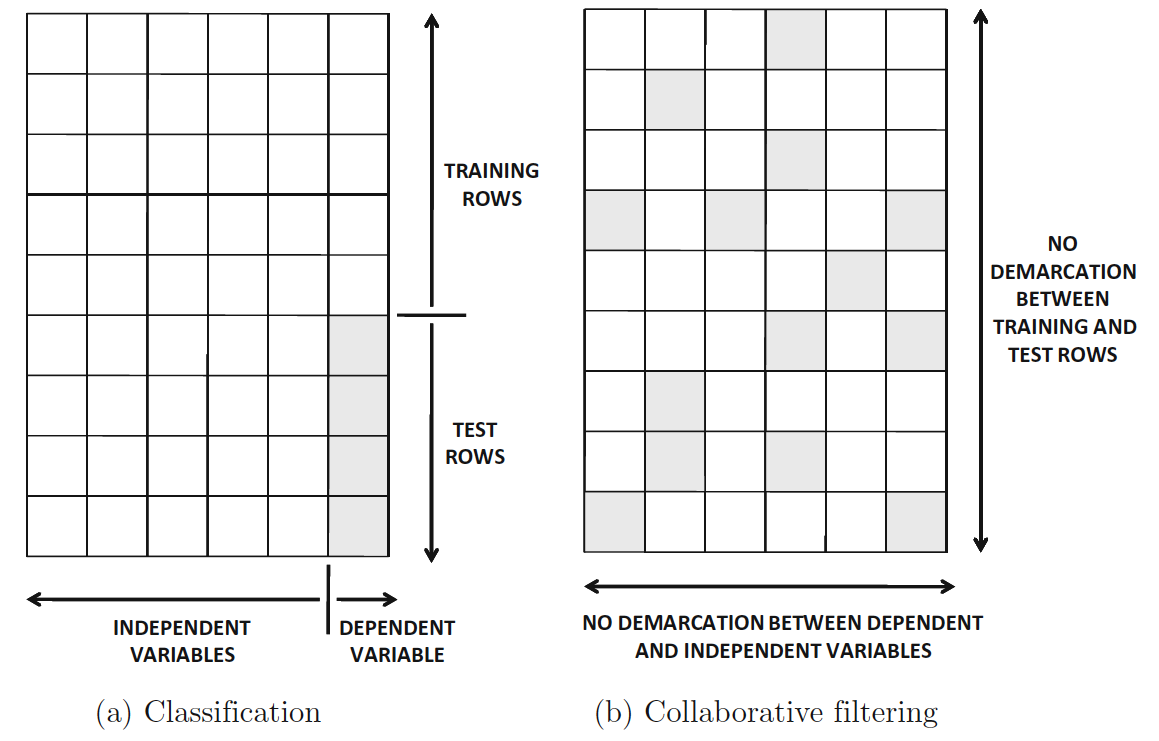
(Hopmans, 2015) (Slika 2.) Razlika izmedju sistema za preporuku zasnovanog na sadradnji u odnosu na onog zadnovanog na sadrzaju

### 3.4.1. Zasnovani na sadržaju

Princip preporuke zasnovane na sadržaju je da predloži članke koji imaju slične informacije o sadržaju odgovarajućim korisnicima. Na primer, u preporuci koja preporučuje poslove podnosiocu posla, sadržaj je lične informacije i želje za radom. Dok preporučuju kandidate za regrutare, opis posla koji postavljaju regrutatori, uključujući i pozadinski opis preduzeća, koristi se kao sadržaj za preporuke.(ovo nista ne valja)

### 3.4.2. Zasnovani na saradnji

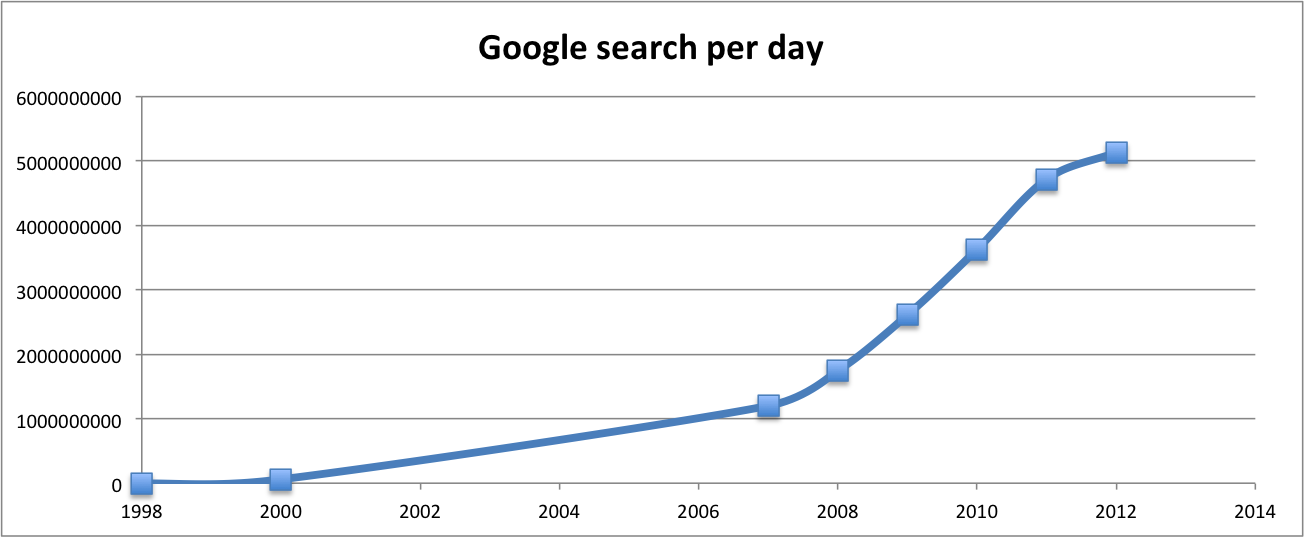
Glavna karakteristika sistema za preporuku zasnovanih na saradnji (“Collaborative filtering recomender systems”) jeste to su oni u potpunosti nezavisni od racunarske reprezenatacije objekata koje oni ustvari preporucuju. To znaci da mogu savrseno dobro da rade i sa objektima poput muzike, slika ili filmova. Sa druge strane postoji i odredjeni nedostaci ovakvog sistema za preporuku. Jedni od najpoznatijih su problem pocetnog nedostatka podataka (“cold-star problem”) i rasutost podataka. (Tondji, 2018)



### 3.4.3. Hibridni sistemi

## 3.5. TF-IDF

Davne 1998 godine Google je vrsio obradu od oko 9800 pretraga dnevno. U 2012 godini ova brojka se popela na 5.13 milijardi pretraga dnevno sto se moze videti na slici 3. (Vembunarayanan, 2013)



Slika 3. Kolicina pretrage na Google.com na godisnjem nivou

Nacin na koji je Google doziveo ovakav rast se krije upravo u koriscenju algoritma slicnom TF-IDF pod nazivom pageRank algoritam. On uzima u obzir koliko je sajt koji je rezultat pretrage relevantan ali takodje uzima u obizir i pretragu korisnika da bi uporedio sve relevantne dokumente i ocenio ih.

Ako primenimo ovu logiku nad tekstom ocenjivanje relevatnosti neke reci zapocinje njenim prisustvom u tom tekstu kao i njenom frekventnoscu unutar samog dokumenta. Medjutim ocenjivanje samo na osnovu broja ponavljanja reci u dokumentu nije dovoljan pokazatelj za njen znacaj tj. vrednost. Na kraju svakom terminu u dokumentu dodeljujemo tezinu koja zavisi od broja ponavljanja unutar samog dokumenta. Ovakav nacin dodeljivanja tezine na osnovu broja ponavljanja se naziva - “ucestalost termina” (“term-frequency”).

Vazno je navesti i da u algoritmu pod nazivom “skup reci” (“Bag of word”) koji pojednostavljuje analizu reci ,u tom smislu da zanemarije semantiku i gramatiku, koja se pojavljuje u prirodnom jeziku kao i izvlacenje informacija. To bi znacilo da dva dokumenta sa slicnim recima jesu slicni jedan drugom bez obzira na raspored tih reci unutar svakoj od dokumenata. Ono sto se postavlja kao pitanje jeste relevantnost svih reci unutar dokumenta, gde kao logican odgovor dobijamo da sve reci nisu podjednako bitne bez obzira na njihovu frekventnost.

Iz svega navedeno vidimo da sama ucestalost termina ima fundamentalni problem: svi termini se postmatraju kao podjednako bitni. Cak iako neki od njih imaju minimalnu diskriminacionu moc za pruzanje relevantnosti. Kao resenje se nudi “inverzna ucestalnost termina” (“inverse document frequency”) koju dobijamo ako sa ukupnim broj dokumenata u kolekciji (N) podelimo usestalost pojaviljavanja termina (t) u toj kolekciji . (Manning, Prabhakar, & Schütze, 2008)

idft = log (N / dft)

Iz ovoga mozemo zakljuciti da je inverzna ucestalost termina (idft) visa ukoliko se termin redje pojavljuje ili niza ukoliko se on pojavljuje cesce u kolekciji.

Sledeci korak jeste da kobinovanjem dft i idft  dobijemo slozenu (mesovitu) tezinu za svaki termin u svakom dokumentu. TF-IDF (tf-idft,d) dodeljuje tezinu terminu t u dokumentu d na sledeci nacin:

tf-idft,d = tft,d × idft

Drugim recima tf-idft,d dodeljuje tezinu terminu t u dokumentu d na nacin na koji:

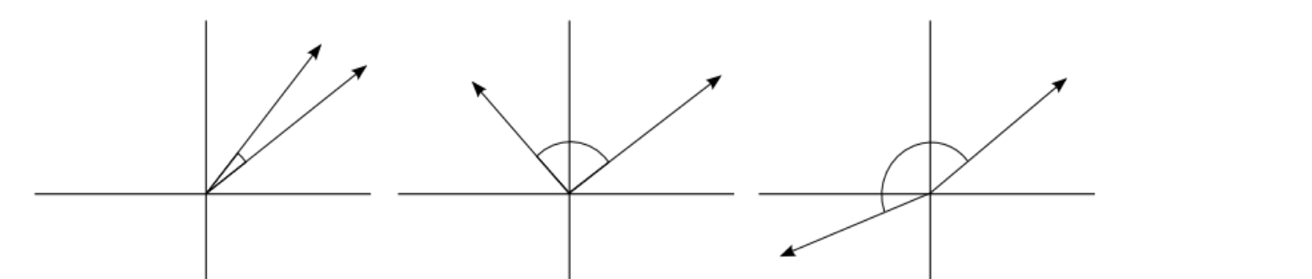
* Termin poseduje najvecu tezinu kada se pojavljuje u velikom broju na malom broju dokumenata unutar kolekcije
* Temin poseduje nizu vrednost ako se redje pojavljuje u dokumentu, ili se pojavljuje u vecem delu dokumenata ( time sto je znacaj tog termina umanjen)
* Termin poseduje najnizu vrednost kada se pojavljuje u skoro svim dokumentima

Sada ukoliko posmatramo svaki dokument kao vektor u kome svaki deo korespondira jednom od termina zajedno sa tezinama koje su dodeljne kao rezultat TF-IDF, takav vektor se moze iskoristi u daljoj analizi kod pronalazenja slicnosti izmedju dokmenata ili u slucaju ovog rada pronalazenja poslovnih ponuda sa potrebim i slicnim vestinama.

## 3.6. Kosinusna slicnost između vektora (Coasine similarity)

Kosinusna slicnost izmedju dva vektora, ili u nasem slucaju dva dokumenta u vektorskom prostoru, jeste mera kojom se izracunava kosinus izmedju tih uglova. Metrika koju uzimamo u obzir jeste orijentacija vektora a ne njegova velicina.

U tom smislu poredjenje vektora se radi u normalizovanim prostoru zato sto ne uzimamo velicinu svakog ponavljanja termina u TF-IDF-u u svakom dokumentu vec samo ugao izmedju tih dokumenata sto je prikazano i na slici ispod (Slika 5.). (Perone, 2013)



(Slika 5.) (Perone, 2013) Oredjivanje slicnosti izmedju dokumenata na osnovu cosinusa njihovih vektorski reprezentacija

U prvom slucaju vidimo da je ugao izmedju vektora jako mali te su takvi dokumenti slicni. Drugi slucaj govori o vektorima koji se nalaze pod uglom od 90 stepeni te izmedju takvih dokumenata ne postoji nikakava slicnost. Na kraju je prikazan slucaj kada se radi o dokumentima cije vektorske reprezentacije govore da su suprotni po prirodi.

## 3.7. Razvoj modela

Kako zelimo da primenimo dva pristupa u resavanju problema i definisanju sistema za preporuku u prvom delu razvoja modela cemo raditi na sistemima zanovanim na sadrzaju a u drugom delu cemo primeniti i podatke sa korisnicima kako bi mogli da izvrsimo tesitranje sistema za poreporuku zasnovanih na kolaboraciji.

Jedan od ključnih koraka bilo koje analize podataka jeste postupak predprocesiranja podataka, zato što radeći to omogućujemo bolju preciznost modela koja umnogome zavisi od kvaliteta podataka.

To pre svega uključuje konsolidaciju podataka u smislu skupljanja, selekcije i spajanja. Zatim čiscenja podataka od nedostajućih vrednosti, ukljanjanja podataka koji ne donose vrednost kao i smanjenje šuma u podacima. Zajedno sa tim delom mozemo da pocemo sa razvojem modela.

Koraci za rešenje ovog praktičnog problema mogu biti formulisani na sledeći način:

1. Analiziranje podataka i utvrđivanje najbitnijih veština koje mogu okarakterisati jednu poziciju koja je otvorena.
2. Analiza i transformacija datih podataka mašinskom obradom prirodnog jezika (NLP - „Natural language processing“) , kao i korišćenjem regularnih izraza
3. Modeliranje seta sa korisnicima u vidu ponderisanja sume svih interakcija i logaritmovanje kako bi se dobila uravnotezena distribucija
4. Razvoj modela kroz testiranje različitih algoritama za klasifikaciju, klasterovanje u vidu TF-IDF, slicnost izmedju vektora.
5. Učenje modela na trening podacima, a zatim testiranje nad podacima nad punim setom podataka kako bi na izlazu videli predviđanje veština koje su potrebne za određeni posao.

# Četvrto poglavlje

## 4.1. Rezultati istraživanja i diskusija

# Peto poglavlje

## Zakljucak

Ovaj rad je izvrsen dvoetapno. U prvom delu smo upoznali sistem regrutacije kako se on obavljao nekada i kako je era interneta uticala da nastanu online sistemi za regrutaciju. Daljim razvojem to je uslovilo nastanak novih metoda i popularizaciju sistema za preporuku u online regrutaciji potencionalnih kadrova.

Prikazali smo razlike izmedju tehnika sistema za preporuku i iz toga primenili sva tri navedena sistema kako bi prikazali njihove prednosti i mane.

Razliticim tehnikama obrade teksta poput prirodne obrade jezika, TF-IDF kao i regularnih izraza smo oblikovali sadrzaj kako bi bolje primenili sisteme za preporuku i time saznali kako i koliko su slicne vestine za slicne poslove.

U drugom delu smo primenom svih tehnika manipulacije i analize teksta izvukli kljucne reci ili u nasem slucaju kljucne vestine koje su najvise opisivale nas dokument i time povecali kvalitet onoga sto dobijamo kao izlaz iz naseg prediktivnog algoritma. U implemetaciji algoritma prvo smo krenuli sa sistemom preporuke zasnovanom na sadrzaju kako bi prikazali kljucne vestine za odredjenje poslove bez ucesca korisnika preko kojeg smo davali ocene. Prikazan je zatim sistem preporuke zasnovan na preporuci i kao zakljucak se predlaze primena hibridnog sistema kako bi se izbegao problem pocetnog nedostatka podataka. Ovakav sistem preporuke pokazao se boljim i rezultate koji je dao daleko su superioriniji od prethodna dva.

Ovaj rad treba da posluzi u buducem istrazivanju i njegov sadrzaj da bude osnova za dalje razvijanje metoda za implementaciju sistema za preporuku kod procesa zaposaljavanja sa apekta poslodavca kao i onih koji traze posao.

# **Reference**

Domeniconi, G., Moro, G., Pagliarani, A., & Pasolini, R. (2016, January). *Job Recommendation From Semantic Similarity of LinkedIn Users’ Skills.* Retrieved from Research gate: https://www.researchgate.net/publication/298211329\_Job\_Recommendation\_From\_Semantic\_Similarity\_of\_LinkedIn\_Users%27\_Skills

Gusdorf, M. (2008). *Recruitment and Selection: Hiring the Right Person.* Retrieved from SHRM Academic Initiatives: https://www.shrm.org/academicinitiatives/universities/TeachingResources/Documents/Recruitment%20and%20Selection%20IM.pdf

Hong, N. V., Nduyen, H., Duong, H. N., & Snasel, V. (2016). n -Gram-Based Text Compression. *Computational Intelligence and Neuroscience*.

Hopmans, T. (2015, November 19). *A recommendation system for blogs: Setting up the prerequisites (part 1)*. Retrieved from Marketing Technologist: https://www.themarketingtechnologist.co/building-a-recommendation-engine-for-geek-setting-up-the-prerequisites-13/

Kingsley, G. (1950). *Human begavior and the principle of least effort: An introduction to human ecology.* Cambrige.

Malinowski, J., Keim, T., Wendt, P., & Weitzel, D. (2006). *Matching People and Jobs: A Bilateral Recommendation Approach.* Retrieved from The College of Information Sciences and Technology: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.8172&rep=rep1&type=pdf

Manning, C. D., Prabhakar, R., & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press*.

Mills, T. (2018, Jul 2). *What Is Natural Language Processing And What Is It Used For?* Retrieved from Forbes: https://www.forbes.com/sites/forbestechcouncil/2018/07/02/what-is-natural-language-processing-and-what-is-it-used-for/#7ee74645d71f

Neves, M. (2016, April 11). Retrieved from Semanticscholar: https://pdfs.semanticscholar.org/presentation/a575/e0cef057241668a53e75f0627189a0e7e92a.pdf

Patel, B., Kakuste, V., & Eirinaki, M. (2017). *CaPaR: A Career Path Recommendation Framework.* Retrieved from IEEE Computer Society: https://www.computer.org/csdl/proceedings/bigdataservice/2017/6318/00/07944917.pdf

Perone, C. S. (2013, September 12). *Machine Learning :: Cosine Similarity for Vector Space Models (Part III)*. Retrieved from Terra Incognita: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Petković, M. (2014). *Organizacija: dizajn, ponašanje, ljudski resursi, promene.* Beograd: Ekonomski fakultet CID.

Prakash, N. M., Lucila, O.-M., & Wendy, C. W. (2011, September). *Natural language processing: an introduction.* Retrieved from https://dx.doi.org/10.1136%2Famiajnl-2011-000464

PromptCloud. (2017). *U.S. Technology Jobs on Dice.com*. Retrieved from Kaggle: https://www.kaggle.com/PromptCloudHQ/us-technology-jobs-on-dicecom/kernels

Resnick, P., & Varian, H. R. (1997, March 3). *Recommender systems.* Retrieved from Communications of the ACM: https://dl.acm.org/citation.cfm?id=245121

Rodríguez, G. (2018). *Introduction to Recommender Systems in 2018*. Retrieved from Tryolabs: https://tryolabs.com/blog/introduction-to-recommender-systems/

SAS. (2018). *Natural Language Processing*. Retrieved from SAS: https://www.sas.com/en\_us/insights/analytics/what-is-natural-language-processing-nlp.html

Tondji, L. N. (2018, February). *Web Recommender System for Job Seeking and Recruiting.* Retrieved from Reseachgate: https://www.researchgate.net/publication/323726564

Trenkle, W. B. (2001). N-Gram-Based Text Categorization. *ResearchGate*. Retrieved from http://odur.let.rug.nl/vannoord/TextCat/textcat.pdf

Vembunarayanan, J. (2013, October 7). *Tf-Idf and Cosine similarity*. Retrieved from Seeking wisdom: https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/