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:
Doc.dr.Miloš Jovanović

Ime i prezime studenta:
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

_____

dr. Milan Vukićević, docent, član komisije

_____

dr. Slađan Babarogić, vanredni profesor, član komisije

_____

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 pomoću 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

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

Prvo poglavlje

1.1. 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 može 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.

1.2. Dosadašnji najvažniji rezultati u zadovoljavanju potreba u predmetnoj oblasti

Potraga za poslom započinje procesom u kome pojedinci traze zaposlenje u sferi koja najviše odgovara njihovim kriterijumima i sposobnistima.
Sa druge strane, pronalaženje odgovarajućeg kandidata je ključni zadatak za odeljenje ljudskih resursa. Oni na osnovu potrebe kompanije otvaraju pozicije, vrše selekciju i na kraju obezbeđuju kvalitetan kadar. Oba problema, traženja posla i traženja zaposlenog, se svode na isti u tom smislu da se radi o kontinuiranom procesu i komunikaciji između njih, koja kao rezultat može obezbediti odgovarajućeg novozaposlenog za kompaniju i posao za pojedinca koji je u potrazi sa poslom. (Domeniconi, Moro, Pagliarani, & Pasolini, 2016)

Informacione tehnologije su u proteklim godinama promenile način na koji ljudi dolaze do zaposlenja kao i način na koji rade uopste. Sprovedeno istraživanje top 1000 kompanija u Nemackoj pokazalo je da je pojava interneta zamenila štampane medijume oglašavanja kao glavni kanal za regrutaciju. Sa oko 78% upraznjenih mesta koja su objavljena na sajtovima kompanija i 49% novootovrenih pozicija koje su oglašene putem internet portala za zapošaljavanje dolazimo do zaključka da su kanali putem interneta daleko nadmašili starije načine oglašavanja štampanim putem.

Takođe istraživanje kaže da se procenat zaposlenja putem interneta podigao za 58% 2004. godine. (Malinowski, Keim, Wendt, & Weitzel, 2006)

Sve veći broj onih koji traže posao to danas čine upravo tako što dele svoja akademska dostignuća i profesionalne informacije putem interenta. U isto vreme kompanije sve brže prihvataju svet u kome se proces regrutovanja dešava “online”. Na osnovu istraživanja “Jobvite”, 68% online ljudi koji traže posao su diplomci ili post-diplomci, dok 94% poslodavaca koriste ili planiraju da krenu da koriste društvene mreže za regrutovanje novozaposlenih.

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

1.3. 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 može 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 mogući kandidat može 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 zapošljavanje, bila je neohodna reorganizacija načina na koji kompanije obavaljaju određene aktivnosti u određenim sferama. Jedna od tih je bila sfera regrutacije.

Postavljanja poslovnih ponuda na internet stranicama kompanija se uglavnom vršilo na delu stranice pod nazivom "Karijera" ("Carrier"). Zainteresovani bi odlazili na te stranice i aplicirali putem neke online prijave. Međutim vremenom su se razvile i platforme specijalno dizajnirane za proces regrutacije tj. zapošaljvanja. Na tim mestima su zainteresovani mogli da naprave svoj profil koji bi zatim popunili sa ključnim informacijima u vezi sa njihovim školovanje, prethodnim iskustvom i veštinama koje poseduju. Nakon toga bi, ukoliko bi se otvorila nova pozicija, imali mogućnost da apliciraju na istu i time steknu šansu u budućem zaposlenju. Nažalost ovakav sistem je bio održiv samo u ranijim fazama interneta. Trenutna situacija je bila takava da bi za određenu poslovnu ponudu stizale hiljade odgovora tj. apliciranja. Iz ovoga se stvorila potreba za sistemima za preporuku.

Sa druge strane čak i traženje posla je nekada mogao biti iscrpljujuća aktivnost. Najčešći pristup bi bio da se uz pomoc par ključnih reči izvrši pretraga na stranici potencionalnog poslodavca. Rezultat pretrage bi vratio listu poslova koji sadrže neku od tih reci. Nažalost 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 ograničeno na obradu ponuda od kandidata preko osnovnih menadžerskih aplikacija. Ali sa ekspanzijom količine podataka na internetu i uopšte podataka u digitalnom formatu i razvojem e-biznisa zahtevalo je određenu reformu načina na koji su kompanije do tog trenutka obavljale proces regrutacije. Platforma za regrutovanje putem interneta jeste jedna od najuspesnijih otkrića putem koje se poslodavci i kandidati otkrivaju. Ovakve platforme su dozivele pravu ekspanziju zbog sve težeg procesa regrutovanja novih potencijala. Za svaki postavljen oglas za posao, hiljade biografija se svakodnevno može poslati za otvorene pozicije. Analogno tome takođe postoji ogroman broj radnih biografija koje polako postaju dostupne online.

Takava ogromna količina dostupnih informacija u vidu poslovnih ponuda, njihovih zahteva kao i radnih biografija je postala odlično mesto za unapređivanje kvaliteta u smislu koliđine poklapanja zahteva i ponude. Naravno taj kvalitet je ostao na niskom nivou zbog pristupa logickog operatora (“true-false”) a time velike količine podataka i šansi za dobro poklapanje ostaju neiskorišćeni. Iz toga se stvorila potreba za korišćenjem sistema za preporuku kako bi poslodovaci uspešno mogli da obrade ogromne količine podataka brzo i efikasno. Svakako rešenje nije moguće ostvariti u kratkom roku i sami sistemi za preporuku predstavljaju izazov u polju istraživanja koje se i dalje razvija.

Kako bi se ta ideja dalje razvijala u ovom radu cemo se baviti nekim problemima i potencionalnim rešenjima za unapređenje sistema za preporuku kod otkrivanja veština potrebnih za sam posao.

Peronalizovani sistemi kao što su sistemi za preporuku su u proteklim godinama privukli pažnju velikog broja istraživača. Od kada se termin “Sistemi za preporuku” prvi put pojavio u izdanju magazina “Communications of the ACM“ vreme i istraživanje je uloženo kako bi se takvi sistemi poboljšali i učinili pouzdanijim. Jedan deo istraživača se fokusirao na spajanje sistema za preporuku na osnovu sadržaja sa sistemom za preporuku na osnovu saradnje kako bi se prevazišao 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 organizacije 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 privlačenje potencionalnih kandidata u određenom vremenskom roku i u broju koji je predviđen sa odgovarajucim kvalifikacijama da se prijave za posao unutar organizacije

Ekonomsko stanje zemlje određuje uslove i značajno može da utiče na mogućnost organizacije da privuče kvalitetan kadar. Ukoliko se radi o stabilnoj ekonomskoj situaciji, sa niskim nivoom nezaposlenisti, tada postoji verovatnoća da će organizacija morati da se takmiči sa drugim sličnim organizacijama kako bi privukla one vredne zapošljavanja. Ako se ipak radi

o slabijoj ekonomiji dolazi se do situacije, gde ukoliko postoji visoka nezaposlenost, veliki broj prijava može stići na otvorenu poziciju sa samo par onih koji zaista ispunjavaju kriterijume poslovnih zahteva novootvorenog radnog mesta.

Način na koji će kompanije oglasiti svoje pozicije varira od angažovanja eksternih firmi koje traze zaposlene za određeni procenat do postavljanja oglasa na sajtu kompanije ili novina. Takođe postoje varijante stažiranja kako bi se procenio kvalitet kadra.

Međutim najznačanije 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 današnje vreme mnoge organizacije postavljaju otvorene pozicije na posebnim lokacijama na interenetu poput “Carrier Builder”,”Monster.com”,”Linedin.com”,”dice.com” i sl. time kompletno prelazeći na elektronski način regrutovanja. To naravno nije bez razloga urađeno, vec je posledica prednosti koje takav sistem sa sobom donosi, poput nižih cena oglašavanja, lakoće i brzine postavljanja kao i šireg skupa ljudi, do kojih je mogće doći, nego što bi to bilo sa starijim sistemima. (Gusdorf, 2008)

Onima koji traže posao ovo takođe olakšava zato što u jednom danu mogu poslati svoje radne biografije i time aplicirati na više mesta. Ovo nažalost stvara probleme u odeljenju ljudskih resursa zato što zbog povećanog obima prijava moraju napraviti selekciju onih koji ne zadovoljavaju kriterijumom potrebe otvorene pozicije. Međutim kako se veštine i kvalifikacije danas sve više razvijaju i postoji citav dijapazon istih postoji uvek šansa 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

Za mnoge ljude, znanje i veštine predstavljaju slične koncepte koji se svakodnevno koriste da bi se opisale kopetencije pojedinca. Naravno istina je da su to jako različiti koncepti, koji imaju ipak neke sličnosti.

Znanje je informacija koju pojedinac dobija kroz različite aktivnosti poput čitanja, slušanja, dodira i sl. Koncept znanja se odnosi na upoznatost sa oblašću i njenim teoretskim konceptima. Takvo znanje se može prenositi sa jedne na drugu osobu.

Veštine sa druge strane nije moguće prenositi zato što su one u osnovi primenjeno znanje u određenoj situaciji. One se razvijaju kroz vežbu i kao takve zavise isključivo od pojedinca u kojoj će meri biti dospeti. I u tome se krije suštinska razlika zato što je znanje teoretsko a vesti predstavljaju praktično primenjeno znanje.

Ovakvo saznanje se lako može primeniti i u procesu zapošljavanja. Pojedinac može biti upoznat sa osnovnim konceptima koji posao zahteva ali ne mora nužno imati sve potrebne veštine . Naravno prilikom razvijanja veština posledično tome se šire i znanja koje individua može imati. Ako uzmemo primer inženjera za avione, on poseduje sva znanja o principu leta i samog aviona međutim nema veštine koje su potebne da bi bio pilot. (Boulet, 2015)

Treće poglavlje

3.1. Problem istraživanja i metodologija

U ovom radu će 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 reči koje drugi ljudi mogu da interpretiraju na različite načine. Jednostavnije rečeno radi se o komunikaciji, ali naravno uvek treba imati na umu da reči mogu imati mnogo dublje značenje u zavisnosti od konteksta. Mašinska obrada prirodnog jezika (Natural Language Processing - NLP) se fokusira na konteksutalni patern pre nego na vokalni način na koje su reči izgovorene (Mills, 2018)

Mašinska obrada prirodnog jezika je u svojoj osnovi forma veštacke inteligencije koja analizira ljudski jezik. Postoje varijateti međutim sve imaju zajedničku osobinu a to je da predstavljaju tehnologiju koja pomaze mašinama da razumeju naš jezik, pa cak i da komunicijaju sa nama koristeci naš jezik. Kao izvor za svoja saznanja ona koristi različite dicipline uključujuci računarske nauke i računarsku lingvistiku u svom zadatku koji je smanjenje jaza između ljudske komunikacije i razumevanja od strane računara. (SAS, 2018)

Kao ljudi mi pricamo jezikom koji je nama poznat međutim jezik računara je mašinski jezik koji je nama nerazumljiv. Na svom najnizem nivou mašine komuniciraju na osnovu jedinica i nula u vidu signala. Od vremena kada su se podaci spustali na magnetnu traku do današnjeg vremena gde većina modernih uređaja poseduje neku vrstu personalnog asistenta koji je takođe, naravno, digitalni.

Upravo je mašinska obrada prirodnog jezika omogućila računarima da čitaju tekstove, razumeju govor, interpretiraju ih i iz toga izvuku delove koji su od značaja.

Svoju istoriju je započela 1950 godine kao mešavina veštacke inteligencije i ligvistike. Njeni koreni vuku iz tehnologije za izvlačenje informacija iz teksta ("Informational retreval – IR"), koja primenjuje statistički zasnovane tehnike da indeksira i pretraži velike količine teksta efikasno.

Odlike prirodnog jezika kao što su veličina, nestruktuirana priroda kao i činjenica da je podlozno interpretaciji dovelo je do problema kada su u pitanju standardni pristupi parsiranju. Mašinska obrada prirodnog jezika mora da omogući izvlačenje sematnike iz teksta uzimajući u obzir delove govora tj. teksta kao što su imenice, glagoli, pridevi i sintaksa rečenice. (Prakash, Lucila, & Wendy, 2011)

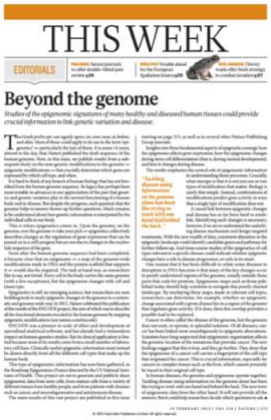
I dok su nadgledano i nenadgledano mašinsko učenje široko rasprostranjeni načini za modeliranje ljudskog jezika takođe postoji potreba za sintaksnom i semantickom razumevanju koji u većem delu mašinskog učenja nije prisutan. Mašinska obrada prirodnog jezika pomaže u otkljanjanju problema višeznačnosti u jeziku i postavljanju numeričke strukture koju kao izvor koriste programi za prepoznavanje govora ili analizu teksta. Ona uključuje različite tehnike intepretiranje ljudskog jezika koja se kreće od statističkih metoda i mašinskog učenja do metoda zanovanih na pravilima i algoritmima.

Mašinska obrada prirodnog jezika podrzumeva tokenizaciju i parsiranje, kao i pravljenje rečenica na osnovu rečnika, morfološke analize i vađenja korena reči ("lemmatization and stemming") (Manning, Prabhakar, & Schütze, 2008). Ovakvim pristupom inverzno iz jezika dolazimo do teksta i na kraju iz teksta do skupa reci.(Slika 3.2.1)

Recnik se sastoji od skupa reci

(http://learnenglish.britishcouncil.org/en/vocabulary-games)

Tekst cini niz reci iz recnika


THIS WEEK

EDITORIALS

Beyond the genome

(http://www.nature.com/polopoly_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf)

Jezik je konstruiran iz skupa svih mogućih tekstova

(http://www.old-engli.sh/language.php)

Slika 3.2.1. (Neves, 2016) Skup reči u nizu predstanja recnik, dok skup svih rečnika čini jezik

3.3. n-gram

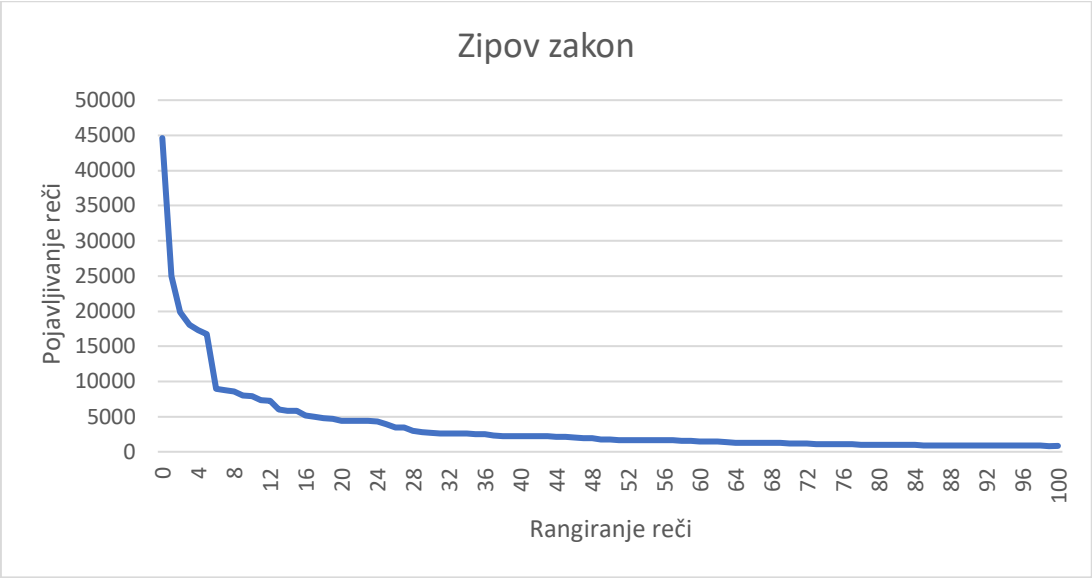
Kategorizacija teksta predstavlja ključni zadatak u obradi dokumenata, čime se stiče mogućnost obrade ogromne količine podataka koji se nalaze u elektronskoj formi. (Trenkle, 2001) Ono što predstavlja realnu poteškoću u obradi teksta u elektronskim dokumentima, jesu tekstualne greške bile ona gramatičkog ili sintaksnog tipa. Da bi sistem teksutalne kategorizacije bio proglašen pouzdanim, za sve različite vrste izvora podataka neophodno je da ipak ima neku dozu tolerancije na oređene greske kao i način da ih prevaziđe.

Dokumenta u elektronskom formatu poticnu iz raznih izvora. Neki su generisani kao deo programa koji je zaduzen za pisanje teksta i kao takvi podređeni 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 što su skenirani dokumenti uz pomoc posebnih softvera za prebacivanje iz papirne u elektronsku formu.

Prilikom skeniranja ili bolje rečeno prepoznavanja teksta sa papira neminovno je da će nastati neka greška 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 dobrodošao.

U ljudskom jeziku neke reči se pojavljuje češće od drugih. Jedan od osnovnih načina da se iskaže takva ideja je danas poznata kao Zip-ov zakon koja je prikazana ispod na grafikonu (Grafikon 3.3.1) (Kingsley, 1950). On kaže da je pojavljivanje n-te najčešće reči u ljudskom jeziku u tekstu inverzno proporcionalno broju n. Ovo znači da najčešća rec u nekom jeziku se pojavljuje dva puta češće od druge reči i tri puta češće od trece itd, implicirajuci da uvek postoje reči koje su na neki način dominatne u tom jeziku. Ovakav zaključak se odnosi i na reči generalno ali i na određene 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 3.3.1. Zip-ov zakon distribucije reči u prirodnom jeziku

n-gram predstavlja kontinuiranu sekvencu od n reči iz neke rečenice ili teksta na osnovu govora. To znači da to može biti veznik, slovo, reč, rečenica, slozenica ali ono što ih sve odlikuje jeste da se posmatraju kao atomska vrednost. N-gram od jedne "stavke" tj. reči se naziva "unigram",od dve "bigram",tri "trigram" itd. (Hong, Nduyen, Duong, & Snasel, 2016)

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

3.4 Sistemi za preporuku

Sistemi za preporuku sve brze postaju defakto način za preporuku u različitim aplikacijama za preporuku prozivoda, usluga i sveukupno informacija korisnicima istih. Mnoge internet aplikacije su se pridružile trendu korišćenja sistema za preporuku da bi obezbedile veci profit time što bi korisnicima smanjile vreme pretrage i dolaženja do idealnog proizvoda ili usluge. Neke od takvih kompanija su "Amazon", "Microsoft","Aliexpress","Netflix" i slično. (FadhelAljunid & Manjaiah, 2017)

Sve ove kompanije su uspešno integrisale sisteme preporuke za komercijalnu upotrebu i time uvećale prodaju a samim tim i prihode prilikom prodaje preko interneta i jos bitnije obezbedile lojalnost kupaca. Druge kompanije su takođe razvijale lokalne ("inhouse") generičke sisteme preporuke a neki od njih su "Net Perceptions", "Epiphany", "Art Technology", "Broad Vision". (Huang, Zeng, & Chen, 2004)

Kompanije koje koriste sisteme za preporuku fokusiraju se na povećanje prihoda na osnovu bolje personalizove ponude i poboljšanja zadovoljstva korisnika. Po pravilu oni ubrzavaju vreme pretrage i olakšavaju da korisnici dođu do sadržaja koji je relevantan za njih. Sa tako personalizovanim sadržajem komapnije dobijaju prednost na trzištu i smanjuje se mogućnost gubitka korisnika od strane konkurencije. (Rodríguez, Introduction to Recommender Systems in 2018, 2018)

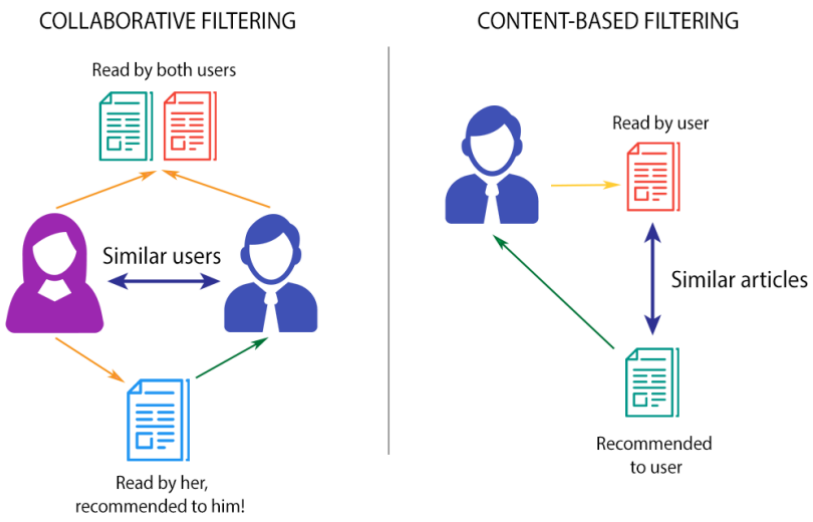
Na osnovu ovoga možemo napraviti klasifikaciju sistema za preporuku na:

- Zasnovani na sadržaju - koji koriste karakteristične informacije.
- Zasnovani na saradnji - koriste "korisnik-element" interakcije.
- Hibridni sistemi za preporuku – kombinuju 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 sadržaju su ograničeni u njihovim mogućnostima u smislu da će preporučene stvari bili više slične 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 sadržaju u otkrivanju skrivenih paterna. Takav sistem preporuke se više koncentriše na korisnika i njegove preference, pre nego na sadržaj onoga sa čime je korisnik imao dodira. (Hopmans, 2015)

Nedostatak sistema za preporuku zasnovanih na saradnji jeste to što su mu za dobre preporuke potrebne veće količine podataka u vidu istorije aktivnosti korisnika da bi otkrio dobre paterne. Tako nešto nije problem kada se radi sistemima zasnovanim na sadržaju koji mogu da daju rezultate sa malo, a nekada i bez ikakve informacije o istoriji podataka i stoga su lakši za implementaciju.

Pojednostavljena reprezentacije razlike ova dva sistema je prikazana na Slici 3.4.1. Na njoj se jasno vidi da u prvom slučaju ukoliko su oba korisnika imala interakciju sa istim ili sličnim sadržajem onda sistem sa određenom sigurnoscu može da preporuči osobi B nešto sa čime je osoba A imala interakciju. U drugom slučaju se vidi da ukoliko je korisnik imao interakciju sa nekim sadržajem, sistem može da mu preporuči drugi sadržaj na osnovu sličnosti između njih.



Slika 3.4.1.(Hopmans, 2015) Razlika između sistema za preporuku zasnovanog na sadradnji u odnosu na onog zadnovanog na sadržaju

3.4.1. Zasnovani na sadržaju

Sistemi koji implementiraju preporuku zasnovanu na sadržaju kreću od analize dokumenata ili desktriptora sa kojima je korisnik imao neku interakciju. Nad tim oni prave profil korisnika sa kojim mogu dalje da rade preporuke sa novim sadržajem. Dalji proces je samo poklapanje atributa korisnika ili sadržaja sa drugim sadržajem.

Prednost koriscenja sistema za preporuku se ogleda u sledecim aspektima:

- Nezavnistnost od korisnika – oni koriste podatke o korisniku samo da bi izgradili profil ali ne zavise od sistema ocenjivanja tog korisnika i drugih kako bi napravile sistem preporuke što je slučaj sistemima preporuke zasnovanih na saradnji.
- Transaprentnost – objašnjenje zašto je sistem preporučio neki sadržaj i kako je došao do toga je sama posledica dektriptora nekog sadržaja koji je uslovio da se taj, slican, sadržaj preporuči, a do njega je lako doći. Za razliku od njih preporuka na osnovu saradnje daje preporuku nepoznatog korisnika sa sličnim preferencama i u tom smislu je nedostupna stvar (“black box”).

- Novi sadržaj – u ovom smislu su sistemi za preporuku na osnovu sadržaja odlični zato što mogu da preporuče novi sadržaj iako on nema ocenu drugih korisnika. To je zato što se ovakav sistem oslanja na sam sadržaj tj. odlike tog sadržaja kako bi ga uporedio i sa njemu sličnim i time napravio preporuku. (Lops, Semeraro, & Gemmis, 2011)

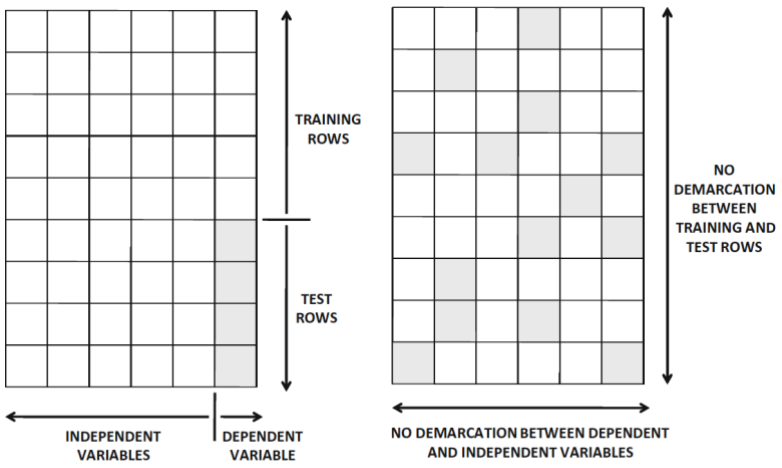
3.4.2. Zasnovani na saradnji

Pristup sistema zasnovanih na saradnji se suštiniski razlikuje od onog koji je pristuan u sistemima zasnovanim na sadržaju. Umesto preporuke elementa (teksta, proizvoda, pesme, filma, posla itd) sličnom onom sa kojim je kojim je korisnik imao dodira u proslosti, sistem preporučuje elemente koji su slični korisnici konzumirali. Ukratko umesto da preporučuje na osnovu sličnosti elementata on daje preporuku na osnovu sličnosti korisnika. (Balabanović & Shoham, 1997)

Glavna karakteristika sistema za preporuku zasnovanih na saradnji ("Collaborative filtering recomender systems") jeste to što su oni u potpunosti nezavisni od računarske reprezenatacije objekata koje oni ustvari preporučuju. To znači da mogu savršeno dobro da rade i sa objektima poput muzike, slika ili filmova. Sa druge strane postoji i određeni nedostaci ovakvog sistema za preporuku. Jedni od najpoznatijih su problem početnog nedostatka podataka ("cold-star problem") i rasutost podataka. (Tondji, 2018)

U ovakvim sistemima preporuke se koriste interacije korisnika da bi filtrirale stvari koje su od interesa. Ako bi vizualizovali takav set interakcija matricom u kojoj bi za svaki par (i,j),(a,b) bi se preslikavao na interakciju između korisnika i_a i elementa j_b.

Jedan način za posmatranje sistema za preporuku zasnovanih na saradnji bi bio da se oni posmatraju kao generalizacije klasifikacije ili regresije. Međutim za razliku od takvih pristupa gde se predviđa promenljiva koja direktno zavisi od drugih promenljivih u sistemima za poreporuku zasnovanim na saradnji ne postoji razlika između jedne promenljive i promenljive na nivou klase. Kao što je spomenuto, ako se problem predstavi kao matrica, ne radi se o predviđanju vrednosti jedne kolone vec bilo koje vrednosti koja postoji u tom sistemu. (Rodríguez, 2018) (Slika 3.4.2.1.)



Slika 3.4.2.1. (Rodríguez, 2018) Poređenje klasifikacije i sistema zasnovanog na saradnji

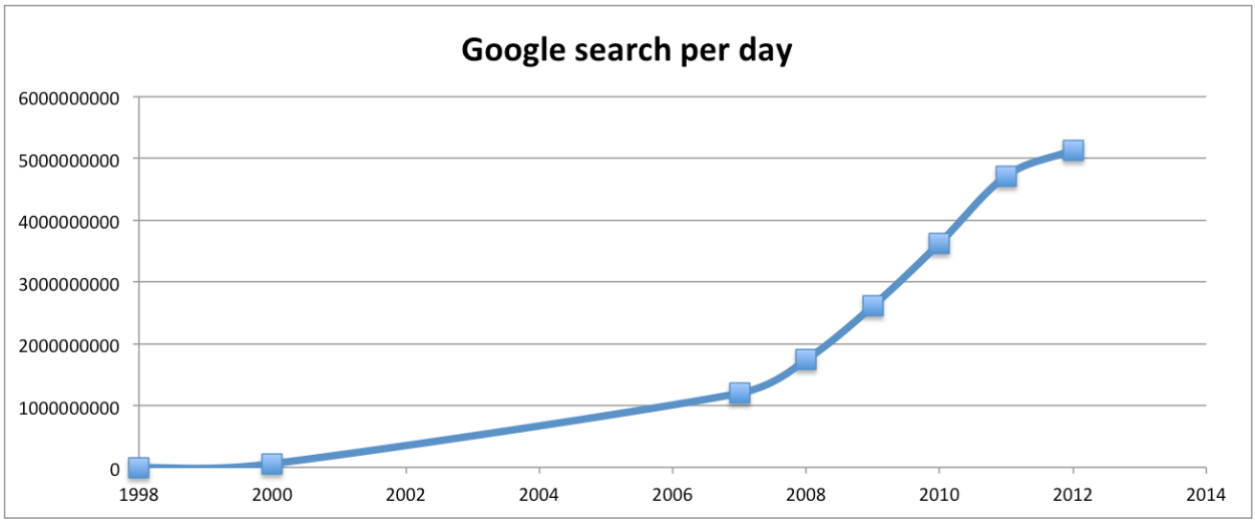
3.4.3. Hibridni sistemi

Skorija istraživanja su pokazala da hibridni pristup sistema za preporuku, koji kombinujući sistem zanovan na saradnji i sadržaju može biti efikasniji od korišćenja tih sistema odvojeno.

Postoji više načina na koje može biti implementiran od kojih je jedan pravljenje odvojenih sistema za preporuku zasnovanih na saradnji i sadržaju i onda ih kombinovati, dodavanjem zasnovanog na sadržaju na sistem zasnovan na saradnji ili obrnuto. Istraživanja su pokazala da u određenim slučajevima kobinacija ovih sistema, umesto odvojenog rada, može da da na svom izlazu mnogo preciznije preporuke. Takođe sa ovakvim sistemom preporuke lakše je prevazići ograničenja koji, odvojeno, imaju da dva sistema preporuke kao što su problem početnog nedostataka podataka ili rasutost podataka o kojima je bilo reči.

3.5. TF-IDF

Davne 1998 godine Google je vršio obradu od oko 9800 pretraga dnevno. U 2012 godini ova brojka se popela na 5.13 milijardi pretraga dnevno što se može videti na slici 3.5.1. (Vembunarayanan, 2013)



Slika 3.5.1. Količina pretrage na Google.com na godišnjem nivou

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

Ako primenimo ovu logiku nad tekstom ocenjivanje relevatnosti neke reči započinje njenim prisustvom u tom tekstu kao i njenom frekventnošću unutar samog dokumenta. Međutim ocenjivanje samo na osnovu broja ponavljanja reči u dokumentu nije dovoljan pokazatelj za njen značaj tj. vrednost (težinu). Krećemo tako što svakom terminu u dokumentu dodeljujemo težinu koja zavisi od broja ponavljanja unutar samog dokumenta. Ovakav način dodeljivanja težine na osnovu broja ponavljanja se naziva - "učestalost termina" ("term-frequency").

Vazno je spomenuti algoritam pod nazivom “skup reci” (“Bag of word”) koji pojednostavljuje analizu reči ,u tom smislu da zanemaruje semantiku i gramatiku, koja se pojavljuje u prirodnom jeziku kao i izvlačenje informacija. To bi značilo da dva dokumenta sa sličnim rečima jesu slični jedan drugom bez obzira na raspored tih reči unutar svakog od dokumenata. Ono što se postavlja kao pitanje jeste relevantnost svih reči unutar dokumenta, gde kao logičan odgovor dobijamo da sve reči nisu podjednako bitne bez obzira na njihovu frekventnost.

Iz svega navedeno se vidi da sama učestalost termina ima fundamentalni problem: svi termini se postmatraju kao podjednako bitni. Čak iako neki od njih imaju minimalnu diskriminacionu moć za pružanje relevantnosti. Kao rešenje se nudi “inverzna učestalnost termina” (“inverse document frequency”) koju se dobija ako sa ukupnim broj dokumenata u kolekciji (N) podelimo učestalost pojavljivanja termina (t) u toj kolekciji . (Manning, Prabhakar, & Schütze, 2008)

$$idf_t = \log (N / df_t)$$

Iz ovoga možemo zaključiti da je inverzna učestalost termina (idf_t) viša ukoliko se termin ređe pojavljuje ili niža ukoliko se on pojavljuje češće u kolekciji.

Sledeci korak jeste da se kobinovanjem df_t i idf_t dobije slozena (mešovita) težina za svaki termin u svakom dokumentu. TF-IDF ($tf\text{-}idf_{t,d}$) dodeljuje težinu terminu t u dokumentu d na sledeci način:

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t$$

Drugim rečima $tf\text{-}idf_{t,d}$ dodeljuje težinu terminu t u dokumentu d na način na koji:

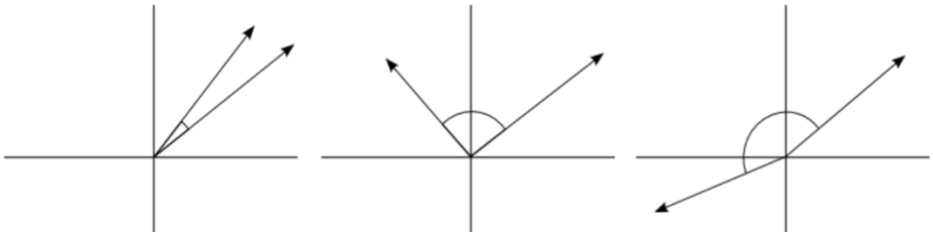
- Termin poseduje najveću težinu kada se pojavljuje u velikom broju na malom broju dokumenata unutar kolekcije.
- Temin poseduje nižu vrednost ako se ređe pojavljuje u dokumentu, ili se pojavljuje u većem delu dokumenata (time što je značaj tog termina umanjen).
- Termin poseduje najnižu vrednost kada se pojavljuje u skoro svim dokumentima.

Sada ukoliko posmatramo svaki dokument kao vektor u kome svaki deo korespondira jednom od termina zajedno sa težinama koje su dodeljne kao rezultat TF-IDF, takav vektor se može iskoristi u daljoj analizi kod pronalazenja sličnosti između dokmenata ili u slučaju ovog rada pronalaženja poslovnih ponuda sa potrebim i sličnim veštinama.

3.6. Kosinusna sličnost između vektora

Kosinusna sličnost između dva vektora ("Coasine similarity"), ili u našem slučaju dva dokumenta u vektorskom prostoru, jeste mera kojom se izračunava kosinus između tih uglova. Metrika koja se uzima u obzir jeste orijentacija vektora a ne njegova veličina.

U tom smislu poređenje vektora se radi u normalizovanim prostoru zato što ne uzimamo veličinu svakog ponavljanja termina u TF-IDF-u u svakom dokumentu već samo ugao između tih dokumenata što je prikazano i na slici ispod (Slika 3.6.1.). (Perone, 2013)



Slika 3.6.1. (Perone, 2013) Oređivanje sličnosti između dokumenata na osnovu cosinusa njihovih vektorski reprezentacija

U prvom slučaju vidimo da je ugao između vektora jako mali te su takvi dokumenti slični. Drugi slučaj govori o vektorima koji se nalaze pod uglom od 90 stepeni te između takvih dokumenata ne postoji nikakava sličnost. Na kraju je prikazan slučaj 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 rešavanju problema i definisanju sistema za preporuku u prvom delu razvoja modela ćemo raditi na sistemima zanovanim na sadržaju 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 možemo 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, sličnost između 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

Podaci nad kojima će biti izvršeno istraživanje su deo takmičenja na "kaggle.com" (Dice, 2017) i oni predstavljaju uzorak od 22 hiljade poslovnih ponuda sa sajta "dice.com" koje je izvršeno od strane "promptcloud.com". Originalni set je sadrzao vise od 4.6 miliona poslovnih ponuda na podrucju Amerike.

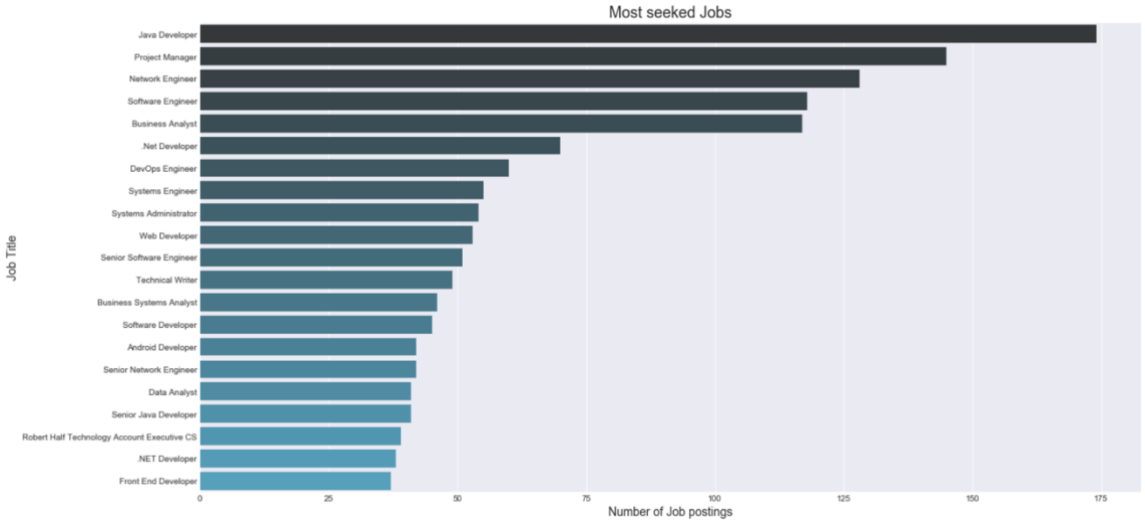
Polja koja su dostavljena u setu podataka su:

- advertiserurl – link ka konkretnom poslu na "dice.com"
- company – ime kompanije koja je otvorila poziciju
- employmenttype_jobstatus – vremenski okvir zaposlenja
- jobdescription – tekstualni opis posla
- joblocation_address – država u kojoj je pozicija otvorena
- jobtitle – naziv pozicije
- postdate – vreme otvaranja pozicije
- shift – smena/ transport do posla…
- skills – veštine koje su potrebne za obavljanje posla
- uniqid – jedinstveni identifikator

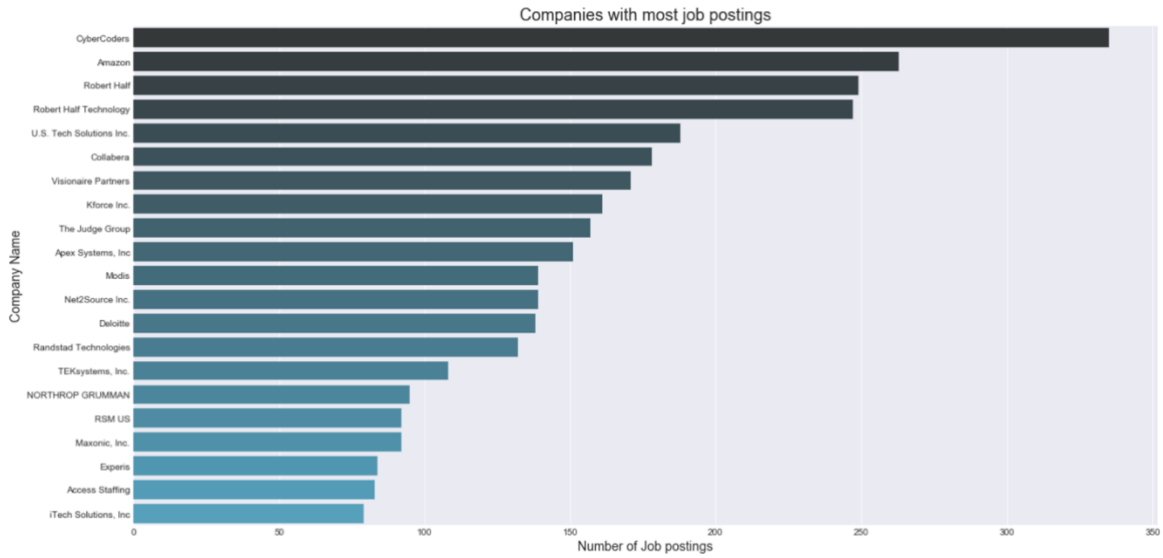
advertiserurl	company	employmenttype_jobstatus	jobdescription	jobid	joblocation_address	jobtitle	postdate
ttps://www.dice.com/jobs/detail/AUTOMATION-TE...	Digital Intelligence Systems, LLC	C2H Corp-To-Corp, C2H Independent, C2H W2, 3 M...	Looking for Selenium engineers...must have sol...	Dice Id : 10110693	Atlanta, GA	AUTOMATION TEST ENGINEER	1 hour ago
https://www.dice.com/jobs/detail/Information-S...	University of Chicago/IT Services	Full Time	The University of Chicago has a rapidly growin...	Dice Id : 10114469	Chicago, IL	Information Security Engineer	1 week ago
https://www.dice.com/jobs/detail/Business-Solu...	Galaxy Systems, Inc.	Full Time	GalaxE.SolutionsEvery day, our solutions affec...	Dice Id : CXGALXYS	Schaumburg, IL	Business Solutions Architect	2 weeks ago
https://www.dice.com/jobs/detail/Java-Develope...	TransTech LLC	Full Time	Java DeveloperFull-time/direct-hireBolingbrook...	Dice Id : 10113627	Bolingbrook, IL	Java Developer (mid level)- FT- GREAT culture,...	2 weeks ago
https://www.dice.com/jobs/detail/DevOps-Engine...	Matrix Resources	Full Time	Midtown based high tech firm has an immediate ...	Dice Id : matrixga	Atlanta, GA	DevOps Engineer	48 minutes ago
https://www.dice.com/jobs/detail/SAP-FICO-Arch...	Yash Technologies	Full Time, Permanant	We are looking for a Senior SAP FICO Architect...	Dice Id : 10111847	Chicago, IL	SAP FICO Architect	2 weeks ago
https://www.dice.com/jobs/detail/Network-Engin...	Noble1	Full Time, Direct Hire	Network Engineer Job Description A Network Eng...	Dice Id : 90884761	Atlanta, GA	Network Engineer	1 hour ago
https://www.dice.com/jobs/detail/Sr.-Web-Appli...	Bluebeam Software, Inc.	Full Time, Full Time	Bluebeam is looking for talented sr. web devel...	Dice Id : 10110132	Chicago, IL	Sr. Web Application Developer (Cloud Team) - C...	2 weeks ago
https://www.dice.com/jobs/detail/Front-End-Dev...	Genesis10	Full Time, Direct Placement	This is a fulltime position for a Javascript d...	Dice Id : gentx001	New York, NY	Front End Developer	7 hours ago
https://www.dice.com/jobs/detail/Application-S...	VanderHouwen & Associates, Inc.	C2H W2, Contract to hire	SummaryOur client is the leading provider of o...	Dice Id : vhassoc	Seattle, WA	Application Support Engineer	7 hours ago

Slika 4.1.1.

Preliminarnom analizom se dolazi do podataka o najtraženijem poslu, kompaniji koja najviše ima otvorenih mesta, kao i državi koja ima najveci broj otvorenih radnih mesta:



Slika 4.1.2. Najtraženiji poslovi koji su u ponudi



Slika 4.1.3. Kompanije sa najvećim brojem otvorenih mesta

	joblocation_address	Locationwise
974	new york, ny	1368
49	atlanta, ga	1245
1249	san francisco, ca	886
1286	seattle, wa	661
132	boston, ma	606

Slika 4.1.4. Države sa najvecim brojem otvorenih mesta

Prvih pet mesta su rezervisana za developere i biznis i projekat menadzere i takve pozicije svojim brojem govore o popularsnosti tj. poražnji za takvim kadrom. U daljem testiranju ćemo prikazati koje su to veštine koje najbolje opisuju jednog java developera. Međutim, pre toga je neophodno izvrsiti predprocesiranje dataset-a.

Kolone koje mogu doneti najveću infomacionu dobit su opis posla i veštine. Ove dve kolone kako su u formi slobodnog unosa neophodno je primeniti:

- regularne izraze – uklanjanjem svih suvisnih karaktera

```
def clean_string(strings):
    result = []
    for value in strings:
        value = value.strip()
        value = re.sub('([!?\',*+.$-/])',' ', value)
        value = re.sub("\d+", "", value)
        result.append(value)
    return result
```

- primenom mašinske obrade prirodnog jezika vrši se uklanjanje reči koje za svrhe ovog istraživanja nemaju nikakvu informacionu dobit. Tekst opisa oglasa ili vestina može da sadrži reči poput “the”,”is”,”are” i sl. koje će biti uklonjenje. Kako ne postoji univerzalni korpus ovakvih reči bice korišćen onaj koji se nalazi na repozitorijumu ntlk (Pythonspot, 2017)

Job_Description_Without_Stopwords	Job_Skills_Without_Stopwords
java developerfull time direct hirebolingbrook...	please see job description
hi established etouch systems technology ser...	need strong core java developer good data expe...
job req #: job description: java lead ecomme...	looking two to three senior java developers ex...
trigyn s direct financial client immediate ope...	see job description
trigyn s direct financial client immediate ope...	see job description
software engineer perform analysis design de...	access analysis analytical analytical skill...
long term contract opportunity years avail...	contract corp to corp contract w full time

Slika 4.1.5. Veštine i opis posla nakon primene mašinske obrade prirodnog jezika

U sledećem koraku iz nestruktuiranog tekst primenom tehnike za izvlačenje informacija iz teksta TF-iDF, takav tekst pretvaramo u vektorsku strukturu, gde je svaka reč jedna pozicija u vektoru. Primenom iste tehnike se odredjuje i vrednost težine svake reči u tekstu na način koji je objašnjen u radu. Ovo radimo kako bi mogli da poredimo sličnosti izmedju ovakvih vektora u prostoru za pronalaženje slicnih poslova ili bolje rečeno sličnih skupova veština koje opisuju slične poslove.

Istu princip je primenjen i na tekstualnom opisu posla kao i na veštinama kako bi dao uvid u to koliko ustvari pojavljivanje jedne reči zaista ne mora da znači da je ona relevatna.

Top skills based on your search using Term Frequency:			Top skills based on your search using the addition of Inverse-Document Frequency:		
	count	index		index	vec_sum
42	533	java developer	135532	web services	5.923553
696	370	years experience	64316	java jee	5.486617
653	296	web services	23841	core java	5.380147
9	263	software development	113703	software development	4.996627
1387	241	java jee	33821	development experience	4.827194
186	236	computer science	45280	experience java	4.499613
431	186	development experience	30186	design patterns	3.787196
151	183	core java	82978	object oriented	3.731795
1585	176	experience java	29840	design development	3.482375
191	174	and or	44755	experience developing	3.481923
2952	162	design patterns	108792	senior java	3.480218
2949	161	object oriented	46534	experience working	3.466404
894	156	application development	19541	communication skills	3.433388
185	147	degree computer	9537	bachelor degree	3.372417
773	145	design development	4853	application development	3.347421
598	139	communication skills	28074	degree computer	3.279836
1160	136	experience working	28076	degree computer science	3.169350
2771	135	experience developing	102076	required skills	3.151272
2485	121	• experience	64006	java development	3.084135
440	115	senior java	46638	experience years	2.814242

Slika 4.1.6. TF-iDF urađen nad opisom posla

Top skills based on your search using Term Frequency:			Top skills based on your search using the addition of Inverse-Document Frequency:		
	count	index		index	vec_sum
205	136	java jee	1146	java jee	34.210629
5	97	core java	1116	java developer	27.880978
182	81	web services	401	core java	21.808200
6	75	java developer	1207	java spring	19.530348
293	61	spring hibernate	2502	spring hibernate	17.592061
24	58	java spring	2898	web services	16.170814
200	51	jee spring	1365	jee spring	14.854855
109	29	java javascript	1208	java sql	9.049881
73	27	developer development	99	angular js	8.050512
194	27	angular js	1140	java javascript	7.704561
211	24	spring mvc	2523	spring mvc	6.101927
710	21	design patterns	1118	java development	5.487163
504	20	java sql	2538	spring struts	5.465156
213	19	html css	532	developer development	5.390719
108	19	jee java	1422	job description	5.312236
225	19	pl sql	1006	html css	5.133063
488	18	object oriented	517	design patterns	4.919533
25	18	spring framework	1180	java oracle	4.684387
176	16	soap rest	2195	restful web	4.468920
282	15	full stack			

Slika 4.1.7. TF-iDF urađen nad veštinama

Kao izlaz imamo “sparse” matricu sledecih dimenzija:

```
<725x3015 sparse matrix of type '<class 'numpy.float64'>'
with 5576 stored elements in Compressed Sparse Row format>
```

Gde nakon primene cosinusne sličnosti za pronalaženje sličnih vektora tj. sličnih poslova na osnovu vestina koje zahtevaju za poziciju “java developer” uz pomoć sistema za preporuku zasnovanih na sadržaju dobijamo sledeći rezultat:

```
0     java developer (mid level)- ft- great culture,...
1        core java developer with distributed computing
2                           java developer (ecommerce)
3                                        java developer
4                                  sr. java developer
5                                        java developer
6                         senior java developer - lead
7                        sr. full stack java developer
8                                        java developer
9                                        java developer
10                    java developer, angularjs, nodejs
11                  java developer (content management)
12                                       java developer
13                                 java developer 12676
14     core java developer with jbpm and multithreading
15                           java developer with android
16                                       java developer
17                     sr. java developer (local to ma)
18                          back-end java developer atg
19                            full stack java developer
Name: jobtitle, dtype: object
```

Slika 4.1.8. Poslovi koji su najsličniji na osnovu veština

Kako bi mogućnost istraživanja bila proširena i na sisteme za preporuku na osnovu saradnje uvodi se jos jedan dataset koji sadrži podatke o interakcijama koje su korisnici imali sa otvorenim pozicijama. Treba napomenuti da je u pitanju sajt tako da korisnici imaju mogucnost da “lajkuju” posao, urade “follow” tj. ostave poziciju memorisanu radi, evenetualnog, kasnije vraćanja na isti, ostave komenar ili samo pogledaju. Svaka od ovih interakcija ostaje zabeležena kao što je prikazano na slici ispod.

	uniq_id	eventType	user_id
0	fffe1767dbc1713944851a0a4f02ec5b	LIKE	101131
1	fffe1767dbc1713944851a0a4f02ec5b	LIKE	100110
2	fffe1767dbc1713944851a0a4f02ec5b	VIEW	100110
3	fffd6e0361c1aecb4d099f6465392a77	FOLLOW	101320
4	fffd6e0361c1aecb4d099f6465392a77	FOLLOW	100320
5	fffd6e0361c1aecb4d099f6465392a77	LIKE	100320
6	fffcf89ea054a29e92204955ca846d63	VIEW	100622
7	fffcf89ea054a29e92204955ca846d63	VIEW	100232
8	fffcf89ea054a29e92204955ca846d63	COMMENT	100232
9	fffc6f328cef53e76f5e5bb78ed23ba9	BOOKMARK	101380

4.1.9. Iterakcije korisnika sa otvorenim poslovnim pozicijama

Kako bi implementirali sistem za preporuku na osnovu saradnje prvo je neophodno da diskretizujemo vrenosti interakcija. Time dobijamo određene vrednosti za određene interakcije.

```
event_type_strength = {
   'VIEW': 1.0,
   'LIKE': 2.0,
   'BOOKMARK': 3.0,
   'FOLLOW': 4.0,
   'COMMENT': 5.0,
}
```

```
user_rating_df['eventStrength'] = user_rating_df['eventType'].apply(lambda x: event_type_strength[x])
```

```
user_rating_df.head(5)
```

	uniq_id	eventType	user_id	eventStrength
0	fffe1767dbc1713944851a0a4f02ec5b	LIKE	101131	2.0
1	fffe1767dbc1713944851a0a4f02ec5b	LIKE	100110	2.0
2	fffe1767dbc1713944851a0a4f02ec5b	VIEW	100110	1.0
3	fffd6e0361c1aecb4d099f6465392a77	FOLLOW	101320	3.0
4	fffd6e0361c1aecb4d099f6465392a77	FOLLOW	100320	3.0

Slika 4.1.10. Diskretizacija interakcija

Da bi sistem za preporuku bio bolje informativan uzimamo samo korisnike koji su imali barem 25 interakcija na sajtu. Takođe uzimamo samo one poslove koji su imali barem 25 interakcija zato što su to poslovi za koje možemo da kažemo da su interesovali korisnike. Dobija se 1501 korisnik i 56472 iterakcije. Zatim agregiramo interakcije sumom težina svih interakcija koje su imali i primenjujemo logaritamsku funkciju kako bi imali uravnoteženu

distribuciju. Ovo radimo iz razloga što korisnici mogu više puta da pogledaju neki oglas ili komentarišu prikazano na slikama 4.3.11-4.3.14.

```
# users: 1501
# users with at least 25 ratings: 1209
```

Slika 4.1.11. Korisnici sa preko 25 interakcija

```
# of interactions: 66000
# of interactions from users with at least 25 interactions: 56472
```

Slika 4.1.12. Interakcije sa preko 25 interakcije od strane korisnika

```
# of unique user/item ratings: 37631
```

	user_id	uniq_id	eventStrength
0	100000	0f350d3a20b61289bd882547210090b4	1.807355
1	100000	29517932e0cc69b7f93196b101ac55fb	3.169925
2	100000	2960929545141233ae4185317727842e	1.807355
3	100000	2b574988cc462c5f5f1d0cfe81db909d	2.321928
4	100000	361514035c6ea8b06d07a285548b4d7a	1.000000
5	100000	397a3f1d0d9366ee899b523b36b2b800	1.807355
6	100000	3d4f800b32b1a0a5388bac60bbf27b8c	2.807355
7	100000	3d8b1489c2b2fbacc2b56f899f529185	1.807355
8	100000	3e4e15ab2e9f0543916bc84a82a66166	2.000000
9	100000	3f50d0d0c793a0ab1470438ae89cb66a	2.807355

Slika 4.1.13.

Evaluacija izuzetno bitna u mašinskom učenju zato što omogućuje poređenje različitih algoritama i parametara za model. Tehnikom unakrsne validacije (“cross-validation”) ćemo omogućiti da model pustimo nad podacima nad kojima nije učio. Korisimo jednostavniju verziju unakrsne validacije pod nazivom “holdout” (Schneider , 1997) , koja uzima nasumičnih 20% koji se koriste za kasniju evaluaciju.

Opet primenjujuci tehniku TF-iDF dobijamo vektore međutim sada su ti vektori napravljeni kao kombinacija polja iz opisa posla i polja sa veštinama. Metrika sa kojom ćemo ocenjivati nas sistem za preporku će biti odziv (“recall”) (Shung, 2018) koja će nam reći da li je korisnik imao interakcije sa top 100 preporuka za tog korisnika.

U rezultatu koji je prikazan na slici vidi visoka preciznost i za top 5 i 10. Ovo znaci da je u za oko 70% poslova koji su imali interakciju u testu u top 5 sa ovim modelom (od nasumičnih 100). Sličan rezultat je i za preciznost kada je top 10 u pitanju.

	_person_id	hits@10_count	hits@5_count	interacted_count	recall@10	recall@5
56	100003	8	7	10	0.800000	0.700000
1037	100138	7	7	10	0.700000	0.700000
351	101190	8	8	10	0.800000	0.800000
14	100772	9	6	10	0.900000	0.600000
107	100559	7	7	10	0.700000	0.700000
329	100642	7	7	9	0.777778	0.777778
933	100824	7	7	9	0.777778	0.777778
124	100672	8	8	9	0.888889	0.888889
310	100411	5	5	9	0.555556	0.555556
311	100190	8	8	9	0.888889	0.888889

Slika 4.1.14. preciznost sistema za preporuku

Na kraju će biti prikazana testiranje time što će biti prikazani podaci koji su rezultat interakcije jednog od korisnika (korisnik: 100622) sa preporukama za poslove koje su rezultat sistema za preporuku zasnovanog na sadržaju prikazanog na slikama 4.1.15-4.1.16

	uniq_id	jobtitle	Job_Skills_Without_Stopwords
20	478c1957e13a5fc94a11f7b486af0579	rsa archer administrator	rsa archer configuring troubleshooting resol...
23	4a39d298eca2e094d9d139c9e932d746	cosmos developer	cosmos sql developer xdi reports xflow sqlize...
14	9ba7b93aa12d3dbd147e8a0ef891e2b5	it service desk coordinator	"technical support" "help desk" "management" "...
10	881484516bb1d9ef6c3aa82eab7ee6c2	windows engineer	sccm citrix xenapp vdi desktop osd windo...
8	bf5ccc527aa02e0f79a6ae45a65a8583	python developer systematic trading	python developer systematic trading
2	0a7ca273ed28fa8fafbd349968b408f3	arcgis consultant	arcgis python perl java matlab geospatial...
12	493d13ad220cddf38d473da8eb34bafa	mongo db dba	mongodb mongodb cluster mongodb ops oracl...
11	04eea29ffc966c49eecbf0b637889e1b	sql database administrator iii	sql database administrator iii
19	2a5a7d54f1fd3a7dfe7c4981592b4fc0	senior software development engineer, java ser...	agile api browser development ecommerce f...
22	e535c9bbc835e147ef0686658da19aa7	technology analyst - us	drupal x php programmer
4	63d5ba94dd574ada4fd532dbfb7d798f	salesforce solution architect	salesforce solution architect
21	de4d117856aed45bad30f5238b2c81b4	business intelligence engineer	algorithms analysis business intelligence h...
6	1ade2d1a9b5aebd3a7cc348606615100	business systems analyst	business systems analyst
1	f97e9d0acc59e2f8ff2dd6195f332dd0	lead systems engineer, exchange and office 365	azure exchange and or exchange office aws
16	228d899c7147c51fe5250a937e9586d0	princ./ lead dsp engineer	dsp c firmware cellular lte rf wcdma ...
24	3c0692e69a60bc9c1ebf6f48facff51a	software development engineer - comixology	android api c# development git ios java ...
17	d391f6e7ea5ea175fb600d12c2a9fbc6	information security analyst (relocation provi...	windows active directory linux cisco tacac ra...
7	34aab8e7965c0bf025a63f51118b343e	senior software development engineer | amazon ...	algorithm analysis architecture business re...
3	2049addd0bd9b530d44f78d2c6cce841	senior it project manager	project management pmp csm agile scrum cyber s...
15	2fa5a91b54fe18256cac3919f1b43309	it manager	experience current technologies products; stra...

Slika 4.1.15. 20 poslova sa kojima je korisnik zaista imao interakciju

	uniq_id	jobtitle	Job_Skills_Without_Stopwords
0	881484516bb1d9ef6c3aa82eab7ee6c2	windows engineer	sccm citrix xenapp vdi desktop osd windo...
1	478c1957e13a5fc94a11f7b486af0579	rsa archer administrator	rsa archer configuring troubleshooting resol...
2	9ba7b93aa12d3dbd147e8a0ef891e2b5	it service desk coordinator	"technical support" "help desk" "management" "...
3	7976f22d91f7500c39d42c618a2025b3	senior windows systems administrator ts/sci (i...	accounting ada database exchange help desk...
4	de4d117856aed45bad30f5238b2c81b4	business intelligence engineer	algorithms analysis business intelligence h...
5	3c0692e69a60bc9c1ebf6f48facff51a	software development engineer - comixology	android api c# development git ios java ...
6	032adaf4209d3051fc5bd4a780b71c70	.net developer	net c# javascript jquery
7	2a5a7d54f1fd3a7dfe7c4981592b4fc0	senior software development engineer, java ser...	agile api browser development ecommerce f...
8	9d0df27cbcd781532296c62f42d33c75	software development engineer, web framework	agile browser development ecommerce founda...
9	c8f0b7fcd0dc9cc4be18392770b839f7	software development engineer, web framework	agile browser development ecommerce founda...
10	1cca7cafb9f656717e4531720570d4eb	backend qa engineer	java performance testng
11	bf8c3ab590a3346c9b19583621f2b531	sql database administrator iii	sql database administrator iii
12	04eea29ffc966c49eecbf0b637889e1b	sql database administrator iii	sql database administrator iii
13	1272f4baa440d7bb0d9c00ef6f030d46	software engineer, catalog systems	algorithm analysis development e commerce ...
14	e535c9bbc835e147ef0686658da19aa7	technology analyst - us	drupal x php programmer
15	977045cacf760cb4bc007cfb11af2511	software development engineer	algorithm analysis css development html j...
16	6606c605a8fe726ce2b0f738146a8842	software development engineer | amazon prime &...	algorithm analysis architecture business re...
17	34aab8e7965c0bf025a63f51118b343e	senior software development engineer | amazon ...	algorithm analysis architecture business re...
18	47be10099ed2a7b856c70f7ac4b1cb3b	windows systems engineer iii - active directory	analysis analytical skills application serve...
19	e050d953f84d5fa318c6a6499ac0cbdf	software development engineer	agile algorithms architecture automated de...

Slika 4.1.16. 20 poslova preporučenih sistemom za preporuku

Zaključak je da je sistem kao preporuku dao poslove koji zaista odgovaraju onima sa kojima je korisnik zaista i imao interakcije.

Peto poglavlje

5.1. Zaključak

Ovaj rad je izvršen 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 između tehnika sistema za preporuku i iz toga primenili sva tri navedena sistema kako bi prikazali njihove prednosti i mane.

Različitim tehnikama obrade teksta poput prirodne obrade jezika, TF-IDF kao i regularnih izraza smo oblikovali sadržaj kako bi bolje primenili sisteme za preporuku i time saznali kako i koliko su slične veštine za slične poslove.

U drugom delu smo primenom svih tehnika manipulacije i analize teksta izvukli ključne reči ili u nasem slučaju ključne veštine koje su najvise opisivale naš dokument i time povećali kvalitet onoga što dobijamo kao izlaz iz naseg prediktivnog algoritma. U implemetaciji algoritma prvo smo krenuli sa sistemom preporuke zasnovanom na sadržaju kako bi prikazali ključne veštine za određenje poslove bez učešća korisnika preko kojeg smo davali ocene. Prikazan je zatim sistem preporuke zasnovan na preporuci i kao zaključak se predlaze primena hibridnog sistema kako bi se izbegao problem početnog nedostatka podataka.

Ovaj rad treba da posluzi u buducem istrazivanju i njegov sadržaj 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

Balabanović, M., & Shoham, Y. (1997, March). *Content-Based, Collaborative Recommendaton*. Retrieved from Communication of the ACM: http://courses.ischool.utexas.edu/donturn/2008/fall/INF_385Q/readings/Balabanovic_Shoham-1997-Fab.pdf

Boulet, G. (2015, October 17). *The Difference Between Knowledge And Skills: Knowing Does Not Make You Skilled*. Retrieved from eLearning industry: https://elearningindustry.com/difference-between-knowledge-and-skills-knowing-not-make-skilled

Dice. (2017). *U.S. Technology Jobs on Dice.com: 22,000 US-based Technology Job Listings*. Retrieved from Kaggle: https://www.kaggle.com/PromptCloudHQ/us-technology-jobs-on-dicecom/home

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

FadhelAljunid, M., & Manjaiah, H. D. (2017). *A Surevey on recommendation systems for social media using big date analytics.* Retrieved from International Journal of Latest Trends in Engineering and Technology: https://www.ijltet.org/journal/151063987410.pdf

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/

Huang, Z., Zeng, D., & Chen, H. (2004). *A Comparative Study of Recommendation Algorithms in E- Commerce Applications.* Retrieved from Semantic Scholar: https://pdfs.semanticscholar.org/b2cc/302b01f4ad174c941b9fb4525e972560a3dc.pdf

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

Lops, P., Semeraro, G., & Gemmis, M. d. (2011, January). *Content-based Recommender Systems: State of the Art and Trends.* Retrieved from Research gate: https://www.researchgate.net/publication/226098747_Content-based_Recommender_Systems_State_of_the_Art_and_Trends

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
Pythonspot. (2017). *NTLK stop words*. Retrieved from Pythonspot: https://pythonspot.com/nltk-stop-words/
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, May 9). *Introduction to Recommender Systems in 2018*. Retrieved from Tryo labs: 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
Schneider , J. (1997, February 7). *Cross validation*. Retrieved from Carnegie Mellon University - School of Computes Science: https://www.cs.cmu.edu/~schneide/tut5/node42.html
Shung, P. K. (2018, March 15). *Accuracy, Precision, Recall or F1?* Retrieved from Towards Data Science: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
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/