ELEKTROTEHNIČKI FAKULTET

UNIVERZITETA U BEOGRADU

**KATEDRA ZA RAČUNARSKU TEHNIKU I INFORMATIKU**

**DIPLOMSKI RAD**

Petar Korda

PRIMENA DATA MINING-A U PSIHO-ONKOLOGIJI, I IMPLEMENTACIJA TEKSTUALNOG ANALIZATORA I ONKOLOŠKOG PREDIKTORA

**BEOGRAD, 2015**

**SADRŽAJ**

[**1. Uvod** 2](#_Toc414378787)

[1.1. Data mining mehanizmi i tehnike 3](#_Toc414378788)

[1.2. Primene 3](#_Toc414378789)

[**2. Opis problema** 5](#_Toc414378790)

[**3. Opis realizovanog rešenja** 8](#_Toc414378791)

[3.1. Određivanje tipova ličnosti koje spadaju u rizičnu/nerizičnu grupu od dobijanja kancera 8](#_Toc414378792)

[3.2. Istraživanje doktora Gosart-Matičeka o psihosocijalnim faktorima u genezi organskih oboljenja 9](#_Toc414378793)

[3.3. Zaključak 13](#_Toc414378794)

[3.4. Određivanje rečnika 13](#_Toc414378795)

[3.5. Implementacija Tekstualnog Analizatora 19](#_Toc414378796)

[3.6. Priprema ulaza za tekstualnu analizu 21](#_Toc414378797)

[3.7. Tekstualna analiza 22](#_Toc414378798)

[3.8. Struktura ulaza i izlaza 24](#_Toc414378799)

[**4. Analiza rezultata** 28](#_Toc414378800)

[4.1. Poređenje sa postojećim istraživanjima 29](#_Toc414378801)

[**5. Zaključak** 31](#_Toc414378802)

# **1. Uvod**

Data mining je u protekloj deceniji postala dobro uspostavljena disciplina u domenu veštačke inteligencije (AI - artificial intelligence) i inžinjeringa znanja (KE - knowledge engineering). Ima korene u statistici i mašinskom učenju (machine learning), ali obuhvata i mnoge druge oblasti računarstva. U poslednje vreme postoji dosta interesovanja oko ove oblasti, iz razloga što su napreci u hardveru obogućili procesorsku moć za obradu ogromnih količina podataka. Za data mining se može reći da je više primena, nego tehnologija, i baš zbog toga se očekuje da će ostati aktuelna i u skorijoj budućnosti.

Poreklo termina data mining datira iz osamdesetih kada je termin počeo da se koristi, uglavnom u istraživačkim krugovima. Data mining možemo definisati kao skup mehanizama i tehnika, realizovanih u softveru sa ciljem da se dođe do skrivenog znanja iz datog skupa podataka. U ranim devedesetim, bio je opšte prihvaćen kao pod proces otkrivanja skrivenih informacija u okviru mnogo većeg procesa otkrivanja znanja u bazama podataka (KDD - Knowledge Discovery in Databases)[[1]](#footnote-2). Najčešće korišćena definicija KDD-a je iz 1996 i pripisuje se Fayyad-u: “Netrivijalan proces identifikovanja validnih, noebičnih, potencijalno korisnih i razumljivih šablona u podacima”. Ostali pod procesi KDD-a su: priprema podataka za obradu i analiza/vizualizacija rezultata. Data mining se inicijalno bavio tabularnim podacima (baze podataka), dok je sada sve veći fokus na primenama nad ne standardnim skupovima podataka (npr. slike, kolekcije dokumenata, tekst, video.. ). Popularnost ovih tehnika je sve veća, i razne korisne primene nalaze se u nauci, ali i u komercijalnim okruženjima, a neki od primera su telekomunikacije, psihologija, finansije, kao i mnogi drugi. Ovo veliko interesovanje za oblast pronalaženja skrivenog znanja pripisujemo sve bržem napretku tehnologije i hardvera, i mogućnosti za rad sa sve većim skupovima podataka koristeći samo personalne računare.

## 1.1. Data mining mehanizmi i tehnike

Data mining tehnike se mogu grubo podeliti na: (1) izvačenje/pronalaženje šablona (pattern extraction), (2) grupisanje podataka (data clustering), (3) klasifikacija/kategorizacija (classification). U nastavku je svaka od ovih tehnika ukratko opisana [1].

1. *Pronalaženje šablona*

Kroz istoriju, data mining je oduvek imao znatan fokus na pronalaženje informacija koje se ponavljaju, tj. šablona u datom skupu podataka. Ovi šabloni mogu da se jave u različitim formama, npr. potrošačke navike kupaca, trendovi i tako dalje. Šablon je bilo koja česta pojava entiteta, događaja, objekata..

Jedna od tehnika pronalaženja šablona jeste traženje asocijativnih pravila (ARM - Association Rule Mining). Ideja kod ove tehnike jeste da se uoče česta ponavljanja/šabloni i relacije u okviru datog skupa podataka, i onda na osnovu njih izvući asocijativna pravila. Asocijativno pravilo govori da ukoliko se određeni entiteti pojavljuju zajedno, onda postoji i određena verovatnoća da se pojave još neki entiteti iz disjunktnog skupa. Primer ovoga jeste sistem za preporuke (ukoliko se korisniku svideo film x, verovatno je da će mu se svideti i film y).

1. *Grupisanje podataka*

Grupisanje podataka je tehinka kategorisanja različitih skupova podataka. Ova tehnika je već dobro uspostavljen oblik data mining-a, a zapravo predstavlja vrstu mašinskog učenja (machine learning). Interesantno je da ne postoji “najbolji” algoritam za grupisanje svih vrsta podataka, već su jedni dobri za jednu vrstu podataka, a drugi za drugu.

1. *Klasifikacija podataka*

Ova tehika se bavi konstrukcijom klasifikatora koji se primenjuju na ranije nepoznate podatke, kako bi se oni grupisali, tj. podelili na određene klase. Klasifikacija ima dodirnih tačaka sa tehnikom grupisanjem podataka, a razlikuju se u tome što klasifikacija zahteva skup podataka iz koga će da uči, i na osnovu toga izgradi određene klase (training data). Zbog toga za klasifikaciju kažemo da je to “nadgledano učenje” (supervised learning), a za grupisanje podataka “ne nadgledano učenje” (unsupervised learning). Klasifikaciju možemo uporediti sa rezonovanjem na osnovu prethodnih iskustava.

Klasifikatori mogu da budu u različitim oblicima: stabla odlučivanja, SVM-ovi itd. Klasifikatori mogu da budu: binarni klasifikatori (podatku se dodeljuje jedna od dve klase), više klasni klasifikatori (podatku se dodeljuje jedna od mogućih klasa, a čiji je broj veći od dva) ili multi klasni klasifikatori (podatku se može dodeliti više od jedne klase).

## 1.2. Primene

Neke od popularnijih primena data mining-a su:

1. *Tekst analiza*

Prirodan sledeći korak od analize tabularnih podataka je bila analiza tekstualnih podataka. Tipična primena jeste kategorizacija velikih količina tekstualnih dokumenata (npr. veb stranica), kao i podtraga za mišljenjem i stavovima ljudi prema nečemu (recimo, prema nekom proizvodu). Još jedna primena jeste kontekstualno sažimanje tekstualnih struktura.

Jedan od glavnih izazova kod tekstualne analize, je kako na najbolji način predstaviti tekst, tj. pripremiti ga za primenu tehnika data mining-a. Jedan od čestih oblika reprezentacije tekstualnih fajlova je bag-of-words reprezentacija, gde je tekst predstavljen kao niz ključnih reči (moguća je i bag-of-phrases reprezentacija, gde se umesto ključnih reči koriste ključne fraze)

1. *Analiza slika*

Kao i kod tekstualne analize, jedan od većih problema jeste u kom formatu predstaviti sliku (i 2d i 3d) tako da je ona pogodna za tehnike data mining-a. Neke od pogodnih reprezentacija su histogrami i grafovi.

Ova oblast je još uvek u povoju, i u određenim implementacijama pokazuje odlične rezultate. Na primer dijagnostika pacijenta na ostovu snimka magnetne rezonance.

1. *Analiza grafova i stabala*

Analiza grafova i stabala se u suštini bavi identifikacijom skupova podataka koji se ponavljaju (šabloni), a to su u ovom slučaju pod grafovi. Ovi šabloni mogu biti pod grafovi koji se često pojavljuju u kolekcijama različitih grafova, ili pod grafovi koji se pojavljuju u jednom velikom grafu.

Popularna primena analize grafova je mining društvenih mreža (social networks mining). Ovde je cilj da se identifikuju grupe ljudi (npr. sličnih interesovanja) u okviru ovih mreža.

Alati i tehnike data mining-a nalaze primene u mnogim aplikacijama koje treba da donose proaktivne i znanjem pokretane odluke i izlaze. Korporacije i finansijske institucije, koriste ove tehnike kako bi povećali prodaju, smanjili troškove i procenili rizike. Farmaceutska istraživanja ih koriste za pronalaženje novih jedinjenja i načina lečenja određenih bolesti. Veliku primenu, data mining ima i u državnoj i bezbednosti uopšte, kao što su predikcija kriminalnih aktivnosti, praćenje i pronalaženje terorista itd. [1] [2]

# 

# **2. Opis problema**

Ono čime se ovaj radi bavi, tj. problem koji će biti opisan, a zatim kasnije i razložen i rešen, se u kratkim crtama može opisati sledećim tekstom:

*Zadatak je napraviti sistem koji će da vrši onkološke predikcije na osnovu ulaza koji je u obliku privatnog mejl sandučeta. Potrebno je formirati “rečnik”, tačnije niz ključnih reči na osnovu kojih će se vršiti tekstualna analiza e-mejlova i dobijati željeni izlaz iz sistema.*

Suština ovog problema jeste zapravo kako povezati jezik koji osoba koristi u pisanju elektronskih pisama sa onkološkim oboljenjima, tačnije sa rizikom od oboljenja od raka. Reči i jezik koji koristimo u svakodnevnom životu odražavaju ko smo u suštini, i ovo nije ništa novo ili iznenađujuće. Jezik je najčešći i najpouzdaniji način kojim ljudi prevode njihove misli i emocije u formu koju i drugi mogu da razumeju. Samim tim, reči i jezik su sama srž komunikacije i psihologije, i predmet izučavanja kognitivnih, kliničkih, socijalnih, ali i mnogih drugih psihologa, sa ciljem da razumeju ljudsko biće. Sa ovim u vezi, potrebno je spomenuti još jednu granu psihologije, a to je psiho-onkologija, koja se bavi presekom psihologije i onkologije, tačnije efektima kancera na psihološko zdravlje pacijenta, ali i socijalnim, bihejvioralnim i psihološkim faktorima koji mogu da utiču na tok bolesti ili predstavljaju rizik od samog oboljenja. Još pre 4000 godina jedan indijski lekar i filozof je rekao sledeće: “Postoje dve vrste oboljenja, telesna i mentalna. Svaka od njih nastaje iz one druge. Mentalni poremećaji rađaju se iz fizičkih, i, slično, fizički poremećaji dolaze iz mentalnih”. Isto mišljenje imao je i Hipokrit, kao i mnogi drugi, a sve se svodi na to da ličnost i oboljenje stoje u veoma bliskoj interakciji i da zanemarivati ličnost znači izostavljati veoma značajan faktor kako za profilaksu, tako i za lečenje [3]. Ako uzmemo primer karcinoma, mnogi sugerišu da postoje ličnosti sklone njemu. Drugim rečima, karcinom se mnogo lakše razvija kod ljudi određenog tipa.

Na osnovu ovoga, problem kojim se bavi ovaj rad možemo sagledati ovako. Bitni faktori problema su:

1. *Profili, tj. tipovi ličnosti koje su sklonije karcinomu, ali i oni koji se nalaze na drugoj strani lestvice, tj. spadaju u nerizičnu grupu.*

Tačnije, ovde je potrebno jasno definisati “klase” ljudi na osnovu njihovog rizika da obole od raka. Dve su klase:

* 1. *Rizični*
  2. *Nerizični*

Koji su to tipovi ličnosti koji spadaju u klasu a, a koji u klasu b. Takođe, bitno je kako izvršiti podelu na tipove ličnosti. Postoji mnoštvo psiholoških “teorija ličnosti”, kao što su: teorija tipa ličnosti A i B, teorija 4 temperamenta, Jungova teorija itd. Pitanje je da li je moguće primeniti (ili koja je najpogodnija) neku od već postojećih teorija koje se bave različitim tipovima ličnosti i primeniti je u kontekstu problema.

1. “Rečnik”, tj. ključne reči i jezik, koji ovi tipovi ličnosti koriste, i to u kontekstu e-mejlova

Ovde postoji nekoliko pod problema:

a. Koji jezik koristiti za rečnik

b. U kom formatu čuvati rečnik (json, xml, itd.)

c. Kako mapirati ključne reči (ili ključne izraze) sa određenim tipom ličnosti

d. Odrediti težinske vrednosti datih ključnih reči (ako je potrebno)

1. Metod tekstualne analize ovih e-mejlova

To jest, kako na osnovu ulaza (e-mejlova) i rečnika dobiti izlaz. Ovde je bitno naglasiti da

treba i jasno definisati:

1. Format ulaza (postoji veliki broj standardnih i nestandardnih formata mejl sandučića, a neki od njih su: mbox, Maildir, MH, itd)
2. Format izlaza (da li da izlaz bude binarni, ili da postoji više mogućnosti za krajnji rezultata)

Takođe, nakon što se jasno definiše format ulaza, potrebno ga je i pripremiti za samu

analizu (takođe se treba videti kako implementirati ovu pripremu). Ako izlaz posmatramo

kao izlaz iz crne kutije, on bi trebao da pokaze da li data osoba (čije sanduče je bilo ulaz

crne kutije) spada u rizičnu ili nerizičnu grupu za oboljenje od kancera. Međutim, da li se

taj rizičan/nerizičan izlaz može finije podeliti unutar samih klasa, tj. da li postoje oni koji

su manje, srednje ili više rizični, ili u obrnutoj situaciji, nerizični.

Nakon što je ulaz pripremljen, i format izlaza poznat, potrebno je razmišljati o samoj

analizi teksta. Da li vršiti najobičnije prebrojavanje nađenih ključnih reči (ili izraza) iz

rečnika, ili raditi neko težinsko prebrojavanje. Takođe, nema svaka reč uvek isto

značenje, već u mnogome zavisi od konteksta u kome se nalazi, pa samim tim neke reči

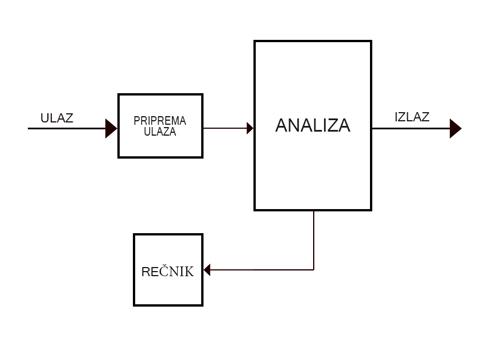
mogu da nose pogrešnu informaciju ukoliko se ne posmatra i okruženje u kome se

nalaze. Treba uzeti u obzir i sinonime i antonime ključnih reči iz rečnika, i kako

pronalaziti i tretirati ove reči. Još neki od problema vezanih za samu tekstualnu analizu

jesu: negacija (kako detektovati da se reč nalazi u njenoj negativnoj formi) i kako odrediti

subjekat na koji se pronađena reč iz rečnika odnosi.



*Slika 1: Moduli sistema za određivanje rizika oboljenja od karcinoma*

*na osnovu privatnih e-mejlova, i njihova povezanost*

Još neka od pitanja su: koju tehnologiju koristiti (postoji mnogo biblioteka za različite programske jezike za procesiranje teksta, kao i samih jezika koji su manje ili više pogodni za ovaj zadatak) i optimizacija.

Ukoliko se ovaj problem posmatra sa višim stepenom apstakcije, u idealnoj situaciji ovaj sistem bi mogao biti primenljiv i na mnoge druge probleme. Postojao bi ulaz, koji je u nekoj formi teksta (ne mora biti e-mejl), rečnik (koji sadrži ključne elemente koji se pretražuju u ulaznim fajlovima, i koji možda imaju određene težinske vrednosti koje su vezane sa problem koji se rešava), sistem za tekstualnu analizu (koji na osnovu rečnika prolazi kroz ulazne fajlove i pretražuje ih) i izlaz (koji može biti format rešenja mnogobrojnih problema, a ne samo ovog).

U sledećem poglavlju biće izneto rešenje konkretnog problema u onkološkoj (ali koji može da se primeni i na druge slične probleme) primeni.

# **3. Opis realizovanog rešenja**

*Zadatak je napraviti sistem koji će da vrši onkološke predikcije na osnovu ulaza koji je u obliku privatnog mejl sandučeta. Potrebno je formirati “rečnik”, tačnije niz ključnih reči na osnovu kojih će se vršiti tekstualna analiza e-mejlova i dobijati željeni izlaz iz sistema.*

U predsojećem tekstu, detaljno će biti objašnjeno rešenje gore navedenog problema, kao i koraci kroz koje se prolazilo u dolasku do istog. Takođe, biće reči i o abstrakciji ovog problema, i primeni sistema na mnogo širi opseg.

Najbitniji koraci u ovom rešenju su:

1. Određivanje tipova ličnosti koje spadaju u rizičnu/nerizičnu grupu od dobijanja kancera
2. Određivanje “rečnika”
3. Implementacija tekstualnog analizatora

U daljem tekstu, svaki od ovih koraka je detaljno opisan.

## 3.1. Određivanje tipova ličnosti koje spadaju u rizičnu/nerizičnu grupu od dobijanja kancera

Uloga tipa ličnosti, kao i drugih psihosocijalnih faktora kao vid rizika od dobijanja raka i drugih bolesti je predmet mnogobrojnih naučnih istraživanja i radova.

Od davnina se pretpostavljalo da psihološki profil osobe ima bitnu ulogu u nastanku oboljenja kod ljudi. Još 200-te godine pre nove ere Galen je uočio da su melanholične žene sklonije kanceru, nego one sangviničnog temperamenta. 1962, Kisen i Eysenck (Ajzenk) su izvršili jedno od prvih modernih studija povezanosti tipa ličnosti i raka i zaključili da su pacijenti sa rakom pluća češće bili ekstravertni. Istraživanja koja su sledila u narednim godinama nisu bila konzistentna i dobijani su različiti rezultati. Od šest retrospekvinih kohort studija[[2]](#footnote-3), četiri su pronašle statistički značajniju povezanost između ekstraverzije, nižih skorova neuroticizma, nižih skorova anksioznosti i rizika od raka dojke i pluća, dok preostale dve studije nisu pronašle nikakvu značajniju vezu izmedju tipa ličnosti, temperamenta i incidence kancera. Od deset prospektivnih kohort studija[[3]](#footnote-4), pet su pronašle statistički značajniju vezu između “Tipa 1” ličnosti (racionalni, principijalni, perfekcionisti, itd.), bespomoćnosti i pojave raka (ne zavisno od tipa raka), kao i anti-emocionalnog ponašanja i raka dojke. Ostalih pet istraživanja nije došlo do značajnijih rezultata, koji pokazuju vezu između ličnosti i pojave raka [4].

Kao i rezultati naučnih istraživanja, su podeljena i ne može se sa sigurnošću tvrditi da postoji veza između tipa ličnosti, temperamenta, kao i drugih psihosocijalnih faktora i rizika od oboljenja od raka.

Rešenje problema koje se ovde opisuje je bazirano na rezultatima istraživanja doktora Grosart-Matičeka (Grossarth-Maticek) koje je vršio u Jugoslaviji i Nemačkoj, a koja za zaključak imaju da ličnost i oboljenje stoje u veoma bliskoj interakciji. U sledećem odeljku će u kratkim crtama biti opisana ova istraživanja i njeni rezultati.

## 3.2. Istraživanje doktora Gosart-Matičeka o psihosocijalnim faktorima u genezi organskih oboljenja

Grosart-Matiček je izveo tri značajne studije. U svakoj od njih ličnost subjekta je ispitana na početku studije duž dve različite dimenzije. Jedna od njih je pomoću upitnika od 109 pitanja podeljena na nekoliko skala. Postojala je, na primer, skala racionalnosti i emocionalnosti. Postojali su i drugi upitnici koje je koristion Grosart-Matiček da bi predvideo pojavu karcinoma. Dva upitnika sa daleko najvećom prediktivnom tačnošću su upitnik koji ispituje racionalno i neemocionalno mišljenje, kao i onaj koji ispituje osećanja bespomoćnosti i beznađa. Takođe, tu je i upitnik koji ispituje iritabilnost, uzbuđenost, ljutnju i kod koga se javila negativna korelacija sa pojavom karcinoma. Deset godina kasnije, urađeno je ispitivanje da bi se videlo ko je iz uzorka umro, i od čega.

U studiji rađenoj u Jugoslaviji u mestu Crvenka, svako drugo domaćinstvo je bilo posećeno i u najstariji ukućanin je ušao u uzorak istraživanja. Subjekti istraživanja su uradili upitnike, a nakon deset godina rezultati su bili sledeći: osobe koje su imale više skorove na upitnicima racionalnosti, neemocionalnog mišljenja, kao i bespomoćnosti i beznađa su češće obolevali od karcinoma.

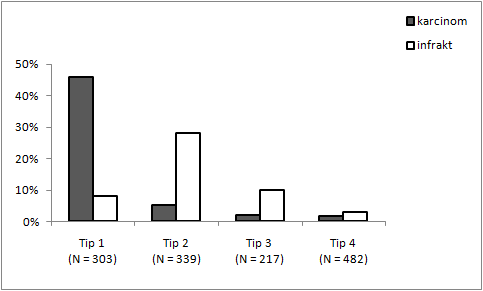
Slični rezultati dobijeni su i u druge dve studije rađene u Hajdelbergu u Nemačkoj. Ovo su takođe bile desetogodišnje prospektivne studije. Prva od njih je rađena na slučajnom uzorku koji je imao specifikovana ograničenja u kompoziciji pola i uzrasta, dok se uzorak u drugom istraživanju razlikovao od prvog po tome što su u njemu ispitanici bili pod jakim stresom. Psiho-socijalni stres je jedna od varijabli koje su se uvek iznova i iznova pojavljivale kad god su lekari razmatrali vanmedicinske faktore geneze kancera. Grosart-Matiček se posebno bavio interpersonalnim stresom - vrstom stresa koji se javlja kada valjena osoba umre, ili napusti drugu osobu. Ovako su dobijena dva uzorka, vrlo slična po uzrastu, polu, navikama pušenja, itd. ali je jedan uzorak normalan, dok je drugi dobijen selekcijom na osnovu činjenice da su subjekti u njemu u stanju visokog stresa.

Grosart-Matiček je pomoću upitnika profilisao ispitanike iz sva tri istraživanja, na sledeći način:

1. *Tip 1*: takozvani kanceru-sklon tip, koji se u životu odlikuje iracionalnošću, lakim predavanjem, osećajem beznađa i bespomoćnosti, osećanjima da se ne može ni boriti niti predati. Takva osoba je bespomoćna kada se suoči sa problemima.
2. *Tip 2*: osoba sklona koronarnim srčanim oboljenjima i odlikuje se osećanjima ljutnje, iritabilnosti, uzbuđenosti jedne ili druge vrste kad god dođe u stresnu situaciju.
3. *Tip 3*: kod ovog tipa osobe postoji tendencija kombinovanja osobina tipa 1 i 2, kroz smenjivanje osobina jednog i drugog tipa.
4. *Tip 4*: normalni autonomni tip ličnosti koja je u stanju da se suoči sa stresom, koja je razvila metode za borbu protiv stresa i koja se ne predaje osećanjima beznađa, već je na putu da uradi nešto efikasno u vezi sa tim.

Slede rezultati dobijeni u sva tri istraživanja. Ovde se mogu videti tipovi 1, 2, 3 i 4, a ispod je broj slučajeva svakog od njih. Takođe, na ipsilon osi može se videti procenat ljudi određenog tipa koji su umrli od karcinoma ili od koronarnih srčanih oboljenja.

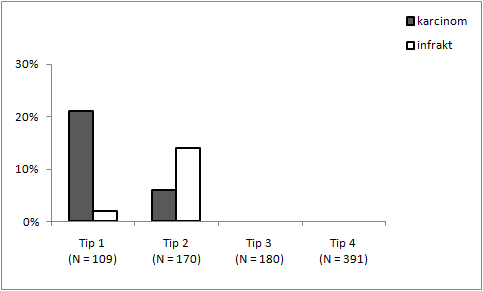
U jugoslovenskom uzorku, od osoba tipa 1, skoro 50% je umrlo od karcinoma, dok je relativno malo njih (manje od 10%) umrlo od koronarnih oboljenja. Kod osoba tipa 2, situacija je potpuno suprotna, a od osoba tipa 3 i 4, sve u svemu, relativno malo njih je umrlo od bilo koje od ovih bolesti. Tipovi 3 i 4 su zdravi tipovi, pri čemu je tip 4 zdraviji od tipa 3.



*Slika 2: Rezultati istraživanja rađenog u Jugoslaviji*

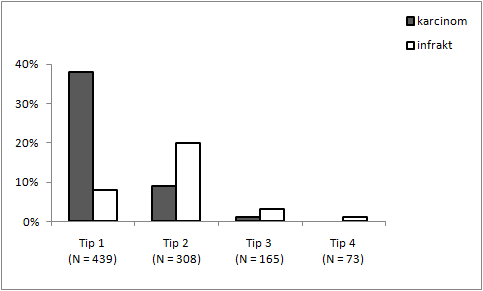
Jugoslovenski uzorak se sastojao od veoma starih subjekata, pa je i očekivano da je dosta njih umrlo nakon deset godina. Slede rezultati, hajndelberškog istraživanja na normalno uzorku (uzorak je mnogo mlađi, pa samim tim postoji manje smrtnih slučajeva).

Rezultati su sledeći: osobe tipa 1 su pre umirale od karcinoma nego od koronarnih srčanih oboljenja, dok je kod osoba tipa 2 obrnuto. Kod osobe tipa 3 i 4 uopšte nije bilo smrtnih slučajeva u periodu od deset godina.



*Slika 3: Rezultati istraživanja rađenog u Hajdelbergu (normalna grupa)*

U stresnoj grupi bilo je 40% više smrtnih slučajeva nego u normalnoj grupi. I opet, osobe tipa 1 su pre umirale od karcinoma. Na osnovu ovoga, zaključuje se da je stres veoma važan činilac u uzrokovanju bolesti i smrti.



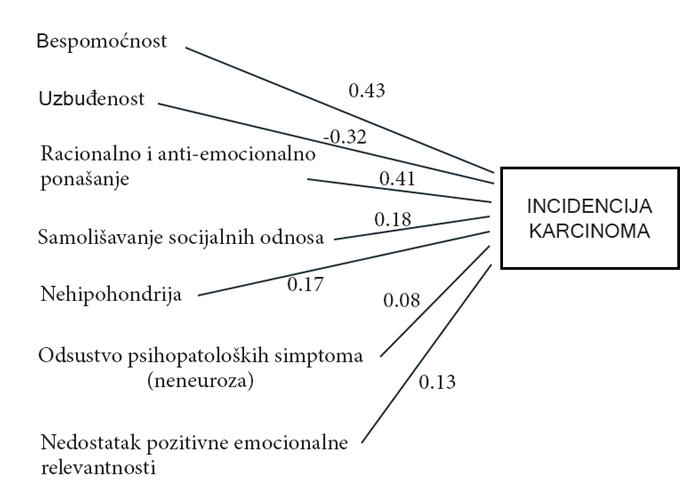
*Slika 4: Rezultati istraživanja rađenog u Hajdelbergu (stresna grupa)*

Kako bi podvrdio rezultate dobijene u ovim istraživanjima, Grosart-Matiček je izvršio još jednu eksperimentalnu studiju. Uveo je svoj koncept bihejvioralne terapije, sa posebnim tretmanima u kojima je pokušavao da promeni ponašanje osobe sa karakteristikama tipa 1 ili 2 u ponašanje osoba sa karakteristikama tipa 4. Tačnije, uradio je sledeće: uzeo je 100 ljudi koji su po svojim osobinama ličnosti bili skloni kanceru. Podelio ih je u dve grupe: kontrolnu i terapijsku (na koju je primenjivao njegovu terapiju), pa ih je pratio kroz period od 13 godina. Rezultati su bili sledeći:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Još uvek živi** | **Umrli od karcinoma** | **Umrli od drugih uzroka** | **Total** |
| **Kontrolna grupa** | 19 | 16 | 15 | 50 |
| **Terapijska grupa** | 45 | 0 | 5 | 50 |
|  | 64 | 16 | 20 | 100 |

*Slika 5: Rezultati ekperimentalnog istraživanja sa kontrolnom i terapijskom grupom*

Ovim je takođe pokazano, da ličnosti tipa 1 imaju veći rizik od oboljenja, i smrti uzrokovane karcinomom, ali i da je uz korekcije u psiho-socijalnom ponašanju i stavu moguće smanjiti ovaj rizik. Ovaj eksperiment ide i u prilog ranijim istraživanjima koja su došla do sličnih zaključaka.



*Slika 6: Osobine ličnosti koje utiču na pojavu karcinoma*

Na slici 6su sa leve strane predstavljene osobine ličnosti koje po istraživanjima doktora Grosart-Matičeka utiču na pojavu i oboljenje od karcinoma, kao i njihove težine (na linijama) koje predstavljaju u kojoj meri jedna osobina utiče na incidenciju karcinoma (pozitivna vrednost ide u prilog razvijanju raka, a negativna obrnuto) [4].

## 3.3. Zaključak

Na osnovu ovih rezultata i istraživanja napravljena je podela na tipove ličnosti podrebne za rešenje problema psiho-onkološkog prediktora. Dve klase ličnosti su korišćene:

1. *Rizični* - imaju veći rizik od dobijanja karcinoma, a poseduju osobine tipa 1 (istraživanje opisano gore), tj.: racionalno i neemocionalno ponašanje i mišljenje, bespomoćnost, predavanje i samolišavanje socialnih odnosa.
2. *Nerizični* - imaju manji rizik od dobijanja karcinoma, a poseduju osobine koje su suprotne onima iz grupe rizičnih: emocionalno ponašanje i mišljenje, neracionalnost, neuroze, uzbuđenost, pozitivni stavovi, ne predaje se lako i bori se sa problemima

## 3.4. Određivanje rečnika

Sledeći korak u rešenju je da se odrede ključne reči, ili izrazi koje bi se koristile za profilisanje ličnosti na rizičnu ili nerizičnu. U stvari, pitanje ovde je koji bi to jezik koristila ličnost koja ima osobine iz rizične grupe (racionalno, bespomoćnost, itd.), a koji ona ličnost iz nerizične grupe (neracionalnost, neuroze, itd.).

Neke od osnovnih pretpostavka na kojima je dalje bazirano rešenje su:

1. Korišćeni jezik je engleski, zbog prirode ulaznih test fajlova
2. Rečnik se sastoji samo od ključnih reči, a ne i izraza (zbog jednostavnosti, i brže pretrage)

Kako nije bilo moguće doći do upitnika koji je koristio doktor Grosart-Matiček za profilisanje ispitanika, urađeno je mapiranje jednog drugog psihološkog test koji se koristi kao test ličnosti. Ovaj test se zove NEO PI R, u formi je tekstualnog upitnika i ispituje nekoliko varijabli kod ispitivanih ličnosti: neuroticizam,ekstravercija, otvorenost, saradljivost, savesnost. Unutar ovih osobina postoji još dublja podela i ona je podeljena u sledećoj tabeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NEUROTICIZAM** | **EKSTRAVERZIJA** | **OTVORENOST** | **SARADLJIVOST** | **SAVESNOST** |
| anksioznost | toplina | fantazija | poverenje | kompetentnost |
| hostilnost | gregarnost | estetika | iskrenost | red |
| depresivnost | asertivnost | osećanja | altruizam | dužnost |
| socijalna nelagodnost | aktivitet | aktivnost | popustljivost | postignuće |
| impulsivnost | potraga za uzbuđenjem | ideje | skromnost | samodisciplina |
| vulnerabilnost | pozitivne emocije | vrednosti | blaga narav | promišljenost |

*Slika 7: Varijable NEO PI R psihološkog testa ličnosti*

Test se sastoji od 240 pitanja, gde se svako pitanje vezuje za neku osobinu, tj. relacija između pitanja i neke od gore spomenutih osobima je više na jedan (jedna osobina ima po 8 pitanja koja je testiraju). Na pitanja se odgovara nekim od brojeva jedan do pet u zavisnosti od toga koliko se na toj skali ispitanik slažee sa pitanjem. Neki od primera pitanja slede:

*Depresivnost*

Retko se osećam usamljen i tužan.

Ponekad se osećam potpuno bezvrednim.

Retko sam tužan ili potišten.

Ponekad imam jaka osećanja grešnosti i krivice.

Obično okrivljujem sebe kad nešto ne ide kako treba.

Imam loše mišljenje o sebi.

Ponekad mi se sve čini sumornim i beznadežnim.

Prečesto se, kada stvari krenu loše, obeshrabrim i dođe mi da se predam.

*Impulsivnost*

Retko preterujem u bilo čemu.

Teško se odupirem svojim željama i porivima.

Dobro odolevam okušenjima.

Dešava mi se da se prejedem.

Retko se prepuštam svojim trenutnim nagonima.

Ponekad se prežderem dok mi ne pozli.

Ponekad nešto uradim bez razmišljnja pa se posle kajem.

U stanju sam da uvek vladam svojim osećanjima.

Ono što bi sada trebalo uraditi jeste vezati svaku od ovih osobina koje se testiraju NEO PI R-om za neku od dve klase rizičnih ili nerizičnih tipova ličnosti (ako je moguće, ako ne samo ignorisati ta pitanja). Nakon što se ovo uradi, ideja je da se za svaku osobinu, na osnovu pitanja dodje do ključnih reči, i njihovih sinonima i antonima, koje najbolje opisuju odgovore, i to u različitim nivoima intenziteta.

Date su tabele koje pokazuju koje se osobine iz NEO PI R-a mogu primeniti u modelu dva tipa ličnosti rizičnih i nerizičnih. Sa leve strane (označeno sa ++++) nalaze se osobine koje su pripisane tipu ličnosti koji ima veći rizik da oboli od raka, a sa desne (označeno sa ----) drugoj grupi, nerizičnih. Prvi red tabele je osobina koja odgovara modelu ličnosti iz istraživanja doktora Grosart-Matičeka, i *grp\_score* koji predstavlja tezinu ove osobine u riziku sa pojavu karcinoma (na osnovu Slike 6).

|  |  |
| --- | --- |
| **RACIONALNO PONASANJE grp\_score = 0.41** | |
| ++++ | ---- |
| *promisljenost* | *impulsivnost* |
|  | *fantazija* |

|  |  |
| --- | --- |
| **ANTIEMOCIONALNO PONASANJE grp\_score = 0.41** | |
| ++++ | ---- |
| *popustljivost* | *hostilnost* |
|  | *impulsivnost* |
|  | *osecanja* |

|  |  |
| --- | --- |
| **BESPOMOCNOST grp\_score = 0.43** | |
| ++++ | ---- |
| *vulnerabilnost* | *asertivnost* |

|  |  |
| --- | --- |
| **UZBUDJENOST grp\_score = -0.41** | |
| ++++ | ---- |
|  | hostilnost |
|  | potraga za uzbudjenjem |

|  |  |
| --- | --- |
| **SAMOLISAVANJE SOCIJALNIH ODNOSA grp\_score = 0.18** | |
| ++++ | ---- |
|  | toplina |

|  |  |
| --- | --- |
| **PSIHOZE grp\_score = -0.8** | |
| ++++ | ---- |
| anksioznost |  |
| socijalna nelagodnost |  |

|  |  |
| --- | --- |
| **POZITIVNE EMOCIJE grp\_score = -0.13** | |
| ++++ | ---- |
| depresivnost | pozitivne emocije |
| blaga narav |  |

*Slika 7: Tabele veze osobina iz testa i osobina iz modela Gosart-Matičeka*

Za svaku od ovih osobina, postupak traženja ključnih reči bio je sledeći:

1. Za svako pitanje pronađene su odgovarajuće reči koje ga opisuju, ali i koje bi neka osoba koja ima tu osobinu koristila u svakodnevnom jeziku kako bi objasnila/opisala kako se oseća. Na primer za pitanje “Često se osećam usamljen i tužan” neke od dobijenih ključnih reči su: *depressed, sad i lonely*. Ili za “Retko se brinem za budućnost”: *concerned, anxious*. Ovo je urađeno za svako pitanje. Treba napomenuti da su za neka pitanja dobijene iste reči, i one su eliminisane ukoliko su bile unutar iste “više” osobine (psihoza je “viša” osobina za anksioznost), tj. unutar iste tablice.
2. Pronađeni su sinonimi i različite moguće forme za reči nađene pod 1. Za *depressed, sad, lonely* neki od sinonima su: *unhappy,melancholic, moody, heartbroken, ripped, hurt* itd. Ovi sinonimi su traženi uz pomoć engleskog rečnika sinonima.
3. Pronađeni su antonimi za reči nađene pod 1. i 2. takođe uz pomoć rečnika. Za primer pod 2. to su reči: *happy, awesome, joyful, blissful*, itd.
4. Sve ključne reči, njeni sinonimi i antonimi dobijeni za jedno pitanje su podeljeni na osnovu nekoliko kriterijuma na dva puta po tri grupe (jedna za sinonime, jedna za antonime). Ove tri grupe, zapravo predstavljaju intenzitet kojim je jedna reč vezana za datu osobinu. Tako na primer imamo reči koji nedvosmisleno opisuju stanje jedne osobe, a imamo i one koje u zavisnosti od konteksta mogu da imaju različita značenja.

Kriterijumi za razvrstavanje u ove tri grupe su: preciznost (koliko precizno reč objašnjava osobinu, ili osećanje), kontekstualnost (da li reč jednosmisleno objašnjava osobinu, ili može da se koristi i u drugim kontekstima) i frekvencija korišćenja (koliko se često neka reč koristi u svakodnevnom govoru).

Grupama su dodeljene težine 1, 0.6 i 0.3 (-1, -0.6 i -0.3 za antonime). Reči u grupi jedan su obično u nedvosmislenoj vezi sa datom sobinom, i precizno je opisuju. To su u slučaju depresivnosti: *depressed, depression, sad, sadness, unhappy, melancholy, melancholic.* U grupi dva su reči koje precizno opisuju osobinu, ali mogu da se koriste u različitim kontekstima (*lonely, pessimistic, weeping, weep, moody, heartbroken* za depresivnost). U poslednjoj grupi se nalaze reči koje ili neprecizno i višesmisleno opisuju osobinu, ili osećanje, ili se u svakodnevnom govoru veoma često koriste (pa bi moglo doći do pogrešnih rezultata). To su za gornji slučaj: *hurt, ripped, empty, insignificant, useless, abandoned, guilty, quit.* Isto ovo, urađeno je i za sve antonime.

Za format rečnika uzet je JSON format fajlova i jedan od ulaza u rečniku, tj. jedna od osobina izgleda ovako:

{

"name":"anxiaty",

"weight":-0.08,

"lists":[

{"w":1, "words":["anxiety", "anxiousness","apprehensive", "concerned", "concerns", "fearful", "fear", "worried", "worry", "tense", "fright", "frighten", "frightened"]},

{"w":0.6, "words":["anxious", "afraid", "distressed", "fidgety", "scared", "uneasy", "unease", "uptight", "antsy"]},

{"w":0.3, "words":["jittery", "nervous", "restless"]},

{"w":-1, "words":["unanxious", "unanxiousness", "unfearful", "unafraid", "unconcerned", "unworried", "unfrightened", "fearless"]},

{"w":-0.6, "words":["courageous", "confident", "calm", "unscared", "dauntless", "undaunted", "untroubled"]},

{"w":-0.3, "words":["unshakable"]}

]

}

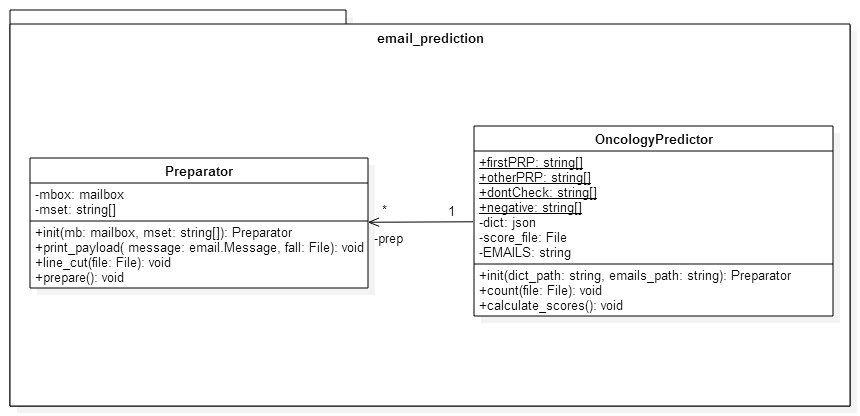
Što se tiče značenja polja, *weight* predstavlja ukupnu težinu/značaj ove osobine u riziku oboljenja od raka (uzeto sa slike 6), i to pozitivne vrednosti idu u prilog rizičnoj grupi ljudi, a negativne u prilog nerizičnoj grupi. Sledi polje lists koje predstavlja niz grupa reči, od kojih svaka ima svoju težinu *w,* i listu reči *words*.

Ceo rečnik je spakovan u jednom fajlu (*dictionary.json*). Ukoliko se želi ubaciti podrška za neki drugi jezik, potrebno je samo napraviti još jedan json fajl koji će imati prikazanu strukturu.

## 3.5. Implementacija Tekstualnog Analizatora

Tekstualni analizator je implementiran u Python-u, kao Bajesov analizator (broje se pojavljivanja elemenata iz rečnika i računa se suma na osnovu njihovih težina). Neke od dodatnih mogućnosti ovog analizatora su: detekcija negacije, detekcija subjekta i vizualizacija rešenja. Još jedna komponenta celog sistema je program koji vrši pripremu ulaza, pre nego što on uđe u analizator.

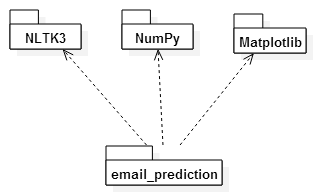
UML dijagram sistema i klasa koje ga čine je dat u nastavku.



*Slika 8: UML dijagram klasa koje čine tekstualni analizator*

Što se tiče eksternih biblioteka za implementaciju tekstualnog analizatora koriste se:

1. NLTK 3 (Natural Language Toolkit) - biblioteka za Python programe koji rade sa ljudskim govorom i svakodnevnim jezikom
2. NumPy - biblioteka za matematička izračunavanja
3. Matplotlib - biblioteka za grafički prikaz rezultata



*Slika 9: Eksterne biblioteke korišćene u analizatoru*

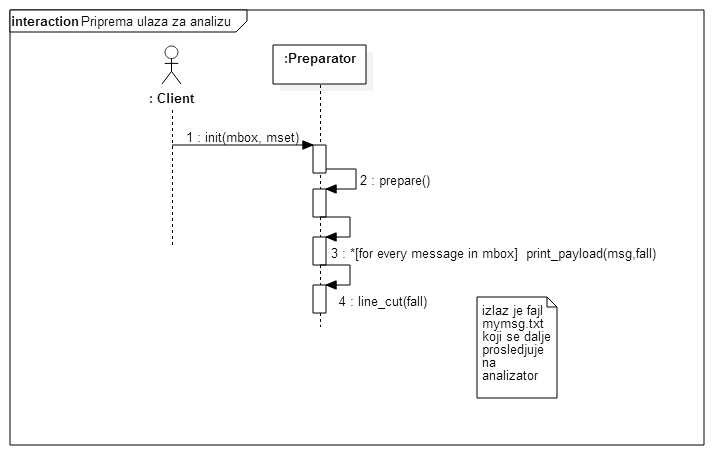
## 3.6. Priprema ulaza za tekstualnu analizu

Pripremu ulaza vrši *Preparator* klasa, i to tako što se inicijalizuje sa mejl sandučetom koji je u formatu .mbox i sa skupom *mset* koji sadrži sve email adrese koje će se posmatrati.

Mbox file format je format za skladištenje kolekcija emejl poruka. Sve poruke se nalaze u jednom velikom tekstualno fajlu, jedna za drugom. Početak svake poruke označen je frazom “From: “ i emejl adresom pošiljaoca. Nakon toga slede razne meta informacije i telo poruke. Poruka se završava znakom za novi red.

Iz .mbox fajla se izvlače samo poruke koje su interesantne za tekstualnu analizu, tj. poruke poslate sa adresa iz skupa *mset*. Ovo sve radi metoda *prepare()* koja u petlji prolazi kroz sve poruke sandučeta i za svaku poruku koja je poslata sa adrese iz skupa *mset* štampa njeno telo poruke (metoda *print\_payload*) u tekstualni fajl. Na kraju se poziva i metoda *line\_cut* koja eliminiše sve duplikate poruka u okviru konverzacija (na primer Gmail ima mogućnost da u okviru poruke čuva i ranije poruke u okviru te konverzacije, tj. poruke koje su predhodile datoj).

Na sledećem dijagramu je prikazana cela interakcija:



*Slika 10: Dijagram sekvence pripreme ulaza za analizu*

## 3.7. Tekstualna analiza

Kao što je ranije rečeno ovde se vrši prebrojavanje elemenata iz teksta koji se nalaze u rečniku, i računa se njihova suma. Nekoliko parametara je još uključeno, a to su:

1. Kontekst - da li se nađena reč u tekstu odnosi da osobu koju analiziramo, ili na nekog drugog, tj. na koji subjekat se reč odnosi. Na primer, ukoliko neko u mejlu opisuje kako se oseća uplašeno, da bi predikcija bila ispravno urađena, bitno je da znamo da li se nađena ključna reč odnosi na subjekat ispitivanja ili nekog drugog. Rečenice: “Ja se osećam uplašeno” i “Izgledao mi je uplašeno” obe sadrže ključnu reč uplašeno, međutim zaključci do kojih dolazimo su potpuno različiti. Zato je bitno, brojiti samo one elemente koji se odnose na ispitivani subjekat.

Ovo je u implementiranom analizatoru izvedeno tako što se kad god se naiđe na ličnu

zamenicu u tekstu (lične zamenice u engleskom su: I, me, my, you, her, she, his, itd.)

proverava se da li se ona nalazi u skupu *firstPRP* (lične zamenice koje se odnose na

prvo lice) ili *otherPRP* (sve ostale), i na osnovu ovoga se fleg *pronoun* postavlja na 1 ili 0

respektivno. Ovaj fleg važi do kraja rečenice, i u zavisnosti od njega zavisi da li se

pronađena ključna reč uračunava u sumu ili ne.

1. Težina - svaka nađena ključna reč ima svoju težinsku vrednost (o ovome je bilo reči o odeljku vezanom za određivanje rečnika), koja se može naći u rečniku. Svaki pronađeni element se množi svojom težinom grupe i težinom samog elementa (*weight\*w*).
2. Negacija - ukoliko se ključna reč u rečenici pojavljuje u negativnom kontekstu, potrebno je njenu ukupnu teženu još pomnožiti sa minus jedan.

Negacija u engleskom jeziku najčešće predstavlja pojavljivanje reči kao što su: not,

don’t, doesn’t, nobody, never, hardly, itd. ispred glagola. Algoritam koji je u

implementiranom tekstualnom analizatoru korišćen je sledeći: ako se ispred pronađene

ključne reči iz rečnika nalazi neka od reči iz skupa *negative* (elementi koji nagoveštavaju

negaciju), i to na razdaljini koja je manja od maksimalne distance (u ovom slučaju ta

distanca je tri reči) onda ključna reč ulazi u sumu pomnožena sa minus jedan.

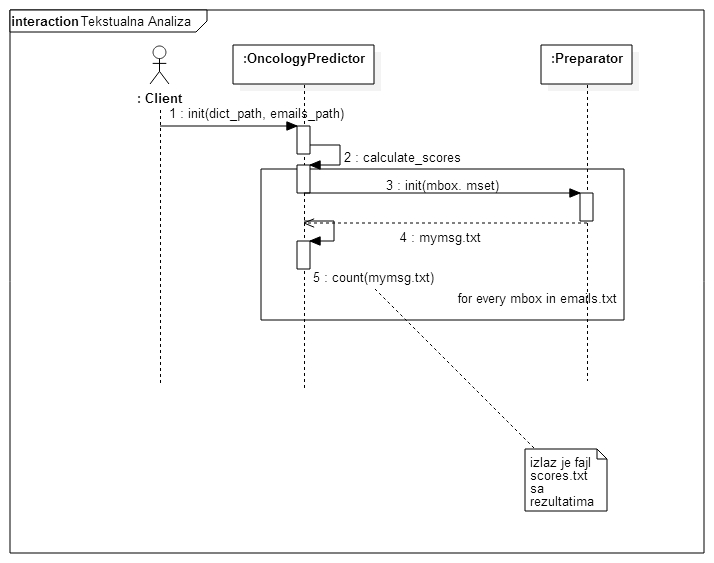
Osnovni princip rada ovog analizatora je:

1. Inicijalizuje se sa putanjom do rečnika i putanjom to *emails.txt* fajla, koji predstavlja spisak svih sandučića koji se analiziraju, zajedno sa pridruženim mejl adresama od interesa.
2. Metoda *calcutate\_scores()* u petlji za svako sanduče i skup mejl adresa (koji se nalaze u fajlu *emails.txt*), instancira prvo objekat klase *Preparator,* koji vrši pripremu ulaza (opisano gore - poziva se metoda *prepare()*), i nakon toga poziva svoju metodu *count()* koja vrši prebrojavanje i analizu.
3. Metoda *count()* za ulaz ima pripremljen tekstualni fajl *“mymsg.txt”* koji je izlaz *Preparator-a*. U petlji se svaka rečenica tokenizuje i proverava se da li se token nalazi i u rečniku, i to uz gore navedena ograničenja. Kako bi se optimizovala pretraga, neke od najčešće korišćenih reči u engleskom jeziku se ne pretražuju u rečniku (a koje nisu od značaja za problem). To su reči iz skupa *dontCheck,* i na primer to su: the, of, in, that, for, with, on, itd. Ukoliko je token pronađen u rečniku, uzima se njegova težina (množi se sa -1 ako je fleg za negaciju 1) i sabira se sa ukupnom sumom i brojač za datu ključnu reč se inkrementira. Kada se prođe kroz ceo ulazni fajl rezultat se upisuje u izlazni fajl *scores.txt* zajedno sa brojačima pronađenih ključnih reči.

Izlaz za jedno sanduče je težinska suma svih pronađenih ključnih reči, i ona može biti pozitivna ili negativna. Pozitivna vrednost označava da osoba čiji su mejlovi bili analizirani spada u rizičnu grupu za oboljevanje od karcinoma. Nasuprot tome, negativna vrednost označava da se osoba nalazi u nerizičnoj grupi. U zavisnosti od toga kolika je ta vrednost, osoba može biti pod manjim ili većim rizikom/nerizikom.

Izlaz je ostavljen kao brojčana vrednost iz razloga što ovaj isti sistem može biti korišćen i za razne druge probleme, te tumačenje ove izlazme sume može biti različito. Jedino što treba uraditi, kako bi se ovaj sistem primenio na neki drugi problem, jeste promena rečnika. U zavisnosti od konteksta, potrebno je definisati ključne reči i njihove težinske vrednosti (slično kao što je urađeno za ovaj problem).

Dijagram interakcije tekstualnog analizatora dat je u nastavku:



*Slika 11: Dijagram sekvence analize teksta*

## 

## 3.8. Struktura ulaza i izlaza

Od ulaznih fajlova potrebno je napraviti *emails.txt,* u kome se nalaze lista mejl sandučića kojima je pridružen skup emejl adresa od interesa. Format ovog fajla je: za svako sanduče za koje se vrši analiza potrebno je imati liniju koja počinje sa “MBOX file: “ i putanju do .mbox fajla, kao i spisak emejl adresa vezanih za to sanduče, a koje su od interesa (jedna adresa po redu). Ovi fajlovi se mogu staviti u folder *mailboxes* kao na slici 13, ali i ne moraju (onda mora da se prosledi putanja do njih kroz komandnu liniju pri pokretanju programa). Primer jednog *emails.txt* sledi:

*emails.txt*

dwMBOX file: /documents/predictor/mailboxes/shankman-j.mbox

a..shankman@enron.com

jeffrey.shankman@enron.com

MBOX file: /documents/predictor/mailboxes/south-s.mbox

steven.south@enron.com

MBOX file: /documents/predictor/mailboxes/dorland-c.mbox

chris.dorland@enron.com

MBOX file: /documents/predictor/mailboxes/kaminski-v.mbox

j.kaminski@enron.com

vince.kaminski@enron.com

j..kaminski@enron.com

kaminski@enron.com

MBOX file: /documents/predictor/mailboxes/crandell-s.mbox

sean.crandall@enron.com

MBOX file: /documents/predictor/mailboxes/fossum-d.mbox

drew.fosskaminski@enron.comum@enron.com

MBOX file: /documents/predictor/mailboxes/sanchez-m.mbox

monique.sanchez@enron.com

MBOX file: /documents/predictor/mailboxes/ward-k.mbox

s..ward@enron.com

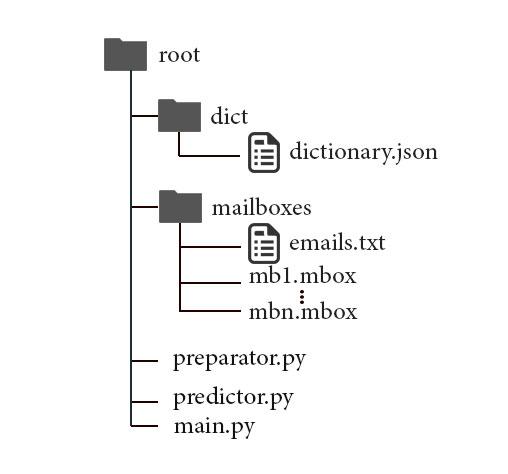
kim.ward@enron.com

MBOX file: /documents/predictor/mailboxes/rapp-b.mbox

bill.rapp@enron.com

Slika 12: Format ulaznog fajla emails.txt

Format rečnika *dictionary.json* je već ranije pokazan u odeljku “Određivanje rečnika”.



*Slika 13: Struktura programskog direktorijuma*

Izlaz iz sistema je fajl *score.txt* koji će biti napravljen u korenom direktorijumu. Format fajla *score.txt* je: za svako analizirano sanduče ispisuje se spisak reči koje su pronađene u mejlovima, zajedno sa brojem pojavljivanja (reči koje su pronađene kao negacije označene su sa ! ispred ključne reči). U poslednjem redu ispisana je ukupna suma za dato sanduče. Primer jednog *score.txt* fajla:

==========================================================

Result for: shankman-j.mbox

Counter({'great': 101, 'excited': 68, 'happy': 66, 'appreciate': 59, 'fun': 50, 'pleased': 45, 'interested': 45, 'love': 38, 'control': 36, 'able': 32, 'strategy': 28, 'focus': 27, 'unable': 26, 'fact': 19, 'reason': 18, 'pressure': 17, 'cool': 15, 'assume': 15, 'confident': 14, 'glad': 14, 'crack': 13, 'prepared': 13, 'responsible': 13, 'attractive': 12, 'concerned': 12, 'responsibility': 11, 'respect': 11, 'care': 11, 'reserved': 11, 'vision': 10, 'enjoyed': 10, 'nice': 9, '!able': 8, 'assist': 8, 'grace': 8, 'sad': 8, 'focused': 7, '!worry': 7, 'balance': 7, 'handle': 6, 'kill': 6, 'planned': 6, 'concerns': 6, 'laughing': 6, 'guilty': 6, 'mood': 6, 'afraid': 5, 'consequence': 5, 'tired': 5, 'yield': 5, 'awesome': 4, 'weak': 4, 'delighted': 4, '!blame': 4, 'feeling': 4, 'passion': 4, 'exposed': 4, 'pleasure': 4, 'idiot': 4, 'charge': 4, 'alone': 3, 'calm': 3, 'intention': 3, 'concentrate': 3, 'fear': 3, 'mistake': 3, 'dream': 3, '!love': 3, 'content': 3, 'collaboration': 3, 'annoyed': 2, 'consequences': 2, 'reality': 2, 'cooperation': 2, 'distressed': 2, 'excitement': 2, 'charity': 2, 'facts': 2, 'oversight': 2, 'silence': 2, 'qualified': 2, 'generous': 2, 'ease': 2, 'competent': 2, 'peace': 2, 'nervous': 2, 'accountable': 2, '!interested': 2, 'shame': 2, 'collapse': 2, 'neglected': 2, 'blame': 2, 'worried': 2, 'abide': 2, 'authority': 2, 'careful': 1, '!alone': 1, 'cooperate': 1, 'preparation': 1, 'vivid': 1, 'eager': 1, 'accused': 1, 'bored': 1, '!control': 1, '!capable': 1, 'unhappy': 1, 'sympathy': 1, 'envision': 1, '!determined': 1, 'argue': 1, '!stable': 1, 'resist': 1, 'anxious': 1, 'honor': 1, 'empty': 1, '!happy': 1, 'error': 1, 'reflect': 1, 'insignificant': 1, 'suitable': 1, 'thrilled': 1, 'forgive': 1, 'fight': 1, '!prepared': 1, 'furious': 1, '!concerned': 1, 'hurt': 1, '!content': 1, 'concentration': 1, '!reason': 1, 'controlled': 1, 'powerful': 1, '!great': 1, 'determined': 1, 'imagine': 1, 'condemn': 1, 'quit': 1, 'worry': 1, 'satisfied': 1, '!worried': 1, 'threatened': 1})

-10.62299999999995

==========================================================

Result for: south-s.mbox

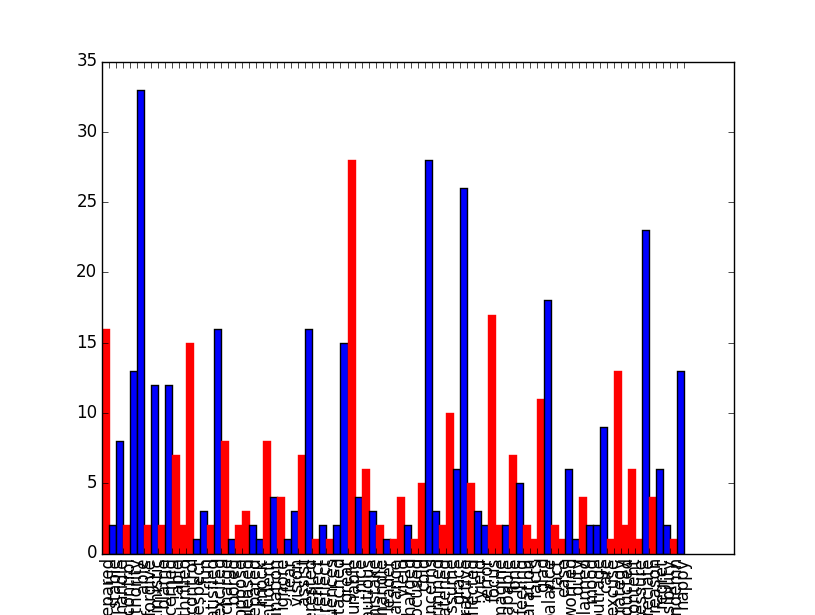
Counter({'control': 6, 'fact': 2, '!able': 2})

-0.024000000000000063

==========================================================

Slika 14: Izlazni fajl score.txt

Takođe, moguće je i vizuelno prikazati podatke.



*Slika 15: Vizuelni prikaz rezultata za jedno analizirano sanduče*

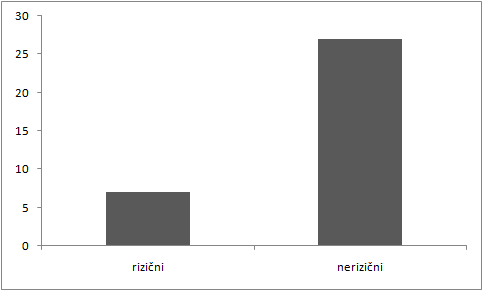
Na x osi nalaze se pronađene ključne reči, a na y broj pojavljivanja u tekstu. Plava boja označava ključne reči koje doprinose profilisanju osobe u nerizičnu grupu, a crvena u rizičnu.

# **4. Analiza rezultata**

Ulazni skup podataka je Enron Corpus, baza preko 600 hiljada mejlova zaposlenih u ovoj kompaniji. Ova kolekcija emejlova je jedna od jedinih, i svakako najvećih baza pravih emejlova, te su i mnogo istraživanja i radovi, koji imaju veze sa procesiranjem jezika, društvenim mrežama, itd. rađeni nad ovim skupom podataka.

Konkretno, uzeto je 34 mejl sandučića iz ove baze, koji su bili pušteni kroz sistem. Posmatrali su se samo sandučići poslatih poruka (Sent) koji su bili vezivani za osobu koja se analizirala. Izlaz sistema se tumačio na dva načina: (1) Negativni rezultat je označavao da se osoba nema rizik da oboli od raka, a pozitivni da se osoba nalazi u rizičnoj grupi. Ovde nije uzeto u obzir gde se na skali nalazi ovo rešenji, niti koliko je uopšte ključnih reči pronađeno. (2) Dodata je još jedna grupa ljudi čiji su rezultati u okolini nule, ili je broj pronađenih ključnih reči mali, pa se ne može doneti validan zaključak. Rezultati koji su dobijeni su sledeći:

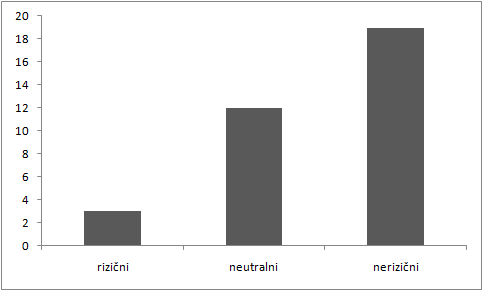
1. Od 34 osobe, 7 se nalazilo u rizičnoj grupi, a 27 u nerizičnoj grupi.



*Slika 16: Rezultati prema prvom testu*

U procentima, ovo znači da oko 20% ovih ljudi ima veći rizik da oboli od raka.

1. Prema drugom modelu tumačenja rezultata, imamo tri grupe ljudi: rizične, nerizične i one koji su neutralni ili je bilo nemoguće proceniti visinu njihovog rizika za oboljenje od karcinoma. U ovom slučaju, od 34 osobe, 19 je bilo nerizičnih, 3 rizičnih i 12 neutralnih.

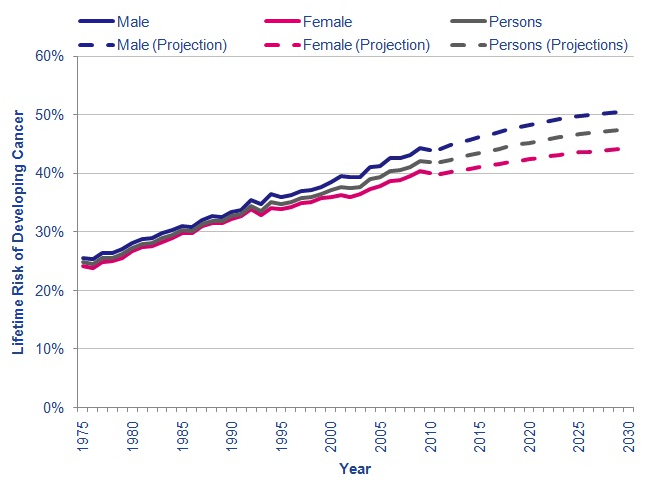


*Slika 17: Rezultati prema drugom modelu*

Ovo znači da 55% ovih nema rizik da oboli od karcinoma, dok je ostatak (45%) ili pod rizikom, ili su u neutralnoj grupi.

## 4.1. Poređenje sa postojećim istraživanjima

Kako bi se ovi rezultati uporedili, za primer je uzeto Britansko istraživanje o riziku ljudi od dobijanja raka. Mnogi od parametara korišćenih u ovom istraživanju, nisu uzeti u obzir, kao što su godine, pol, različite vrste kancera i drugi. Rezultati ovog istraživanja govore da otprilike 43% ljudi oboli od neke vrste raka, tj. otprilike jedna u dve osobe. Ovo je najviše poklapa sa rezultatima dobijenim prema drugom modelu, jer od svih ljudi 55% nije imalo rizik od oboljevanja, što ostavlja 43% osoba koje su u druge dve grupe.



*Slika 18: Britansko istraživanje o riziku dobijanja raka*

Slični rezultati dobijeni su i u istraživanju Američke udruženja za maligne bolesti (American Cancer Society). Prema njemu 43% ljudi ima povećan rizik da oboli od bilo koje vrste raka, dok 22% ima veći rizik da od karcinoma i umre [5].

# 

# **5. Zaključak**

Bez pogovora se može reći da reči, fraze i jezički sklopovi koje ljudi koriste u njihovom svakodnevnom životu, stoje u jakoj vezi sa unutrašnjim psihološkim stanjem te osobe, i odražavaju njihove misli, emocije, namere i motivacije, a kako ove stavke predstavljaju srž ljudskog bića i sa mnogim drugim procesima koji se odvijaju u nama. Sa ovim u vezi, dolazi se do zaključka da je kompjuterska tekstualna analiza, data mining i njene primene u detekciji emocija, misli i stavova iz teksta, jedna od gorućih tema današnjice sa ogromnim značajem i beskrajnim primenama.

U ovom radu dat je jedan primer sistema koji na osnovu analize e-mejl sandučića, a u opštijem slučaju teksta uopšte, vrši onkološke predikcije. Mnoga istraživanja su pokazala, ono što se od davnina sumnjalo, a to je da unutrašnje stanje osobe, njene emocije i reakcije mogu da budu veliki faktor rizika za oboljevanje. Ovaj rad se zasniva na jednom od takvih istraživanja, koje pokazuje da postoje određeni rizični i nerizični tipovi ličnosti, koji su skloniji da obole od karcinoma ili su u neku ruku zaštićeni od istog.

Sistem za sada radi jednostavno prebrojavanje ključnih reči iz rečnika koje se traže u tekstu, i na osnovu konaćnog skora profiliše osobu u rizičnu ili nerizičnu grupu. Rečnik je skup ključnih reči kojima su dodeljene određene težine na osnovu toga koliko jak “signal” predstavljaja da je osoba jedan ili drugi tip ličnosti.

Ono što je sledeći korak, i unapređenje ovog sistema, je dodavanje prepoznavanja skupova reči, tj. fraza, jer jedna reč može imati različita značenja u zavisnosti od konteksta u kome se nalazi. Takođe, kako bi predikcije bile tačnije, potrebno je ubaciti još mnoge druge faktore, kao što su: pol osobe, njena lokacija, zanimanje, godine, jer oni predstavljaju veoma važne informacije za predikciju pojave karcinoma. Pored ovoga, potrebno je dodati prepoznavanje različitih formta jedne iste reči, takozvano *stemm-ovanje*. Na primer u engleskom jeziku, reči *connection, connections, connecting* bi bile prepoznate kao jedna ista reč, bez potrebe da se u rečnik dodaje svaka od gore navedenih. Naravno, i sama struktura izlaza bi trebala da se finije podeli na nekoliko klasa rizika. Kako bi se ovaj sistem podigao na još veći nivo, bilo bi potrebno ubaciti neki algoritam mašinskog učenja, gde bi program sam, na osnovu ulaznih trening podataka, naučio sam da profiliše ljude u rizične grupe na osnovu teksta.

Svi dobijeni rezultati bi trebali biti nekako provereni. To bi najbolje bilo uraditi praćenjem jedne grupe ljudi tokom nekog perioda. Prvo bi se njihovi e-mejl sandučići pustili kroz onkološki prediktor, koji bi ih profilisao u jednu ili drugu grupu. Nakon toga, trebalo bi videti da li će neko od njih stvarno oboleti od karcinoma u budućnosti. Još jedan, možda jednostavniji način, bio bi da se uzme grupa ljudi koje već imaju karcinom i da se njihovi e-mejl sandučići puste kroz sistem, i da se vidi u kojoj meri se rezultati prediktora poklapaju sa istinom.

Potrebno je i pronaći druge primene ovog sistema, jer je njegova struktura i implementacija takva da može da se koristi i za druge probleme. Jedina izmena koja se treba načiniti, jeste izmena rečnika, kao i možda implementacija programa za pripremu podataka, ukoliko se radi sa nekim drugim formatima. Tako bi se na primer, mogli analizirati blogovi i društvene mreže sa ciljem da se odredi koje osobe imaju rizik da obole od bolesti X, ili koje osobe su spremne da kupe proizvod Y, na osnovu jezika koji one koriste u svom izražavanju.

Tehnologija i inovacije u oblasti tekst analize i data mininga, postaju ozbiljna konkurencija tradicionalnim sociološkim i psihološkim metodama. Za vreme koje je potrebno da se ovim metodama izanalizira jedna osoba, kompjuterski je moguće proći kroz hiljade dokumenata, transkripta i drugih formi teksta i uraditi detaljnu analizu. Internet nam omogućava ogromnu bazu podataka i beskrajno skriveno znanje i informacije uskladištene na njemu, pa nije ni čudno što stojimo na pragu jedne nove tehnološke ere, ere data mininga, sa posebnim akcentom na tekstualnoj analizi, i kompjuterskoj analizi svakodnevnog jezika.

# **6. LITERATURA**

[1] Frans Coenen, Data mining: past, present and future. *The Knowledge Engineering Review*, Vol. 26:1, 25–29. & *Cambridge University Press*, 2011 doi:10.1017/S0269888910000378

[2] Jiban K Pal, Usefulness and applications of data mining in extracting information from different perspectives. *Annals of Library and Information Studies*, Vol. 58, March 2011, pp. 7-16

[3] Yla R. Tausczik, James W. Pennebaker, *The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology* 29(1) 24–54, 2010 SAGE Publications doi: 10.1177/0261927X09351676

[4] Grossarth-Maticek, Ronald, Bastiaans, Jan and Kanazir, Dusan, *Psychosocial Factors as Strong Predictors of Mortality from Cancer, Ischaemic Heart Disease and Stroke: The Yugoslav Prospective study*. *Journal of Psychosomatic Research* 29(2):167- 176, 1985

[5] SEER Cancer Statistics Review 1975-2011, *National Cancer Institute*

1. Danas se KDD ne odnosi samo na baze podataka, već na podatke uopšte (Knowledge Discovery in Data) [↑](#footnote-ref-2)
2. Zaključci se donose na osnovu prošlosti, a osobe koje su bile predmet istraživanja se dalje ne prate [↑](#footnote-ref-3)
3. Predmeti istraživanja se prate tokom određenog vremenskog perioda, i onda se donose zaključci [↑](#footnote-ref-4)