Analiza tehnika traženja proteinskih motiva

Đurić, Antonija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Antonija Đurić

ANALIZA TEHNIKA TRAŽENJA PROTEINSKIH MOTIVA

Diplomski rad

Voditelj rada:
doc. dr.sc. Pavle Goldstein

Zagreb, veljača 2018.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem roditeljima i bratu na podršci i ljubavi koju su mi pružili tijekom studiranja. Hvala mentoru doc.dr.sc. Palvu Goldsteinu na savjetima, strpljenju i pomoći pri izradi diplomskog rada.

Sadržaj

Uvod

U današnje vrijeme postoji puno rječnika gdje su zapisane riječi koje se koriste i pomoću tih riječi možemo razlučiti o čemu se u pojedinom tekstu govori. U tekstovima o biologiji javljat će se više bioloških pojmova, dok će u matematici zastupljenije biti riječi kao što su *teorem*, *algebra*, *aritmetika*. Gledajući biološke nizove kao što su proteini ili nizovi DNA, kod kojih ne postoje rječnici, teško je "vidjeti" o čemu govore pa nas je zanimalo kako bi bilo analizirati i klasificirati tekstove za koje nemamo nikakve rječnike. Tu dolazimo do teme i cilja ovog diplomskog rada.

Biološki nizovi su nizovi bez separatora u odgovarajućem biološkom alfabetu. Kod bioloških nizova imamo malo informacija, ne znamo o čemu govore pa bi teško bilo analizirati krajnje rezultate. Kao što smo spomenuli, jedan primjer biološkog niza su proteini. Proteini su nizovi amino-kiselina koje karakterizira specifičan supstitucijski uzorak. Umjesto analize takvih nizova, mi ćemo analizirati tekstove u prirodnom jeziku zbog toga što kod njih znamo što očekivati pa će lakše biti protumačiti rezultate. Kako bi tekstovi u prirodnom jeziku što više sličili nizovima proteina, uklonit ćemo sve separatore i brojeve te će ostati samo nizovi slova. Jedina informacija koju ćemo imati o tekstovima u prirodnom jeziku jest kojoj klasi pripadaju.

Ovaj diplomski rad podijeljen je u 5 poglavlja. U prvom poglavlju ukratko ćemo spomenuti i objasniti pojmove iz vjerojatnosti i statistike koje ćemo koristiti kasnije u radu. Drugo poglavlje pobliže će nam objasniti što je *klasifikacija* i kako se provodi. U trećem poglavlju opisat ćemo način na koji smo došli do traženih motiva, odnosno do našeg željenog rječnika. Objasnit ćemo sam postupak iteracije kojim dobijemo motive te definirati na koji način motivu pridružimo njegovu vrijednost. U četvrtom poglavlju uspoređivati ćemo dobivene profile. Pokušat ćemo vidjeti koliko dobro naši profili mogu odrediti kojeg je sadržaja određeni tekst. Naposljetku, u zadnjem poglavlju iznijet ćemo sve kriterije koje smo koristili pri odabiru motiva te analizirati dobivene rezultate.

Poglavlje 1

Pojmovi iz vjerojatnosti i statistike

Ovo poglavlje bavit će se definicijama pojmova koji će nam se javljati kasnije u samom diplomskom radu. U odjeljka (1.1) definirat ćemo pojmove vezane uz vjerojatnosti, a oni su preuzeti iz izvora [4].

1.1 Vjerojatnost

Definicija 1.1.1. *Pod **slučajnim pokusom** podrazumijevamo takav pokus čiji **ishodi**, odnosno **rezultati** nisu jednoznačno određeni uvjetima u kojima izvodimo pokus. Rezultate slučajnog pokusa nazivamo **događajima**.*

Definicija 1.1.2. *Neka je A događaj vezan uz neki slučajni pokus. Pretpostavimo da smo taj pokus ponovili n puta i da se u tih n ponavljanja događaj A pojavio točno n_A puta. Tada broj n_A zovemo **frekvencija** događaja A, a broj $\frac{n_A}{n}$ **relativna frekvencija** događaja A.*

Definicija 1.1.3. *Osnovni objekt u teoriji vjerojatnosti je neprazan skup Ω koji zovemo **prostor elementarnih događaja** i koji reprezentira skup svih ishoda slučajnih pokusa. Točke ω iz skupa Ω zvat ćemo **elementarni događaji**.*

Definicija 1.1.4. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) jest σ-**algebra skupova** (na Ω) ako je:*

(F1) $\emptyset \in \mathcal{F}$

(F2) $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$

(F3) $A_i \in \mathcal{F}, i \in \mathbb{N} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.1.5. *Neka je $\mathcal{F}$ σ-algebra na Ω. Uređen par $(\Omega, \mathcal{F})$ se zove **izmjeriv prostor**.*

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest **vjerojatnost** ako vrijedi:*

(P1) $\mathbb{P}(\Omega) = 1$ (normiranost vjerojatnosti)

(P2) $\mathbb{P}(A) \geqslant 0$, $A \in \mathcal{F}$ (nenegativnost vjerojatnosti)

(P3) $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ te $A_i \cap A_j = \emptyset$ za $i \neq j \Longrightarrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ (prebrojiva ili σ-aditivnost vjerojatnosti)

*Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ gdje je $\mathcal{F}$ σ-algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se **vjerojatnosni prostor**.*

*Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Elemente σ-algebre zovemo **događaji**, a broj $\mathbb{P}(A)$, $A \in \mathcal{F}$ se zove **vjerojatnost događaja** A.*

Definicija 1.1.7. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definiramo funkciju $\mathbb{P} : \mathcal{F} \to [0, 1]$ ovako:*

$$\mathbb{P}_A(B) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \ B \in \mathcal{F}$$

*Lako je provjeriti da je $\mathbb{P}_A$ vjerojatnost na $\mathcal{F}$ i nju zovemo **vjerojatnost od B uz uvjet A**.*

Definicija 1.1.8. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A_i \in \mathcal{F}$, $i \in I$ proizvoljna familija događaja. Kažemo da je to **familija nezavisnih događaja** ako za svaki konačan podskup različitih indeksa $i_1, i_2, \ldots, i_k$ vrijedi*

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j})$$

Označimo sa $\mathcal{B}$ σ-alegebru generiranu familijom svih otvorenih skupova na skupu realnih brojeva $\mathbb{R}$. $\mathcal{B}$ zovemo σ-**algebra skupova na** $\mathbb{R}$, a elemente σ-alegebre $\mathcal{B}$ zovemo **Borelovi skupovi**.

Definicija 1.1.9. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je **slučajna varijabla** (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, odnosno $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.1.10. *Neka je X slučajna varijabla na Ω. **Funkcija distribucije** od X je funkcija $F_X = F : \mathbb{R} \to [0, 1]$ definirana sa:*

$$F(x) = \mathbb{P}\{X \leqslant x\} = \mathbb{P}\{\omega : X(\omega) \leqslant x\}, \ x \in \mathbb{R}.$$

Definicija 1.1.11. *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i neka je F_X njezina funkcija distribucije. Kažemo da je X **apsolutno neprekidna** ili, kraće,*

neprekidna slučajna varijabla *ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ takva da je:*

$$F_X(x) = \int_{\infty}^{x} f(t)d\lambda(t),\ x \in \mathbb{R} \tag{1.1}$$

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.1) zove **funkcija gustoće vjerojatnosti od** X, to jest od njezine funkcije distribucije F_X ili, kraće, **gustoća od** X.

1.2 Primjeri slučajnih varijabli

Gama distribucija. Eksponencijalna distribucija

Neka je $\alpha > 0, \beta > 0$ i $\Gamma(x) = \int_0^{\infty} e^{-t}t^{x-1}dt,\ x > 0$ gama funkcija. Neprekidna slučajna varijabla ima **gama distribuciju** s parametrima α i β ako joj je funkcija gustoće f dana s:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}e^{\frac{-x}{\beta}}, & x > 0 \\ 0 & , x \leqslant 0 \end{cases} \tag{1.2}$$

Ako je $\alpha = 1$ i $\beta = \frac{1}{\lambda}$, tada kažemo da X ima **eksponencijalnu distribuciju** s parametrom λ. Funkcija gustoće eksponencijalne distribucije s parametrom λ je

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0 & , x \leqslant 0 \end{cases} \tag{1.3}$$

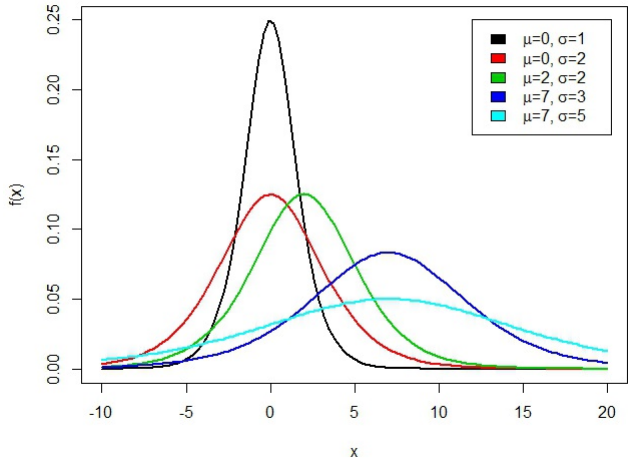
Logistička distribucija

Neka je $\mu, \beta \in \mathbb{R}, \beta > 0$. Neprekidna slučajna varijabla X ima **logističku distribuciju** s parametrima μ, β ako joj je funkcija gustoće dana s:

$$f(x) = \frac{e^{\frac{x-\mu}{\beta}}}{\beta\left(1 + e^{\frac{x-\mu}{\beta}}\right)^2},\ x \in \mathbb{R} \tag{1.4}$$

Neka je $p, q > 0$. Slučajna varijabla X ima **generaliziranu logističku distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{e^{pq}}{(1+e^{y})^{p+q}},\ x \in \mathbb{R} \tag{1.5}$$



Slika 1.1: Funkcije gustoće logističke distribucije s različitim parametrima

1.3 Teorija ekstremnih vrijednosti

Ovaj odjeljak bavit će se pojmovima koji će nam biti bitni kod analize distribucije maksimalnih ocjena. Pojmovi su detaljnije obrađeni u izvorima [2] i [5] odakle su i preuzeti.

Gumbelova distribucija

Neka su $\mu \in \mathbb{R}$ i $\beta > 0$. Neprekidna slučajna varijabla X ima **Gumbelovu distribuciju** sa parametrima μ i β ako joj je funkcija gustoće dana s:

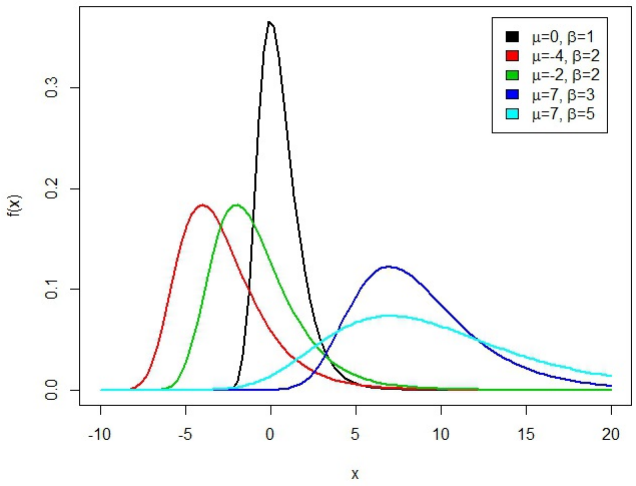
$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}}, \ x \in \mathbb{R}. \tag{1.6}$$

Neka je $p > 0$. Slučajna varijabla X ima **generaliziranu Gumbelovu distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \frac{1}{\Gamma(p)} e^{-px} e^{e^{-px}}, \ x \in \mathbb{R}. \tag{1.7}$$

Korolar 1.3.1. *Neka su X_1 i X_2 nezavisne generalizirane Gumbel distribuirane slučajne varijable s parametrima p i q, respektivno. Tada slučajna varijabla $Y = X_1 - X_2$ ima generaliziranu logističku distribuciju s parametrima p i q.*

Slika 1.1 i Slika 1.2 preuzete su iz izvora [2].



Slika 1.2: Funkcije gustoće Gumbelove distribucije s različitim parametrima

Fréchetova distribucija

Neka su $\alpha > 0, \beta > 0$ i $\mu \in \mathbb{R}$. Slučajna varijabla X ima **Fréchetovu distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{x-\mu}\right)^{\alpha+1} e^{-\left(\frac{\beta}{x-\mu}\right)^{\alpha}}, \ x \in \mathbb{R}. \tag{1.8}$$

Weibullova distribucija

Neka su $\alpha > 0$, $\beta > 0$. Slučajna varijabla X ima **Weibullovu distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \begin{cases} \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & x \geqslant 0 \\ 0 & , x < 0 \end{cases} \tag{1.9}$$

Teorem 1.3.2. *Neka su $X_1, X_2, \ldots, X_n$ jednako distribuirane slučajne varijable i neka je $M_n = max\{X_1, X_2, \ldots, X_n\}$. Ako postoji $a_n > 0$ i $b_n \in \mathbb{R}$ tako da je $\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leqslant x\right) = F(x)$, gdje je F nedegenerirana distribucija, tada granična distribucija F pripada Gumbelovoj, Fréchetovoj ili Weibullovoj distribuciji.*

1.4 Osjetljivost i specifičnost testa

U ovom odjeljku definirat ćemo pojmove osjetljivost i specifičnost testa. Ti pojmovi trebat će nam kasnije kod analize uspješnosti metode. Same definicije preuzete su iz izvora [2].

Osjetljivost testa (stopa stvarno pozitivnih) mjeri proporciju pozitivnih elemenata uzorka ispravno prepoznatih testom u odnosu na ukupni broj pozitivnih elemenata, dok specifičnost testa (stopa stvarno negativnih) mjeri proporciju negativnih elemenata uzorka ispravno prepoznatih testom u odnosu na ukupni broj negativnih.

$$\text{osjetljivost} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih} + \text{broj lažno negativnih}}$$

$$\text{specifičnost} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih} + \text{broj lažno pozitivnih}}$$

Osim same osjetljivosti i specifičnosti testa, korisno je definirati i veličine kojima opisujemo učinkovitost testa. Pozitivno prediktivna vrijednost (PPV) mjeri u kojem postotku pozitivno identificirani elementi zaista jesu stvarno pozitivni, s druge strane negativna prediktivna vrijednost (NPV) mjeri postotak negativno identificiranih elemenata koji zaista jesu negativni.

$$\text{pozitivna prediktivna vrijednost} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih} + \text{broj lažno pozitivnih}}$$

$$\text{negativna prediktivna vrijednost} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih} + \text{broj lažno negativnih}}$$

Rezultate često prikazujemo tablicom.

		Predviđeno stanje		
		pozitivno stanje	negativno stanje	
Stvarno stanje	pozitivno stanje	stvarno pozitivno (TP)	lažno negativno (FN)	osjetljivost
	negativno stanje	lažno pozitivno (FP)	stvarno negativno (TN)	specifičnost
		PPV	NPV	

Tablica 1.1: Veličine uspješnosti testa

Poglavlje 2

Klasifikacija

Klasifikacija dokumenata je problem koji se javlja svugdje, npr. u bibliotekarstvu, informatici i računarstvu. Zadatak je pridružiti dokument jednoj ili više postojećih klasa. Klasa je skup objekata koji imaju iste karakteristike. Dokumenti koje klasificiramo mogu biti tekstualni, slike, glazba, ... Svaka vrsta dokumenata ima svoje probleme klasifikacije. Dokumenti mogu biti klasificirani prema raznim atributima kao što su tema, vrsta datoteke, godine, autor. U ovom diplomskom radu bavit ćemo se klasifikacijom tekstualnih dokumenata po sadržaju.

Kada spomenemo pojam "tekst" odmah uz njega vežemo i pojam "rječnik" kao skup riječi od koji je sastavljen tekst. U slučaju kada imamo rječnik klasifikacija je standardni proces u kojem se obično koristi bag-of-words model ([1],[3]). U tom modelu, tekst se analizira kao vreća (eng. *bag*) riječi, zanemaruje se gramatika i poredak riječi. Time ćemo se detaljnije baviti u daljnjim poglavljima.

Tekstovi koje ćemo analizirati u ovom diplomskom radu biti će tekstovi bez separatora. Naime, biološki nizovi kao što su proteini i nizovi DNA, su nizovi bez separatora u odgovarajućem biološkom alfabetu. Kako nećemo imati separatora, to povlači činjenicu da nećemo imati ni rječnik. Da bi proveli sam postupak klasifikacije morat ćemo prvo napraviti vlastiti rječnik. Zbog toga ćemo proći tekstom kako bi pronašli česte stringove (nizove slova).

Ovdje ćemo definirati i dva pojma koja ćemo kasnije dosta koristiti. Prvi pojam koji ćemo definirati je *dokument*. Pod tim pojmom smatrat ćemo jedan red slova bez separatora. Cijeli skup *dokumenata* nazivat ćemo *kolekcijom*.

Poglavlje 3

Traženje motiva i računanje profila

3.1 Traženje motiva

Kako bi u ovom diplomskom radu mogli provesti klasifikaciju tekstova, bitno nam je doći do vlastitog rječnika, odnosno do nekog popisa stringova (niza slova) koje često pronalazimo u određenim tekstovima. Taj postupak zvat ćemo traženje motiva. Pod pojmom motiv podrazumijevamo niz slova određene duljine u tekstu koji dozvoljava specifične promjene. Prije nego analizom dođemo do motiva trebamo uzeti neki početni string, određene duljine. Takav string nazivat ćemo upitom.

Potrebno je upitom proći kroz svaki dokument i pronaći podnizove duljine m koji su najsličniji polaznom upitu duljine m. Tu ćemo još uvesti pojam ocjena sličnosti. U našem slučaju ocjena sličnosti će biti broj koji će nam govoriti na koliko mjesta se naš početni string, odnosno upit, poklapa sa svakim podnizom. Postoje razne definicije ocjena sličnosti, ali kako mi ne znamo što imamo u kolekcijama i na koji način da uspoređujemo upit s podnizovima odlučili smo se za najjednostavniju ocjenu sličnosti.

Prvo ćemo objasniti na koji način upitom prolazimo kroz dokument da bi pronašli najsličniji podniz. Neka je x dokument duljine n te je y upit duljine m tako da vrijedi $m < n$. Naš upit najprije usporedimo s prvim podnizom jednake duljine zatim se pomičemo za jedno mjesto u desno te ponovno uspoređujemo upit sa sljedećim podnizom. Takav postupak radimo dok ne dođemo do zadnjeg podniza u dokumentu, odnosno dok ne dođemo na mjesto $n - m + 1$, nakon toga prelazimo na sljedeći dokument i radimo istu usporedbu. Grafički to izgleda ovako:

$$
\begin{array}{cccccccccc}
x_1 & x_2 & x_3 & \dots & x_{m-1} & x_m & x_{m+1} & x_{m+2} & \dots & x_n \\
y_1 & y_2 & y_3 & \dots & y_{m-1} & y_m & & & &
\end{array}
$$

$$
\begin{array}{cccccccccc}
x_1 & x_2 & x_3 & x_4 & \dots & x_m & x_{m+1} & x_{m+2} & \dots & x_n \\
 & y_1 & y_2 & y_3 & \dots & y_{m-1} & y_m & & &
\end{array}
$$

$$
\begin{array}{cccccccccc}
 & & & \dots & & & & & & \\
 & & & & & \dots & & & & \\
 & & & & & & & \dots & &
\end{array}
$$

$$
\begin{array}{cccccccccc}
x_1 & x_2 & x_3 & x_4 & \dots & x_{n-m+1} & x_{n-m+2} & x_{n-m+3} & \dots & x_n \\
 & & & & & y_1 & y_2 & y_3 & \dots & y_m
\end{array}
$$

Dok tako prolazimo upitom kroz sve dokumente, istovremeno računamo i ocjenu sličnosti. Što će ocjena biti veća, to je poklapanje bolje. Primjerom ćemo pokazati na koji način uspoređujemo dva niza.

$$
\begin{array}{cccccccccc}
c & o & r & r & e & l & a & t & i & o \\
c & o & r & e & a & l & a & n & i & o
\end{array}
$$

Ovdje vidimo da imamo 7 preklapanja. Isti postupak radimo na cijeloj kolekciji. Dakle, prolazimo upitom po svakom dokumentu i računamo ocjene sličnosti za svaki podniz. Na kraju pamtimo maksimalnu ocjenu u svakom retku jer ćemo pomoću nje odrediti koji će nam podnizovi biti od važnosti.

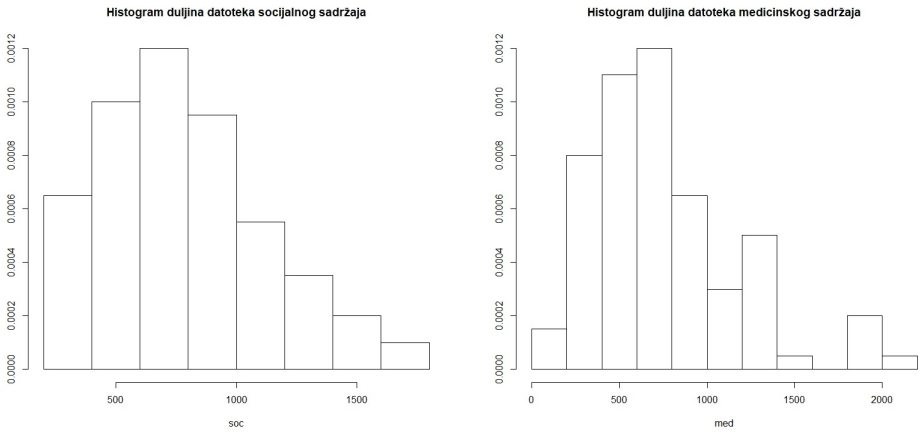
3.2 Značajnost ocjene sličnosti

Kada smo prošli kroz cijelu kolekciju i popisali sve maksimalne ocjene sličnosti, da bi dobili najbolje pogotke, trebamo pogledati distribuciju maksimalnih ocjena. Međutim, dovoljno je pogledati desni rep distribucije jer će se naši maksimumi nalaziti upravo u tom repu.

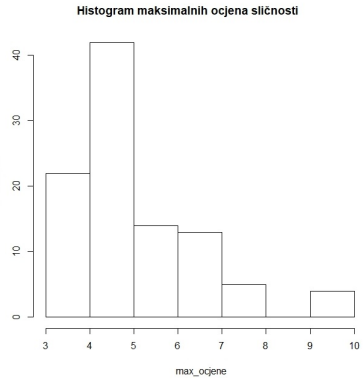
Ako pretpostavimo jednaku distribuiranost dokumenata, tada bi iz teorije ekstremnih vrijednosti slijedilo da maksimalne ocjene prate Gumbelovu distribuciju. Naši dokumenti, međutim, nisu jednake duljine pa ne možemo doći do takvog zaključka. Pogledamo li histograme duljina svih dokumenata u obje kolekcije, naslućujemo da bi oni mogli pratiti Gumbelovu distribuciju (Slika 3.1).

Prema Korolaru (1.3.1) imamo da razlika dvije Gumbel distribuirane slučajne varijable prati logističku distribuciju. Preostaje pogledati histogram maksimalnih ocjena (Slika 3.2).

Utvrdili smo da su maksimalne ocjene po nizovima logistički distribuirane pa treba vidjeti na koji način ćemo odlučiti koje su nam od tih ocjena statistički značajne. Dakle, kao



Slika 3.1: Histogrami duljina datoteka obje kolekcije



Slika 3.2: Histogram maksimalnih ocjena

što smo prije spomenuli, gledamo desni rep distribucije jer se u njemu nalaze maksimalne ocjene. Preostaje nam odrediti prag koji će nam određivati koji podniz, s pripadajućom ocjenom sličnosti, će se smatrati dobrim pogotkom. Zanimat će nas podnizovi s ocjenom sličnosti većom ili jednakom pragu. Sam prag definirat ćemo formulom

$$prag = \mu + skala \cdot \beta,$$

gdje je μ očekivanje neprekidne slučajne varijable s logističkom distribucijom, β parametar logističke distribucije, a *skala* proizvoljan pozitivan broj. Odabir dobre *skale* veoma je bitan. Ukoliko bi uzeli preveliku *skalu* kao rezultat dobili bi uzorke koji su slični sami

sebi, u suprotnom, ako bi uzeli prenisku *skalu* dobili bi velik broj uzoraka s jako malom međusobnom sličnosti. U ovom diplomskom radu koristili smo *skale* 7 i 9, ali to ćemo detaljnije opisati u poglavlju (5).

Nakon što smo izračunali prag, prolazimo cijelom kolekcijom i ispisujemo podnizove koji su imali ocjenu sličnosti veću ili jednaku od praga.

3.3 Iteracija

Postupak traženja motiva provodi se iterativnim putem. Najprije se odredi upit. U prvoj iteraciji on je jedan niz slova, a kasnije se može sastojati i od više nizova slova. Nakon što se odredi upit traže se podnizovi slični njemu, na način kako je to opisano u odjeljku (3.1). Najbolje podnizove određujemo pomoću praga, čiji način određivanja smo opisali u odjeljku (3.2), i njih spremamo kao nove upite. Nakon toga, kolekcija se ponovno pretražuje s novim upitom. Iterativni proces će stati kada nema promjene u listi pogodaka ili kada se postigne zadani broj iteracija.

Kada se završi cijeli proces iteracije za jedan početni upit, u dokumentu se pomičemo za jedno mjesto u desno, uzima se novi upit i ponovno se vrti ista iteracija. Postupak se ponavlja dok nismo iskoristili sve upite u polaznoj kolekciji.

3.4 Profili

Podnizove koje smo prihvatili kao relevantne zovemo *motivi*, a stohastički (vjerojatnosni) opis motiva zovemo *profil*. Profil je niz distribucija, po jedna za svaki stupac motiva. Ako se motiv sastoji od nizova duljine 10, profil tog motiva bit će onda zadan sa deset vjerojatnosnih distribucija. Za i-ti stupac motiva izračunat ćemo relativnu frekvenciju svih slova te ćemo profil definirati kao vektore $f_i = (f_{i1}, f_{i2}, \ldots, f_{i26)})$ za $i = 1, 2, \ldots, m$ gdje m duljina motiva te je f_{ij} vjerojatnost da se u i-tom stupcu motiva pojavi j-to slovo. Postoji mogućnost da dobijemo da je vjerojatnost nekog slova u stupcu jednaka 0. Kako bi izbjegli takav slučaj radimo sljedeće:

$$g_{ij} = \frac{f_{ij} + 0.01}{1.26},\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, 26 \tag{3.1}$$

gdje su $g_i = (g_{i1}, g_{i2}, \ldots, g_{i26})$ za $i = 1, 2, \ldots, m$, novi vektori distribucija. Preostaje nam definirati vektor $q = (q_1, q_2, \ldots, q_{26})$ gdje je q_i vjerojatnost pojavljivanja slova i u cijeloj kolekciji.

Sada nas zanima omjer vjerojatnosti iz (3.1) i vjerojatnosti pojavljivanja slova u cijeloj kolekciji (q), pa računamo *log-odds ratio*:

$$h_{ij} = \log\left(\frac{g_{ij}}{q_j}\right),\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, 26 \tag{3.2}$$

te vektori $h_i = (h_{i1}, h_{i2}, \ldots, h_{i26})$ za $i = 1, 2, \ldots, m$, predstavljaju tražene profile. *Log-odds ratio* radimo jer imamo malo informacija pa ćemo ovako bolje saznati koliko je prisutnost ili odsutnosti slova u motivu povezana s prisutnošću ili odsutnošću slova u cijeloj kolekciji. Drugim riječima, u formuli (3.2) mjerimo koliko je vjerojatnije da se dani niz slova iz kolekcije pojavio "iz profila" nego slučajno.

Poglavlje 4

Bag-of-profiles reprezentacija

U ovom poglavlju odgovorit ćemo na pitanje koje nas zanima od početka – možemo li bez poznavanja rječnika odrediti kojeg je sadržaja određeni tekst, to jest kojoj klasi taj tekst pripada.

U prethodnim poglavljima opisali smo načine na koje smo došli do motiva te na kraju i do profila. Sada preostaje vidjeti koliko će dobro ti profili opisati neke nove kolekcije. Pod pojmom opisati dokument smatramo da smo za taj dokument odredili bag-of-profiles reprezentaciju. Prvo trebamo izračunati ocjene po kojima ćemo uspoređivati koja kolekcija je bolje opisana. Ocjenu dokumenta iz kolekcije, u odnosu na profil, računamo tako da zbrajamo odgovarajuće vrijednosti matrice profila za svaku poziciju.

U sljedećem odjeljku matematički ćemo opisati kako dobijemo tražene ocjene.

4.1 Ocjene dokumenta u odnosu na profile

Sada ćemo opisati na koji način dobijemo bag-of-profiles reprezentaciju. Neka je $P = \{P_1, P_2, \ldots, P_m\}$ skup profila, $D = \{D_1, D_2, \ldots, D_r\}$ skup dokumenata te ćemo sa a_{ij} označiti ocjenu dokumenta D_i u odnosu na profil P_j. Svaki profil iz skupa P je također skup sačinjen od l vektora pa ćemo stoga definirati $P_j = \{p_{j0}, p_{j1}, \ldots, p_{j(l-1)}\}$ za $j = 1, 2, \ldots, m$. U našem će slučaju uvijek vrijediti da je $l = 10$.

Uzet ćemo sada profil P_j te neka je $x^{(k)} = x_k x_{k+1} \ldots x_{k+l-1}$ string na k-toj poziciji. Definiramo evaluaciju profila P_j na sljedeći način:

$$s_k = \sum_{h=0}^{l-1} \log \frac{\mathbb{P}(x_{k+h}|p_{jh})}{\mathbb{P}(x_{k+h}|q)}, \tag{4.1}$$

gdje je $k = 1, 2, .., n - l + 1$, $j = 1, 2, \ldots, m$, a n duljina dokumenta D_i za $i = 1, 2, \ldots, r$.

Ocjenu a_{ij} dobit ćemo tako da uzmemo najveću ocjenu dobivenu formulom (4.1), odnosno

$$a_{ij} = \max_{k=0,1,\ldots,n-l+1} s_k,\ i = 1, 2, \ldots, r,\ j = 1, 2, \ldots, m.$$

Na ovaj način izračunat ćemo težine svakog dokumenta u odnosu na profile. Zbirku dokumenata prikazujemo sljedećom matricom:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{21} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix},$$

gdje reci predstavljaju dokumente, a stupci predstavljaju profile.

Kako bi vidjeli koliko dobro dobivene težine opisuju kolekciju te kako bi mogli uspoređivati kolekcije međusobno, računamo sljedeće norme za svaki red matrice:

$$\|a_i\|_1 = |a_{i1}|+|a_{i2}|+\ldots+|a_{im}|,\ (l_1\ norma)$$
$$\|a_i\|_2 = \sqrt{a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2},\ (l_2\ norma)$$
$$\|a_i\|_\infty = \max\{|a_{i1}|, |a_{i2}|, \ldots, |a_{im}|\},\ (l_\infty\ norma).$$

Za analizu smo uzeli dvije testne kolekcije, također bez separatora. Jedna je kolekcija medicinske tematike, druga socijalne tematike. Obje kolekcije imaju po 50 dokumenata. Za svaku od tih kolekcija pokušali smo vidjeti koji ih profili najbolje opisuju. Radili smo usporednu analizu. Kolekciju medicinske tematike najprije smo opisali profilima dobivenim iz medicinske kolekcije, a zatim profilima socijalne kolekcije. Na isti način kolekciju socijalne tematike najprije smo opisali socijalnim profilima, a nakon toga medicinskim motivima. Za svaku kolekciju dobili smo po dvije bag-of-profiles reprezentacije. Kao što smo već spomenuli, da bi lakše uspoređivali koji profili bolje opisuju kolekcije, izračunali smo norme. Radi preciznije analize računali smo sve tri navedene norme, ali će u daljnjim tablicama biti prikazani samo rezultati dobiveni računanjem l_1 *norme*.

Sada ćemo analizirati rezultate dobivene opisivanjem medicinske kolekcije. Tu ćemo imati dva slučaja. Prvi slučaj je kada imamo podjednak broj motiva medicinske i socijalne tematike, a drugi slučaj će biti kada će nam se broj motiva jako razlikovati, ali imamo jednakost *skale*. Razloge zbog kojih nam se javljaju ova dva slučaja objasniti ćemo bolje u poglavlju (5).

Prvo smo testirali varijantu kada imamo podjednak broj motiva iz obje kolekcije. Dobili smo sljedeće rezultate koje možemo vidjeti u Tablici 4.1. U tablici smo za svaki dokument istaknuli koji ga profil bolje opisuje.

Medicinski profili	Socijalni profili	Medicinski profili	Socijalni profili
4207.3438527	1392.2400617	261.2953139	**995.3935884**
111.4519204	12.6449145	267.7725532	**2001.9222285**
859.2441300	**1780.1445204**	451.8624759	**1287.5986710**
175.7607768	**736.0149676**	**98.5028866**	46.5924143
402.7377790	**452.1147154**	52.4352925	**442.1695318**
165.2212318	**519.0600153**	**579.0122710**	189.4797137
634.5692119	**2457.5387155**	1159.6015719	**1467.5154981**
1150.5106590	256.3653111	1132.8585256	**1818.6856625**
1156.2123340	528.6630712	1217.3558461	**3104.0138015**
238.6809494	**453.1202559**	1694.7105296	**2423.1600626**
8.1912717	**1222.5401211**	891.8921391	**1521.9526390**
924.8446741	**1293.3316006**	**1840.2290624**	1457.9783177
4.3910407	**1006.5941673**	**1047.6939234**	94.8377411
454.1139825	**839.8865063**	**972.0035760**	507.4245281
581.3828215	**2537.5596499**	**1548.2768984**	961.0956288
712.8602375	139.3282804	520.0696461	**522.6172285**
414.2302594	**1523.0412606**	790.5266387	**2900.1023327**
816.6342508	**3067.1391450**	**1315.6235762**	1182.1614299
1435.8799498	599.3147737	1109.1395747	**1950.7042294**
1198.6854179	421.6962989	**641.4693034**	606.9550941
1429.8655110	**2101.7063230**	**1372.6018516**	1086.8932617
381.9756960	**585.2055948**	1155.7072145	**2093.1673010**
713.8888127	**1491.5291841**	**930.5907977**	867.8955062
147.1041128	**583.6495594**	**1101.5907936**	552.8051997
2033.8354436	10.4495394	**1676.4183786**	1364.6330642

Tablica 4.1: Medicinska kolekcija opisana medicinskim i nereduciranim socijalnim profilima

Gledajući rezultate nismo u mogućnosti odrediti kojoj bi klasi pripadala kolekcija. Otprilike oba skupa profila podjednako dobro opisuju kolekciju. To nas dovodi do zaključka da je problem u odabranim motivima. Da bi dobili podjednak broj motiva, morali smo spuštati *skalu* pri određivanju praga. Na taj način pokupili smo dio motiva koji nisu nužno vezani uz socijalnu temu pa dobro opisuju i medicinsku.

Kako vidimo da ne dobivamo željene rezultate, napravit ćemo sada analizu drugog slučaja. Analizirat ćemo medicinsku kolekciju opisanu medicinskim profilima i socijalnim profilima, ali uz uvjet jednakosti *skale*. Rezultati koje smo dobili nalaze se u Tablici 4.2. Ovdje smo kao i u prvoj tablici istaknuli profile koji bolje opisuje kolekciju. Jasno možemo

vidjeti da medicinski profili puno bolje opisuju medicinsku kolekciju od socijalnih profila. Vidimo da su svega 3 dokumenta bolje opisana socijalnim profilima.

Medicinski profili	Socijalni profili	Medicinski profili	Socijalni profili
4207.3438527	3.8492511	261.2953139	**423.1465212**
111.4519204	0.0000000	**267.7725532**	131.4545345
859.2441300	131.4545345	**451.8624759**	134.7695881
175.7607768	0.0000000	**98.5028866**	0.0000000
402.7377790	0.0000000	52.4352925	**214.0720916**
165.2212318	19.8788910	**579.0122710**	91.0295904
634.5692119	135.6071610	**1159.6015719**	0.0000000
1150.5106590	0.0000000	**1132.8585256**	131.4545345
1156.2123340	0.0000000	**1217.3558461**	0.0000000
238.6809494	1.8630996	**1694.7105296**	91.0295904
8.1912717	**578.3058868**	**891.8921391**	235.6122543
924.8446741	38.1124528	**1840.2290624**	473.5961772
4.3910407	0.0000000	**1047.6939234**	0.0000000
454.1139825	131.4545345	**972.0035760**	24.7896819
581.3828215	131.4545345	**1548.2768984**	93.0734565
712.8602375	0.0000000	**520.0696461**	0.0000000
414.2302594	224.5671128	**790.5266387**	109.7957187
816.6342508	388.8323747	**1315.6235762**	0.0000000
1435.8799498	0.0000000	**1109.1395747**	544.0562127
1198.6854179	28.1929503	**641.4693034**	0.0000000
1429.8655110	169.4756950	**1372.6018516**	0.0000000
381.9756960	115.2223849	**1155.7072145**	496.1550169
713.8888127	224.6069206	**930.5907977**	0.0000000
147.1041128	7.6928096	**1101.5907936**	0.0000000
2033.8354436	0.0000000	**1676.4183786**	252.4457890

Tablica 4.2: Medicinska kolekcija opisana medicinskim i reduciranim socijalnim profilima

Isti postupak napravit ćemo sa socijalnom kolekcijom. Radi lakšeg razlikovanja socijalnih profila uvest ćemo pojmove *reducirani* i *nereducirani* socijalni profili. *Reduciranim* socijalnim profilima smatrat ćemo profile dobivene uz uvjet jednakosti *skale*, a pod *nereduciranim* socijalnim profilima smatrat ćemo socijalne profile dobivene uz uvjet da motiva medicinske i socijalne kolekcije ima podjednako.

U prvom slučaju kada smo kolekciju opisali medicinskim i nereduciranim socijalnim profilima, u Tablici 4.3 se jasno vidi da je kolekcija socijalnog sadržaja. Vidimo da medicinski motivi loše opisuju zadanu kolekciju.

Socijalni profili	Medicinski profili
1844.4724095	51.7690291
3230.9083782	12.6195488
3630.2109939	108.3528941
820.8523931	523.4721746
509.0904033	24.0319681
1778.1706241	117.6303496
4109.2376688	0.0000000
3636.8775266	691.6318689
2866.6091779	228.7477966
2044.6922638	2.9465469
1097.6934134	0.9992560
1866.7005127	302.5027872
1045.9552938	52.5103265
889.7220503	103.3236280
630.5727630	0.0000000
4266.9072092	210.9614486
1864.7780464	0.0000000
4218.4749819	0.4944767
2136.6400463	226.5498146
2147.5523737	0.0000000
4576.3456165	533.3814034
1315.2251305	230.1227848
1907.5955830	207.0066058
53.1605687	0.0000000
1.7116771	0.0000000

Socijalni profili	Medicinski profili
2476.5439141	0.0000000
1151.5513871	0.0000000
1946.2293574	204.0253878
823.6547983	21.1919740
3173.1980993	59.2365448
1716.6305310	0.0000000
2939.6050920	363.1544311
1597.5151424	618.1993331
2563.6961045	231.5727117
2839.9864136	125.4409072
1743.9289061	414.9983132
4701.2645176	755.1453015
2363.1757086	13.3508946
3101.4464761	83.1654965
982.1954441	207.1246956
1173.3487425	155.8979856
3314.3549983	22.4900780
1127.9567470	15.7490245
1447.9622333	157.0638824
2463.1558053	44.3660186
647.0574946	0.0000000
785.5255807	0.0000000
7316.5397352	1350.6219911
2537.5213920	284.9451610
791.2947261	0.0000000

Tablica 4.3: Socijalna kolekcija opisana nereduciranim socijalnim i medicinskim profilima

Preostaje nam još vidjeti drugi slučaj, kako će socijalnu kolekciju opisati reducirani socijalni profili i medicinski profili. Rezultate možemo vidjeti u Tablici 4.4. Vidimo da ne bi mogli odrediti kojeg je sadržaja kolekcija. Oba skupa profila dobro opisuju podjednak broj dokumenata, iako vidimo da ovdje imamo i slučaj gdje je norma jednaka 0 za oba skupa profila.

Neke od standardnih mjera za analiziranje rezultata su mjere evaluacije, odnosno odaziv (eng. *recall*) i preciznost (eng. *precision*). Te mjere ćemo definirati na sljedeći način,

$$recall_i = \frac{r_i}{r_n}$$

Socijalni profili	Medicinski profili	Socijalni profili	Medicinski profili
0.0000000	**51.7690291**	**531.7853396**	0.0000000
0.0000000	**12.6195488**	**109.9055774**	0.0000000
215.9270973	108.3528941	109.9055774	**204.0253878**
0.0000000	**523.4721746**	0.0000000	**21.1919740**
0.0000000	**24.0319681**	**613.6232485**	59.2365448
0.0000000	**117.6303496**	**901.5737165**	0.0000000
763.9568328	0.0000000	218.2752609	**363.1544311**
0.0000000	**691.6318689**	0.0000000	**618.1993331**
392.3642224	228.7477966	**796.7703991**	231.5727117
36.5134840	2.9465469	**1128.6443787**	125.4409072
127.7949665	0.9992560	**489.6339938**	414.9983132
76.1250200	**302.5027872**	**766.2279507**	755.1453015
4.5264590	**52.5103265**	**34.7682237**	13.3508946
124.9492375	103.3236280	**666.1829730**	83.1654965
442.5158349	0.0000000	114.7492502	**207.1246956**
347.7459660	210.9614486	**157.5514320**	155.8979856
0.0000000	0.0000000	**524.1130557**	22.4900780
70.1917517	0.4944767	**90.9833272**	15.7490245
637.2897518	226.5498146	**332.3077914**	157.0638824
439.0539080	0.0000000	**974.4981942**	44.3660186
321.8256406	**533.3814034**	**0.4664436**	0.0000000
698.9922991	230.1227848	0.0000000	0.0000000
3.1463352	**207.0066058**	1213.3159458	**1350.6219911**
0.0000000	0.0000000	106.2592120	**284.9451610**
0.0000000	0.0000000	0.0000000	0.0000000

Tablica 4.4: Socijalna kolekcija opisana reduciranim socijalnim i medicinskim profilima

$$precision_i = \frac{r_i}{i},$$

gdje je r_i broj relevantnih dokumenata između i najviše rangiranih dokumenata, a r_n je ukupan broj relevantnih dokumenata u zbirci dokumenata. Kolekciju medicinskih dokumenata i kolekciju socijalnih dokumenata spojili smo u jednu kolekciju. Tu novu kolekciju opisali smo medicinskim profilima i izračunali norme. Norme smo sortirali te smo istaknuli norme dobivene kod socijalne kolekcije. Rezultati su prikazani u Tablici (4.5).

Gledajući sortirane norme u Tablici 4.5 vidimo da u prvih 50 najbolje opisanih dokumenata imamo 10 socijalnih dokumenata. Za dobivene rezultate dobijemo sljedeće veličine

Dokument	l_1 norma
M	4207,3438527
M	2033,8354436
M	1840,2290624
M	1694,7105296
M	1676,4183786
M	1548,2768984
M	1435,8799498
M	1429,8655110
M	1372,6018516
S	**1350,6219911**
M	1315,6235762
M	1217,3558461
M	1198,6854179
M	1159,6015719
M	1156,2123340
M	1155,7072145
M	1150,5106590
M	1132,8585256
M	1109,1395747
M	1101,5907936
M	1047,6939234
M	972,0035760
M	930,5907977
M	924,8446741
M	891,8921391

Dokument	l_1 norma
M	859,2441300
M	816,6342508
M	790,5266387
S	**755,1453015**
M	713,8888127
M	712,8602375
S	**691,6318689**
M	641,4693034
M	634,5692119
S	**618,1993331**
M	581,3828215
M	579,0122710
S	**533,3814034**
S	**523,4721746**
M	520,0696461
M	454,1139825
M	451,8624759
S	**414,9983132**
M	414,2302594
M	402,7377790
M	381,9756960
S	**363,1544311**
S	**302,5027872**
S	**284,9451610**
M	267,7725532

Tablica 4.5: Kolekcije opisane medicinskim profilima

uspješnosti testa:

$$recall_{50} = \frac{r_{50}}{r_{100}} = \frac{40}{50} = 0.8 \tag{4.2}$$

$$precision_{50} = \frac{r_{50}}{50} = \frac{40}{50} = 0.8. \tag{4.3}$$

Rezultati pokazuju da medicinski profili vrlo dobro opisuju medicinsku kolekciju.

Na isti način smo novu kolekciju opisali socijalnim profilima. Najprije smo novu kolekciju opisali reduciranim socijalnim profilima te smo najboljih 50 rezultata prikazali u Tablici 4.6. Dokumente medicinskog sadržaja smo istaknuli.

Dokument	l_1 norma
S	1213,3159458
S	1128,6443787
S	974,4981942
S	901,5737165
S	796,7703991
S	766,2279507
S	763,9568328
S	698,9922991
S	666,1829730
S	637,2897518
S	613,6232485
M	**578,3058868**
M	**544,0562127**
S	531,7853396
S	524,1130557
M	**496,1550169**
S	489,6339938
M	**473,5961772**
S	442,5158349
S	439,0539080
M	**423,1465212**
S	392,3642224
M	**388,8323747**
S	347,7459660
S	332,3077914

Dokument	l_1 norma
S	321,8256406
M	**252,4457890**
M	**235,6122543**
M	**224,6069206**
M	**224,5671128**
S	218,2752609
S	215,9270973
M	**214,0720916**
M	**169,4756950**
S	157,5514320
M	**135,6071610**
M	**134,7695881**
M	**131,4545345**
M	**131,4545345**
M	**131,4545345**
M	**131,4545345**
M	**131,4545345**
S	127,7949665
S	124,9492375
M	**115,2223849**
S	114,7492502
S	109,9055774
S	109,9055774
M	**109,7957187**
S	106,2592120

Tablica 4.6: Kolekcije opisane reduciranim socijalnim profilima

Vidimo da je u 50 najbolje opisanih dokumenata njih 21 medicinskih. Sada ćemo izračunati veličine uspješnosti testa:

$$recall_{50} = \frac{r_{50}}{r_{100}} = \frac{29}{50} = 0.58 \tag{4.4}$$

$$precision_{50} = \frac{r_{50}}{50} = \frac{29}{50} = 0.58. \tag{4.5}$$

Uspoređujući dobivene rezultate s rezultatima (4.2) i (4.3), možemo primijetiti da medicinski profili puno bolje opisuju medicinske dokumente nego socijalni profili socijalne dokumente.

Preostaje nam još pogledati što se dogodi kada novu kolekciju opišemo nereduciranim socijalnim motivima, motivima koje smo dobili smanjivanjem *skale*. Rezultati su prikazani u Tablici 4.7 te smo istaknuli dokumente medicinskog sadržaja.

Dokument	l_1 norma
S	7316,5397352
S	4701,2645176
S	4576,3456165
S	4266,9072092
S	4218,4749819
S	4109,2376688
S	3636,8775266
S	3630,2109939
S	3314,3549983
S	3230,9083782
S	3173,1980993
M	**3104,0138015**
S	3101,4464761
M	**3067,1391450**
S	2939,6050920
M	**2900,1023327**
S	2866,6091779
S	2839,9864136
S	2563,6961045
M	**2537,5596499**
S	2537,5213920
S	2476,5439141
S	2463,1558053
M	**2457,5387155**
M	**2423,1600626**

Dokument	l_1 norma
S	2363,1757086
S	2147,5523737
S	2136,6400463
M	**2101,7063230**
M	**2093,1673010**
S	2044,6922638
M	**2001,9222285**
M	**1950,7042294**
S	1946,2293574
S	1907,5955830
S	1866,7005127
S	1864,7780464
S	1844,4724095
M	**1818,6856625**
M	**1780,1445204**
S	1778,1706241
S	1743,9289061
S	1716,6305310
S	1597,5151424
M	**1523,0412606**
M	**1521,9526390**
M	**1491,5291841**
M	**1467,5154981**
M	**1457,9783177**
S	1447,9622333

Tablica 4.7: Kolekcije opisane nereduciranim socijalnim profilima

Uspoređujući Tablicu 4.6 i Tablicu 4.7 vidimo da su rezultati nešto bolji. U 50 najbolje opisanih dokumenata, 17 dokumenata je medicinskog sadržaja. Veličine uspješnosti testa su sljedeće :

$$recall_{50} = \frac{r_{50}}{r_{100}} = \frac{33}{50} = 0.66 \tag{4.6}$$

$$precision_{50} = \frac{r_{50}}{50} = \frac{33}{50} = 0.66. \tag{4.7}$$

Zaključujemo da socijalni profili loše opisuju socijalne dokumente.

Za kraj nam ostaje ispitati osjetljivost i specifičnost testa koje spominjemo u odjeljku (1.4). Analizu ćemo napraviti na istoj kolekciji na kojoj smo određivali odaziv i preciznost. Kako se kolekcija sastoji od 100 dokumenata, odnosno 50 dokumenata medicinskoga sadržaja i 50 dokumenata socijalnog sadržaja postaviti ćemo 50 kao prag između pozitivnih i negativnih pogodaka. Rezultati su prikazani u sljedećim tablicama.

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
Medicinski profili	pozitivni	TP= 40	FN= 10	osjetljivost= 0.8
	negativni	FP= 10	TN= 40	specifičnost= 0.8
		PPV= 0.8	NPV= 0.8	

Tablica 4.8: Osjetljivost i specifičnost medicinskih profila u kolekciji

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
Socijalni profili	pozitivni	TP= 29	FN= 21	osjetljivost= 0.58
	negativni	FP= 21	TN= 29	specifičnost= 0.58
		PPV= 0.58	NPV= 0.58	

Tablica 4.9: Osjetljivost i specifičnost reduciranih socijalnih profila u kolekciji

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
Socijalni profili	pozitivni	TP= 33	FN= 17	osjetljivost= 0.66
	negativni	FP= 17	TN= 33	specifičnost= 0.66
		PPV= 0.66	NPV= 0.66	

Tablica 4.10: Osjetljivost i specifičnost nereduciranih socijalnih profila u kolekciji

Vidimo da su rezultati jednaki rezultatima odaziva i preciznosti, pa ostajemo pri zaključku da medicinski profili bolje opisuju medicinsku kolekciju nego socijalni profili socijalnu kolekciju.

Poglavlje 5

Kriteriji odabira motiva i analiza rezultata

Kako je cilj ovog diplomskog rada klasifikacija teksta prvo smo trebali izabrati kojeg će sadržaja biti tekstovi na kojima ćemo raditi analize. Bilo je poželjno izabrati što različitije tekstove pa smo se odlučili za tekstove medicinskog i socijalnog sadržaja. Tekstove smo morali prilagoditi određenim kriterijima. Za početak smo uklonili sve separatore i brojeve koji su se nalazili u tekstu kako bi dobili nizove slova. Također smo tekstove prilagodili tako da nam jedan sažetak (eng. *abstract*) članka čini jedan dokument, odnosno jedan red nam predstavlja jedan sažetak članka. U tako prilagođenim kolekcijama, socijalnog i medicinskog sadržaja, uzeli smo podskupove podjednakih duljina na kojima smo trenirali motive. Kolekcije na kojima smo trenirali su imale po 100 dokumenata različitih duljina. Na tim kolekcijama proveli smo korake koje smo objasnili u poglavlju (3).

Prvo smo trebali odlučiti koje duljine će biti početni upit. Odlučili smo se za duljinu 10. Sljedeći uvjet koji smo morali odrediti je *skala* iz odjeljka (3.2). Odabir skale je veoma bitan. Ukoliko bi uzeli prenisku skalu dobili bi veliki broj loših pogodaka koji ne moraju nužno biti vezani uz sadržaj kolekcije iz koje smo ih vadili. S druge strane, ukoliko bi skala bila prevelika, postoji mogućnost da postavimo previsok prag i odbacimo dio pogodaka koji bi bili usko vezani uz sadržaj kolekcije.

Na početku smo uzeli da je *skala* jednaka 9 te smo ispisivali motive koji su imali 5 i više sličnih stringova. Razlog je taj što smo se poveli razmišljanjem da kada se neki string više puta pojavi u kolekciji tada njegovo pojavljivanje nije slučajno. Uspoređujući dobivene rezultate, za obje kolekcije, vidjeli smo da medicinskih motiva ima deset puta više od socijalnih. Kako nam je jedna od želja bila podjednak broj motiva iz svake kolekcije, odlučili smo da ćemo smanjiti *skalu* kada pretražujemo motive za socijalnu kolekciju, u nadi da ćemo dobiti podjednak broj motiva. *Skalu* smo spustili na 7 te smo dobili motive socijalne kolekcije čiji broj je bio podjednak broju motiva dobivenih iz medicinske kolek-

cije. U analizi smo koristili skup motiva dobivenih iz medicinske kolekcije te oba skupa motiva dobivenih iz socijalne kolekcije, pod različitim uvjetima.

Analizirajući rezultate u poglavlju (4), vidimo da je bitno imati veliki broj profila, ali i da su ti profili dovoljno specifični za pojedini sadržaj. Medicinski profili donekle zadovoljavaju ta dva uvjeta što se vidi i po rezultatima. S druge strane, kada gledamo uzorak reduciranih socijalnih profila vidimo da loše opisuju i socijalnu kolekciju i medicinsku, dok nereducirani socijalni profili podjednako dobro opisuju obje kolekcije pa zaključujemo da nisu dovoljno specifični, odnosno i da ih ima premalo. Metodu i način na koji bi dobili profile koji bi zadovoljavali oba uvjeta nismo uspjeli dokučiti.

Bibliografija

[1] N. Cristianini i J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge university press, 2000.

[2] A. Medved, *Lokalno poravnanje i prepoznavanje motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2016.

[3] G. Salton i C. Buckley, *Term-weighting approaches in automatic text retrieval*, In Information Processing and Management, Volume 24, Issue 5 (1988).

[4] N. Sarapa, *Teorija vjerojatnosti*, Školska Knjiga, Zagreb, 2002.

[5] S. Vrbančić, *Lokalno poravnanje i prepoznavanje motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2014.

Sažetak

Cilj ovog diplomskog rada bio je provesti postupak klasifikacije tekstova u prirodnom jeziku, ali bez poznavanja rječnika. Za analizu smo odabrali dvije kolekcije različitih sadržaja. Jednu kolekciju čine sažeci članaka medicinskog sadržaja, a drugu sažeci članaka socijalnog sadržaja. Kolekcije smo modificirali tako što smo im uklonili separatore i brojeve kako bi nizovi u tim kolekcijama bili što sličniji biološkim nizovima.

Na tako uređenim kolekcijama proveli smo postupak traženja motiva za svaku kolekciju zasebno. Ti motivi će predstavljati naš rječnik. Svakom skupu motiva smo pridružili odgovarajući skup profila.

Na samom kraju analizirali smo koliko dobro dobiveni profili opisuju testne kolekcije. Analizom smo došli do zaključka da prvenstveno moramo dobro poraditi na odabiru samih motiva. Motivi moraju biti dovoljno specifični za određeni sadržaj, te ih mora biti i u velikom broju. Ukoliko zadovoljimo ta dva uvjeta sama klasifikacija biti će puno bolje odrađena. Primjer toga su nam rezultati dobiveni pomoću medicinskih motiva. Vidimo da smo dobili zadovoljavajuće rezultate iako je samo testiranje rađeno na prilično malom uzorku.

Summary

This thesis is concerned with classification of documents in a natural language, but without a dictionary. Furthermore, all separators and numerals are removed from training data, hence we are dealing with finite sequences in a Latin alphabet. We have selected two sets of abstracts of articles from medical and sociological databases. On such a collection of modified texts, motif scanning techniques have been applied, to each dataset separately. This yielded two sets of profiles, one for each collection.

In the last section, we discuss classification results that we obtained. In particular, we conclude that the profile set has to be large enough to describe each document, but specific enough to avoid misclassification.

Životopis

Rođena sam 10.11.1993. godine u Zagrebu. Osnovnoškolsko obrazovanje započela sam 2000. u Osnovnoj školi Samobor u Samoboru te 2002. upisujem Glazbenu školu Ferdo Livadić, također u Samoboru. Nakon toga, od 2008. do 2012. godine pohađam Gimnaziju Lucijana Vranjanina u Zagrebu i Glazbenu školu Ferdo Livadić. Po završetku srednjoškolskog obrazovanja, 2012. godine upisujem Preddiplomski studij matematike, nastavnički smjer, na Prirodoslovno-matematičkom fakultetu u Zagrebu. 2015. završavam Preddiplomski studij te iste godine upisujem Diplomski sveučilišni studij Matematička statistika na istom fakultetu.