Simultano klasteriranje dokumenata i riječi

Petra Sočo, Jelena Zaninović

22. prosinca 2020.

- Klasteriranje skupa podataka je grupiranje elemenata na temelju sličnosti.
- Konkretni problem: klasteriranje dokumenata i riječi
- Dualnost između ta dva procesa $\longrightarrow$ ima ih smisla pokušati simultano provesti.
- Ideja: Riječi i dokumente organizirati kao bipartitni graf $\longrightarrow$ problem particioniranja bipartitnog grafa.
- "Algoritam": dekompozicija singularnih vrijednosti pripadajuće matrice i provođenje k-means algoritma na jednodimenzionalnom skupu podataka.

Model bipartitnog grafa

- Težinski, neusmjereni graf je uređeni par $G = (\mathcal{V}, E)$ zadan skupom vrhova $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$ i bridova $\{i, j\}$, gdje svaki brid ima težinu E_{ij}.
- Definiramo **matricu susjedstva** M na sljedeći način:

$$M_{ij} = \begin{cases} E_{ij} & \exists \; brid \; \{i, j\}, \\ 0 & inače. \end{cases}$$

- Za proizvoljnu 2-particiju $\mathcal{V}_1$ i $\mathcal{V}_2$ skupa $\mathcal{V}$ definiramo rez particije:

$$cut(\mathcal{V}_1, \mathcal{V}_2) := \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}.$$

Prethodna definicija u slučaju k-particije:

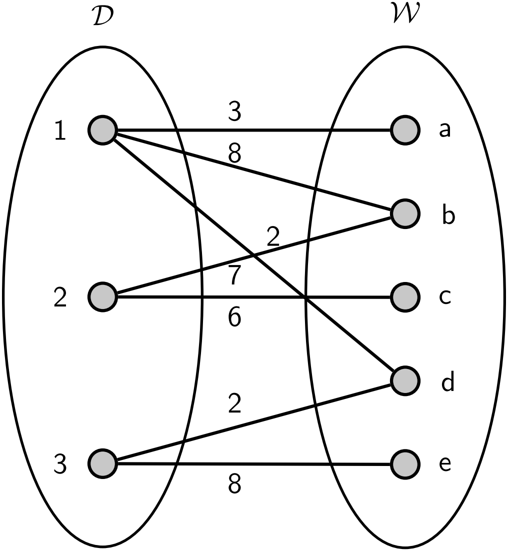
$$cut(\mathcal{V}_1, \ldots, \mathcal{V}_k) := \sum_{i<j} M_{ij}.$$

- Neusmjereni, bipartitni graf je uređena trojka $G = (\mathcal{D}, \mathcal{W}, E)$ gdje su $\mathcal{D} = \{d_1, \ldots, d_n\}$ i $\mathcal{W} = \{w_1, \ldots, w_m\}$ dva skupa vrhova (dokumenti i riječi redom), a $E = \{\{d_i, w_j\} : d_i \in \mathcal{D}, w_j \in \mathcal{W}\}$ je skup bridova.
- Brid $\{d_i, w_j\}$ postoji ako se riječ w_j pojavljuje u dokumentu d_i.
- Jačina te veze izražena je pozitivnom težinom E_{ij} koju postavljamo na brid koji ih povezuje. Matrica susjedstva bipartitnog grafa M ima oblik

$$\mathrm{M} = \begin{bmatrix} 0 & \mathrm{A} \\ \mathrm{A}^T & 0 \end{bmatrix},$$

gdje je A dimenzije $m \times n$ i vrijedi $A_{ij} = E_{ij}$.

- Prvih m redaka i stupaca odgovara riječima, a zadnjih n odgovara dokumentima.


$\mathcal{D}$
$\mathcal{W}$
1
2
3
a
b
c
d
e
3
8
2
7
6
2
8

- Klasteriranje riječi generira klastere dokumenata i obratno:

$$\{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \longleftrightarrow \{\mathcal{W}_1, \ldots, \mathcal{W}_k\}$$

- Riječ w_i pripada klasteru $\mathcal{W}_m$ ako je njena veza s klasterom $\mathcal{D}_m$ jača od veza s ostalim klasterima. Klastere riječi $\mathcal{W}_m$ možemo karakterizirati s:

$$\mathcal{W}_m = \{w_i : \sum_{j \in \mathcal{D}_m} A_{ij} \geq \sum_{j \in \mathcal{D}_l} A_{ij},\ \forall l = 1, \ldots, k\},\ m = 1, \ldots, k.$$

- Slično, iz klastera riječi $\mathcal{W}_1, \ldots, \mathcal{W}_k$ može se dobiti:

$$\mathcal{D}_m = \{d_j : \sum_{i \in \mathcal{W}_m} A_{ij} \geq \sum_{i \in \mathcal{W}_l} A_{ij},\ \forall l = 1, \ldots, k\},\ m = 1, \ldots, k.$$

- Tražimo particiju grafa u kojoj će težine koje povezuju različite klastere biti što manje moguće $\longrightarrow$ minimizacija reza:

$$cut(\mathcal{W}_1 \cup \mathcal{D}_1, \ldots, \mathcal{W}_k \cup \mathcal{D}_k) = \min_{\mathcal{V}_1, \ldots, \mathcal{V}_k} cut(\mathcal{V}_1, \ldots, \mathcal{V}_k),$$

gdje je $\mathcal{V}_1, \ldots, \mathcal{V}_k$ neka k-particija grafa.

Uloga svojstvenog problema u particioniranju grafa

- Naći optimalnu 2-particiju grafa $\longrightarrow$ naći podskupove $\mathcal{V}_1^*$ i $\mathcal{V}_2^*$ takve da je $cut(\mathcal{V}_1, \mathcal{V}_2)$ minimiziran.
- Neka je graf $G = (\mathcal{V}, E)$ zadan s n vrhova i m bridova kojima su pridružene težine E_{ij}
- Definiramo prvo $n \times m$ matricu incidencije I_G: stupac matrice I_G pridružen bridu $\{i,j\}$ ima na i-tom i j-tom mjestu $\sqrt{E_{ij}}$, $-\sqrt{E_{ij}}$ redom, dok su na ostalim mjestima nule.
- Laplaceova matrica L grafa G dana je s:

$$L_{ij} = \begin{cases} \sum_k E_{ik} & i = j, \\ -E_{ij} & i \neq j,\ \exists\ brid\ \{i,j\}, \\ 0 & inače. \end{cases}$$

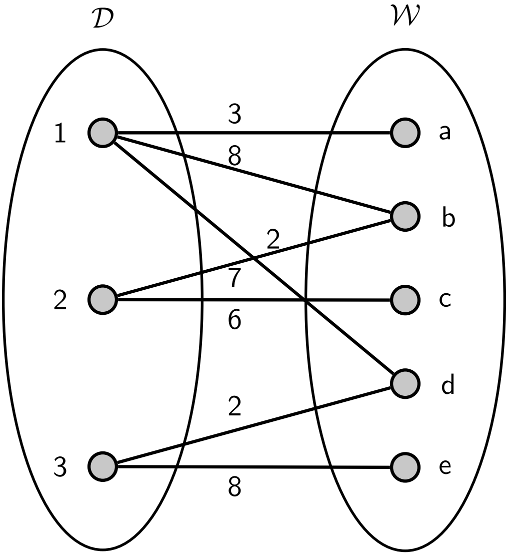
Teorem

Laplaceova matrica $\mathrm{L} = \mathrm{L}_G$ *grafa* G *ima sljedeća svojstva:*

1. $\mathrm{L} = \mathrm{D} - \mathrm{M}$, *gdje je* M *matrica incidencije, a* D *dijagonalna matrica takva da* $D_{ii} = \sum_k E_{ik}$.
2. $\mathrm{L} = \mathrm{I}_G \mathrm{I}_G{}^T$.
3. L *je simetrična pozitivno semi-definitna matrica.*
4. *Neka je* $\mathrm{e} = [1, \ldots, 1]^T$. *Tada je* $\mathrm{Le} = 0$.
5. *Ako graf* G *ima* c *komponenti povezanosti, tada* L *ima* c *svojstvenih vrijednosti jednakih* 0.
6. *Za proizvoljni vektor* x, $\mathrm{x}^T \mathrm{L} \mathrm{x} = \sum_{\{i,j\} \in E} E_{ij}(x_i - x_j)^2$.
7. *Za proizvoljni vektor* x *i skalare* α *i* β *vrijedi*

$$(\alpha \mathrm{x} + \beta \mathrm{e})^T \mathrm{L} (\alpha \mathrm{x} + \beta \mathrm{e}) = \alpha^2 \mathrm{x}^T \mathrm{L} \mathrm{x}.$$

- 3. tvrdnja $\implies$ sve svojstvene vrijednosti matrice L su realne i nenegativne
- 4. tvrdnja $\implies$ $(0, \mathsf{e})$ je jedan svojstveni par matrice L.


$\mathcal{D}$
$\mathcal{W}$
1
2
3
a
b
c
d
e
3
8
2
7
6
2
8

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 8 & 7 & 0 \\ 0 & 6 & 0 \\ 2 & 0 & 2 \\ 0 & 0 & 8 \end{bmatrix}, \quad M = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix},$$

$$L = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & -3 & 0 & 0 \\ 0 & 15 & 0 & 0 & 0 & -8 & -7 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 & -6 & 0 \\ 0 & 0 & 0 & 4 & 0 & -2 & 0 & -2 \\ 0 & 0 & 0 & 0 & 8 & 0 & 0 & -8 \\ -3 & -8 & 0 & -2 & 0 & 13 & 0 & 0 \\ 0 & -7 & -6 & 0 & 0 & 0 & 13 & 0 \\ 0 & 0 & 0 & -2 & -8 & 0 & 0 & 10 \end{bmatrix}, \quad D = diag(L).$$

- Nadalje pretpostavljamo da graf G ima jednu komponentu povezanosti.
- Za proizvoljnu 2-particiju $\mathcal{V}_1, \mathcal{V}_2$ definiramo vektor p kao vektor particije na sljedeći način

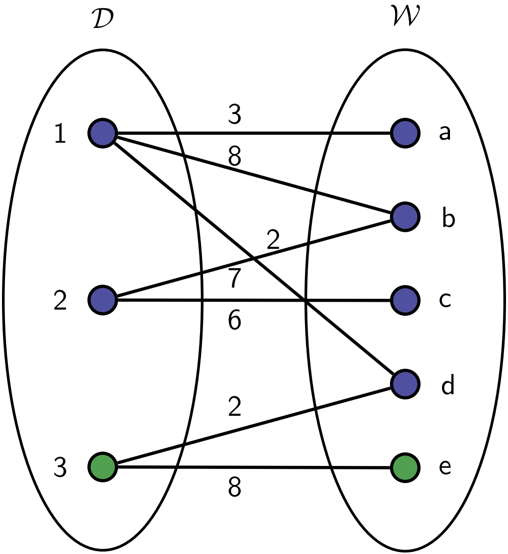
$$p_i := \begin{cases} +1 & i \in \mathcal{V}_1 \\ -1 & i \in \mathcal{V}_2. \end{cases}$$

Teorem

Neka je L *Laplaceova matrica grafa G i* p *vektor particije. Tada za Rayleighov koeficijent vrijedi*

$$\frac{\mathrm{p}^T \mathrm{L} \mathrm{p}}{\mathrm{p}^T \mathrm{p}} = \frac{4}{n} cut(\mathcal{V}_1, \mathcal{V}_2).$$

Primjer particije



- Želimo "balansirane" particije!
- Svakom vrhu pridružujemo težinu $w(i)$ i definiramo dijagonalnu $n \times n$ matricu W sastavljenu od tih težina.
- Za podskup vrhova $\mathcal{V}_l$ definiramo težinu podskupa kao:

$$w(\mathcal{V}_l) = \sum_{i \in \mathcal{V}_l} w(i) = \sum_{i \in \mathcal{V}_l} W_{ii}.$$

- 2-particija je "balansirana" ako su težine podskupova približno jednake $\longrightarrow$ nova funkcija cilja:

$$\mathcal{Q}(\mathcal{V}_1, \mathcal{V}_2) = \frac{cut(\mathcal{V}_1, \mathcal{V}_2)}{w(\mathcal{V}_1)} + \frac{cut(\mathcal{V}_1, \mathcal{V}_2)}{w(\mathcal{V}_2)} \tag{1}$$

Lema

Neka su L *i* W *Laplaceova i matrica težina vrhova nekog grafa G. Tada generalizirani vektor particije* q *definiran s*

$$q_i := \begin{cases} +\sqrt{\dfrac{\eta_2}{\eta_1}} & i \in \mathcal{V}_1 \\ -\sqrt{\dfrac{\eta_1}{\eta_2}} & i \in \mathcal{V}_2. \end{cases}$$

zadovoljava $\mathrm{q}^T \mathrm{W} \mathrm{e} = 0$ *i* $\mathrm{q}^T \mathrm{W} \mathrm{q} = w(\mathcal{V})$. *Gdje je* $\eta_1 = w(\mathcal{V}_1)$ *i* $\eta_2 = w(\mathcal{V}_2)$.

Teorem

$$\frac{\mathrm{q}^T \mathrm{L} \mathrm{q}}{\mathrm{q}^T \mathrm{W} \mathrm{q}} = \frac{cut(\mathcal{V}_1, \mathcal{V}_2)}{w(\mathcal{V}_1)} + \frac{cut(\mathcal{V}_1, \mathcal{V}_2)}{w(\mathcal{V}_2)}$$

- Desna strana izraza u prethodnom teoremu je upravo funkcija koju želimo minimizirati (1) pa pažnju sada možemo usmjeriti na traženje "optimalnog" generaliziranog vektora particije q koji je dan prethodnom lemom.
- Sljedeći teorem omogućuje da diskretni problem nalaženja optimalnog generaliziranog vektora q zamijenimo kontinuiranim.

Teorem

Minimum problema

$$\min_{q\neq 0} \frac{q^T L q}{q^T W q}, \textit{ uz uvjet } q^T W e = 0$$

se postiže za svojstveni vektor druge najmanje svojstvene vrijednosti generaliziranog svojstvenog problema

$$Lz = \lambda W z. \tag{2}$$

Kako pridružiti težine vrhovima?

- $w(i) = 1$ za svaki vrh i pa funkcija iz (1) ima sljedeći oblik

$$\mathcal{Q}(\mathcal{V}_1, \mathcal{V}_2) = \frac{cut(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_1|} + \frac{cut(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_2|}.$$

- Takivm odabirom težina ćemo za vrhove dobiti 2-particije koje će biti "balansirane" u smislu veličina.
- Drugi pristup je da težinu vrha mjerimo pomoću težina koje "izlaze" iz njega, tj. stavit ćemo $w(i) = \sum_k E_{ik}$.
- U tom slučaju će matrica W svojstvenog problema (2) biti jednaka matrici D iz pethodnog teorema.

Veza s dekompozicijom singularnih vrijednosti

- Tražimo drugi najmanji svojstveni vektor generaliziranog svojstvenog problema $\mathsf{L}\mathsf{z} = \lambda \mathsf{D}\mathsf{z}$.
- U slučaju bipartitnog grafa:

$$\mathsf{L} = \begin{bmatrix} \mathsf{D}_1 & -\mathsf{A} \\ -\mathsf{A}^T & \mathsf{D}_2 \end{bmatrix}, \quad \mathsf{D} = \begin{bmatrix} \mathsf{D}_1 & 0 \\ 0 & \mathsf{D}_2 \end{bmatrix}$$

gdje je $D_1(i,i) = \sum_j A_{i,j}$, $D_2(j,j) = \sum_i A_{i,j}$.

$$\Longrightarrow \begin{bmatrix} \mathsf{D}_1 & -\mathsf{A} \\ -\mathsf{A}^T & \mathsf{D}_2 \end{bmatrix} \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} = \lambda \begin{bmatrix} \mathsf{D}_1 & 0 \\ 0 & \mathsf{D}_2 \end{bmatrix} \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \qquad (3)$$

- Pretpostavimo da su D_1 i D_2 regularne

$$\begin{aligned} D_1^{1/2}x - D_1^{-1/2}Ay &= \lambda D_1^{1/2}x, \\ -D_2^{-1/2}A^T x + D_2^{1/2}y &= \lambda D_2^{1/2}y. \end{aligned}$$

- Neka su $u = D_1^{1/2}x$ i $v = D_2^{1/2}y$

$$\begin{aligned} D_1^{-1/2}AD_2^{-1/2}v &= (1-\lambda)u, \\ D_2^{-1/2}A^T D_1^{-1/2}u &= (1-\lambda)v, \end{aligned}$$

- Prethodne jednadžbe definiraju dekompoziciju singularnih vrijednosti (SVD) normalizirane matrice $A_n = D_1^{-1/2} A D_2^{-1/2}$.
- Svojstveni vektor druge (najmanje) svojstvene vrijednosti iz (3) $\longrightarrow$ lijevi i desni singularni vektor koji pripadaju drugoj (najvećoj) singularnoj vrijednosti od A_n

$$A_n v_2 = \sigma_2 u_2, \quad A_n^T u_2 = \sigma_2 v_2 \tag{4}$$

gdje je $\sigma_2 = 1 - \lambda_2$.

- Desni singularni vektor $v_2 \rightsquigarrow$ biparticija dokumenata; lijevi singularni vektor $u_2 \rightsquigarrow$ biparticija riječi.

Biparticioniranje

- Drugi svojstveni vektor matrice L je dan s

$$z_2 = \begin{bmatrix} {D_1}^{-1/2} u_2 \\ {D_2}^{-1/2} v_2 \end{bmatrix}. \tag{5}$$

- Tražimo bi-modalne vrijednosti m_1 i m_2 koje ćemo pridružiti riječima i dokumentima kako bismo direktno čitali rješenje.
- Jedan od pristupa je tražiti m_j takve da minimiziraju funkciju

$$\sum_{j=1}^{2} \sum_{z_2(i) \in m_j} (z_2(i) - m_j)^2.$$

- Prethodni izraz je upravo funkcija cilja kakvu minimizira klasični k-means algoritam.

Algoritam biparticije

1. Za dani A, izračunati $A_n = D_1^{-1/2} A D_2^{-1/2}$,
2. Izračunati druge singularne vektore u_2 i v_2 matrice A_n i definirati z_2 kao u (5),
3. Provesti *k-means* algoritam na jednodimenzionalnim podacima z_2.

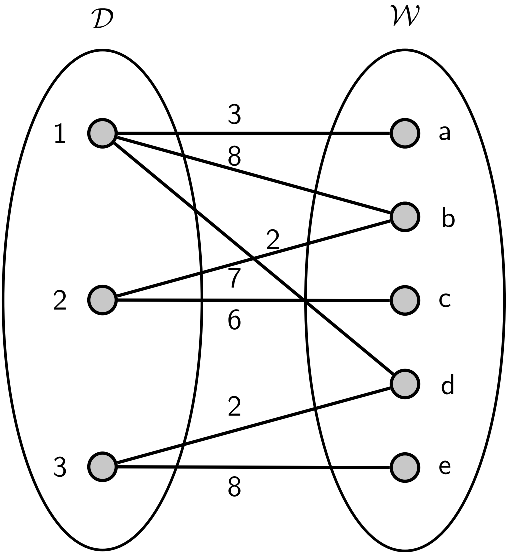
k-particioniranje

- Iskoristimo $l = \lceil \log_2 k \rceil$ singularnih vektora $\mathrm{u}_2, \ldots, \mathrm{u}_{l+1}$ i $\mathrm{v}_2, \ldots, \mathrm{v}_{l+1}$ i definiramo
$$\mathrm{Z} = \begin{bmatrix} \mathrm{D}_1{}^{-1/2}\mathrm{U} \\ \mathrm{D}_2{}^{-1/2}\mathrm{V} \end{bmatrix}. \tag{6}$$
gdje je $\mathrm{U} = [\mathrm{u}_2, \ldots, \mathrm{u}_{l+1}]$ i $\mathrm{V} = [\mathrm{v}_2, \ldots, \mathrm{v}_{l+1}]$.
- Tražimo l-dimenzionalne točke m_j, $j = 1, \ldots, k$ koje ćemo pridružiti dokumentima i riječima. Stoga, minimiziramo funkciju
$$\sum_{j=1}^{k} \sum_{\mathrm{Z}(i) \in m_j} ||\mathrm{Z}(i) - \mathrm{m}_j||^2,$$
gdje je redak $\mathrm{Z}(i)$, $i = 1, \ldots, d + w$ dokument ili riječ.

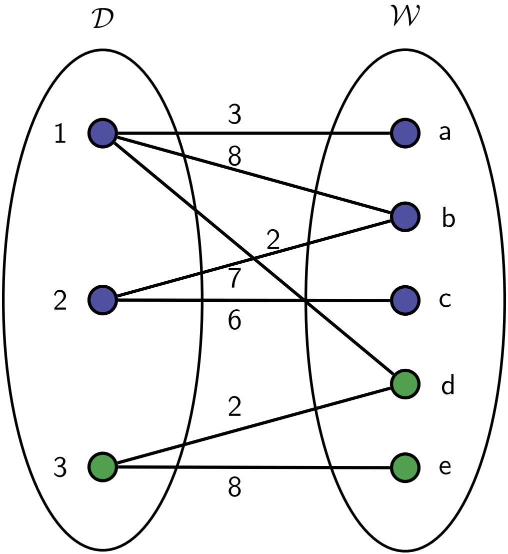
Algoritam k-particije

1. Za dani A, izračunati $A_n = D_1^{-1/2} A D_2^{-1/2}$,
2. Izračunati $l = \lceil \log_2 k \rceil$ singularnih vektora $u_2, \dots, u_{l+1}$ i $v_2, \dots, v_{l+1}$ matrice A_n i definirati matricu Z kao u (6),
3. Provesti *k-means* algoritam na l-dimenzionalnim podacima Z.

Testni primjer



Rezultat



Priprema podataka

- Manji skupovi dokumenata pripremljni su manualno, a veći su preuzeti iz "TMDB 5000 Movie" (www.kaggle.com/tmdb/tmdb-movie-metadata) i "BBC News Summary" (www.kaggle.com/pariza/bbc-news-summary)
- Tekstovi su obrađeni koristeći MATLAB-ov *Text Analytics Toolbox*:
  - tekstovi $\longrightarrow$ niz riječi (*tokenizedDocument*);
  - odstraniti smo najčešće riječi (*RemoveStopWords*);
  - izbrisati interpunkcijske znakove (*removePunctuation*);
  - izvaditi korijene riječi (*normalizeWords*).

Primjer obrađenog dokumenta:
childhood hour other seen other saw bring passion common spring same sourc taken sorrow awaken heart joi same tone lovd lovd alon childhood dawn ...

- Tako uređeni dokumenti su kasnije reprezentirani *bag-of-words* modelom, a pritom je zanemaren poredak riječi i sačuvan podatak o broju pojavljivanja svake riječi.
- Primjerice, iz modela ćemo znati da se u gornjem tekstu "childhood" javlja 2 puta, ali ne znamo da se "hour" javlja prije "spring".
- Dobivene podatke ćemo analizirati mjerom *term frequency-inverse document frequency*.

- Term frequency-inverse document frequency je mjera kojom određujemo težinu riječi u skupu dokumenata.
- Jedna varijanta formule za riječ r, dokument d i skup dokumenata D je

$$tf(r,d) = f_{r,d} \ , \quad idf(r,D) = \ln \frac{N}{n_r}$$

$$tfidf(r,d,D) = tf(r,d) \times idf(r,D)$$

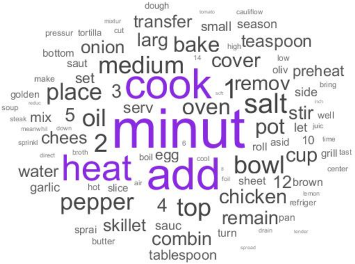
gdje je $f_{r,d} =$ "broj ponavljanja r u d", $N = card(D)$ i $n_r = \{d \in D : r \in d\}$.

- "težina" riječi u dokumentu raste s brojem pojavljivanja u tom dokumentu, a opada s brojem dokumenata u kojima se pojavljuje.

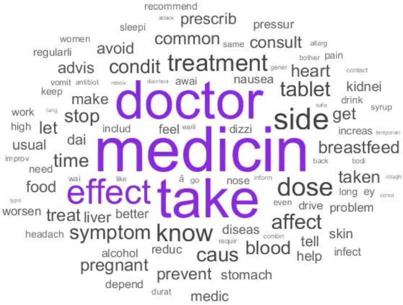
Primjer 1

	recepti	*lijekovi*	*politika*
$\mathcal{D}_0$:	50	0	0
$\mathcal{D}_1$:	0	50	0
$\mathcal{D}_2$:	0	0	50
$\mathcal{W}_0$:	minut, add, cook, heat, salt		
$\mathcal{W}_1$:	medicin, take, doctor, effect, side		
$\mathcal{W}_2$:	mr, govern, blair, labour, ministr		

Primjer 1


dough
cauliflow
transfer
small
season
pressur
tortilla
cut
larg
bake
teaspoon
bottom
onion
medium
cover
low
saut
oliv
preheat
make
set
cook
remov
bring
golden
place
3
1
side
soup
salt
mix
5
oil
serv
oven
stir
well
steak
minut
pot
let
chees
2
roll
asid
10
time
sprinkl
broth
boil
egg
cup
grill
tast
heat
add
bowl
center
water
sheet
12
brown
garlic
hot
slice
chicken
pepper
4
top
refriger
remain
pan
sprai
skillet
sauc
butter
combin
turn
tablespoon


question
like
student
support
interest
wale
council
problem
email
scottish
brown
immigr
month
school
told
home
ban
chang
pension
come
peopl
labour
howard
first
scotland
anim
blunkett
blair
law
sai
alreadi
made
common
elect
new
mr
women
campaign
time
make
michael
secretari
minist
work
includ
year
claim
public
govern
plan
take
offici
foreign
tori
pai
up
bill
democrat
nation
mp
lord
parti
report
leader
back
right
offic
campbel
liber
educ
committe
polit
announc
review
chancellor
parliament


recommend
prescrib
sleepi
pressur
women
common
consult
regularli
avoid
pain
advis
condit
treatment
heart
vomit
awai
nausea
keep
tablet
kidnei
make
drink
work
doctor
syrup
stop
side
get
high
let
includ
feel
dizzi
increas
usual
dai
medicin
breastfeed
improv
time
need
food
nose
dose
taken
effect
take
long
ey
even
drive
problem
worsen
treat
liver
better
affect
headach
symptom
know
diseas
skin
tell
alcohol
reduc
caus
blood
help
infect
pregnant
prevent
stomach
depend
durat
medic

Primjer 2

	sport	*business*	*tech*
$\mathcal{D}_0$:	1	48	1
$\mathcal{D}_1$:	49	0	1
$\mathcal{D}_2$:	0	2	48
$\mathcal{W}_0$:	year, growth, new, compani		
$\mathcal{W}_1$:	olymp, world, athlet, athen		
$\mathcal{W}_2$:	mobil, peopl, gadget, user		

Primjer 2


help mitsubishi
food rais foreign china
financi
mini month 2005
euro need
up produc
bank economi demand
oper share quarter time
growth
job market busi
india
japan rover car
firm price deal
uk
forecast last year oil fall 2004
annual
govern new profit fuel
presid cut
passeng cost back
airlin sale compani rise
account yuko
take rose still group
product report expect
debt offer
invest revenu econom
countri world
flight deficit
compens


event
birmingham marathon
greek miss paula two final
jone sport kenteri indoor second
collin
british radcliff team
last
back
drug take world face medal
athen
month thanou iaaf
jump
time new
olymp ban win award
feder
test record countri
make
cross won
best
race
dope year athlet cont
gold run
decid game
european champion sprinter
charg
well against
championship
balco
chepkemei great


compani
technolog blogger
email version
develop servic
gadget
around make microsoft
problem
digit report
just
like chip uk
up call
month
internet
user net
player data
patent site
work year mobil mr custom
centr get oper
phone
first
world ink new help
laptop
need secur million
firm peopl
open mani time
three
system
current
comput network
game music
number
program softwar
laser
3g

Primjer 3a

	Christmas	*E.A.Poe*	*Cats*
$\mathcal{D}_0$:	3	0	0
$\mathcal{D}_1$:	0	3	0
$\mathcal{D}_2$:	0	0	2
$\mathcal{W}_0$: snow, let, christma, merri, dai			
$\mathcal{W}_1$: sea, love, annabel, lee, kingdom			
$\mathcal{W}_2$: macav, cat, rum, tum, curious			

Primjer 3b

	Christmas	*E.A.Poe*	*Cats*
$\mathcal{D}_0$:	3	1	0
$\mathcal{D}_1$:	0	2	0
$\mathcal{D}_2$:	0	0	2

$\mathcal{W}_0$: snow, let, christma, merri, dai

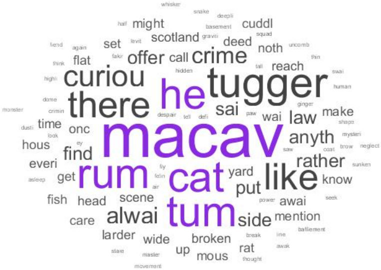
$\mathcal{W}_1$: sea, love, annabel, lee, kingdom

$\mathcal{W}_2$: macav, cat, rum, tum, curious

Primjer 3b


card
bell togeth upon treetop
art goodnight home allow
star yourself ooh goodby fate
olden fire hate
light white happi bough
onc
highest wai littl let merri long
yore know
hear dai
troubl snow heart
thy love everi
on realli oh glisten near
final christma tight just
year warm
write faith go dear kiss dream hang
children
listen storm still bright dy shine
golden gather christmas
trembl slowli friend
sleigh like


childhood sepulchr
darl live other
far down
same chill heaven ago
cloud love beauti life
lee sea mani wind
lovd maiden
angel
came annabel bring
demon
reason know child never
soul kingdom


might cuddl
scotland deed
set noth
flat offer call crime reach
curiou tugger
there he
sai make
time wai law
onc macav anyth
hous
find rather
everi
get rum cat yard like know
put
fish head scene awai
care alwai tum side mention
larder wide broken
up rat
mous

Primjer 3c

	Christmas	*E.A.Poe*	*Cats*
$\mathcal{D}_0$:	3	2	0
$\mathcal{D}_1$:	0	1	0
$\mathcal{D}_2$:	0	0	2

$\mathcal{W}_0$: snow, christma, love, sea, annabel

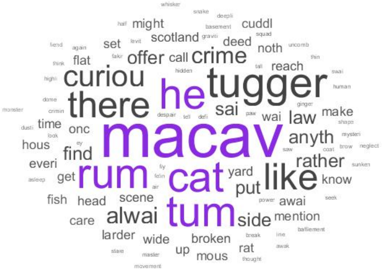
$\mathcal{W}_1$: childhood, other, same, lovd, hour

$\mathcal{W}_2$: macav, cat, rum, tum, curious

Primjer 3c


sleigh
came
beam
togeth
write
on
children
highest
just
awai
happi
goodnight
reason
art
chill
troubl
white
darl
heart
glisten
olden
faith
wind
fate
dream
annabel
beauti
cloud
mani
kiss
shine
sepulchr
year
sea
oh
let
far
love
down
bell
listen
home
near
dy
hate
go
star
snow
wai
heaven
card
warm
littl
lee
dai
slowli
hear
friend
fire
yore
child
trembl
light
know
christma
ooh
golden
soul
storm
final
tight
allow
like
kingdom
merri
bright
hang
never
ago
still
onc
thy
ey
dear
long
yourself
everi
live
gather
goodby
realli
bough
maiden
christmas
treetop
upon
angel


same
childhood
lovdother


might
cuddl
scotland
set
deed
noth
flat
offer
call
crime
reach
curiou
he
tugger
there
sai
wai
law
make
time
onc
macav
anyth
hous
find
everi
rather
get
rum
cat
yard
put
like
know
fish
head
scene
awai
care
alwai
tum
side
mention
larder
wide
broken
up
mous
rat

Primjer 4

	sport	*tech*
$\mathcal{D}_0$:	172	400
$\mathcal{D}_1$:	328	0
$\mathcal{W}_0$:	olymp, world, athlet, athen	
$\mathcal{W}_1$:	year, new, mobil, peopl	

	sport	*tech*
$\mathcal{D}_0$:	4	0
$\mathcal{D}_1$:	23	400
$\mathcal{D}_2$:	471	0
$\mathcal{D}_3$:	2	0
$\mathcal{W}_0$:	harrier, ac, stade, treviso	
$\mathcal{W}_1$:	peopl, game, mobil, phon	
$\mathcal{W}_2$:	game, win, plai, against	
$\mathcal{W}_3$:	republ, ireland, franc, faro	

Primjer 4


gloucest eal sotherton
parma birchfield citi sale
payn viadana yannick level rougeri
barnet green biarritz francai player
perpignan julien centr plai clermont wing
rhy essex l stade s gb replac intern
p go first m holm capt mauro
woodford harrier jone great
dal team aurelien
c return k ladi southal
calvisano d ac r j treviso john
bourgoin jauzion toulousain g turner laharragu
middlesex trafford belgrav maso agen
shaftesburi bergamasco
wigglesworth


launch look
million accord
technolog number devic
current
first download pc system appl
uk net mobil make develop
take player two year call
wai servic new user go
show consum
data up music internet
like peopl mani
market tv mr just
comput game time help
research releas digit gadget
network firm offer set next home microsoft
compani site get phone onlin
content video world
work search softwar
secur inform broadband report


championship
season get injuri titl
coach ireland match
good england ad
unit olymp cup
wale arsen won set up nation
second go first last made
make win best
week world game year start
william time rugbi
like six chelsea
side franc well told face
come try plai two club
think minut player goal
race miss new victori
scotland
open leagu against just point
team three
champion take back manag
final next great
intern test


final
switzerland meet
dai
ireland cypru africa
bafana
franc v israel
winner group first match
republ faro
mauritiu south
seychel plai island
madagascar

Primjer 5

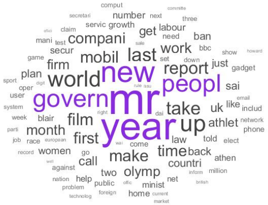
	sport	*entertainment*	*politics*	*business*	*tech*
$\mathcal{D}_0$:	4	19	0	0	1
$\mathcal{D}_1$:	0	1	0	0	0
$\mathcal{D}_2$:	0	2	0	0	0
$\mathcal{D}_3$:	45	23	48	50	49
$\mathcal{D}_4$:	1	5	2	0	0

$\mathcal{W}_0$: best, award, film, book, year, winner
$\mathcal{W}_1$: la, fenic, viotti, opera, director, includ
$\mathcal{W}_2$: christian, andersen, han, prize, author, booker
$\mathcal{W}_3$: mr, year, new, govern, peopl, world
$\mathcal{W}_4$: film, famili, border, ballet, white, year

Primjer 5


award
best
film
book
winner
prize
year
actor
star
won
includ
name
categori
theatr
world
nomin
produc
role
second
2004
novel
win
histori
children
director
richard
actress
british
shortlist
olymp
record
honour
fiction
meet
judi
paul
critic
berlin
hope
four
franc
great
ceremoni
oscar
favourit
new
plai
five
two
made
up
boi
game
mari
nation
first
well
end
last
interact
whitbread
field
indoor
leigh
run
support
six
london
vote
aviat

la
fenic
viotti
opera
traviata
massenet
director
includ
product

andersen
christian
han
prize
booker
author
uk
mcewan
anderson
muriel
canada
jarr
award
intern
spark
bloom
atwood
margaret
british

year
mr
new
govern
peopl
world
take
up
report
compani
mobil
last
work
firm
film
month
first
make
time
back
countri
olymp
call
women
plan
athlet
uk
sai
just
gadget
get
labour
ban
number
growth
servic
secur
blair
law
told
elect
phone
network
includ
like
two
minist
home
public
help
net
against
million
athen
need
bbc
comput
claim
mani
game
sport
user
system
week
job
race
record
go
come
nation
problem
foreign
technolog
market
current
secretari
next
three
show
email
inform

border
film
famili
mr
ballet
year
white
spend
london
director
first
new
york
ossi
apolog
davi
star
chang
right
line
up
lee
made
festiv
friend
servic
reopen
music
becam
opera
life
later
cont
blair
mark
collect
rail
maguir
campaign
conlon
scottish
activist
ag