

Vektorski model

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Vektorski model

- Težinski faktori vezani za pojedine termine u odnosu na dokumente i upite su pozitivne ali ne celobrojne vrednosti
- I upit ima težinske faktore
- Ima rangiranja
- Ima parcijalnog poklapanja upita i dokumenta
- I upit i dokument se predstavljaju kao n -dimenzionalni vektor (n je broj termova u rečniku)
- Ugao koji zaklapaju vektori je obrnuto srazmeran relevantnosti dokumenta za postavljeni upit

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između
- Dobro za korisnike-eksperte sa preciznim razumevanjem svojih potreba i sadržaja kolekcije

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između
- Dobro za korisnike-eksperte sa preciznim razumevanjem svojih potreba i sadržaja kolekcije
- Dobro za aplikacije - one lako mogu da obrade hiljade rezultata

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između
- Dobro za korisnike-eksperte sa preciznim razumevanjem svojih potreba i sadržaja kolekcije
- Dobro za aplikacije - one lako mogu da obrade hiljade rezultata
- Nije dobro za većinu korisnika

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između
- Dobro za korisnike-eksperte sa preciznim razumevanjem svojih potreba i sadržaja kolekcije
- Dobro za aplikacije - one lako mogu da obrade hiljade rezultata
- Nije dobro za većinu korisnika
- Većina korisnika nije u stanju da piše Bulove upite (ili jeste, ali ih mrzi)

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između
- Dobro za korisnike-eksperte sa preciznim razumevanjem svojih potreba i sadržaja kolekcije
- Dobro za aplikacije - one lako mogu da obrade hiljade rezultata
- Nije dobro za većinu korisnika
- Većina korisnika nije u stanju da piše Bulove upite (ili jeste, ali ih mrzi)
- Većina korisnika ne želi da pregleda hiljade pogodaka

Rangiranje rezultata pretrage

- Do sada svi upiti su bili Bulovi
 - dokumenti ili odgovaraju upitu, ili ne odgovaraju - nema između
- Dobro za korisnike-eksperte sa preciznim razumevanjem svojih potreba i sadržaja kolekcije
- Dobro za aplikacije - one lako mogu da obrade hiljade rezultata
- Nije dobro za većinu korisnika
- Većina korisnika nije u stanju da piše Bulove upite (ili jeste, ali ih mrzi)
- Većina korisnika ne želi da pregleda hiljade pogodaka
- Ovo posebno važi za pretragu na webu

Problem Bulovih upita: sve ili ništa

- Bulovi upiti često rezultuju sa malo ($=0$) ili previše ($1000+$) pogodaka

Problem Bulovih upita: sve ili ništa

- Bulovi upiti često rezultuju sa malo ($=0$) ili previše ($1000+$) pogodaka
- Upit 1: “standard user dlink 650” \rightarrow 200,000 hits

Problem Bulovih upita: sve ili ništa

- Bulovi upiti često rezultuju sa malo ($=0$) ili previše ($1000+$) pogodaka
- Upit 1: “standard user dlink 650” \rightarrow 200,000 hits
- Upit 2: “standard user dlink 650 no card found”: 0 hits

Problem Bulovih upita: sve ili ništa

- Bulovi upiti često rezultuju sa malo ($=0$) ili previše ($1000+$) pogodaka
- Upit 1: “standard user dlink 650” \rightarrow 200,000 hits
- Upit 2: “standard user dlink 650 no card found”: 0 hits
- Potrebna je veština da se napiše upit koji će vratiti razuman broj pogodaka

Problem Bulovih upita: sve ili ništa

- Bulovi upiti često rezultuju sa malo ($=0$) ili previše ($1000+$) pogodaka
- Upit 1: “standard user dlink 650” \rightarrow 200,000 hits
- Upit 2: “standard user dlink 650 no card found”: 0 hits
- Potrebna je veština da se napiše upit koji će vratiti razuman broj pogodaka
- Sa rangiranim skupom pogodaka nije važno koliko je velik rezultat

Ocenjivanje kao osnova rangiranja

- Želimo da nađemo dokumente koji su najkorisniji za korisnika (sortiran rezultat)

Ocenjivanje kao osnova rangiranja

- Želimo da nađemo dokumente koji su najkorisniji za korisnika (sortiran rezultat)
- Kako možemo da rangiramo dokumente u odnosu na upit?

Ocenjivanje kao osnova rangiranja

- Želimo da nađemo dokumente koji su najkorisniji za korisnika (sortiran rezultat)
- Kako možemo da rangiramo dokumente u odnosu na upit?
- Dodelimo ocenu (score) – recimo iz $[0, 1]$ – svakom dokumentu

Ocenjivanje kao osnova rangiranja

- Želimo da nađemo dokumente koji su najkorisniji za korisnika (sortiran rezultat)
- Kako možemo da rangiramo dokumente u odnosu na upit?
- Dodelimo ocenu (score) – recimo iz $[0, 1]$ – svakom dokumentu
- Ocena je mera koliko se dokument i upit „poklapaju“ (match)

Ocene poklapanja upita i dokumenta

- Treba nam način za dodelu ocene svakom paru upit/dokument

Ocene poklapanja upita i dokumenta

- Treba nam način za dodelu ocene svakom paru upit/dokument
- Počnimo od upita sa jednim termom

Ocene poklapanja upita i dokumenta

- Treba nam način za dodelu ocene svakom paru upit/dokument
- Počnimo od upita sa jednim termom
- Ako se term ne pojavljuje u dokumentu, ocena bi trebalo da bude 0

Ocene poklapanja upita i dokumenta

- Treba nam način za dodelu ocene svakom paru upit/dokument
- Počnimo od upita sa jednim termom
- Ako se term ne pojavljuje u dokumentu, ocena bi trebalo da bude 0
- Što češće se term pojavljuje u dokumentu, ocena je veća

Ocene poklapanja upita i dokumenta

- Treba nam način za dodelu ocene svakom paru upit/dokument
- Počnimo od upita sa jednim termom
- Ako se term ne pojavljuje u dokumentu, ocena bi trebalo da bude 0
- Što češće se term pojavljuje u dokumentu, ocena je veća
- Razmotrićemo neke varijante kako ovo uraditi

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa
- Neka su A i B skupovi (bar jedan je neprazan)

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa
- Neka su A i B skupovi (bar jedan je neprazan)
- Jaccard-ov koeficijent:

$$\text{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa
- Neka su A i B skupovi (bar jedan je neprazan)
- Jaccard-ov koeficijent:

$$\text{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- $\text{jaccard}(A, A) = 1$

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa
- Neka su A i B skupovi (bar jedan je neprazan)
- Jaccard-ov koeficijent:

$$\text{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- $\text{jaccard}(A, A) = 1$
- $\text{jaccard}(A, B) = 0$ if $A \cap B = \emptyset$

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa
- Neka su A i B skupovi (bar jedan je neprazan)
- Jaccard-ov koeficijent:

$$\text{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- $\text{jaccard}(A, A) = 1$
- $\text{jaccard}(A, B) = 0$ if $A \cap B = \emptyset$
- A i B ne moraju imati isti broj elemenata.

Prvi pokušaj: Jaccard-ov koeficijent

- Uobičajena mera preklapanja dva skupa
- Neka su A i B skupovi (bar jedan je neprazan)
- Jaccard-ov koeficijent:

$$\text{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- $\text{jaccard}(A, A) = 1$
- $\text{jaccard}(A, B) = 0$ if $A \cap B = \emptyset$
- A i B ne moraju imati isti broj elemenata.
- Uvek se dodeljuje broj između 0 i 1.

Jaccard-ov koeficijent: primer

- Koja je ocena upit/dokument dobijena pomoću Jaccard-ovog koeficijenta za:
 - Upit: "ides of March"
 - Dokument: "Caesar died in March"

Šta nije dobro kod Jaccard-ovog koeficijenta?

- Ne uzima u obzir frekvenciju terma (koliko puta se term pojavljuje)
- Retki termovi su informativniji od čestih; Jaccard ovo ne uzima u obzir
- Treba nam bolji način za normalizaciju dužine
- Kasnije ćemo koristiti $|A \cap B| / \sqrt{|A \cup B|}$ (cosine) ...
- ... umesto $|A \cap B| / |A \cup B|$ (Jaccard) za normalizaciju dužine

Podsećanje: binarna matrica incidencije

	Ivanović	Milosavljević	Gostojić	Zarić	...
	D.	B.	S.	M.	...
digitalan	1	1	1	1	
lucene	1	1	1	1	
dokument	1	1	1	1	
obrazovanje	0	0	1	0	
pretraga	1	1	1	1	
multimedijalan	0	1	1	1	
evaluacija	0	0	0	0	
...					

Svaki dokument je prikazan pomoću binarnog vektora $\in \{0, 1\}^{|V|}$.

Podsećanje: binarna matrica incidencije

	Ivanović	Milosavljević	Gostojić	Zarić	...
	D.	B.	S.	M.	...
digitalan	1	1	1	1	
lucene	1	1	1	1	
dokument	1	1	1	1	
obrazovanje	0	0	1	0	
pretraga	1	1	1	1	
multimedijalan	0	1	1	1	
evaluacija	0	0	0	0	
...					

Svaki dokument je prikazan pomoću **binarnog vektora** $\in \{0, 1\}^{|V|}$.

Od sada ćemo koristiti brojačku matricu

	Ivanović	Milosavljević	Gostojić	Zarić	...
	D.	B.	S.	M.	...
digitalan	8	46	7	5	
lucene	11	68	3	2	
dokument	41	953	105	204	
obrazovanje	0	0	1	0	
pretraga	11	56	10	30	
multimedijalan	0	96	1	7	
evaluacija	0	0	0	0	
...					

Svaki dokument je prikazan pomoću vektora broja pojavljivanja $\in \mathbb{N}^{|V|}$.

Od sada ćemo koristiti brojačku matricu

	Ivanović	Milosavljević	Gostojić	Zarić	...
	D.	B.	S.	M.	...
digitalan	8	46	7	5	
lucene	11	68	3	2	
dokument	41	953	105	204	
obrazovanje	0	0	1	0	
pretraga	11	56	10	30	
multimedijalan	0	96	1	7	
evaluacija	0	0	0	0	
...					

Svaki dokument je prikazan pomoću **vektora broja pojavljivanja** $\in \mathbb{N}^{|V|}$.

Model „vreće sa rečima“

- Ne uzimamo u obzir **redosled** reči u dokumentu

Model „vreće sa rečima“

- Ne uzimamo u obzir **redosled** reči u dokumentu
- *John is quicker than Mary* i *Mary is quicker than John* su prikazani na isti način

Model „vreće sa rečima“

- Ne uzimamo u obzir **redosled** reči u dokumentu
- *John is quicker than Mary* i *Mary is quicker than John* su prikazani na isti način
- Ovo se zove **model „vreće sa rečima“** (bag of words)

Model „vreće sa rečima“

- Ne uzimamo u obzir **redosled** reči u dokumentu
- *John is quicker than Mary* i *Mary is quicker than John* su prikazani na isti način
- Ovo se zove **model „vreće sa rečima“** (bag of words)
- Korak nazad: pozicioni indeks razlikuje ova dva dokumenta

Model „vreće sa rečima“

- Ne uzimamo u obzir **redosled** reči u dokumentu
- *John is quicker than Mary* i *Mary is quicker than John* su prikazani na isti način
- Ovo se zove **model „vreće sa rečima“** (bag of words)
- Korak nazad: pozicioni indeks razlikuje ova dva dokumenta
- Čuvanje informacije o poziciji ćemo ostaviti za drugi put

Model „vreće sa rečima“

- Ne uzimamo u obzir **redosled** reči u dokumentu
- *John is quicker than Mary* i *Mary is quicker than John* su prikazani na isti način
- Ovo se zove **model „vreće sa rečima“** (bag of words)
- Korak nazad: pozicioni indeks razlikuje ova dva dokumenta
- Čuvanje informacije o poziciji ćemo ostaviti za drugi put
- Za sada koristimo model „vreće sa rečima“

Frekvencija terma tf

- Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d .

Frekvencija terma tf

- Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d .
- Hoćemo da koristimo tf kada računamo upit/dokument ocene. Kako?

Frekvencija terma tf

- Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d .
- Hoćemo da koristimo tf kada računamo upit/dokument ocene. Kako?
- Sirova frekvencija terma nije ono što nam treba

Frekvencija terma tf

- Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d .
- Hoćemo da koristimo tf kada računamo upit/dokument ocene. Kako?
- Sirova frekvencija terma nije ono što nam treba
- Dokument sa 10 pojava jednog terma je relevantniji od dokumenta sa jednom pojavom istog terma

Frekvencija terma tf

- Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d .
- Hoćemo da koristimo tf kada računamo upit/dokument ocene. Kako?
- Sirova frekvencija terma nije ono što nam treba
- Dokument sa 10 pojava jednog terma je relevantniji od dokumenta sa jednom pojavom istog terma
- Ali nije 10 puta relevantniji

Frekvencija terma tf

- Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d .
- Hoćemo da koristimo tf kada računamo upit/dokument ocene. Kako?
- Sirova frekvencija terma nije ono što nam treba
- Dokument sa 10 pojava jednog terma je relevantniji od dokumenta sa jednom pojavom istog terma
- Ali nije 10 puta relevantniji
- Relevantnost ne raste proporcionalno sa frekvencijom terma

Logaritmska težina frekvencije

- Logaritmska težina frekvencije terma t u d definiše se kao

$$w_{t,d} = \begin{cases} 1 + \log_{10} \text{tf}_{t,d} & \text{if } \text{tf}_{t,d} > 0 \\ 0 & \text{inače} \end{cases}$$

Logaritmska težina frekvencije

- Logaritmska težina frekvencije terma t u d definiše se kao

$$w_{t,d} = \begin{cases} 1 + \log_{10} \text{tf}_{t,d} & \text{if } \text{tf}_{t,d} > 0 \\ 0 & \text{inače} \end{cases}$$

- $0 \rightarrow 0$, $1 \rightarrow 1$, $2 \rightarrow 1.3$, $10 \rightarrow 2$, $1000 \rightarrow 4$, itd.

Logaritmska težina frekvencije

- Logaritmska težina frekvencije terma t u d definiše se kao

$$w_{t,d} = \begin{cases} 1 + \log_{10} \text{tf}_{t,d} & \text{if } \text{tf}_{t,d} > 0 \\ 0 & \text{inače} \end{cases}$$

- $0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 10 \rightarrow 2, 1000 \rightarrow 4$, itd.
- Ocena za par upit/dokument: suma po termovima t za q i d :
$$\text{ocena} = \sum_{t \in q \cap d} (1 + \log \text{tf}_{t,d})$$

Logaritmska težina frekvencije

- Logaritmska težina frekvencije terma t u d definiše se kao

$$w_{t,d} = \begin{cases} 1 + \log_{10} \text{tf}_{t,d} & \text{if } \text{tf}_{t,d} > 0 \\ 0 & \text{inače} \end{cases}$$

- $0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 10 \rightarrow 2, 1000 \rightarrow 4$, itd.
- Ocena za par upit/dokument: suma po termovima t za q i d :
$$\text{ocena} = \sum_{t \in q \cap d} (1 + \log \text{tf}_{t,d})$$
- Ocena je 0 akko nijedan term iz upita nije prisutan u dokumentu

Frekvencija dokumenta

- Retki termini su informativniji od čestih

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan
 - → Želimo **veliku težinu za retke termine** kao što je digitalizacija

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan
 - → Želimo **veliku težinu za retke termine** kao što je digitalizacija
- Razmotrimo term koji je **čest** u kolekciji (npr. visoko, teško, analiza)

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan
 - → Želimo **veliku težinu za retke termine** kao što je digitalizacija
- Razmotrimo term koji je **čest** u kolekciji (npr. visoko, teško, analiza)
 - Dokument koji sadrži ovaj term je verovatno relevantniji od onog koji ga ne sadrži, ali to nije siguran indikator relevantnosti

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan
 - → Želimo **veliku težinu za retke termine** kao što je digitalizacija
- Razmotrimo term koji je **čest** u kolekciji (npr. visoko, teško, analiza)
 - Dokument koji sadrži ovaj term je verovatno relevantniji od onog koji ga ne sadrži, ali to nije siguran indikator relevantnosti
 - → **Za česte termine** želimo pozitivne težine za reči kao što su visoko, teško i analiza, ali **manje težine** nego za retke termine

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan
 - → Želimo **veliku težinu za retke termine** kao što je digitalizacija
- Razmotrimo term koji je **čest** u kolekciji (npr. visoko, teško, analiza)
 - Dokument koji sadrži ovaj term je verovatno relevantniji od onog koji ga ne sadrži, ali to nije siguran indikator relevantnosti
 - → **Za česte termine** želimo pozitivne težine za reči kao što su visoko, teško i analiza, ali **manje težine** nego za retke termine
- Koristićemo frekvenciju dokumenta da uzmemo to u obzir prilikom računanja ocene

Frekvencija dokumenta

- Retki termini su informativniji od čestih
- Razmotrimo term koji je **redak** u kolekciji (npr. digitalizacija)
 - Dokument koji sadrži ovaj term je verovatno relevantan
 - → Želimo **veliku težinu za retke termine** kao što je digitalizacija
- Razmotrimo term koji je **čest** u kolekciji (npr. visoko, teško, analiza)
 - Dokument koji sadrži ovaj term je verovatno relevantniji od onog koji ga ne sadrži, ali to nije siguran indikator relevantnosti
 - → **Za česte termine** želimo pozitivne težine za reči kao što su visoko, teško i analiza, ali **manje težine** nego za retke termine
- Koristićemo frekvenciju dokumenta da uzmemo to u obzir prilikom računanja ocene
- Frekvencija dokumenta je **broj dokumenata u kolekciji u kojima se pojavljuje dati term**

idf težina

- df_t je frekvencija dokumenta, odn. broj dokumenata u kojima se pojavljuje term t

idf težina

- df_t je frekvencija dokumenta, odn. broj dokumenata u kojima se pojavljuje term t
- df je inverzna mera informativnosti terma

idf težina

- df_t je frekvencija dokumenta, odn. broj dokumenata u kojima se pojavljuje term t
- df je inverzna mera informativnosti terma
- Definišemo idf težinu terma t kao:

$$idf_t = \log_{10} \frac{N}{df_t}$$

idf težina

- df_t je frekvencija dokumenta, odn. broj dokumenata u kojima se pojavljuje term t
- df je inverzna mera informativnosti terma
- Definišemo idf težinu terma t kao:

$$idf_t = \log_{10} \frac{N}{df_t}$$

- idf je mera informativnosti terma

idf težina

- df_t je frekvencija dokumenta, odn. broj dokumenata u kojima se pojavljuje term t
- df je inverzna mera **informativnosti** terma
- Definišemo **idf težinu** terma t kao:

$$idf_t = \log_{10} \frac{N}{df_t}$$

- idf je mera **informativnosti** terma
- Koristićemo $\log N/df_t$ umesto N/df_t da „ublažimo“ efekat idf -a

idf težina

- df_t je frekvencija dokumenta, odn. broj dokumenata u kojima se pojavljuje term t
- df je inverzna mera informativnosti terma
- Definišemo idf težinu terma t kao:

$$idf_t = \log_{10} \frac{N}{df_t}$$

- idf je mera informativnosti terma
- Koristićemo $\log N/df_t$ umesto N/df_t da „ublažimo“ efekat idf-a
- Koristimo logaritmovanje i za frekvenciju terma i za frekvenciju dokumenta

Primeri za idf

Izračunati idf_t koristeći formulu: $idf_t = \log_{10} \frac{1,000,000}{df_t}$

term	df_t	idf_t
XMIRS	1	6
digitalizacija	100	4
nedelja	1000	3
analiza	10.000	2
ispod	100.000	1
i	1.000.000	0

Uticaj idf-a na rangiranje

- idf utiče na rangiranje samo ako upit ima bar dva terma

Uticaj idf-a na rangiranje

- idf utiče na rangiranje samo ako upit ima bar dva terma
- Na primer, u upitu “digitalizacija dokumenata”, idf težina povećava relativnu težinu za digitalizacija i smanjuje relativnu težinu za dokumenata

Uticaj idf-a na rangiranje

- idf utiče na rangiranje samo ako upit ima bar dva terma
- Na primer, u upitu “digitalizacija dokumenata”, idf težina povećava relativnu težinu za digitalizacija i smanjuje relativnu težinu za dokumenata
- idf nema uticaja na rangiranje rezultata upita sa jednim termom

Frekvencija kolekcije vs. Frekvencija dokumenta

Reč	cf	df
osiguranje	10440	3997
pokušati	10422	8760

- Frekvencija kolekcije terma t je broj pojavljivanja t u kolekciji

Frekvencija kolekcije vs. Frekvencija dokumenta

Reč	cf	df
osiguranje	10440	3997
pokušati	10422	8760

- Frekvencija kolekcije terma t je broj pojavljivanja t u kolekciji
- Koja reč je bolji term za upit (i trebalo bi da dobije veću težinu?)

Frekvencija kolekcije vs. Frekvencija dokumenta

Reč	cf	df
osiguranje	10440	3997
pokušati	10422	8760

- Frekvencija kolekcije terma t je broj pojavljivanja t u kolekciji
- Koja reč je bolji term za upit (i trebalo bi da dobije veću težinu?)
- Ovaj primer sugerše da je df bolji za težine nego cf:

Frekvencija kolekcije vs. Frekvencija dokumenta

Reč	cf	df
osiguranje	10440	3997
pokušati	10422	8760

- Frekvencija kolekcije terma t je broj pojavljivanja t u kolekciji
- Koja reč je bolji term za upit (i trebalo bi da dobije veću težinu?)
- Ovaj primer sugerše da je df bolji za težine nego cf:
- Želimo da manji broj dokumenata koji sadrži osiguranje dobije veći značaj u odnosu na gomilu dokumenata koji sadrže pokušati pri upitu koji sadrži ova dva terma

Frekvencije terma, kolekcije i dokumenta

Veličina	Simbol	Definicija
frekv. terma	$tf_{t,d}$	broj pojavljivanja t u d
frekv. dokumenta	df_t	broj dokumenata u kolekciji koji sadrže t
frekv. kolekcije	cf_t	ukupan broj pojavljivanja t u kolekciji

Frekvencije terma, kolekcije i dokumenta

Veličina	Simbol	Definicija
frekv. terma	$tf_{t,d}$	broj pojavljivanja t u d
frekv. dokumenta	df_t	broj dokumenata u kolekciji koji sadrže t
frekv. kolekcije	cf_t	ukupan broj pojavljivanja t u kolekciji

- Veza između df i cf ?

Frekvencije terma, kolekcije i dokumenta

Veličina	Simbol	Definicija
frekv. terma	$tf_{t,d}$	broj pojavljivanja t u d
frekv. dokumenta	df_t	broj dokumenata u kolekciji koji sadrže t
frekv. kolekcije	cf_t	ukupan broj pojavljivanja t u kolekciji

- Veza između tf i cf ?

tf-idf težine

- tf-idf težina terma je proizvod njegove tf težine i njegove idf težine

tf-idf težine

- tf-idf težina terma je proizvod njegove tf težine i njegove idf težine
-

$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

tf-idf težine

- tf-idf težina terma je proizvod njegove tf težine i njegove idf težine



$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

- Najpoznatija težina u IR

tf-idf težine

- tf-idf težina terma je proizvod njegove tf težine i njegove idf težine



$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

- Najpoznatija težina u IR
- Poznata i kao: tf.idf, tf x idf

Rezime: tf-idf

- Dodeli tf-idf težinu svakom termu t za svaki dokument d :
$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

Rezime: tf-idf

- Dodeli tf-idf težinu svakom termu t za svaki dokument d :
$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$
- N je ukupan broj dokumenata

Rezime: tf-idf

- Dodeli tf-idf težinu svakom termu t za svaki dokument d :
$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$
- N je ukupan broj dokumenata
- Raste sa brojem pojavljivanja u dokumentu

Rezime: tf-idf

- Dodeli tf-idf težinu svakom termu t za svaki dokument d :
$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$
- N je ukupan broj dokumenata
- Raste sa brojem pojavljivanja u dokumentu
- Raste sa retkošću terma u kolekciji

Binarna \rightarrow brojačka \rightarrow težinska matrica

	Ivanović	Milosavljević	Gostojić	Zarić	...
	D.	B.	S.	M.	...
digitalan	0.25	1.51	0.12	0.05	
lucene	0.75	3.2	0.28	0.18	
dokument	0.12	2.3	0.51	0.83	
obrazovanje	0	0	0.5	0	
pretraga	0.25	1.8	0.24	0.26	
multimedijalan	0	8.25	0.12	0.23	
evaluacija	0	0	0	0	
...					

Svaki dokument je predstavljen vektorom realnih vrednosti tf-idf težina $\in \mathbb{R}^{|V|}$

Binarna \rightarrow brojačka \rightarrow težinska matrica

	Ivanović	Milosavljević	Gostojić	Zarić	...
	D.	B.	S.	M.	...
digitalan	0.25	1.51	0.12	0.05	
lucene	0.75	3.2	0.28	0.18	
dokument	0.12	2.3	0.51	0.83	
obrazovanje	0	0	0.5	0	
pretraga	0.25	1.8	0.24	0.26	
multimedijalan	0	8.25	0.12	0.23	
evaluacija	0	0	0	0	
...					

Svaki dokument je predstavljen **vektorom realnih vrednosti** tf-idf težina $\in \mathbb{R}^{|V|}$

Dokumenti kao vektori

- Svaki dokument se reprezentuje kao vektor realnih brojeva sa tf-idf težinama $\in \mathbb{R}^{|V|}$

Dokumenti kao vektori

- Svaki dokument se reprezentuje kao vektor realnih brojeva sa tf-idf težinama $\in \mathbb{R}^{|V|}$
- Tako imamo $|V|$ -dimenzionalni vektorski prostor

Dokumenti kao vektori

- Svaki dokument se reprezentuje kao vektor realnih brojeva sa tf-idf težinama $\in \mathbb{R}^{|V|}$
- Tako imamo $|V|$ -dimenzionalni vektorski prostor
- Termovi su **ose** prostora

Dokumenti kao vektori

- Svaki dokument se reprezentuje kao vektor realnih brojeva sa tf-idf težinama $\in \mathbb{R}^{|V|}$
- Tako imamo $|V|$ -dimenzionalni vektorski prostor
- Termovi su ose prostora
- Dokumenti su tačke ili vektori u ovom prostoru

Dokumenti kao vektori

- Svaki dokument se reprezentuje kao vektor realnih brojeva sa tf-idf težinama $\in \mathbb{R}^{|V|}$
- Tako imamo $|V|$ -dimenzionalni vektorski prostor
- Termovi su ose prostora
- Dokumenti su tačke ili vektori u ovom prostoru
- Visoka dimenzionalnost: desetak miliona dimenzija kada se primeni na web pretraživač

Dokumenti kao vektori

- Svaki dokument se reprezentuje kao vektor realnih brojeva sa tf-idf težinama $\in \mathbb{R}^{|V|}$
- Tako imamo $|V|$ -dimenzionalni vektorski prostor
- Termovi su ose prostora
- Dokumenti su tačke ili vektori u ovom prostoru
- Visoka dimenzionalnost: desetak miliona dimenzija kada se primeni na web pretraživač
- Vrlo retki vektori - većina vrednosti je 0

Upiti kao vektori

- Ideja 1: upite, kao i dokumente, predstaviti kao vektore u vektorskom prostoru

Upiti kao vektori

- Ideja 1: upite, kao i dokumente, predstaviti kao vektore u vektorskom prostoru
- Ideja 2: rangirati dokumente prema njihovoj blizini sa upitom

Upiti kao vektori

- Ideja 1: upite, kao i dokumente, predstaviti kao vektore u vektorskom prostoru
- Ideja 2: rangirati dokumente prema njihovoj blizini sa upitom
- blizina = sličnost

Upiti kao vektori

- Ideja 1: upite, kao i dokumente, predstaviti kao vektore u vektorskom prostoru
- Ideja 2: rangirati dokumente prema njihovoj blizini sa upitom
- blizina = sličnost
- blizina \approx negativno rastojanje

Upiti kao vektori

- Ideja 1: upite, kao i dokumente, predstaviti kao vektore u vektorskom prostoru
- Ideja 2: rangirati dokumente prema njihovoj blizini sa upitom
- blizina = sličnost
- blizina \approx negativno rastojanje
- Podsećanje: ovo radimo da prevaziđemo problem „ili jesi ili nisi“ Bulovog modela

Kako formalno opisati sličnost u vektorskom prostoru?

- Prvi pokušaj: rastojanje između dve tačke

Kako formalno opisati sličnost u vektorskom prostoru?

- Prvi pokušaj: rastojanje između dve tačke
- (= rastojanje između krajnjih tačaka dvaju vektora)

Kako formalno opisati sličnost u vektorskom prostoru?

- Prvi pokušaj: rastojanje između dve tačke
- (= rastojanje između krajnjih tačaka dvaju vektora)
- Euklidsko rastojanje?

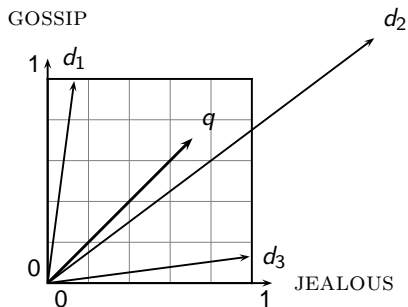
Kako formalno opisati sličnost u vektorskom prostoru?

- Prvi pokušaj: rastojanje između dve tačke
- (= rastojanje između krajnjih tačaka dvaju vektora)
- Euklidsko rastojanje?
- Euklidsko rastojanje je loša ideja ...

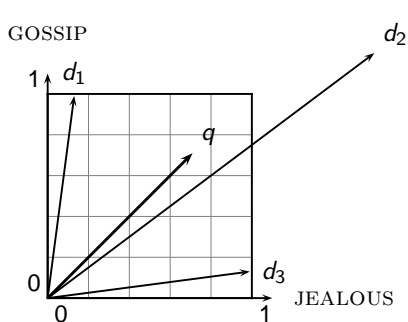
Kako formalno opisati sličnost u vektorskom prostoru?

- Prvi pokušaj: rastojanje između dve tačke
- (= rastojanje između krajnjih tačaka dvaju vektora)
- Euklidsko rastojanje?
- Euklidsko rastojanje je loša ideja ...
- ...jer je **veliko** za vektore **različitih dužina**.

Zašto je rastojanje loša ideja



Zašto je rastojanje loša ideja



Euklidsko \vec{q} i \vec{d}_2 je veliko iako je distribucija termova u upitu q i dokumentu d_2 vrlo slična

Ugao umesto rastojanja

- Rangiraćemo dokumente prema uglu koji zaklapaju sa upitom

Ugao umesto rastojanja

- Rangiraćemo dokumente prema uglu koji zaklapaju sa upitom
- Eksperiment: uzmimo dokument d i dodajmo ga još jednom na njegov kraj; nazovimo to d'

Ugao umesto rastojanja

- Rangiraćemo dokumente prema uglu koji zaklapaju sa upitom
- Eksperiment: uzmimo dokument d i dodajmo ga još jednom na njegov kraj; nazovimo to d'
- „Semantički“ d i d' imaju isti sadržaj

Ugao umesto rastojanja

- Rangiraćemo dokumente prema uglu koji zaklapaju sa upitom
- Eksperiment: uzmimo dokument d i dodajmo ga još jednom na njegov kraj; nazovimo to d'
- „Semantički“ d i d' imaju isti sadržaj
- Ugao između dokumenata je 0, što odgovara maksimalnoj sličnosti

Ugao umesto rastojanja

- Rangiraćemo dokumente prema uglu koji zaklapaju sa upitom
- Eksperiment: uzmimo dokument d i dodajmo ga još jednom na njegov kraj; nazovimo to d'
- „Semantički“ d i d' imaju isti sadržaj
- Ugao između dokumenata je 0, što odgovara maksimalnoj sličnosti
- Euklidsko rastojanje između d i d' je veliko (u svakom slučaju > 0)

Kosinus umesto ugla

- Sledeće dve stvari su ekvivalentne

Kosinus umesto ugla

- Sledeće dve stvari su ekvivalentne
 - Rangiraj dokumente prema **uglu** između upita i dokumenta u rastućem redosledu

Kosinus umesto ugla

- Sledeće dve stvari su ekvivalentne
 - Rangiraj dokumente prema **uglu** između upita i dokumenta u rastućem redosledu
 - Rangiraj dokumente prema **cos**(query,document) u opadajućem redosledu

Kosinus umesto ugla

- Sledeće dve stvari su ekvivalentne
 - Rangiraj dokumente prema **uglu** između upita i dokumenta u rastućem redosledu
 - Rangiraj dokumente prema **cos**(query,document) u opadajućem redosledu
- Kosinus je monotono opadajuća funkcija ugla u intervalu $[0^\circ, 180^\circ]$

Normalizacija dužine

- Kako da izračunamo kosinus?

Normalizacija dužine

- Kako da izračunamo kosinus?
- Vektor se može normalizovati deljenjem svake komponente njegovom dužinom – ovde koristimo L_2 normu:

$$||x||_2 = \sqrt{\sum_i x_i^2}$$

Normalizacija dužine

- Kako da izračunamo kosinus?
- Vektor se može normalizovati deljenjem svake komponente njegovom dužinom – ovde koristimo L_2 normu:

$$||x||_2 = \sqrt{\sum_i x_i^2}$$

- Ovo premešta sve vektore u jediničnu sferu ...

Normalizacija dužine

- Kako da izračunamo kosinus?
- Vektor se može normalizovati deljenjem svake komponente njegovom dužinom – ovde koristimo L_2 normu:

$$||x||_2 = \sqrt{\sum_i x_i^2}$$

- Ovo premešta sve vektore u jediničnu sferu ...
- ...jer je nakon normalizacije: $||x||_2 = \sqrt{\sum_i x_i^2} = 1.0$

Normalizacija dužine

- Kako da izračunamo kosinus?
- Vektor se može normalizovati deljenjem svake komponente njegovom dužinom – ovde koristimo L_2 normu:
$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$
- Ovo premešta sve vektore u jediničnu sferu ...
- ...jer je nakon normalizacije: $\|x\|_2 = \sqrt{\sum_i x_i^2} = 1.0$
- Kao rezultat, i kratki i dugački dokumenti imaju težine istog reda veličine

Normalizacija dužine

- Kako da izračunamo kosinus?
- Vektor se može normalizovati deljenjem svake komponente njegovom dužinom – ovde koristimo L_2 normu:
$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$
- Ovo premešta sve vektore u jediničnu sferu ...
- ...jer je nakon normalizacije: $\|x\|_2 = \sqrt{\sum_i x_i^2} = 1.0$
- Kao rezultat, i kratki i dugački dokumenti imaju težine istog reda veličine
- Efekat na dva dokumenta d i d' (d dodat na samog sebe) sa prethodnog primera: imaju **identične vektore** nakon normalizacije

Kosinusna sličnost upita i dokumenta

$$\cos(\vec{q}, \vec{d}) = \text{sim}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i je tf-idf težina terma i u upitu

Kosinusna sličnost upita i dokumenta

$$\cos(\vec{q}, \vec{d}) = \text{sim}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i je tf-idf težina terma i u upitu
- d_i je tf-idf težina terma i u dokumentu

Kosinusna sličnost upita i dokumenta

$$\cos(\vec{q}, \vec{d}) = \text{sim}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i je tf-idf težina terma i u upitu
- d_i je tf-idf težina terma i u dokumentu
- $|\vec{q}|$ i $|\vec{d}|$ su dužine \vec{q} i \vec{d} .

Kosinusna sličnost upita i dokumenta

$$\cos(\vec{q}, \vec{d}) = \text{sim}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i je tf-idf težina terma i u upitu
- d_i je tf-idf težina terma i u dokumentu
- $|\vec{q}|$ i $|\vec{d}|$ su dužine \vec{q} i \vec{d} .
- Ovo je kosinusna sličnost \vec{q} i \vec{d} ili, ekvivalentno, kosinus ugla između \vec{q} i \vec{d} .

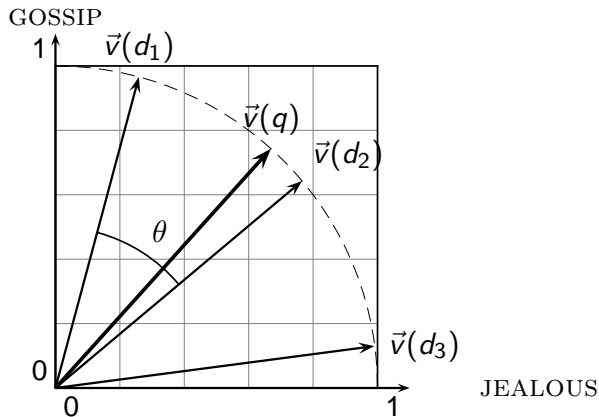
Kosinus za normalizovane vektore

- Za normalizovane vektore kosinus je jednak skalarnom proizvodu

Kosinus za normalizovane vektore

- Za normalizovane vektore kosinus je jednak skalarnom proizvodu
- $\cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_i q_i \cdot d_i$ (ako su \vec{q} i \vec{d} normalizovani)

Ilustracija kosinusne sličnosti



Kosinus: primer

Pretpostavimo da
imamo tri disertacije
u kolekciji:

ID: Ivanović D.

MB: Milosavljević B.

GS: Gostojić S.

Kosinus: primer

Pretpostavimo da
imamo tri disertacije
u kolekciji:

ID: Ivanović D.

MB: Milosavljević B.

GS: Gostojić S.

frekv. terma (broj)

term	ID	MB	GS
indeksiranje	5	58	0
dokument	41	953	105
obrazovanje	0	0	1
multimedijalan	0	96	1

Kosinus: primer

frekv. terma (broj)

term	ID	MB	GS
indeksiranje	5	58	0
dokument	41	953	105
obrazovanje	0	0	1
multimedijalan	0	96	1

Kosinus: primer

frekv. terma (broj)

log frekv.

term	ID	MB	GS
indeksiranje	5	58	0
dokument	41	953	105
obrazovanje	0	0	1
multimedijalan	0	96	1

term	ID	MB	GS
indeksiranje	1,7	2,76	0
dokument	2,61	3,98	3,02
obrazovanje	0	0	1
multimedijalan	0	2,98	1

Kosinus: primer

frekv. dok

term	idf
indeksiranje	0,176
dokument	0
obrazovanje	0,477
multimedijalan	0,176

Kosinus: primer

frekv. dok

 $tf_i df$

term	idf
indeksiranje	0,176
dokument	0
obrazovanje	0,477
multimedijalan	0,176

term	ID	MB	GS
indeksiranje	0,29	0,49	0
dokument	0	0	0
obrazovanje	0	0	0,48
multimedijalan	0	0,52	0,18

Kosinus: primer

 $tf_i df$

term	ID	MB	GS
indeksiranje	0,29	0,49	0
dokument	0	0	0
obrazovanje	0	0	0,48
multimedijalan	0	0,52	0,18

Kosinus: primer

 $tf_i df$

term	ID	MB	GS
indeksiranje	0,29	0,49	0
dokument	0	0	0
obrazovanje	0	0	0,48
multimedijalan	0	0,52	0,18

 $tf_i df$

& normalizacija

term	ID	MB	GS
indeksiranje	1	0,69	0
dokument	0	0	0
obrazovanje	0	0	0,94
multimedijalan	0	0,73	0,35

Kosinus: primer

 $tf_i df$

term	ID	MB	GS
indeksiranje	0,29	0,49	0
dokument	0	0	0
obrazovanje	0	0	0,48
multimedijalan	0	0,52	0,18

 $tf_i df$

& normalizacija

term	ID	MB	GS
indeksiranje	1	0,69	0
dokument	0	0	0
obrazovanje	0	0	0,94
multimedijalan	0	0,73	0,35

- Razmotrimo kako bismo odgovorili na upit: ... multimedijalnih ... indeksiranje ... multimedijalnog , pri čemu ... predstavljaju delove upita koji će nakon pretprocesiranja upita biti izbačeni.

Kosinus: primer

$tf_i df$
& normalizacija

term	ID	MB	GS	upit
indeksiranje	1	0,69	0	0,62
dokument	0	0	0	0
obrazovanje	0	0	0,94	0
multimedijalan	0	0,73	0,35	0,79

Kosinus: primer

tf_idf
& normalizacija

term	ID	MB	GS	upit
indeksiranje	1	0,69	0	0,62
dokument	0	0	0	0
obrazovanje	0	0	0,94	0
multimedijalan	0	0,73	0,35	0,79

- $\cos(\text{ID}, \text{upit}) = 1 * 0,62 + 0 * 0 + 0 * 0 + 0 * 0,79 \approx 0,62$

Kosinus: primer

tf_idf
& normalizacija

term	ID	MB	GS	upit
indeksiranje	1	0,69	0	0,62
dokument	0	0	0	0
obrazovanje	0	0	0,94	0
multimedijalan	0	0,73	0,35	0,79

- $\cos(\text{ID}, \text{upit}) = 1 * 0,62 + 0 * 0 + 0 * 0 + 0 * 0,79 \approx 0,62$
- $\cos(\text{MB}, \text{upit}) = 0,69 * 0,62 + 0 * 0 + 0 * 0 + 0,73 * 0,79 \approx 1$

Kosinus: primer

tf_idf
& normalizacija

term	ID	MB	GS	upit
indeksiranje	1	0,69	0	0,62
dokument	0	0	0	0
obrazovanje	0	0	0,94	0
multimedijalan	0	0,73	0,35	0,79

- $\cos(\text{ID}, \text{upit}) = 1 * 0,62 + 0 * 0 + 0 * 0 + 0 * 0,79 \approx 0,62$
- $\cos(\text{MB}, \text{upit}) = 0,69 * 0,62 + 0 * 0 + 0 * 0 + 0,73 * 0,79 \approx 1$
- $\cos(\text{GS}, \text{upit}) = 0 * 0,62 + 0 * 0 + 0,94 * 0 + 0,35 * 0,79 \approx 0,28$

Kosinus: primer

tf_idf
& normalizacija

term	ID	MB	GS	upit
indeksiranje	1	0,69	0	0,62
dokument	0	0	0	0
obrazovanje	0	0	0,94	0
multimedijalan	0	0,73	0,35	0,79

- $\cos(\text{ID}, \text{upit}) = 1 * 0,62 + 0 * 0 + 0 * 0 + 0 * 0,79 \approx 0,62$
- $\cos(\text{MB}, \text{upit}) = 0,69 * 0,62 + 0 * 0 + 0 * 0 + 0,73 * 0,79 \approx 1$
- $\cos(\text{GS}, \text{upit}) = 0 * 0,62 + 0 * 0 + 0,94 * 0 + 0,35 * 0,79 \approx 0,28$
- Koja disertacija najbolje odgovara upitu?

Izračunavanje kosinusne ocene

CosineScore(q)

```
1  float Scores[ $N$ ] = 0
2  float Length[ $N$ ]
3  for each query term  $t$ 
4  do calculate  $w_{t,q}$  and fetch postings list for  $t$ 
5      for each pair( $d, tf_{t,d}$ ) in postings list
6      do Scores[ $d$ ] +=  $w_{t,d} \times w_{t,q}$ 
7  Read the array Length
8  for each  $d$ 
9  do Scores[ $d$ ] = Scores[ $d$ ]/Length[ $d$ ]
10 return Top  $K$  components of Scores[]
```


Komponente tf-idf težina

Frekv. terma		Frekv. dokumenta		Normalizacija	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-df_t}{df_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^\alpha$, $\alpha < 1$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$				

Komponente tf-idf težina

Frekv. terma		Frekv. dokumenta		Normalizacija	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-df_t}{df_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^\alpha$, $\alpha < 1$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$				

Najbolja poznata kombinacija komponenti težine

Komponente tf-idf težina

Frekv. terma		Frekv. dokumenta		Normalizacija	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N - df_t}{df_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^\alpha$, $\alpha < 1$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$				

Default: bez težina

tf-idf primer

- Mogu da se koriste različite težine za upite i dokumente

tf-idf primer

- Mogu da se koriste različite težine za upite i dokumente
- Notacija: $qqq.ddd$

tf-idf primer

- Mogu da se koriste različite težine za upite i dokumente
- Notacija: `qqq.ddd`
- Primer: `ltn.lnc`

tf-idf primer

- Mogu da se koriste **različite težine** za upite i dokumente
- Notacija: `qqq.ddd`
- Primer: `ltn.lnc`
- upit: logaritamski tf, idf, bez normalizacije

tf-idf primer

- Mogu da se koriste **različite težine** za upite i dokumente
- Notacija: `qqq.ddd`
- Primer: `ltn.lnc`
- upit: logaritamski tf, idf, bez normalizacije
- dokument: logaritamski tf, bez idf, kosinusna normalizacija

Rezime: Rangiranje rezultata u vektorskom modelu

- Predstavi svaki dokument kao tf-idf vektor

Rezime: Rangiranje rezultata u vektorskom modelu

- Predstavi svaki dokument kao tf-idf vektor
- Predstavi upit kao ponderisani tf-idf vektor

Rezime: Rangiranje rezultata u vektorskom modelu

- Predstavi svaki dokument kao tf-idf vektor
- Predstavi upit kao ponderisani tf-idf vektor
- Izračunaj kosinusnu sličnost između upita i svakog dokumenta

Rezime: Rangiranje rezultata u vektorskom modelu

- Predstavi svaki dokument kao tf-idf vektor
- Predstavi upit kao ponderisani tf-idf vektor
- Izračunaj kosinusnu sličnost između upita i svakog dokumenta
- Rangiraj dokumente prema sličnosti

Rezime: Rangiranje rezultata u vektorskom modelu

- Predstavi svaki dokument kao tf-idf vektor
- Predstavi upit kao ponderisani tf-idf vektor
- Izračunaj kosinusnu sličnost između upita i svakog dokumenta
- Rangiraj dokumente prema sličnosti
- Prikaži najboljih K (npr. $K = 10$) dokumenata korisniku