# OSNOVE TEXT MINING-A

Jelena Jovanovic

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

# PREGLED PREDAVANJA

- Šta je Text Mining (TM)?
- Zašto je TM značajan?
- Domeni primene TM-a
- Složenost nestrukturiranog teksta – izvor izazova za TM
- *Bag-of-words* predstava teksta
- *Vector Space Model*
  - Pristupi za pre-procesiranje teksta
  - Metrike za procenu značajnosti termina
  - Kosinusna sličnost

# ŠTA JE TEXT MINING (TM)?

- Primena računarskih metoda i tehnika u cilju *ekstrakcije relevantnih informacija* iz teksta
- *Automatsko otkrivanje značenja / znanja* sadržanih u tekstu
- Primena računara za *otkrivanje novih, nepoznatih informacija i znanja*, kroz *automatizovanu ekstrakciju informacija* iz velikog broja različitih *nestrukturiranih* tekstualnih sadržaja

# ZAŠTO JE TM ZNAČAJAN?

- Nestrukturirani tekstualni sadržaji su opšte prisutni:
  - knjige,
  - finansijski i razni drugi poslovni izveštaji,
  - različita poslovna dokumentacija i prepiska,
  - novinski članci,
  - blogovi,
  - wiki,
  - poruke na društvenim mrežama,
  - …
- Procene su da 80% svih podataka čine nestrukturirani sadržaji

# ZAŠTO JE TM ZNAČAJAN?

- Da bi se taj obim tekstualnih sadržaja efektivno i efikasno koristio, potrebne su metode koje će omogućiti
  - automatizovanu ekstrakciju informacija iz nestrukturiranog teksta
  - analizu i sumiranje ekstrahovanih informacija
- Istraživanja i praksa u domenu TM-a usmereni su na razvoj novih, usavršavanje i primenu ovakvih metoda

# DOMENI PRIMENE TM-A

- Klasifikacija dokumenata*
- Klasterizacija / organizacija dokumenata
- Sumarizacija dokumenata
- Vizuelizacija korpusa (najčešće u svrhe lakšeg pretraživanja dokumenata)
- Predikcije (npr. predviđanje cene akcija na osnovu analize novinskih članaka i finansijskih izveštaja)
- Generisanje preporuka (novinskih članaka, filmova, knjiga, proizvoda generalno, …)

*Termin *dokument* se odnosi na bilo koji tekstualni sadržaj koji čini jednu zaokruženu logičku celinu: blog post, novinski članak, tweet, status update, poslovni dokument, …

# SLOŽENOST NESTRUKTURIRANOG TEKSTA

- Generalno, razumevanje nestrukturiranih sadržaja (tekst, slike, video) je jednostavno za ljude, ali veoma složeno za računare
- U slučaju teksta, problemi su uslovljeni time što je prirodni jezik:
  - pun višesmislenih reči i izraza
  - zasnovan na korišćenju konteksta za definisanje i prenos značenja
  - pun fuzzy, probabilističkih izraza
  - baziran na zdravorazumskom znanju i rezonovanju
  - pod uticajem je i sam utiče na interakcije među ljudima

# DODATNI IZAZOVI ZA TM

- Primena tehnika m. učenja zahteva veliki broj anotiranih dokumenata za formiranje skupa za trening, što je vrlo skupo
  - Takav trening set je potreban za klasifikaciju dokumenata, kao i ekstrakciju entiteta, relacija, događaja
- Visoka dimenzionalnost problema: dokumenti su opisani velikim brojem atributa, što otežava primenu m. učenja
  - Najčešće atribute čine ili svi termini ili na određeni način filtrirani termini iz kolekcije dokumenata koji su predmet analize

# BAG OF WORDS MODEL

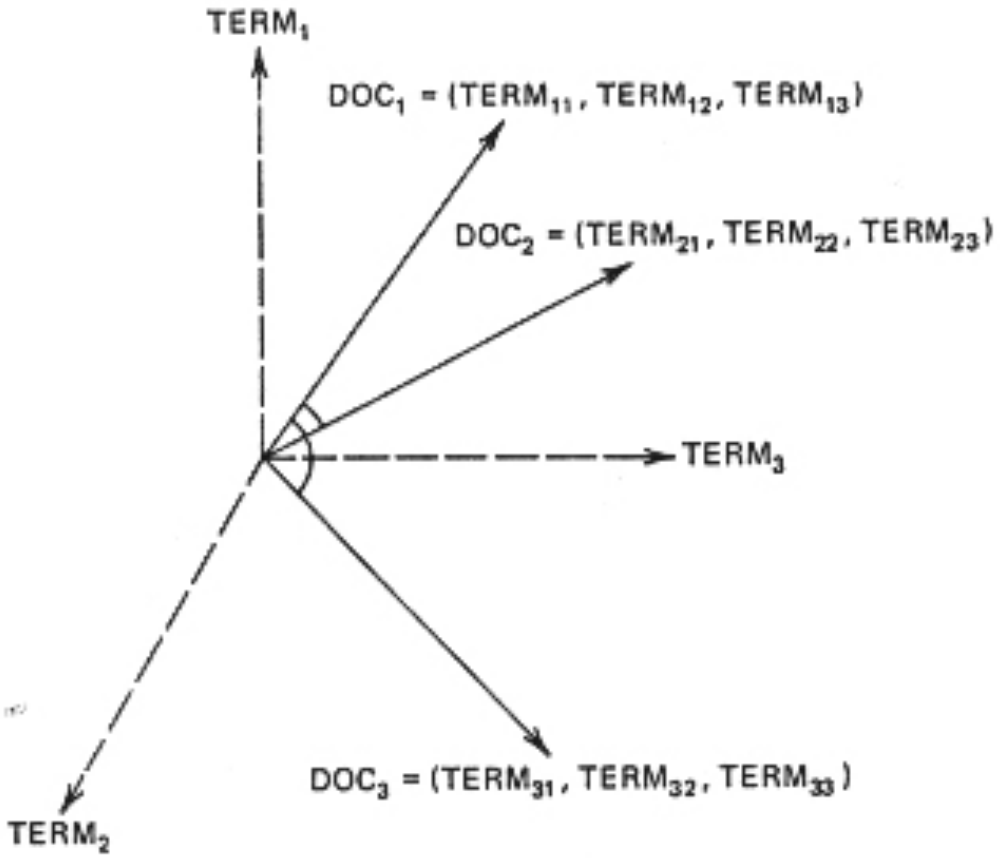
- Tekst se tretira kao prost skup reči
- Pristup zasnovan na sledećim (nerealnim) pretpostavkama:
  - reči su međusobno nezavisne,
  - redosled reči u tekstu je nebitan
- Iako je zasnovan na nerealnim pretpostavkama i vrlo jednostavan, ovaj pristup se pokazao kao vrlo efektan i intenzivno se koristi u TM-u

# BAG OF WORDS MODEL

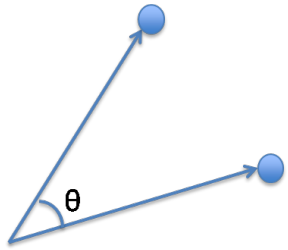

| | |
|---|---|
| aardvark | 0 |
| about | 2 |
| all | 2 |
| Africa | 1 |
| apple | 0 |
| anxious | 0 |
| ... | |
| gas | 1 |
| ... | |
| oil | 1 |
| … | |
| Zaire | 0 |

Reči se izdvajaju iz dokumenata i koriste za formiranje ‘rečnika’ datog korpusa; Zatim se svaki dokument iz korpusa predstavlja kao vektor učestanosti pojavljivanja reči (iz formiranog rečnika) u datom dokumentu

# BAG OF WORDS MODEL:
## PROCENA SLIČNOSTI DOKUMENATA




$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$




Izvor:
http://www.cs.uni.edu/~okane/source/ISR/isr.html

# VECTOR SPACE MODEL

- Generalizacija Bag of Words modela
  - umesto fokusa isključivo na pojedinačne reči, fokus je na *termine*, pri čemu termin može biti jedna reč ili niz reči
  - umesto da se kao mera relevantnosti termina za dati dokument koristi isključivo učestanost pojavljivanja termina u tekstu, koriste se i drugi oblici procene relevantnosti (težine) termina (više o tome kasnije)

# VECTOR SPACE MODEL

- Ako korpus* sadrži $n$ termina ($t_i$, $i=1,n$), dokument $d$ iz tog korpusa biće predstavljen vektorom:

$$d = \{w_1, w_2, \ldots, w_n\}$$

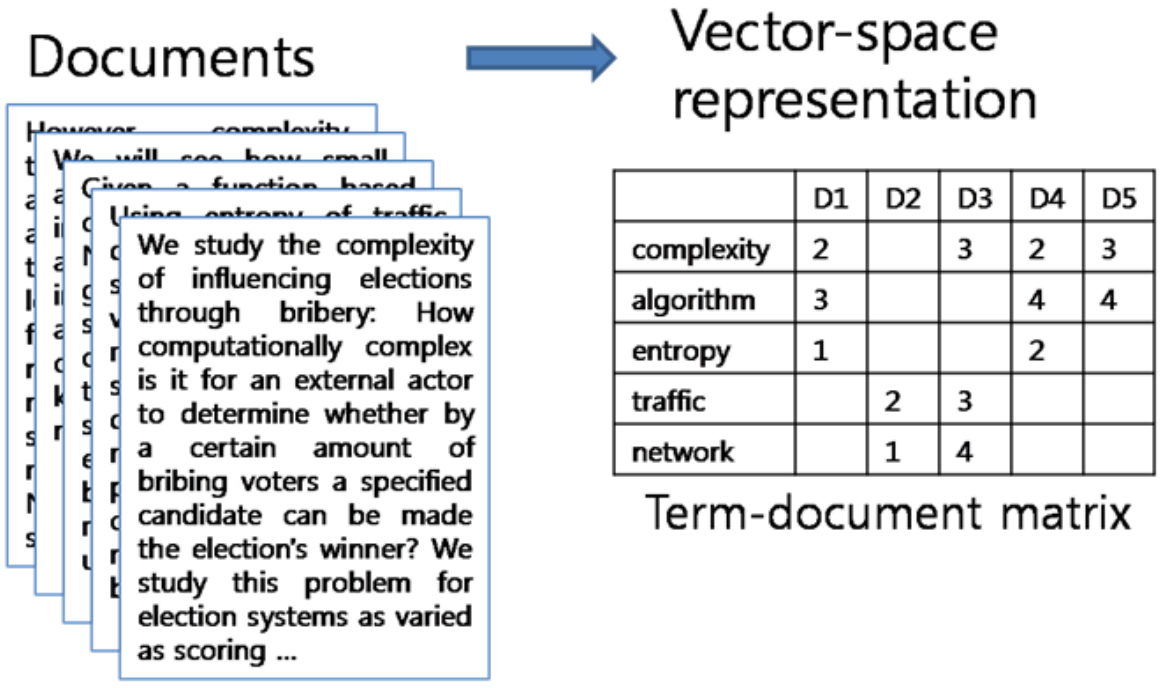
gde su $w_i$ težine pridružene terminima $t_i$

- Ovako kreirani vektori predstavljaju osnovu za formiranje matrice termina i dokumenata (*Term Document Matrix*)

**korpus* je kolekcija dokumenata koji su predmet analize

# VSM: Term Document Matrix

- Term Document Matrix (TDM) je matrice dimenzija *m* x *n* u kojoj:
  - Redovi (*i=1,m*) predstavljaju termine iz korpusa
  - Kolone (*j=1,n*) predstavljaju dokumente iz korpusa
  - Polje *ij* predstavlja težinu termina *i* u kontekstu dokumenta *j*




| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |

Izvor slike:
http://mlg.postech.ac.kr/research/nmf

# VSM: PRE-PROCESIRANJE TEKSTA

- Pre kreiranja TDM matrice, vrši se tzv pre-procesiranje dokumenata iz korpusa
- Razlog: redukovati skup reči na one koje su potencijalno najznačajnije za dati korpus
- Pre-procesiranje (najčešće) obuhvata:
  - normalizaciju teksta
  - odbacivanje termina sa veoma malom i/ili veoma velikom učestanošću u korpusu
  - odbacivanje tzv stop-words
  - svođenje reči na koreni oblik: stemming ili lematizacija

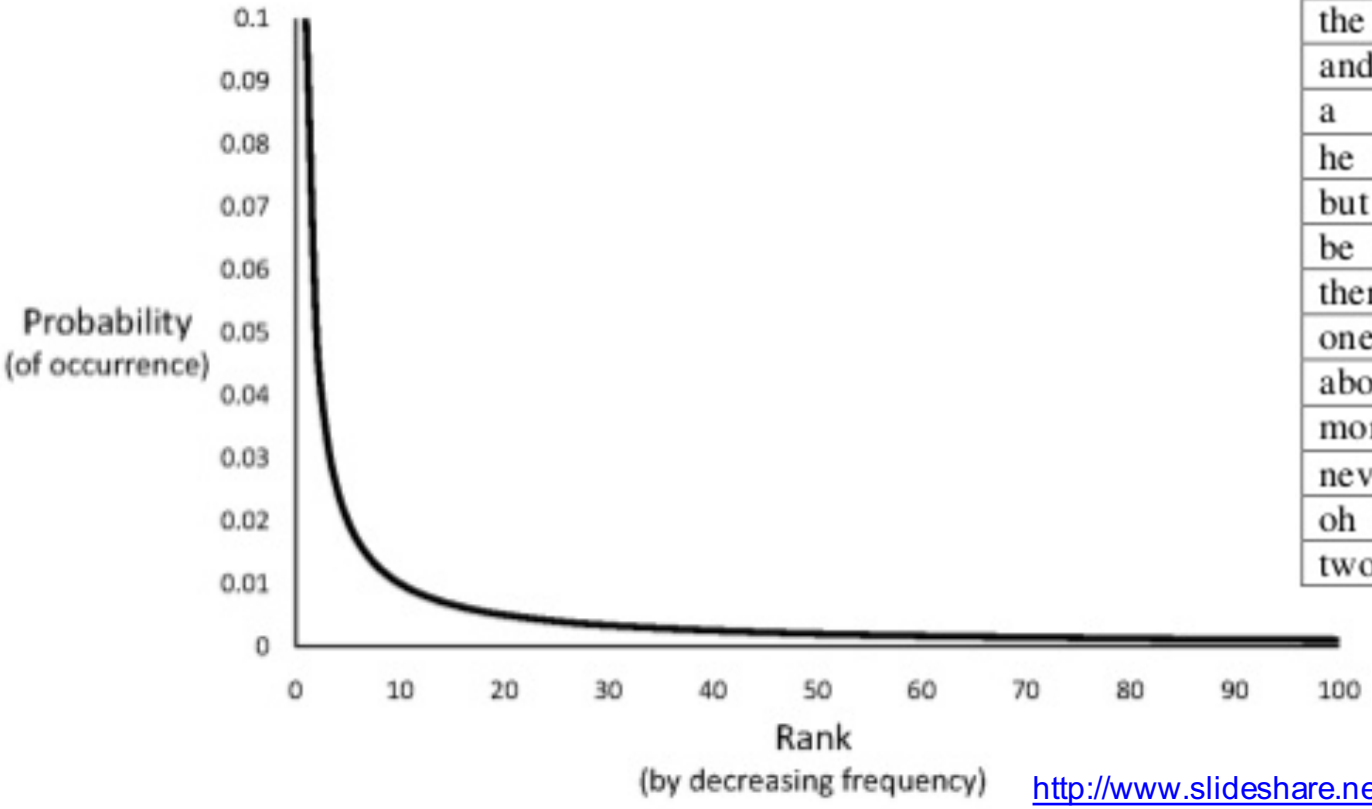
# NORMALIZACIJA TEKSTA

- Cilj: transformisati različite oblike jednog istog termina u osnovni, ‘normalizovani’ format
  - Npr.: Apple, apple, APPLE -> apple

    Intelligent Systems, Intelligent systems, Intelligent-systems -> intelligent systems
- Pristup:
  - Primena jednostavnih pravila:
    - Obrisati sve znake interpunkcije (tačke, crtice, zareze,…)
    - Prebaciti sve reči da budu napisane malim slovima
  - Primena rečnika, npr. WordNet, za zamenu sinonima zajedničkom klasom/konceptom
    - Npr. “automobile, car” -> vehicle

# ELIMINISANJE TERMINA SA SUVIŠE VELIKOM / MALOM UČESTANOŠĆU

- Empirijska zapažanja (u brojnim korpusima):
  - Veliki broj reči ima veoma malu frekvencu pojavljivanja
  - Jako mali broj reči se veoma često pojavljuje u tekstu




| Word | Freq f | Rank r |
|---|---|---|
| the | 3332 | 1 |
| and | 2972 | 2 |
| a | 1775 | 3 |
| he | 877 | 10 |
| but | 410 | 20 |
| be | 294 | 30 |
| there | 222 | 40 |
| one | 172 | 50 |
| about | 158 | 60 |
| more | 138 | 70 |
| never | 124 | 80 |
| oh | 116 | 90 |
| two | 104 | 100 |

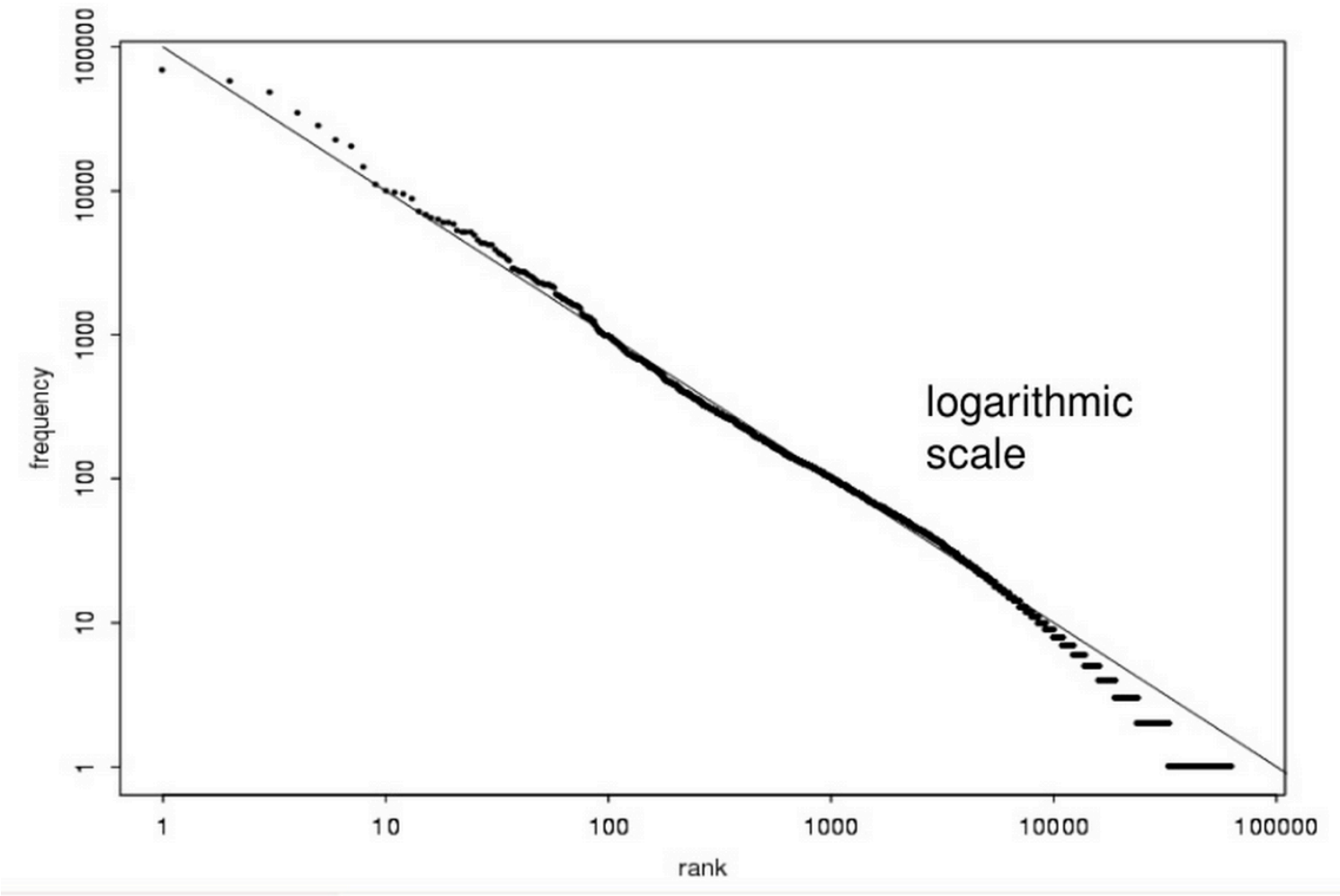
Source: http://www.slideshare.net/cowdung112/info-2402-irtchapter4

# UČESTANOST TERMINA U TEKSTU

Formalizovano *Zipf*-ovim zakonom:

- Frekvenca bilo koje reči u datom korpusu je obrnuto proporcijalna njenom rangu u tabeli frekvencija (tog korpusa)
- Odnosno, proizvod frekvence i ranga termina je približno konstantan na nivou korpusa

# ILUSTRACIJA ZIPF-OVOG PRAVILA




Word frequency in the Brown Corpus of American English text

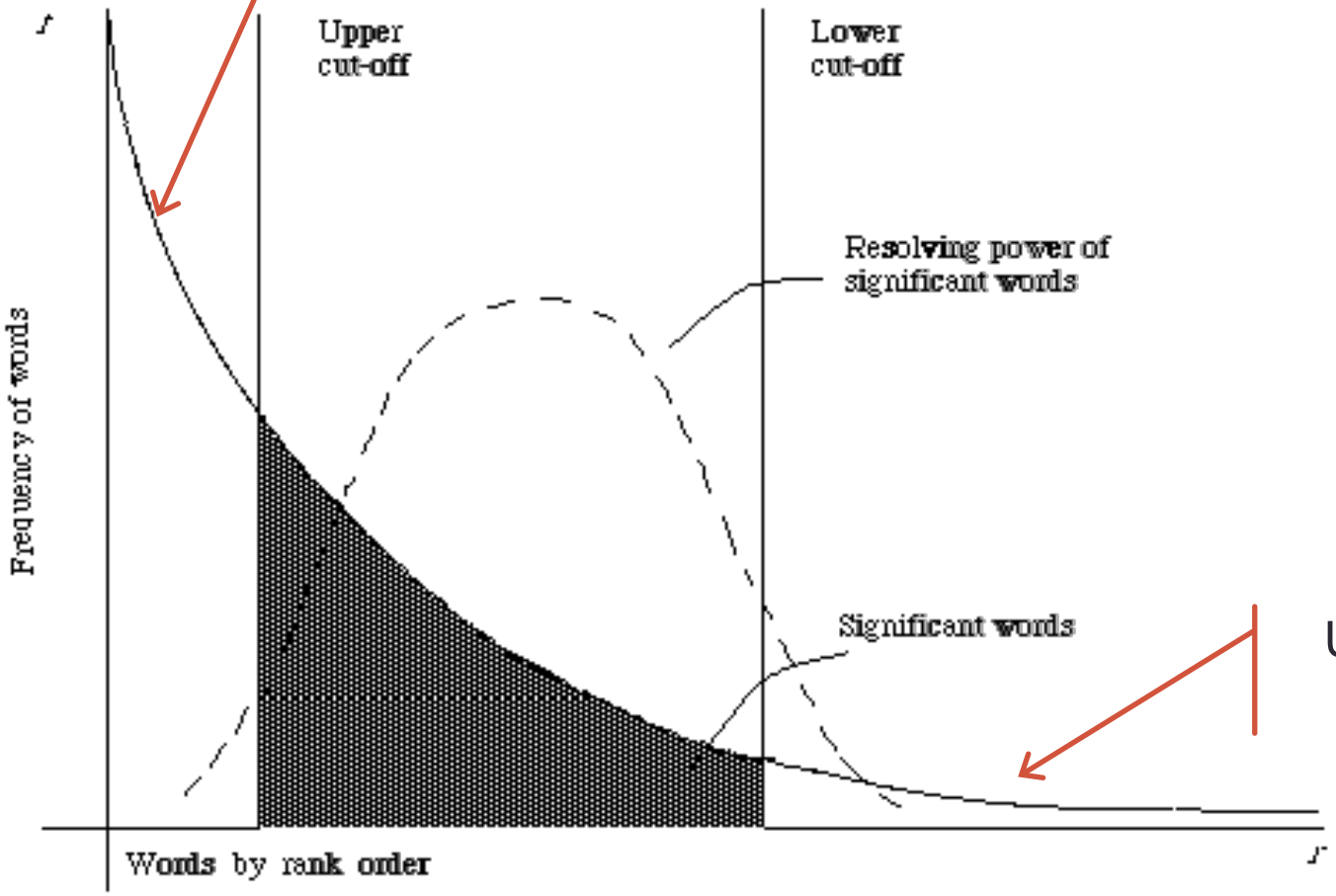
Izvor: http://nlp.stanford.edu/fsnlp/intro/fsnlp-slides-ch1.pdf

# IMPLIKACIJE ZIPF-OVOG ZAKONA

- Reči pri vrhu tabele frekventnosti pretstavljaju značajan procenat svih reči u korpusu, ali su semantički (gotovo) beznačajne
  - Primeri: the, a, an, we, do, to
- Reči pri dnu tabele frekventnosti čine najveći deo vokabulara datog korpusa, ali se vrlo retko pojavljuju u dokumentima
  - Primer: dextrosinistral, juxtapositional
- Ostale reči su one koje najbolje reprezentuju korpus i treba ih uključiti u VSM model

# IMPLIKACIJE ZIPF-OVOG ZAKONA

Ukloniti reči koje ne nose značenje




Ukloniti veoma malo zastupljene reči

Izvor slike:
http://www.dcs.gla.ac.uk/Keith/Chapter.2/Ch.2.html

# STOP-WORDS

- Alternativni / komplementarni pristup za eliminisanje reči koje nisu od značaja za analizu korpusa
- Stop-words su reči koje (same po sebi) ne nose informaciju
- Procenjuje se da čine 20-30% reči u (bilo kom) korpusu
- Ne postoji univerzalna stop-words lista
  - pregled često korišćenih listi: http://www.ranks.nl/stopwords
- Potencijalni problem pri uklanjanju stop-words:
  - gubitak originalnog značenja i strukture teksta
  - primeri: “this is not a good option” -> “option”

    “to be or not to be” -> null

# LEMATIZACIJA I STEMOVANJE

- Dva pristupa za smanjenje varijabiliteta reči izvučenih iz nekog teksta, kroz svođenje reči na njihov osnovni / koreni oblik
- Stemovanje (*stemming*) koristi heuristiku i statistička pravila za odsecanje krajeva reči (tj poslednjih nekoliko karaktera), gotovo bez razmatranja lingvističkih karakteristika reči
  - Npr., argue, argued, argues, arguing -> argu
- Lematizacija (lemmatization) koristi morfološki rečnik i primenjuje morfološku analizu reči, kako bi svela reč na njen osnovni oblik (koren reči definisan rečnikom) koji se naziva lema
  - Npr., argue, argued, argues, arguing -> argue

# VSM: ODREĐEVANJE TEŽINA TERMINA

- Postoje različiti pristupi za dodelu težina terminima u VSM-u
- Jednostavni i široko korišćeni pristupi:
  - Binarne težine
  - Term Frequency (TF)
  - Inverse Document Frequency (IDF)
  - TF-IDF

# VSM: BINARNE TEŽINE

- Težine uzimaju vrednosti 0 ili 1, zavisno od toga da li je dati termin prisutan u razmatranom dokumentu ili ne

Primer:

- Doc1: Text mining is to identify useful information.
- Doc2: Useful information is mined from text.
- Doc3: Apple is delicious.

| | text | information | identify | mining | mined | is | useful | to | from | apple | delicious |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Doc1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Doc2 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |

# VSM: TERM FREQUENCY

- Term Frequency (TF) predstavlja broj pojavljivanja datog termina u razmatranom dokumentu
- Ideja: što se termin češće pojavljuje u dokumentu, to je značajniji za taj dokument

TF(t) = c(t,d)

*c(t,d)* - broj pojavljivanja termina *t* u dokumentu *d*

# VSM: Inverse Document Frequency

- Ideja: dodeliti veće težine neuobičajenim terminima tj. onima koji nisu toliko prisutni u korpusu
- IDF se određuje na osnovu kompletnog koprusa i opisuje korpus kao celinu, a ne pojedinačne dokumente
- Izračunava se primenom formule:

$$IDF(t) = 1 + \log(N/df(t))$$

N – broj dokumenata u korpusu

df(t) – broj dokumenata koji sadrži termin *t*

# VSM: TF–IDF

- Ideja: vrednovati one termine koji nisu uobičajeni u korpusu (relativno visok IDF), a pri tome imaju nezanemarljiv broj pojavljivanja u datom dokumentu (relativno visok TF)
- Najviše korišćena metrika za ‘vrednovanje’ termina u VSM-u
- Izračunava se primenom formule:

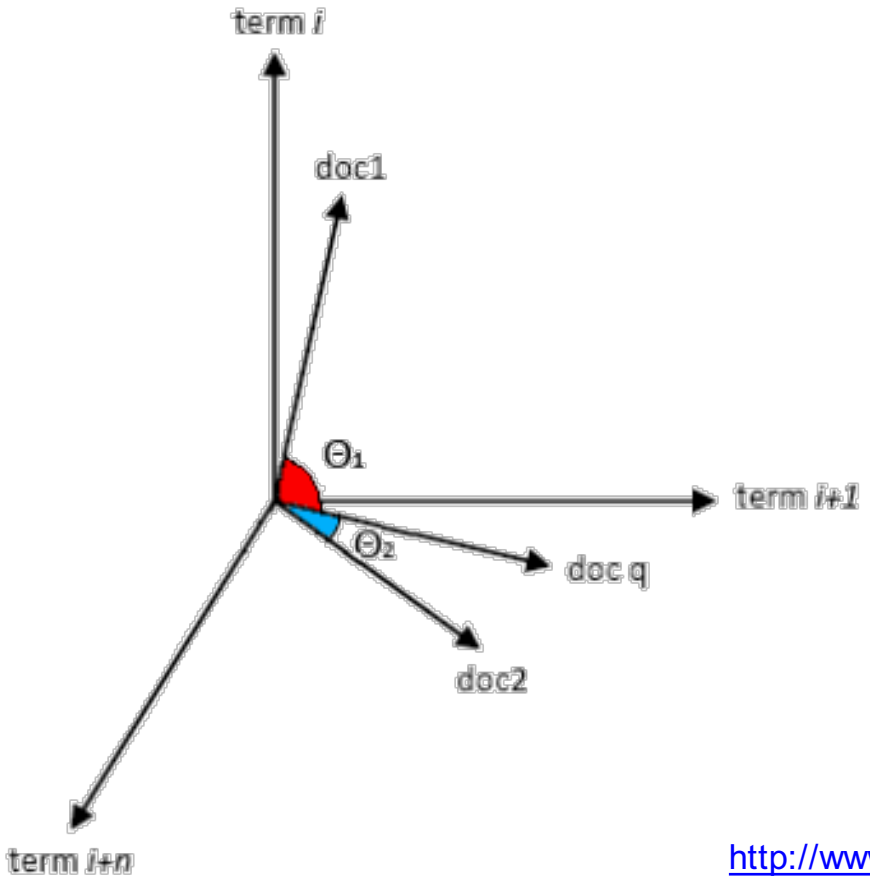
$$\text{TF-IDF}(t) = \text{TF}(t) \times \text{IDF}(t)$$

- Postoje različiti načini za kombinovanje TF i IDF metrika; najjednostavniji način:

$$\text{TF-IDF}(t) = \text{tf}(t) * \log(N/\text{df}(t))$$

# VSM: PROCENA SLIČNOSTI DOKUMENATA

- Osnovno pitanje: koju metriku koristiti za procenu sličnosti dokumenata tj vektora kojima su dok. predstavljeni?
- Najčešće korišćena metrika: kosinusna sličnost (*cosine similarity*)

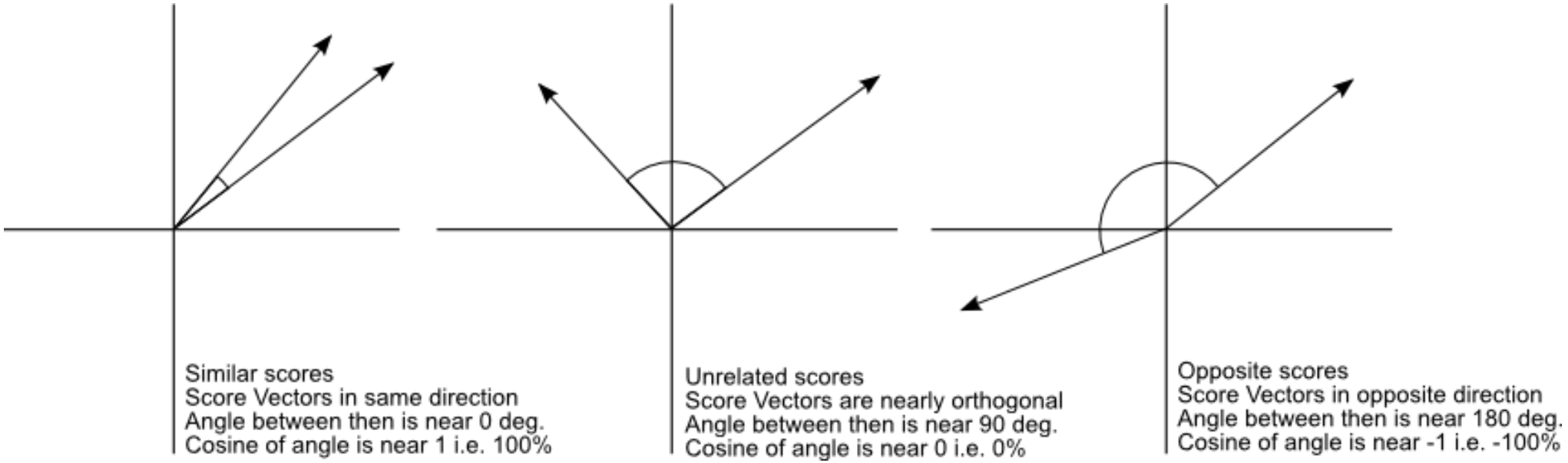



Izvor slike:
http://www.ascilite.org.au/ajet/ajet26/ghauth.html

# KOSINUSNA SLIČNOST

$$\cos(d_i,d_j) = V_i \times V_j / (||V_i|| \, ||V_j||)$$

$V_i$ i $V_j$ su vektori koji odgovaraju dokumentima $d_i$ i $d_j$




Izvor slike:
http://blog.christianperone.com/?p=2497

# PREDNOSTI I NEDOSTACI VSM MODELA

- Prednosti
  - Intuitivan
  - Jednostavan za implementaciju
  - U studijama i praksi se pokazao kao vrlo efektan
- Nedostaci
  - Nerealna pretpostavka nezavisnosti termina u tekstu
  - Zahteva puno angažovanja oko usklađivanja (*tuning*) parametara modela:
    - izbor metrike za računanje težine termina
    - izbor metrike za računanje sličnosti vektora (dokumenata)

# PROCESIRANJE TEKSTA U JAVA-I

Najviše korišćeni Java frameworks za procesiranje i analizu teksta:

- Stanford CoreNLP: http://nlp.stanford.edu/software/corenlp.shtml
- Apache OpenNLP: http://opennlp.apache.org/
- LingPIPE: http://alias-i.com/lingpipe/
- GATE: http://gate.ac.uk/

# ZAHVALNICE

Ovi slajdovi su delimično zasnovani na:

- Predavanjima na temu VSM modela pripremljenim za Text Mining kurs @ Uni. of Virginia (link)
- Prezentaciji “Introduction to Text Mining” preuzetoj sa SlideShare.net (link)

(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3I