

# OSNOVE TEXT MINING-A

---

Jelena Jovanovic

Email: [jeljov@gmail.com](mailto:jeljov@gmail.com)

Web: <http://jelenajovanovic.net>

# ŠTA JE TEXT MINING (TM)?

- Primena računarskih metoda i tehnika u cilju *ekstrakcije relevantnih informacija* iz teksta
- 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

# DOMENI PRIMENE TM-A

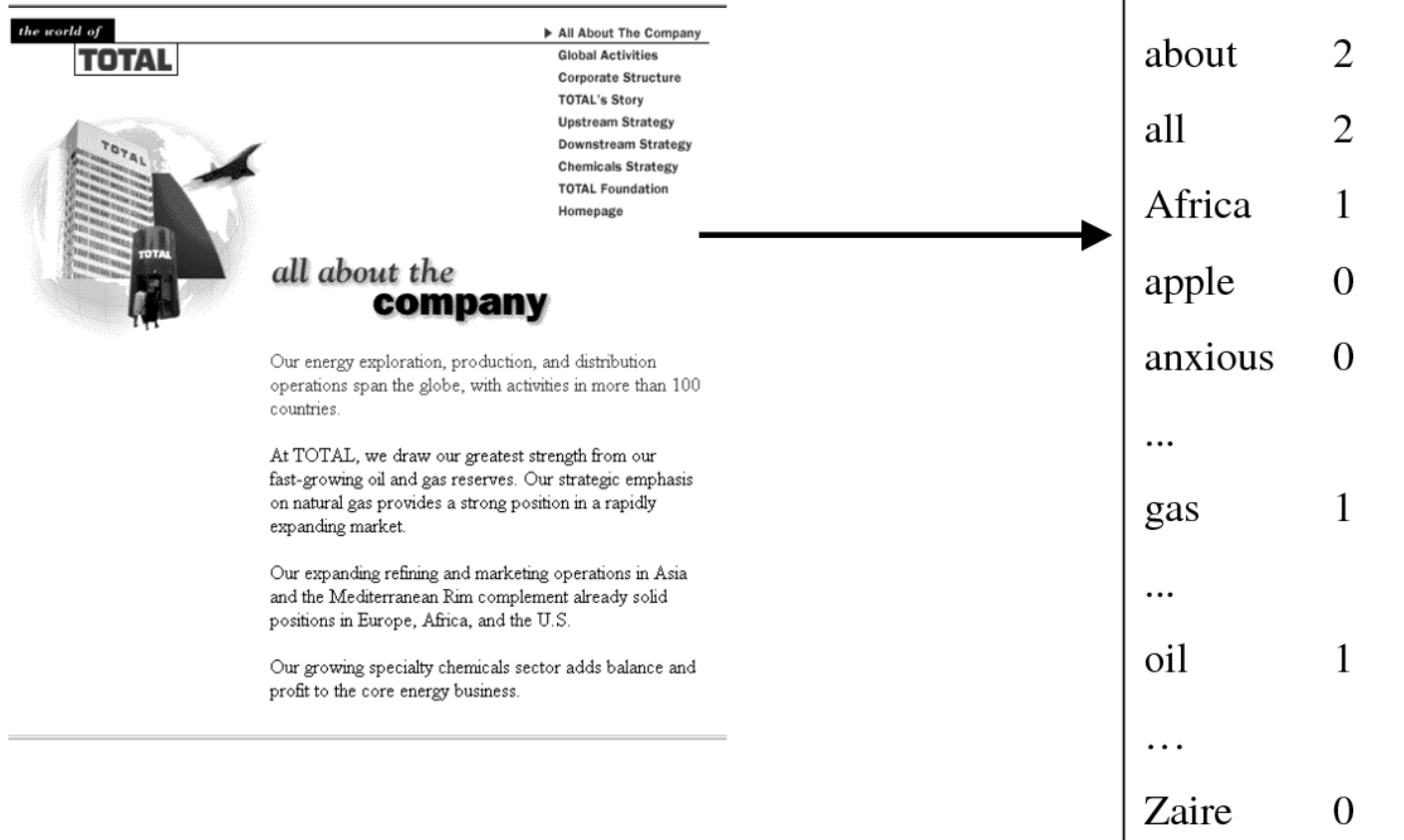
- Klasifikacija dokumenata\*
- Klasterizacija / organizacija dokumenata
- Sumarizacija dokumenata
- Vizuelizacija korpusa (najčešće u svrhe lakšeg pretraživanja dokumenata)
- Predikcije (npr. predviđanje cene akcija na osnovu analize novinskih članaka i finansijskih izveštaja)
- Generisanje preporuka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

# 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



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

# VECTOR SPACE MODEL

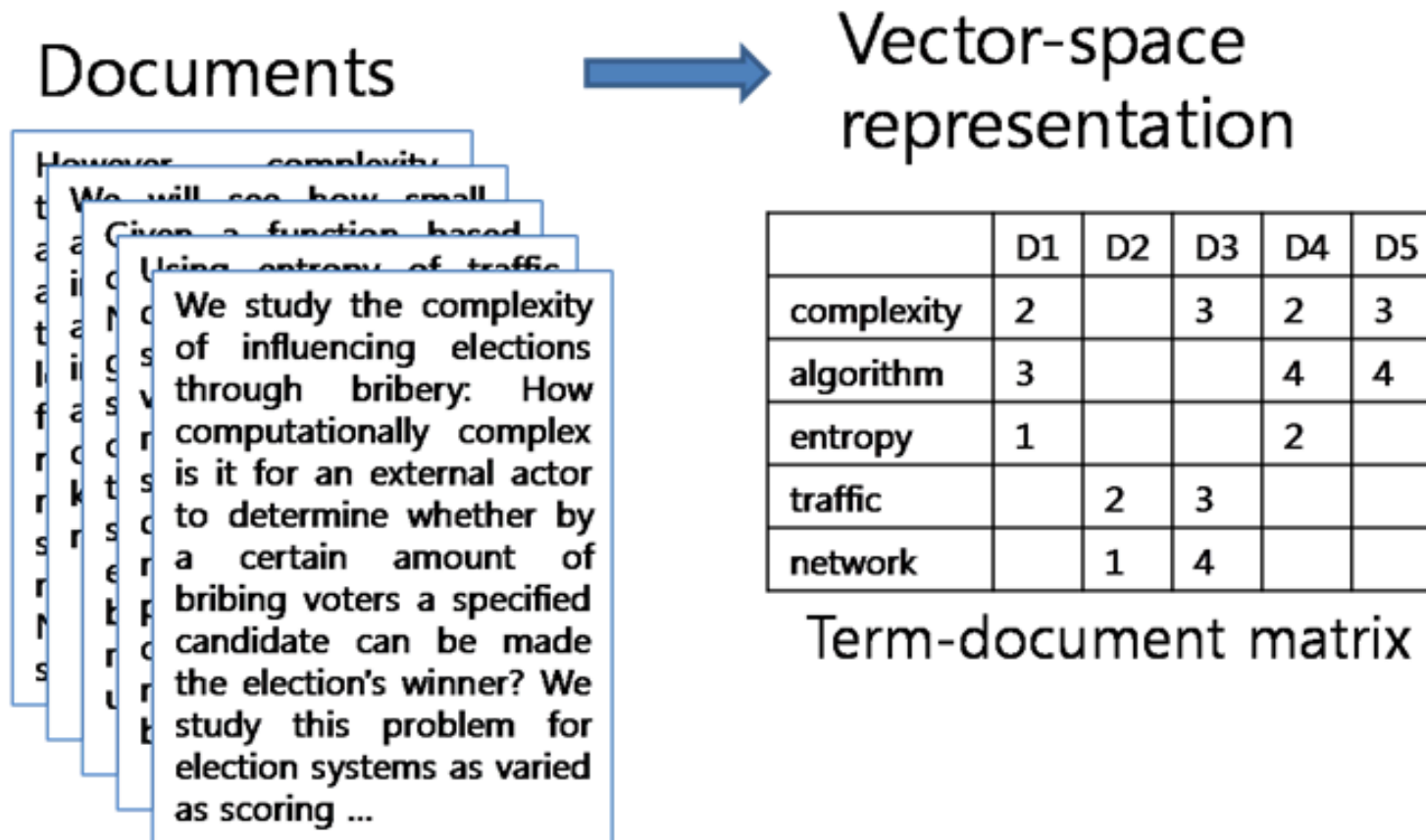
- Generalizacija Bag of Words modela
  - umesto fokusa isključivo na pojedinačne reči, fokus je na *termine*, pri čemu termin može biti jedna reč ili niz reči
  - umesto da se kao mera relevantnosti termina za dati dokument koristi isključivo učestanost pojavljivanja termina u tekstu, koriste se i drugi 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:  $d = \{w_1, w_2, \dots, w_n\}$ , gde su  $w_i$  težine pridružene terminima  $t_i$
- Ovako kreirani vektori predstavljaju osnovu za formiranje matrice termina i dokumenata (*Term Document Matrix*)



# VSM: TERM DOCUMENT MATRIX



# 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: transformirati 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

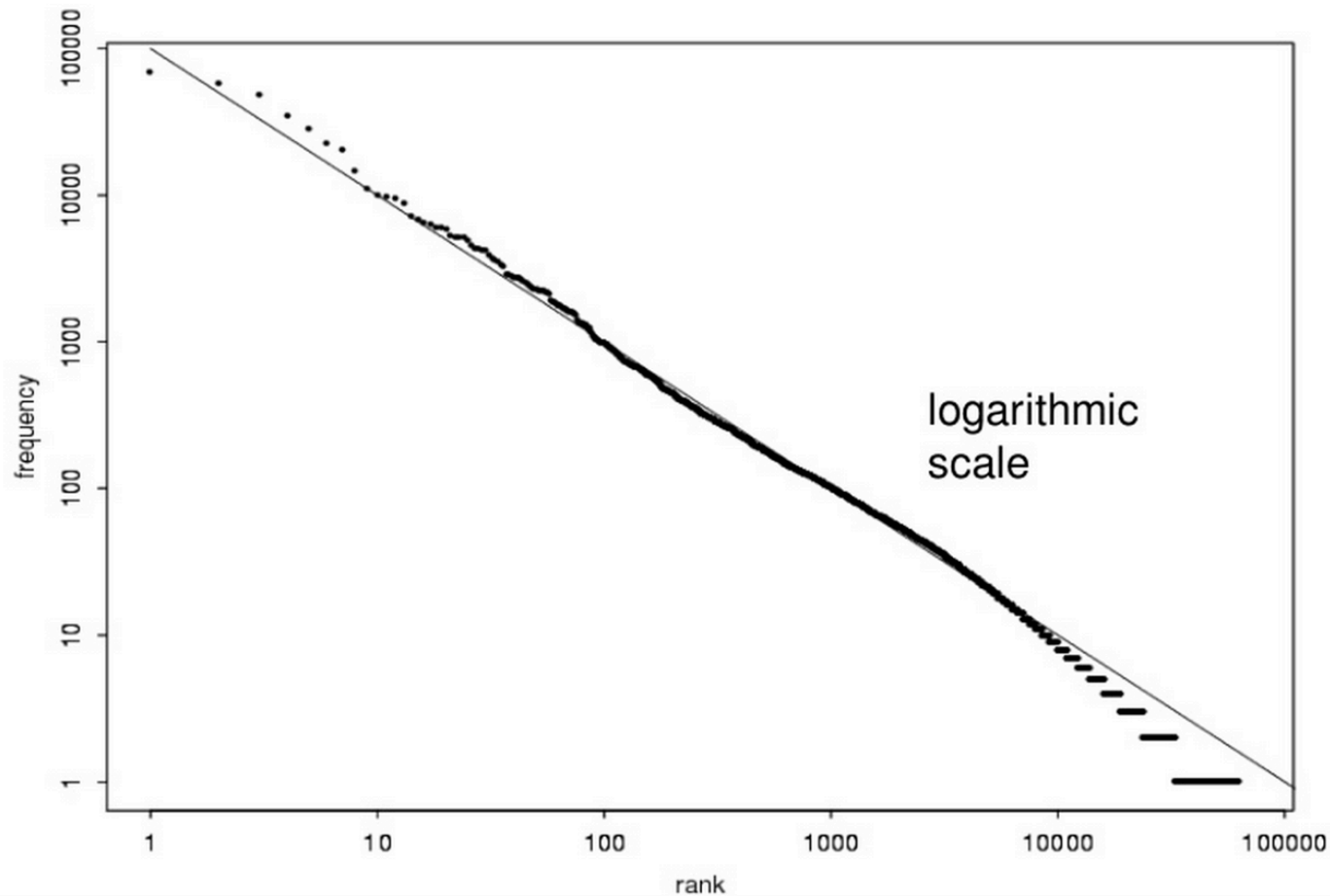
# STOP-WORDS

- 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

# 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
- Formalizovano *Zipf*-ovim zakonom:  
frekvencija bilo koje reči u datom korpusu je obrnuto proporcijalna njenom rangu u tabeli frekvencija (tog korpusa)

# ILUSTRACIJA ZIPF-OVOG PRAVILA



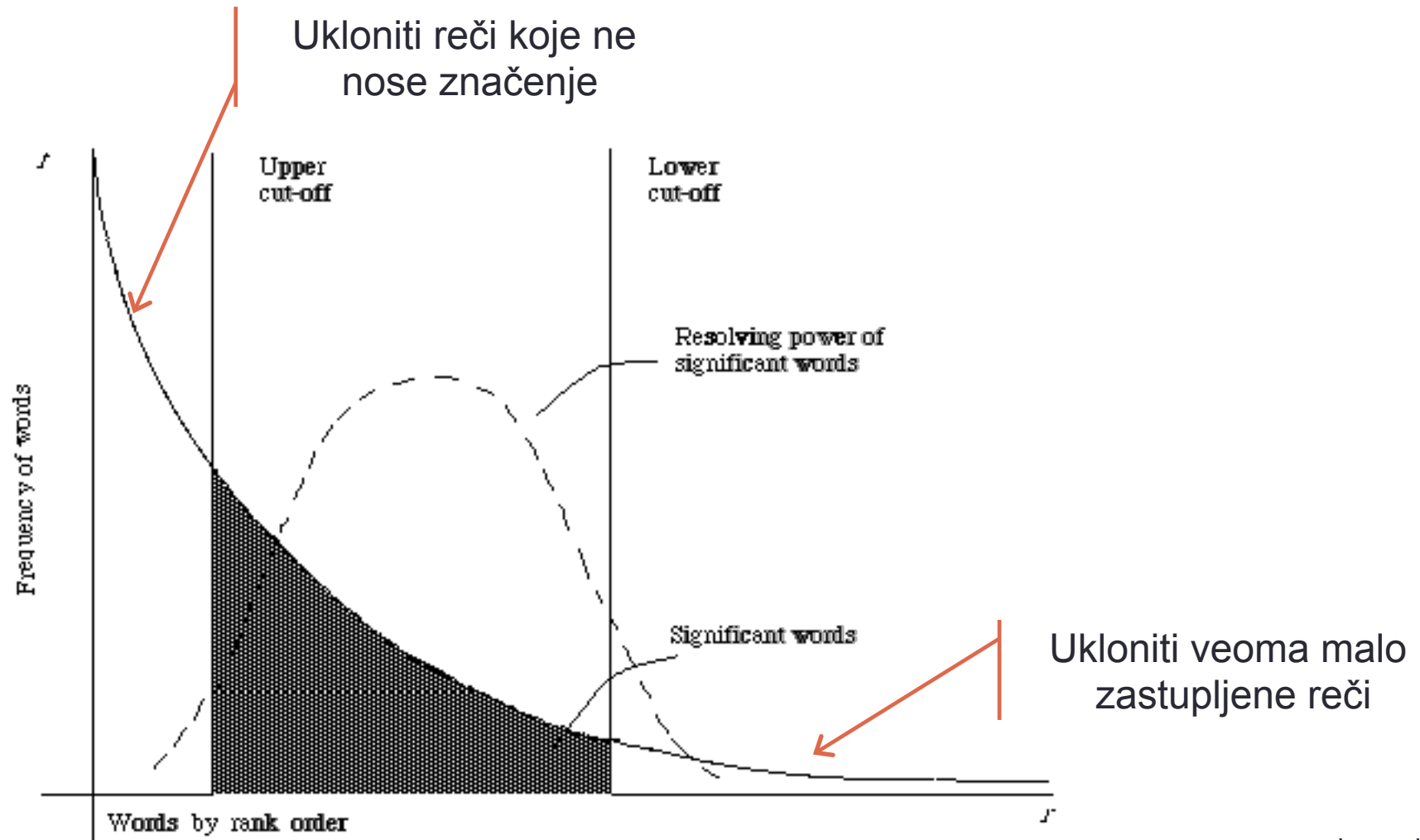
Word frequency in the [Brown Corpus](#) of American English text

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

# IMPLIKACIJE ZIPF-OVOG ZAKONA

- Reči pri vrhu tabele frekventnosti predstavljaju 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
- Ostale reči su one koje najbolje reprezentuju korpus i treba ih uključiti u VSM model

# IMPLIKACIJE ZIPF-OVOG ZAKONA





# LEMATIZACIJA I STEMOMANJE

- 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:

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

$N$  – 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:

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

# 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](#))