Osnove Text Mining-a

Nikola Milikić

Email: nikola.milikic@gmail.com

Web: http://nikola.milikic.info

Autorstvo

- Originalna prezentacija je kreirana od strane:

prof. Jelene Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Šta je Text Mining (TM)?

- Šta je Text Mining (TM) i zašto je značajan?
- Izazovi za TM
- Osnovni koraci TM procesa
  - Preprocesiranje teksta
  - Transformacija teksta i kreiranje atributa
  - *Data mining* nad transformisanim tekstom
- Pregled korisnih linkova i materijala

Šta je Text Mining (TM)?

- Primena računarskih metoda i tehnika u cilju *ekstrakcije relevantnih informacija* iz teksta
- *Automatsko otkrivanje paterna* sadržanih u tekstu
- *Otkrivanje novih, nepoznatih informacija i znanja*, kroz *automatizovanu ekstrakciju informacija* iz velikog broja *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,
  - …

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, usavršavanje i primenu ovakvih metoda

Oblasti primene TM-a

- Klasifikacija dokumenata*
- Klasterizacija / organizacija dokumenata
  - najčešće u svrhe lakšeg pretraživanja dokumenata
- Sumarizacija dokumenata
- Predikcije
  - npr. predviđanje cena akcija na osnovu analize novinskih članaka i sadržaja razmenjenih na društvenim mrežama
- Reputation management
- Generisanje preporuka
  - preporuke 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, …

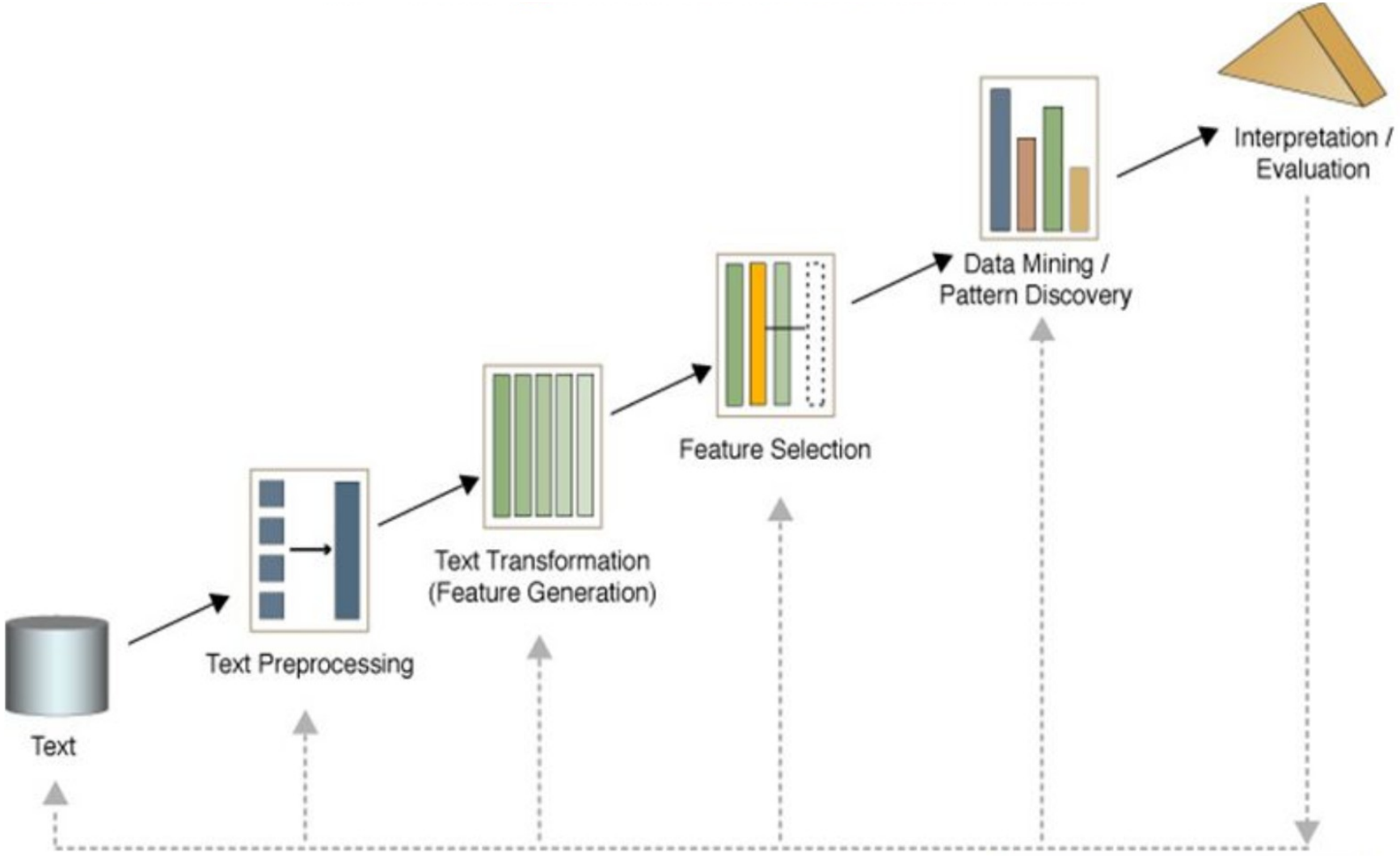
Izazov za TM: Složenost nestrukturiranog teksta

- 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

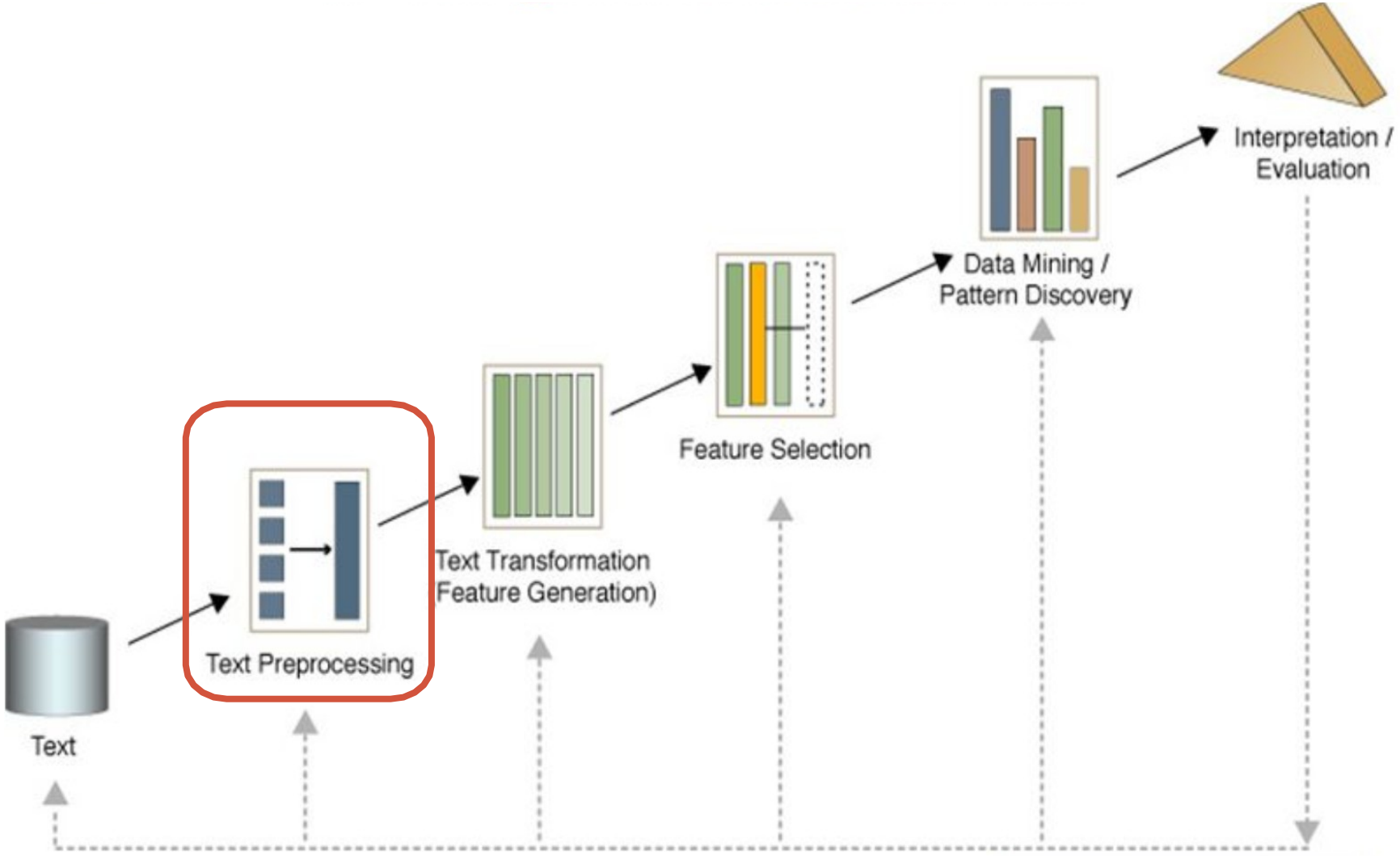
- Primena tehnika m. učenja zahteva veliki broj anotiranih dokumenata za formiranje skupa za trening, što je vrlo skupo
  - Takav trening set je potreban za klasifikaciju dokumenata, kao i ekstrakciju entiteta, relacija, događaja
- Visoka dimenzionalnost problema: dokumenti su opisani velikim brojem 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

Osnovni koraci TM procesa



Izvor slike: http://slideplayer.info/slide/10731087/

Preprocesiranje teksta



Preprocesiranje teksta

- Svrha: redukovati skup reči na one koje su potencijalno najznačajnije za dati korpus
- Neophodan prvi korak bilo kog oblika analize teksta

Preprocesiranje teksta

Najčešće obuhvata:

- normalizaciju teksta
- odbacivanje tzv. stop-words
- POS tagging
- svođenje reči na koreni oblik
- odbacivanje termina sa veoma malom i/ili velikom učestanošću u korpusu

Normalizacija teksta

- Cilj: transformisati različite oblike jednog istog termina u osnovni, ‘normalizovani’ oblik
- Na primer:
  - Apple, apple, APPLE → apple
  - Intelligent Systems, Intelligent-systems, Inteligent systms → intelligent systems

Normalizacija teksta

Pristup:

- Primena jednostavnih pravila:
  - Obrisati sve znake interpunkcije (tačke, crtice, zareze,…)
  - Prebaciti sve reči da budu napisane malim slovima
- Eliminisanje sintaksnih grešaka (misspellings)
  - Primenom raspoloživih misspelling korpusa
- Primena rečnika, npr. WordNet, za zamenu srodnih termina zajedničkim opštijim terminom / konceptom
  - Npr. automobile, car → vehicle

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

POS tagging

Anotacija reči u tekstu tagovima koji ukazuju na vrstu reči (Part of Speech - POS): imenica, zamenica, glagol, …

Primer:

"And now for something completely different"

[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'), ('completely', 'RB'), ('different', 'JJ')]

RB -> Adverb; NN -> Noun, singular or mass; IN -> Preposition, …

Kompletna lista Penn Treebank POS tagova (standardni skup POS tagova za engleski jezik) je raspoloživa ovde.

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 am, are,

  is → be

Eliminisanje termina sa suviše velikom/malom učestanošću

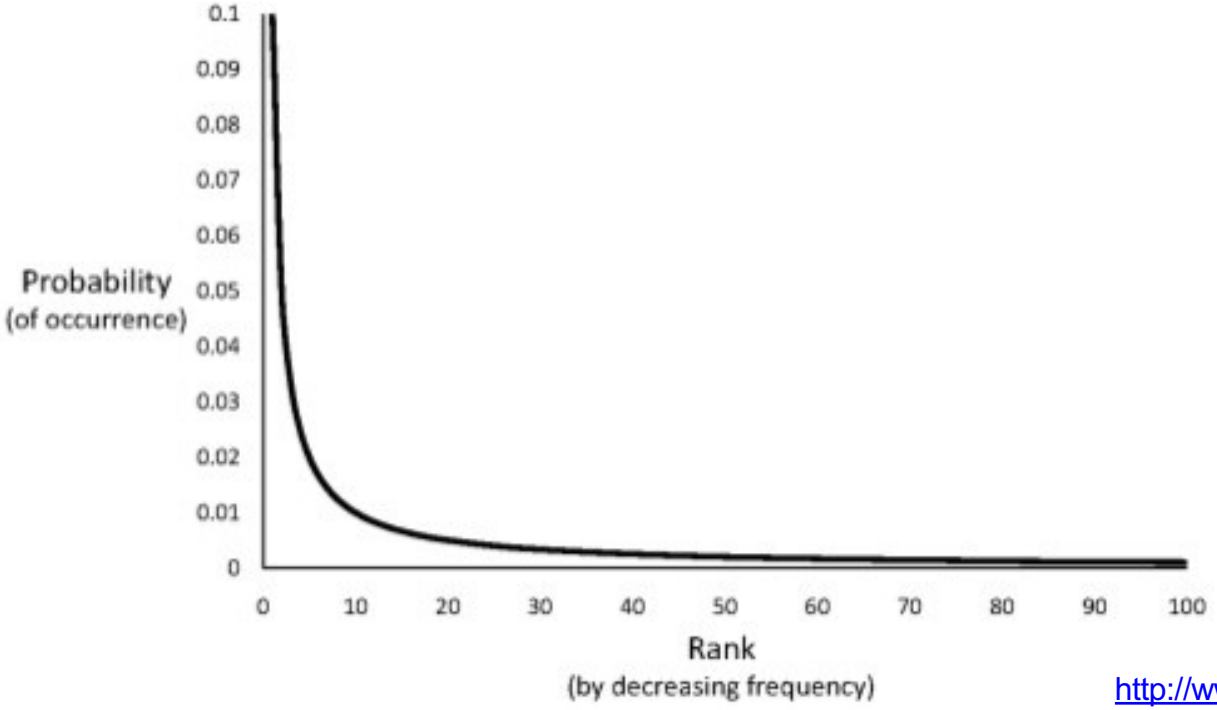
Empirijska zapažanja (u brojnim korpusima):

- Veliki broj reči ima veoma malu frekvencu pojavljivanja
- Jako mali broj reči se veoma često pojavljuje u tekstu

Učestanost termina u tekstu

Empirijska zapažanja formalizovana *Zipf*-ovim pravilom:

Frekvenca bilo koje reči u velikom korpusu je obrnuto proporcijalna njenom rangu u tabeli frekvencija (tog korpusa).



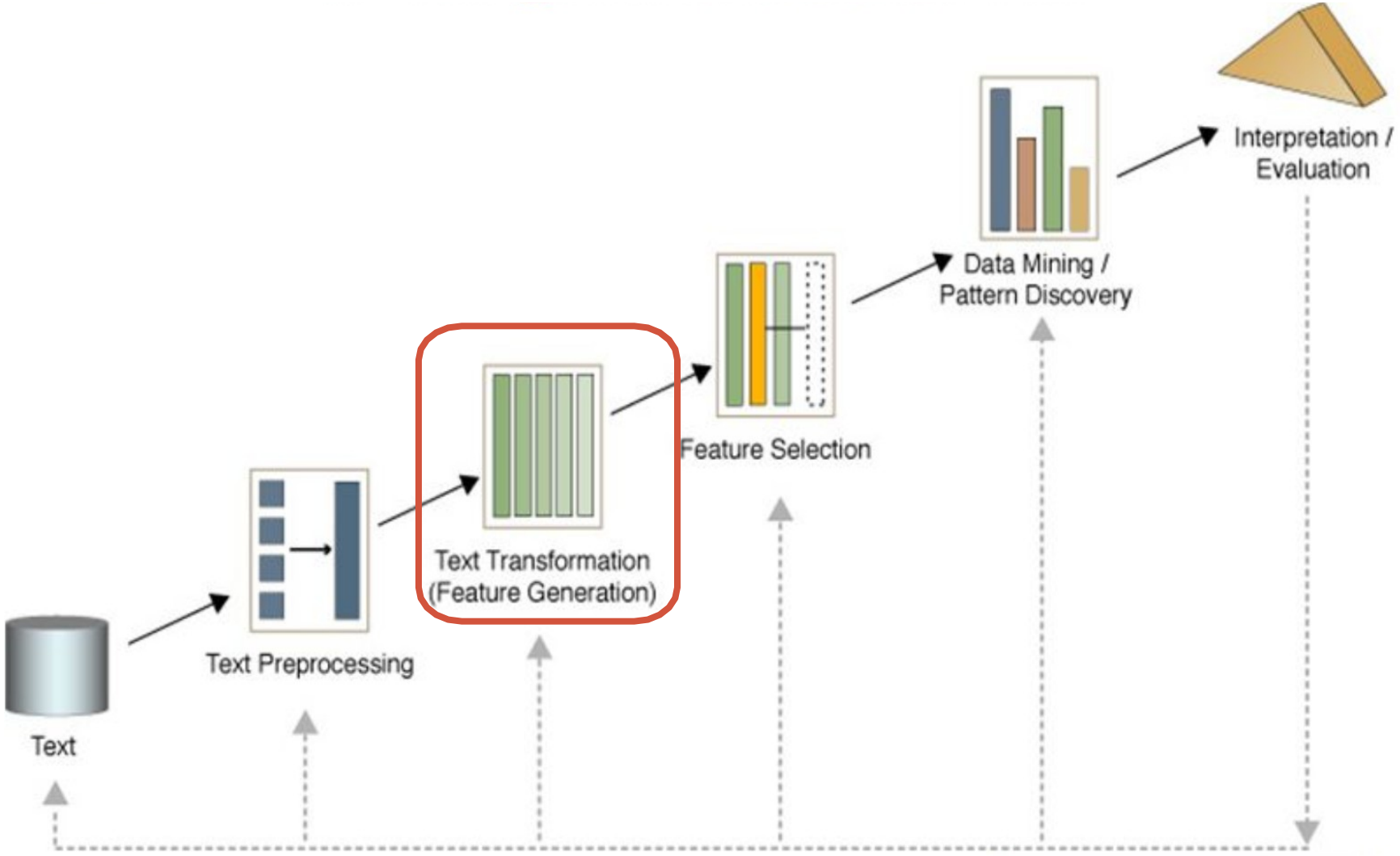
Word	Freq f	Rank r
the	3332	1
and	2972	2
a	1775	3
he	877	10
but	410	20
be	294	30
there	222	40
one	172	50
about	158	60
more	138	70
never	124	80
oh	116	90
two	104	100

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

Implikacije Zipf-ovog zakona

- Reči pri vrhu tabele frekventnosti pretstavljaju značajan procenat svih reči u korpusu, ali su semantički (gotovo) beznačajne
  - Primeri: the, of, 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 model

Transformacija teksta (Feature Creation)



Transformacija teksta

Postupak transformacije nestrukturiranog teksta u strukturirani format pogodan za korišćenje u okviru:

- Statističkih metoda i tehnika
  - npr. topic modeling
- Metoda i tehnika mašinskog učenja
  - klasifikacija, klasterizacija
- Drugih metoda ekstrakcije informacija iz teksta
  - npr. na grafu zasnovane metode za ekstrakciju ključnih termina

Bag of Words (BOW) & Vector Space Model (VSM)

Bag of Words model

- Predstavlja tekst kao prost skup (“vreću”) reči.
- Pristup zasnovan na sledećim pretpostavkama:
  - reči su međusobno nezavisne,
  - redosled reči u tekstu je nebitan.
- Iako je zasnovan na nerealnim pretpostavkama i vrlo je 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

**korpus* je kolekcija dokumenata koji su predmet analize

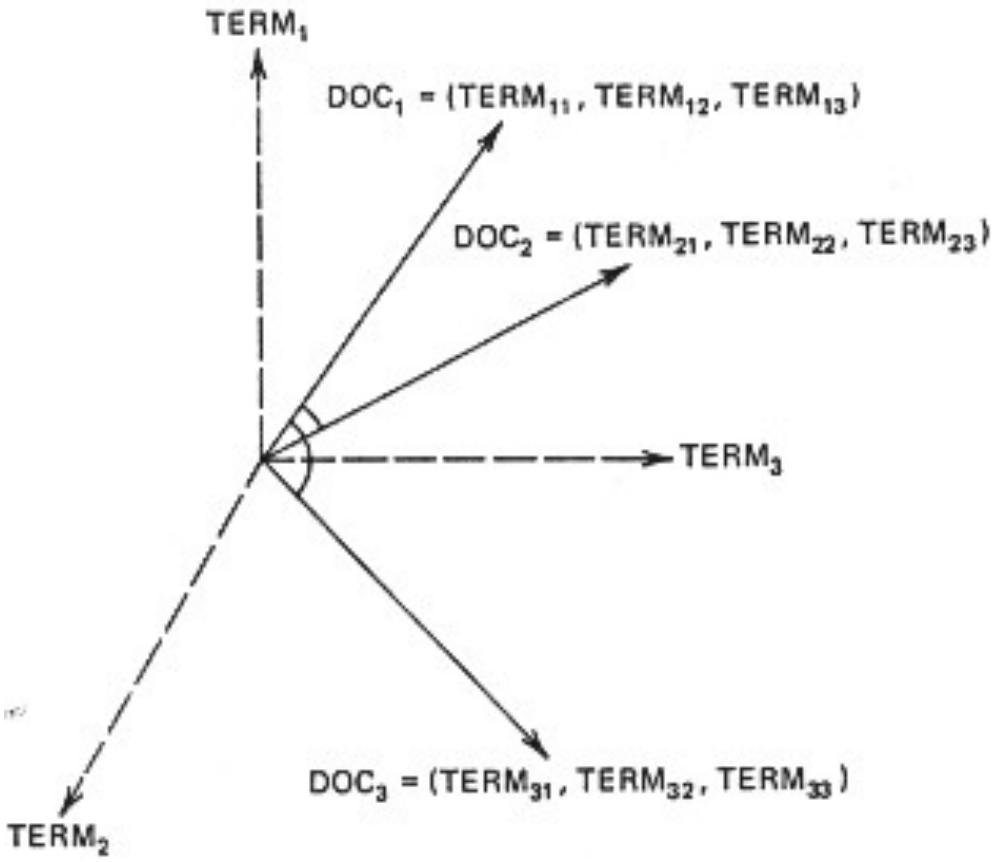
Bag of Words model

Primer:

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

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	2	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

Bag of Words model



Dokumenti kao vektori u n-dimenzionalnom prostoru reči iz korpusa, gde je n broj jedinstvenih reči u korpusu

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

Vector Space Model (VSM)

- Generalizacija Bag of Words modela
  - umesto fokusa isključivo na pojedinačne reči, fokus je na *termine* (n-grams), pri čemu termin može biti jedna reč (uni-grams) ili niz reči (bi-grams, tri-grams,…)
  - 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 (VSM)

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

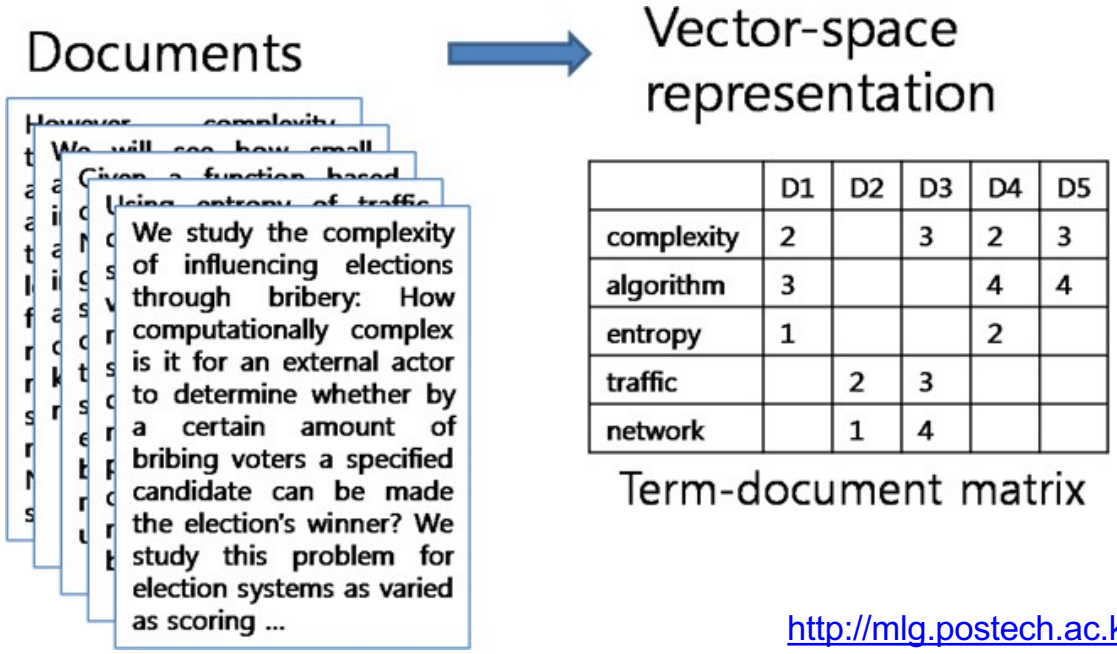
$$d = \{ w_1, w_2, \ldots, w_n \}$$

, gde su w_i težine pridružene terminima t_i koje odražavaju značajnost tih termina za dati dokument

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

VSM: Term Document Matrix

- Term Document Matrix (TDM) je matrica 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: Procena značajnosti termina

- Postoje različiti pristupi za procenu značajnosti termina, tj. dodelu težina terminima u TDM matrici
- Jednostavni i široko korišćeni pristupi:
  - Binarne težine
  - Term Frequency (TF)
  - Inverse Document Frequency (IDF)
  - TF-IDF

Procena značajnosti termina: Binarne težine

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

Procena značajnosti termina: 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

Procena značajnosti termina: 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:

$$IDF(t) = log(N / df(t))$$

N – broj dokumenata u korpusu

$df(t)$ – broj dokumenata koji sadrže termin t

Procena značajnosti termina: 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:

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

Procena značajnosti termina: Normalizacija

- Vrednosti (težine) u TDM matrici se najčešće normalizuju kako bi se neutralizao efekat različite dužine dokumenata
- Tipični pristupi normalizaciji TDM matrice:
  - svaki element (težina) matrice se podeli normom (vektora) termina koji odgovara tom elementu
  - najčešće se koriste Euklidska (L2) ili Menhetn (L1) norma

$$\|x\|_e = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

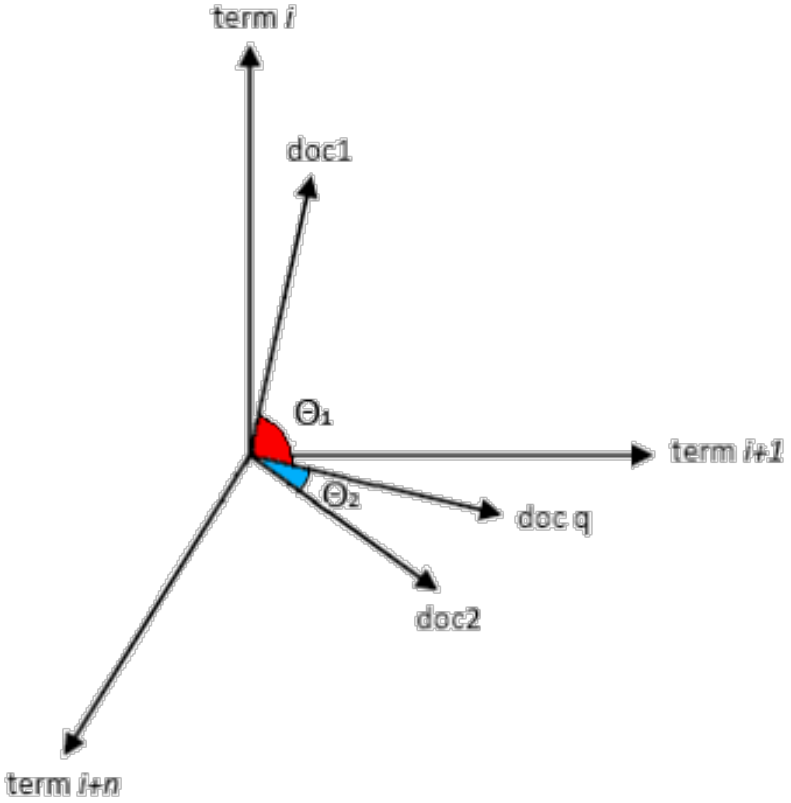
Prednosti i nedostaci VSM modela

Prednosti:

- Intuitivan
- Jednostavan za implementaciju
- U studijama i praksi se pokazao kao vrlo efektan
- Posebno pogodan za procenu sličnosti dokumenata (osnova za klasterizaciju)

VSM: Procena sličnosti dokumenata

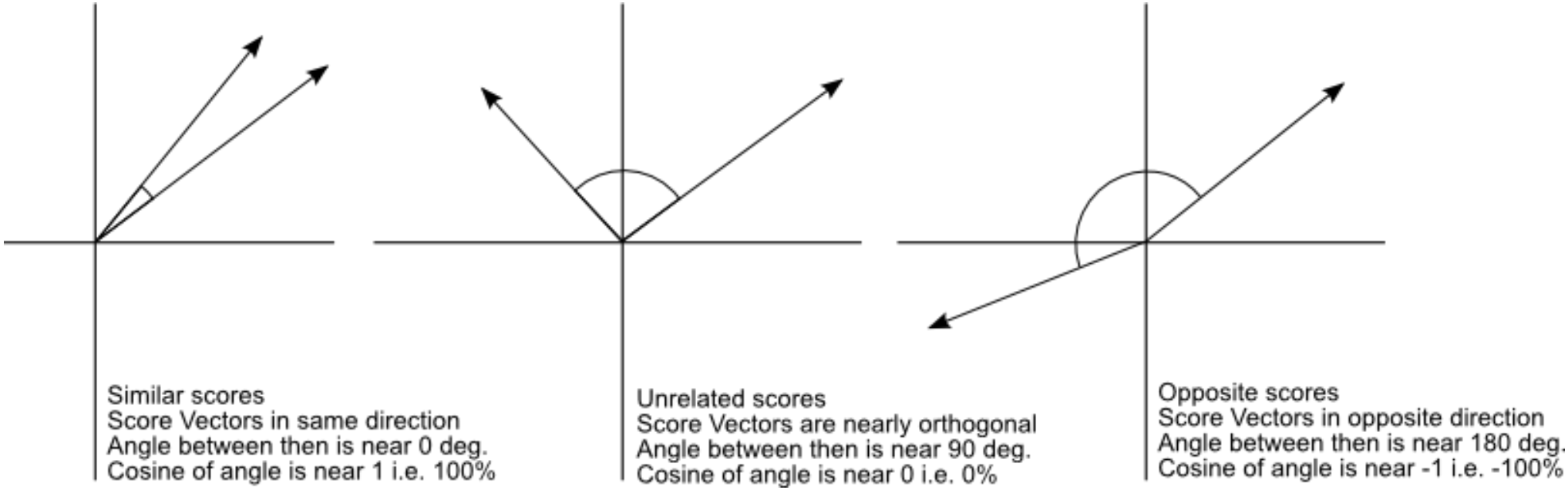
- Predstava dokumenata u formi vektora omogućuje procenu sličnosti dokumenata primenom vektorskih metrika
- Najčešće korišćena metrika: kosinusna sličnost (cosine similarity)



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

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

Procena sličnosti dokumenata: Kosinusna sličnost



S obzirom da su u VSM-u elementi vektora pozitivne vrednosti, sličnost se kreće u opsegu [0,1].

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

Prednosti i nedostaci VSM modela

Nedostaci:

- Nerealna pretpostavka nezavisnosti termina u tekstu
- Zahteva dosta angažovanja oko:
  - selekcije atributa (termina)
  - izbora najbolje metrike za procenu značajnosti (težine) termina
- Ograničenost na reči i izraze kao atribute (features)

Dodatne mogućnosti za kreiranje skupa atributa

Zavisno od konkretnog zadatka i vrste teksta, pored već pomenutih, mogu se definisati i koristiti brojni drugi atributi.

Na primer, za zadatak prepoznavanja ključnih izraza i entiteta u tekstu, obično se koriste sledeći atributi:

- dužina reči
- prisutnost velikih slova
- vrsta reči (POS)
- prisutnost specijalnih znakova
- pozicija reči u rečenici
- vrsta reči u okruženju
- …

Dodatne mogućnosti za kreiranje skupa atributa

Primer: klasifikacija kratkih segmenata teksta prema tome da li nude proizvod koji je na akciji (deal) ili ne.

Two example sentences with probability of being in a deal Web page.

Sentences	Deal probability [0,1.0]
Buy unlimited vouchers as a gift Package. Includes a 7" Google android 2.3 tablet with a 30 pin USB switch adaptor, charger and user manual lightweight and easy to use. Perfect idea for people on the go. Makes a great gift!	0.9998
Challenges address the conceptualization how e-business related knowledge is captured, represented, shared, and processed by humans and intelligent software.	0.0248

Cuzzola, J. et al. (2015) "Automated classification and localization of daily deal content from the Web," *Applied Soft Computing*, 31, 241-256.

Dodatne mogućnosti za kreiranje skupa atributa

Primer: klasifikacija kratkih segmenata teksta prema tome da li nude proizvod koji je na akciji ili ne

Korišćen skup atributa:

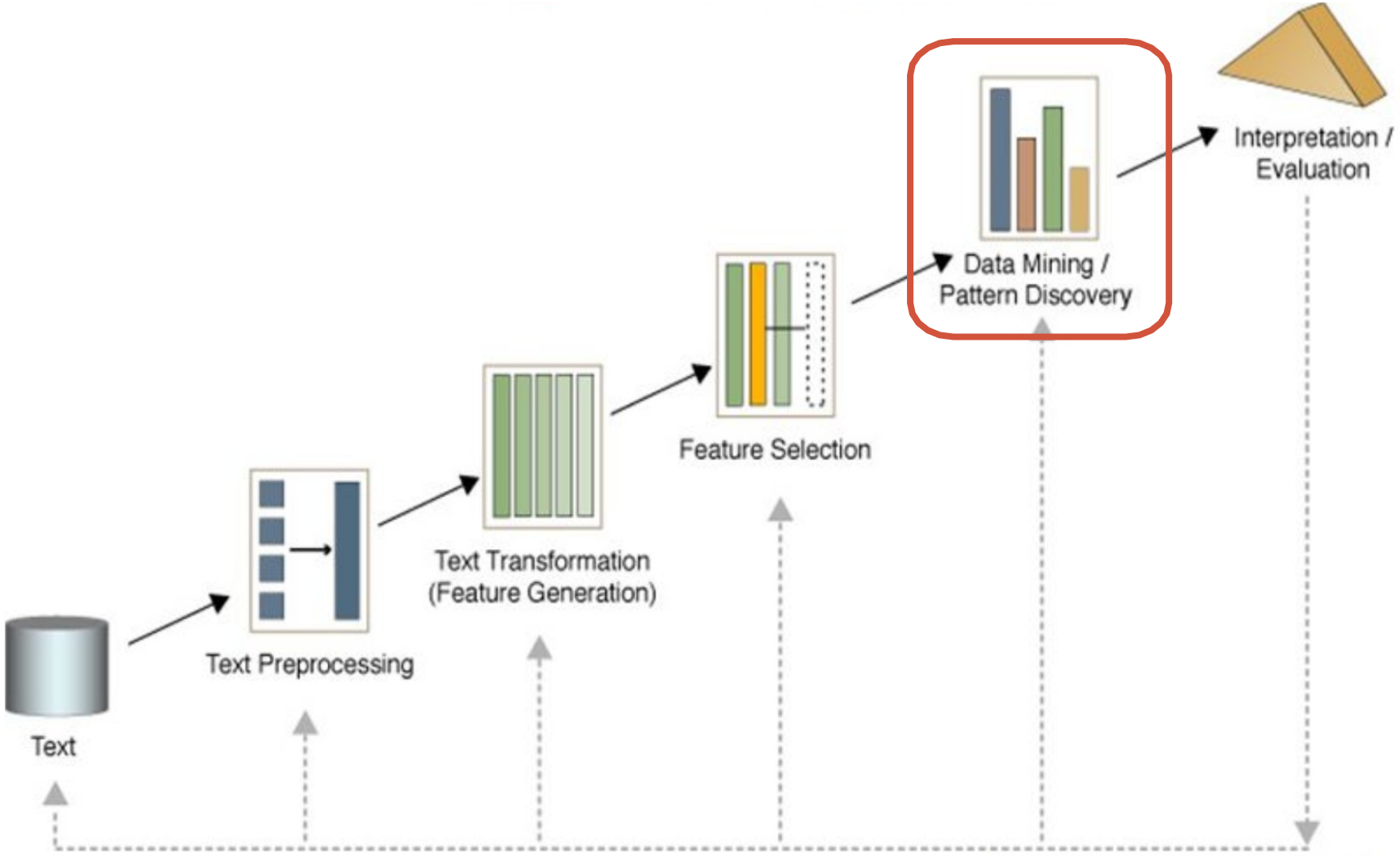
Name	Description
Words	Frequently appearing words in the sentence block and their synonyms and other related words obtained via WordNet
ner_dateI	Number of dates identified through named entity recognition (NER)
ner_organizationI	Number of organization instances identified through NER
ner_timeI	Number of time-related instances identified through NER
ner_locationI	Number of location instances identified through NER
ner_percentageI	Number of percentages identified through NER
ner_moneyI	Number of money values identified through NER
ner_personI	Number of person instances identified through NER
sym_dollarAvgI	The average dollar value identified through POS tagging
sym_percentAvgI	The average percentage identified through POS tagging
sym_CD_posI	Count of numerical values identified through POS tagging
sym_SYM_posI	Count of symbols identified through POS tagging

Cuzzola, J. et al. (2015) "Automated classification and localization of daily deal content from the Web," *Applied Soft Computing*, 31, 241-256.

Dodatne mogućnosti za kreiranje skupa atributa

Dobri saveti za kreiranje / selekciju atributa za potrebe klasifikacije teksta su dati u okviru Categorizing Customer Emails thread-a, Data Science foruma.

Data Mining / Pattern Discovery



Data mining, pattern detection

Tekst predstavljen u strukturiranom formatu, tipično u formi TDM matrice, predstavlja ulaz za različite algoritme mašinskog učenja.

- Klasifikacija
  - Tematska kategorizacija teksta
  - Detektovanje spam-a
  - Analiza sentimenta (polaritet teksta)
  - Prepoznavanje entiteta u tekstu i entity linking
  - …
- Klasterizacija
  - Detektovanje tema teksta
  - Tematska organizacija (grupisanje) dokumenata u korpusu
  - …

Primer klasifikacije teksta

Sentiment analysis

- Klasifikacija teksta prema generalno iskazanom mišljenju / stavu / osećanju (sentimentu) na pozitivne i negativne
- Primer prepoznavanja sentimenta u komentarima filmova*
  - korpus: tekstovi komentara preuzeti sa IMDB.com sajta i raspoloživi u okviru Large Movie Review Dataset-a
  - features: unigrami, bigrami, trigrami
  - korišćena metoda: logistička regresija; dodeljuje svakoj instanci vrednost u intervalu [0,1]; vrednosti > 0.5 se smatraju pozitivnim

*Diplomski rad Marka Stakića, FON, Septembar 2016

Primer klasifikacije teksta

Primer prepoznavanja sentimenta u komentarima filmova

It's the movie equivalent of that rare sort of novel where you find yourself checking to see how many pages are left and hoping there are more, not fewer. (film Pulp Fiction, 1994)

Izlaz klasifikatora (log. regresija): **0.9516**

As someone who's watched more bad movies than you can imagine, I'm mostly immune to the so-bad-it's-good aesthetic, though I can see how, viewed in a theater at midnight after a few drinks, this might conjure up its own hilariously demented reality. (film The Room, 2003)

Izlaz klasifikatora (log. regresija): **0.0111**

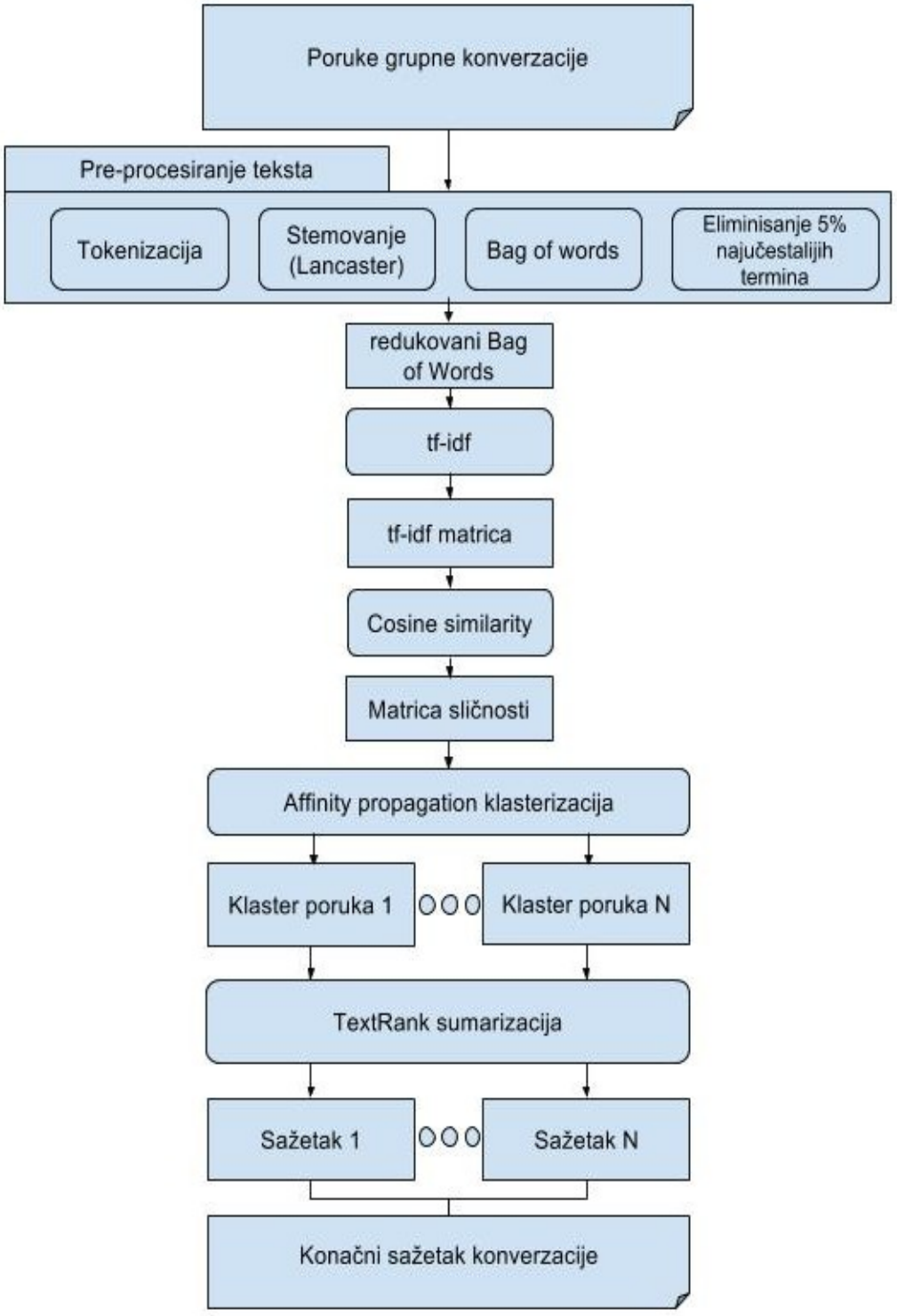
Primer kombinacije više različitih metoda ekstrakcije informacija iz teksta

Sumarizacija poruka online grupnih chat diskusija*

- Kombinuje:
  - predstavljanje teksta chat poruka u BoW formi
  - korišćenje TF-IDF metrike za procenu značajnosti reči
  - klasterizaciju poruka na osnovu procenjene sličnosti (cosine sim.)
  - sumarizaciju klastera poruka primenom TextRank algoritma

*Master rad Stefana Stanojevića, FON, Nov 2017

Sumarizacija poruka online grupnih chat diskusija



Relevantni linkovi

Softverski paketi za TM

- R paketi:
  - tm
  - tidytext
- Java frameworks:
  - Stanford CoreNLP
  - Apache OpenNLP
  - LingPIPE

Preporuke

Knjige:

- J. Silge & D. Robinson. *Text Mining with R – A Tidy Approach*. O'Reilly, 2017. E-verzija raspoloživa: http://tidytextmining.com/
- G.S. Ingersoll, T.S. Morton, A.L. Farris. *Taming Text*. Manning Pub., 2013. (sa primerima u Javi)

Interesantni projekti:

- NELL - Never Ending Language Learner (website) (NYT article) (video lecture)
- SentiStrength - automatic sentiment analysis of social web texts
- TextRank in R – R paket za sumarizaciju teksta i ekstrakciju ključnih izraza primenom TextRank algoritma (sadrži i primere primene)

Preporuke

Interesantni projekti i alati:

- Synesketch – open source biblioteka za prepoznavanje emocija u tekstu i vizuelizaciju prepozatih emocija
- TagMe – open source biblioteka i RESTful servisi za entity linking - prepoznavanje entiteta u tekstu i njihovo linkovanje sa odgovarajucim konceptima iz baze znanja (Wikipedia)

Online kurs:

- Introduction to TextAnalytics with R (on YouTube)

Preporuke

Javno dostupni izvori podataka:

- Project Gutenberg – dobar izvor javno dostupnih tekstova za analizu
  - GutenbergR – R paket za jednostavan pristup sadržajima iz Gutenberg kolekcije
- Local News Research Project data sets
- Datasets for single-label text categorization
- Anonymized discussion forum threads from 60 Coursera MOOCs
- Dataset of personal attacks in Wikipedia 'talk' pages (discussions around article content)

Osnove Text Mining-a

Nikola Milikić

Email: nikola.milikic@gmail.com

Web: http://nikola.milikic.info