OSNOVE TEXT MINING-A

Jelena Jovanovic

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

PREGLED PREDAVANJA

- Šta je Text Mining (TM) i zašto je značajan?
- Izazovi za TM
- Osnovni koraci TM procesa
  - Preprocesiranje teksta
  - Transformacija teksta i kreiranje atributa
  - Selekcija 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,
  - …

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, …

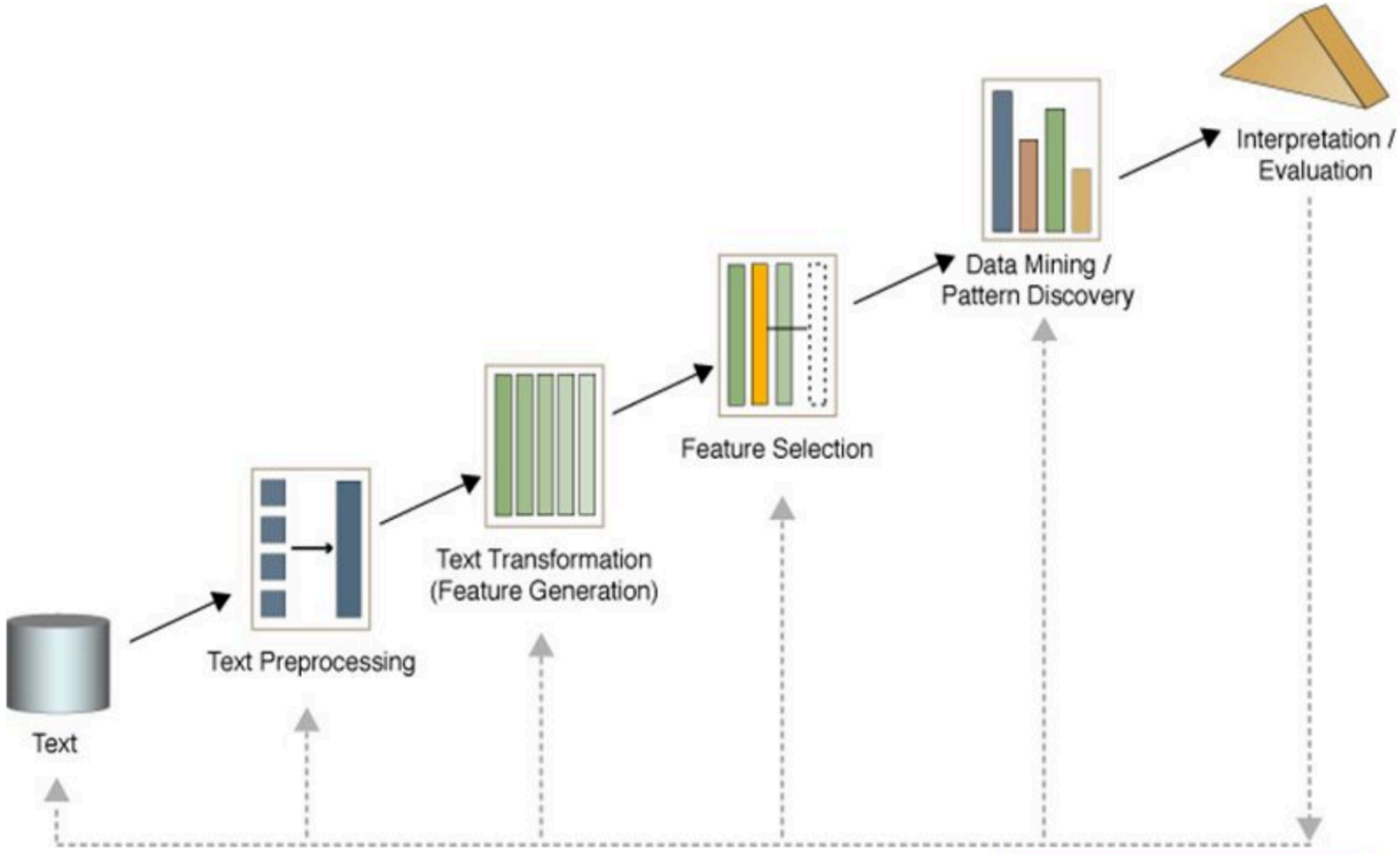
IZAZOV ZA TM: SLOŽENOST NESTRUKTURIRANOG TEKSTA

- 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

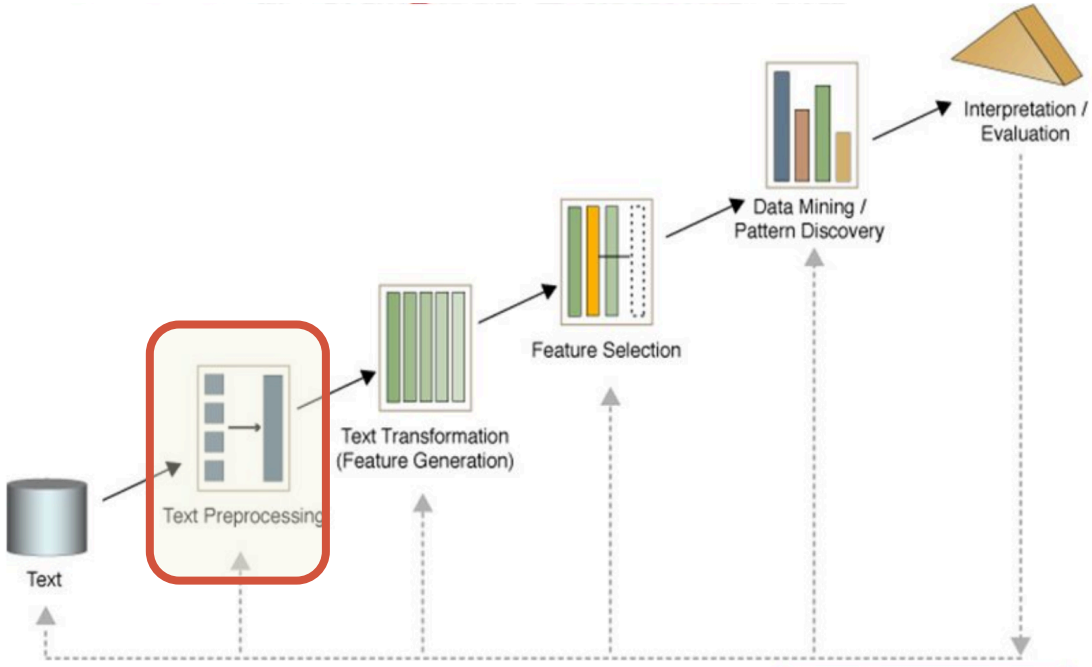
DODATNI IZAZOVI ZA TM

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

NORMALIZACIJA TEKSTA

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

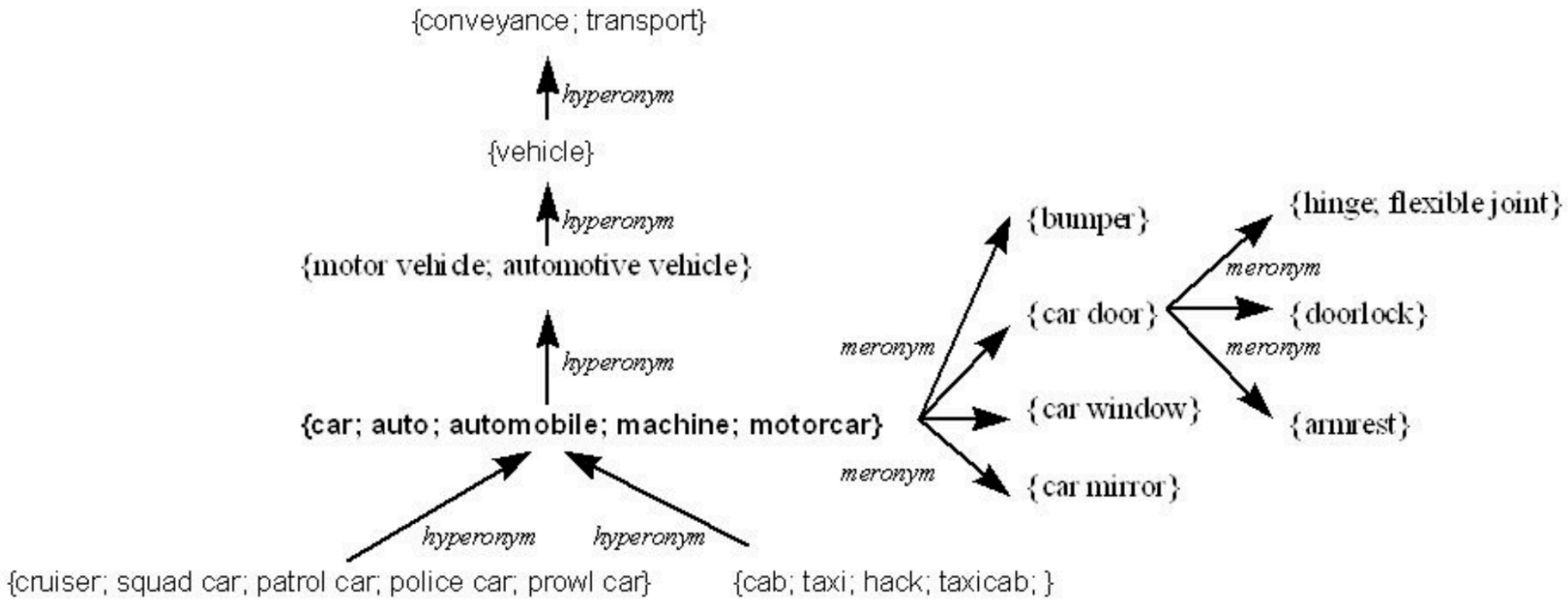
NORMALIZACIJA TEKSTA

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

WORDNET

Ilustracija strukture WordNet-a



Izvor: http://slideplayer.com/slide/5196/

ELIMINISANJE TERMINA SA SUVIŠE VELIKOM / MALOM UČESTANOŠĆU

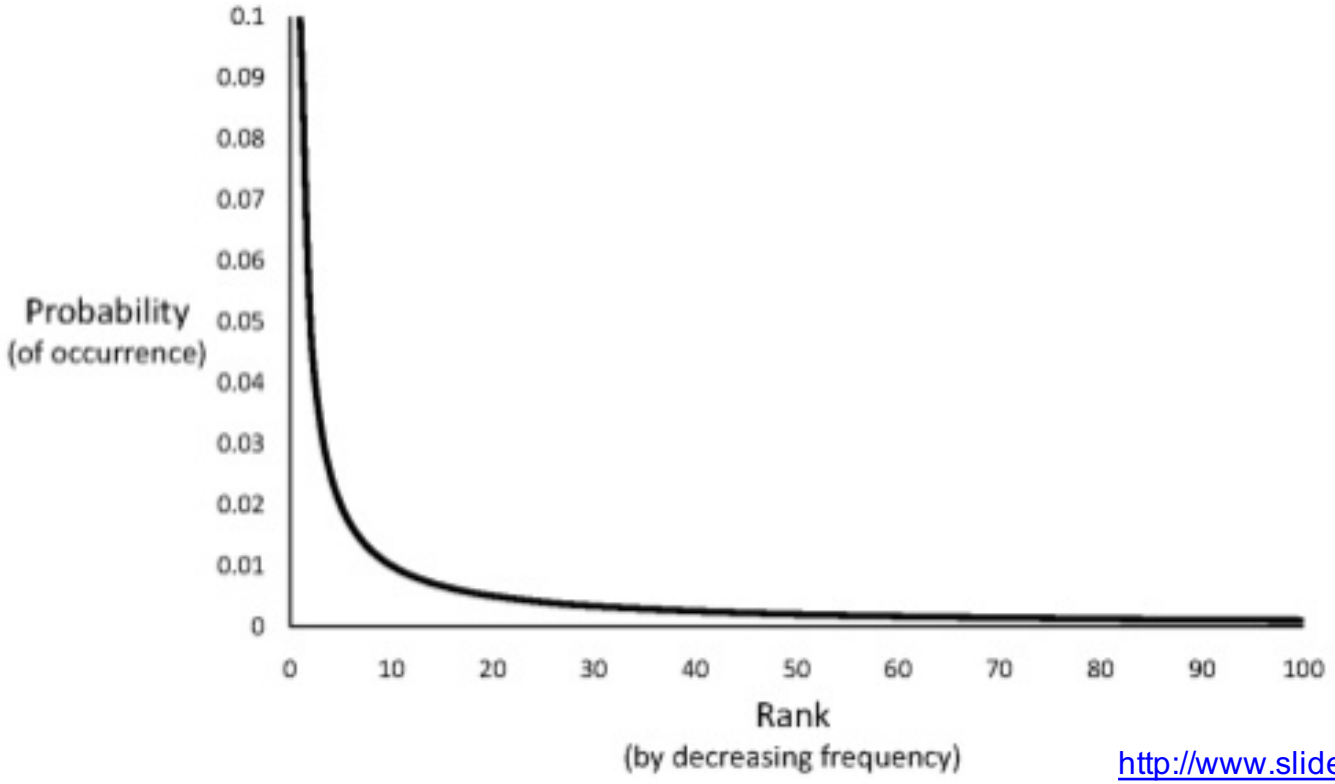
Empirijska zapažanja (u brojnim korpusima):

- 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:

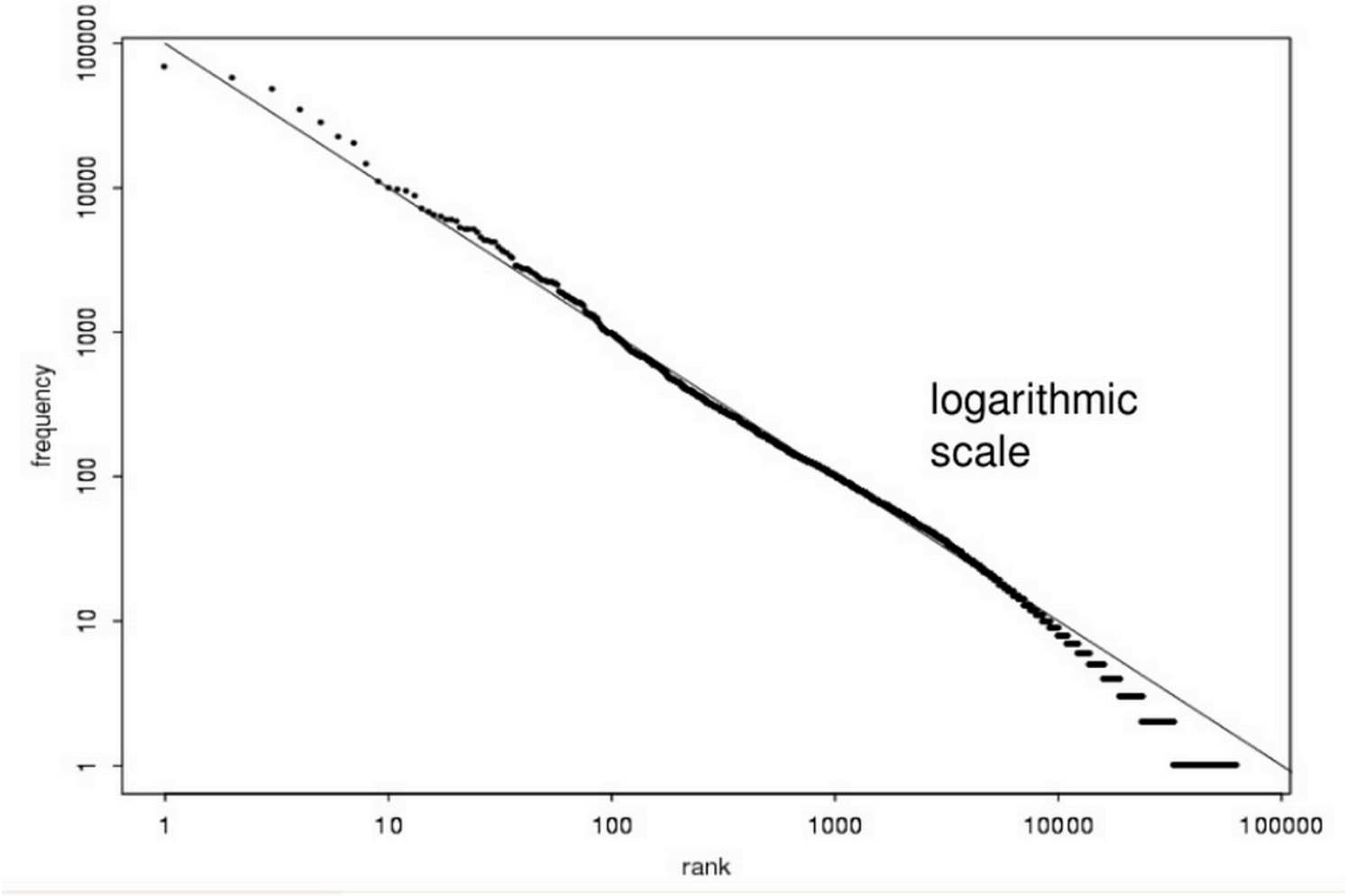
Frekvenca bilo koje reči u velikom korpusu je obrnuto proporcijalna njenom rangu u tabeli frekvencija (tog korpusa)



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

ILUSTRACIJA ZIPF-OVOG PRAVILA



Učestanost termina u Brown-ovom korpusu
(poznati korpus savremenog (američkog) engleskog jezika iz 1961)

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, 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

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')]

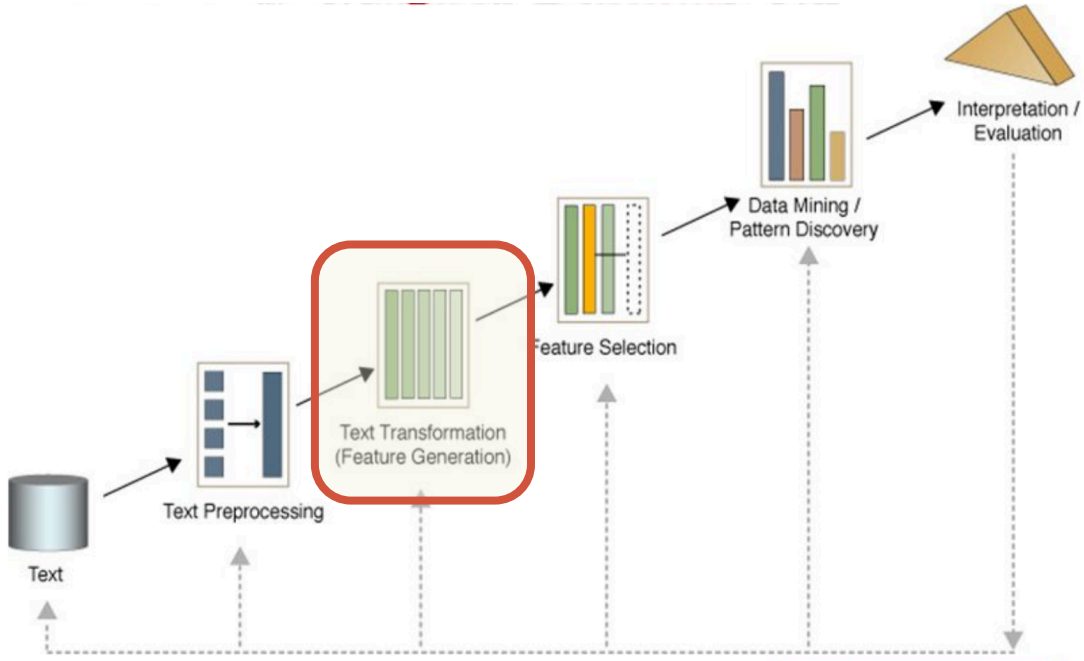
RB -> Adverb; NN -> Noun, singular or mass; IN -> Preposition, …

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



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 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.
- 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

VECTOR SPACE MODEL

- 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)

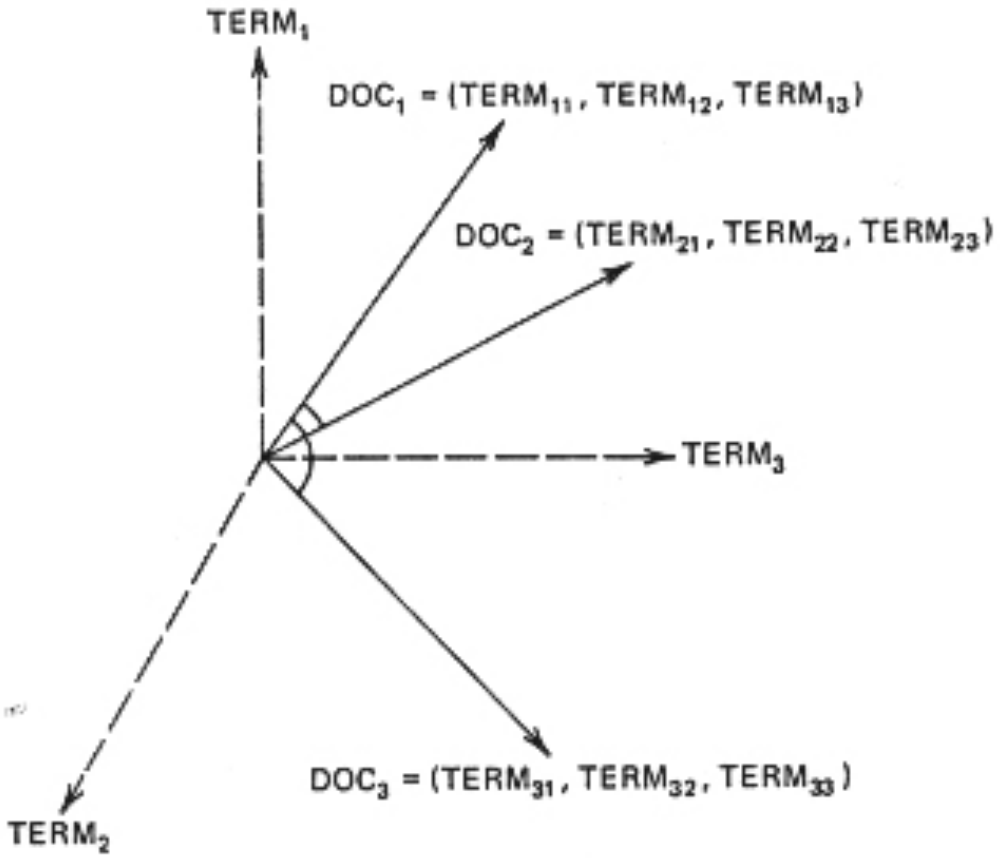
VECTOR SPACE MODEL (VSM)

- 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

BAG OF WORDS MODEL

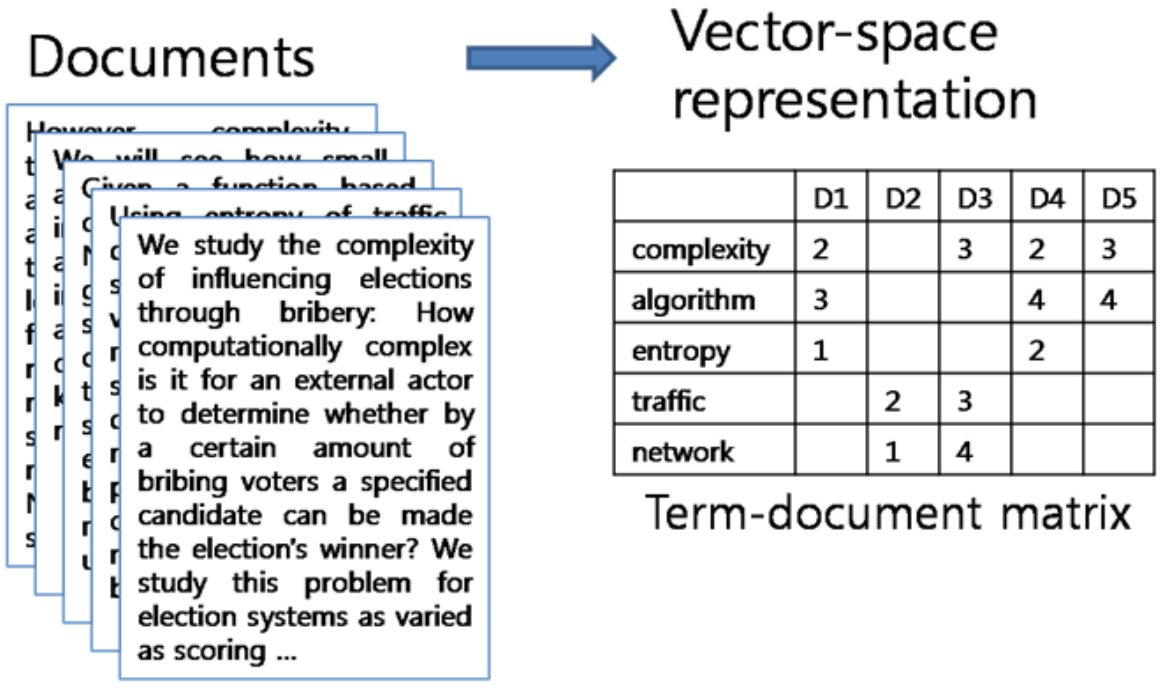


Dokumenti kao vektori u *n*-dimenzionalnom prostoru termina iz korpusa, gde je *n* broj jedinstvenih termina u korpusu

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

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		

Term-document matrix

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: NORMALIZACIJA TF TEŽINA

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

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

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

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

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

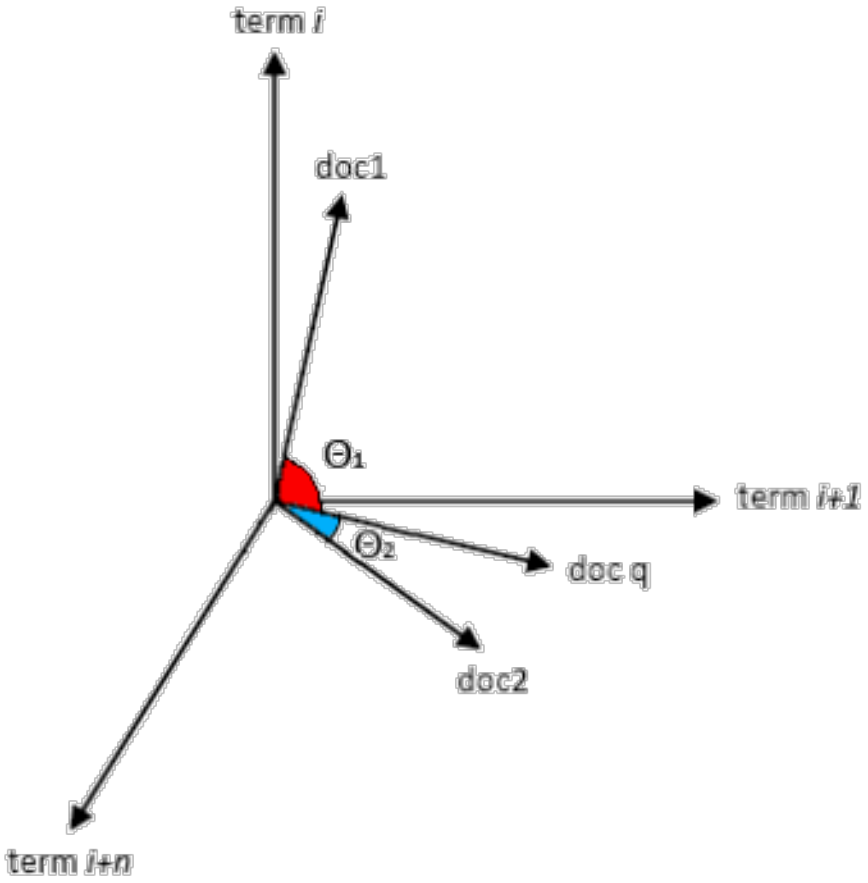
PREDNOSTI I NEDOSTACI VSM MODELA

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

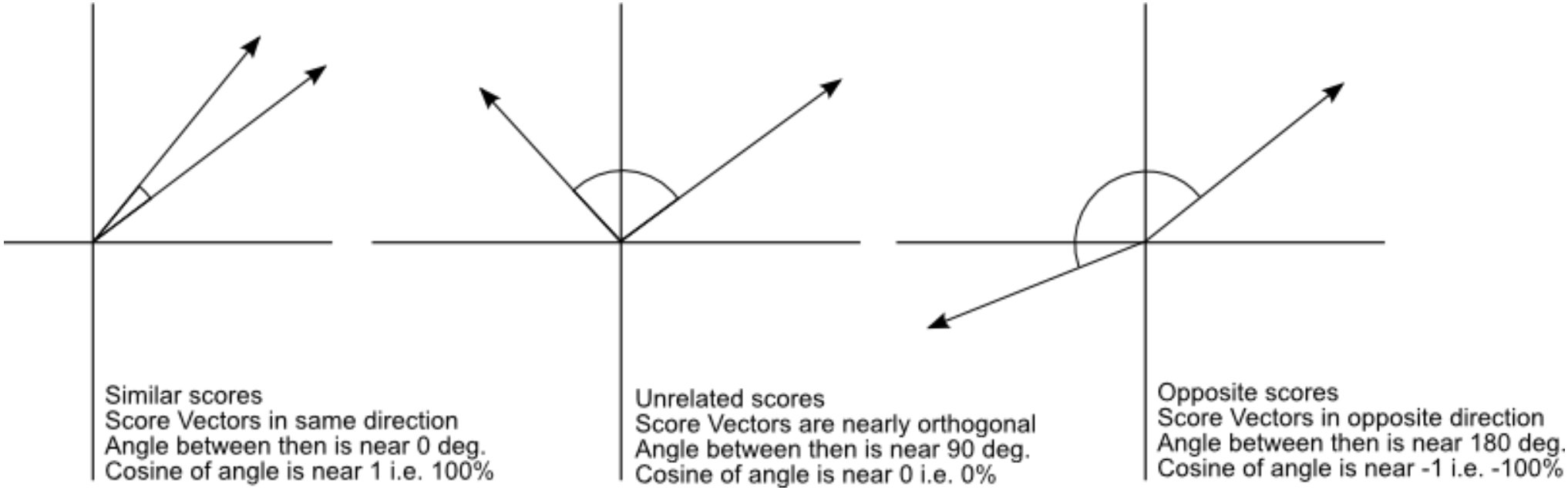
- Predstava dokumenata u formi vektora omogućuje procenu sličnosti dokumenata primenom vektorskih metrika
- Najčešće korišćena metrika: kosinusna sličnost (*cosine similarity*)



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

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 izbora najbolje metrike za procenu značajnosti termina
- Ograničenost na reči i izraze kao atribute (features)

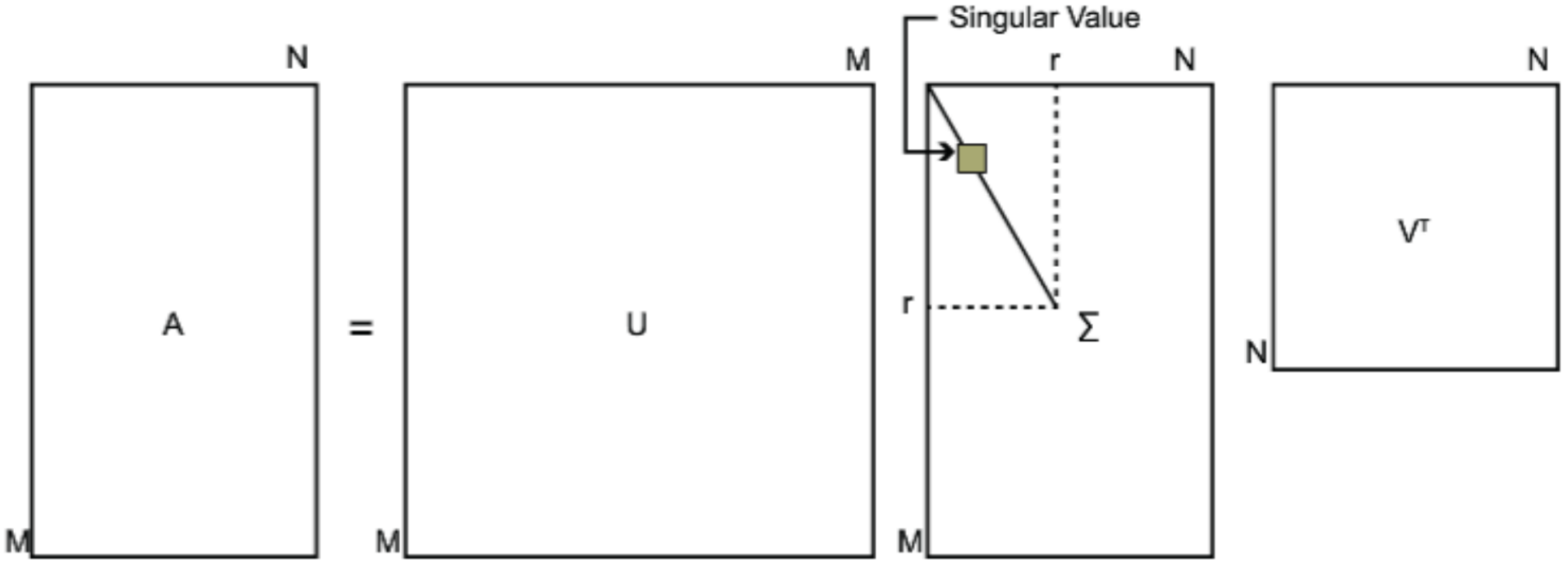
PREGLED DODATNIH PRISTUPA ZA KREIRANJE ATRIBUTA (FEATURES)

PROŠIRENJA KLASIČNOG VSM MODELA

- Adresiraju ograničenost na reči i izraze klasičnih VSM modela
- Umesto termina, elementi vektora mogu biti:
  - Vrednosti dobijene primenom metoda za redukciju prostora atributa, kao što je *Singular Vector Decomposition* (SVD)
  - Entiteti, detektovani u tekstu primenom *entity linking* metoda
  - Teme, detektovane u tekstu primenom *topic modelling* metoda

SINGULAR VALUE DECOMPOSITION (SVD)

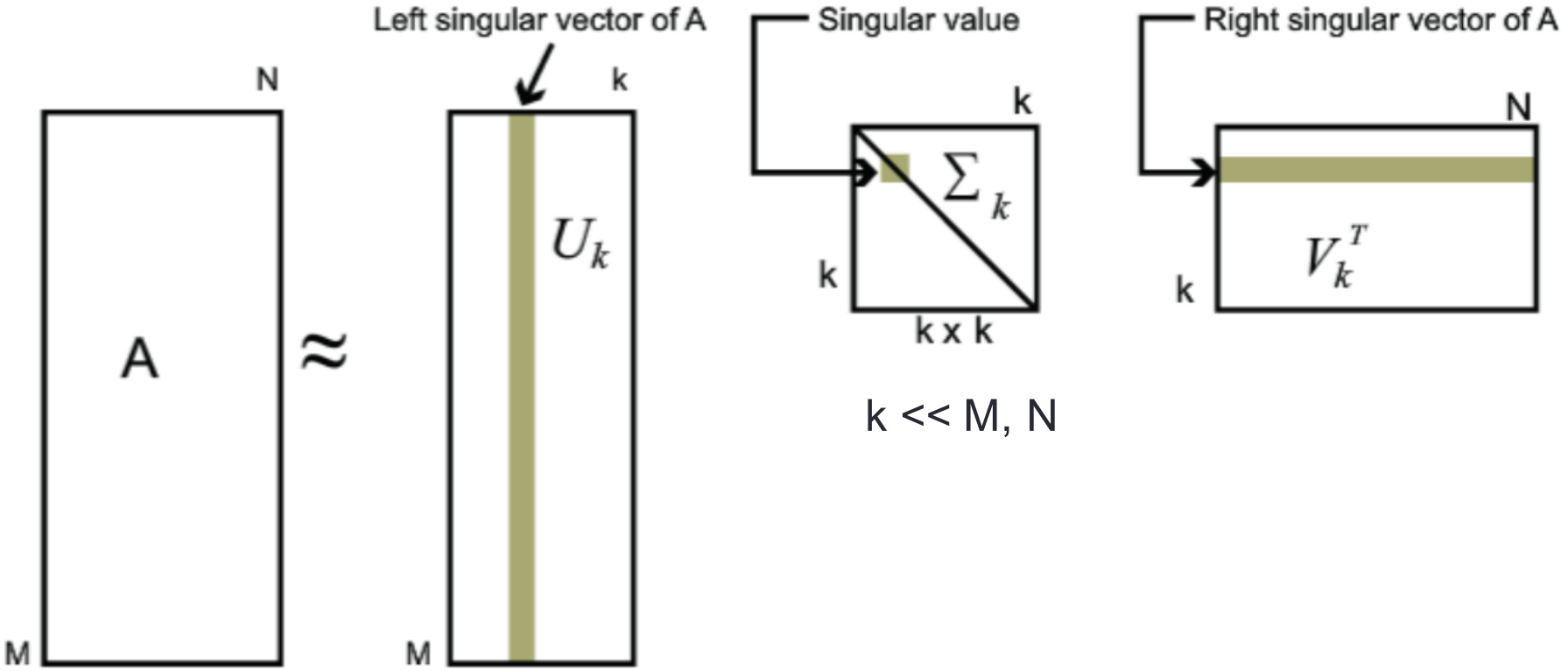
Aproksimacija matrice kroz redukciju ranga



U kontekstu TM-a, A je TDM matrica, sa M termina i N dokumenata

Izvor slike: https://rpubs.com/dsmilo/DATA643-Project3

SINGULAR VALUE DECOMPOSITION (SVD)



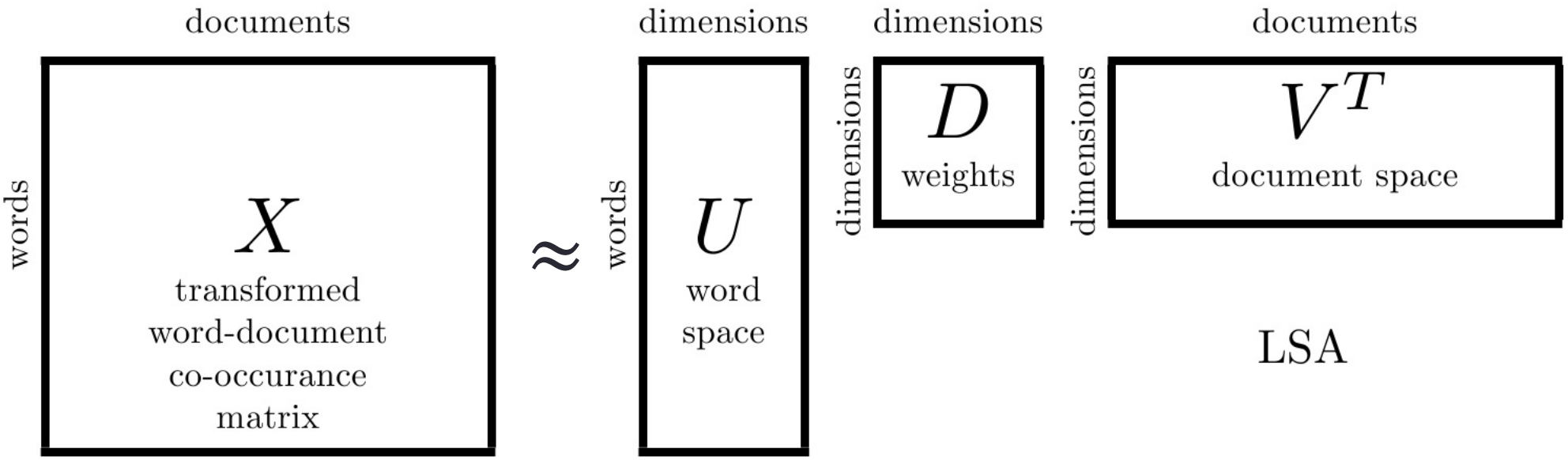
U kontekstu TM-a,

- U_k (*left singular vector*) predstavlja relaciju termina i izvedenih 'koncepata'
- V_k (*right singular vector*) predstavlja relaciju dokumenata i 'koncepata'

Izvor slike: https://rpubs.com/dsmilo/DATA643-Project3

LATENT SEMANTIC ANALYSIS (LSA)

LSA ≈ SVD primenjen na TDM matricu u cilju detekcije “tema” ili “koncepata” u datom korpusu



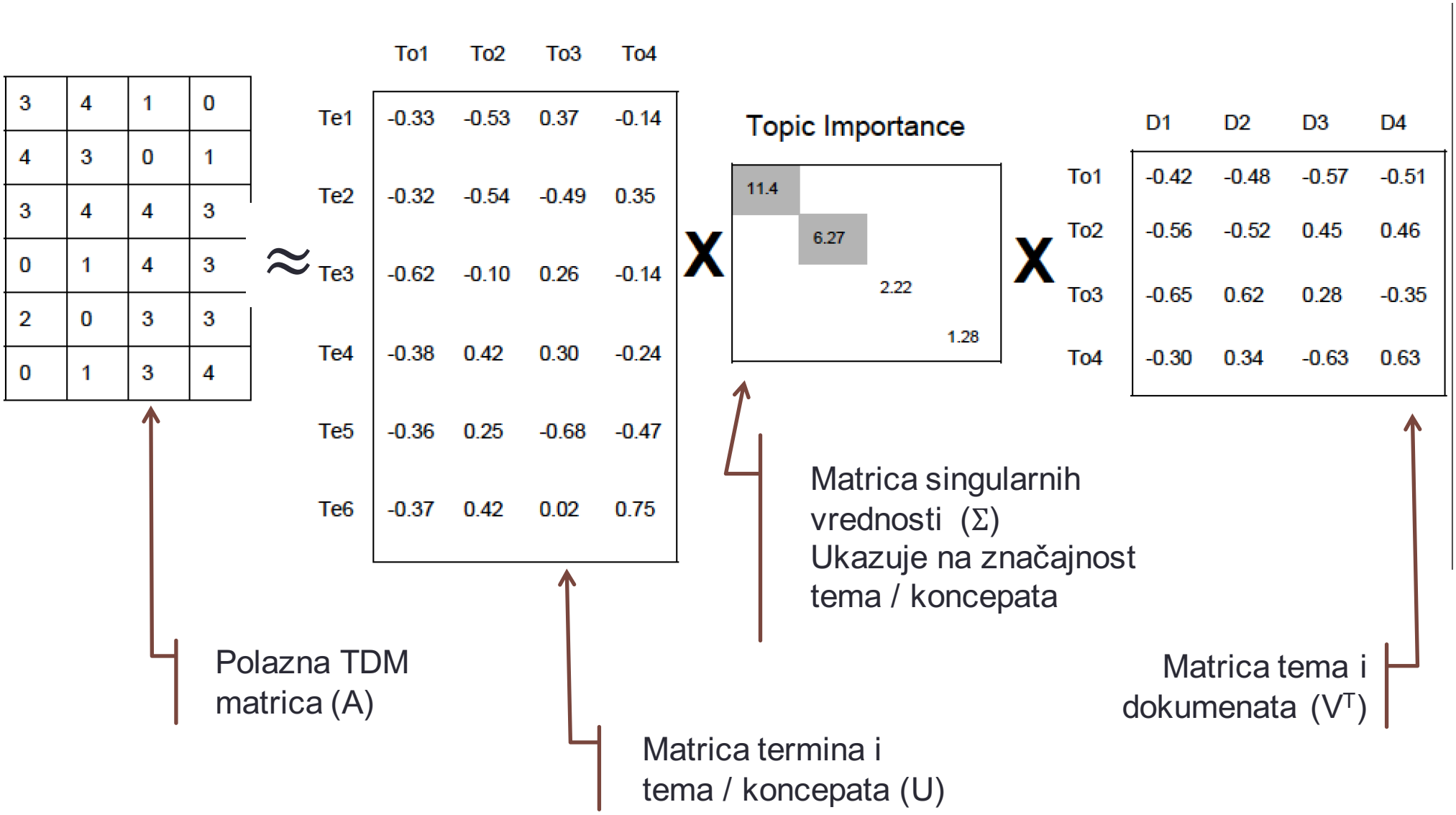
LATENT SEMANTIC ANALYSIS (LSA)

Primer: pretpostavimo da imamo sledeću TDM

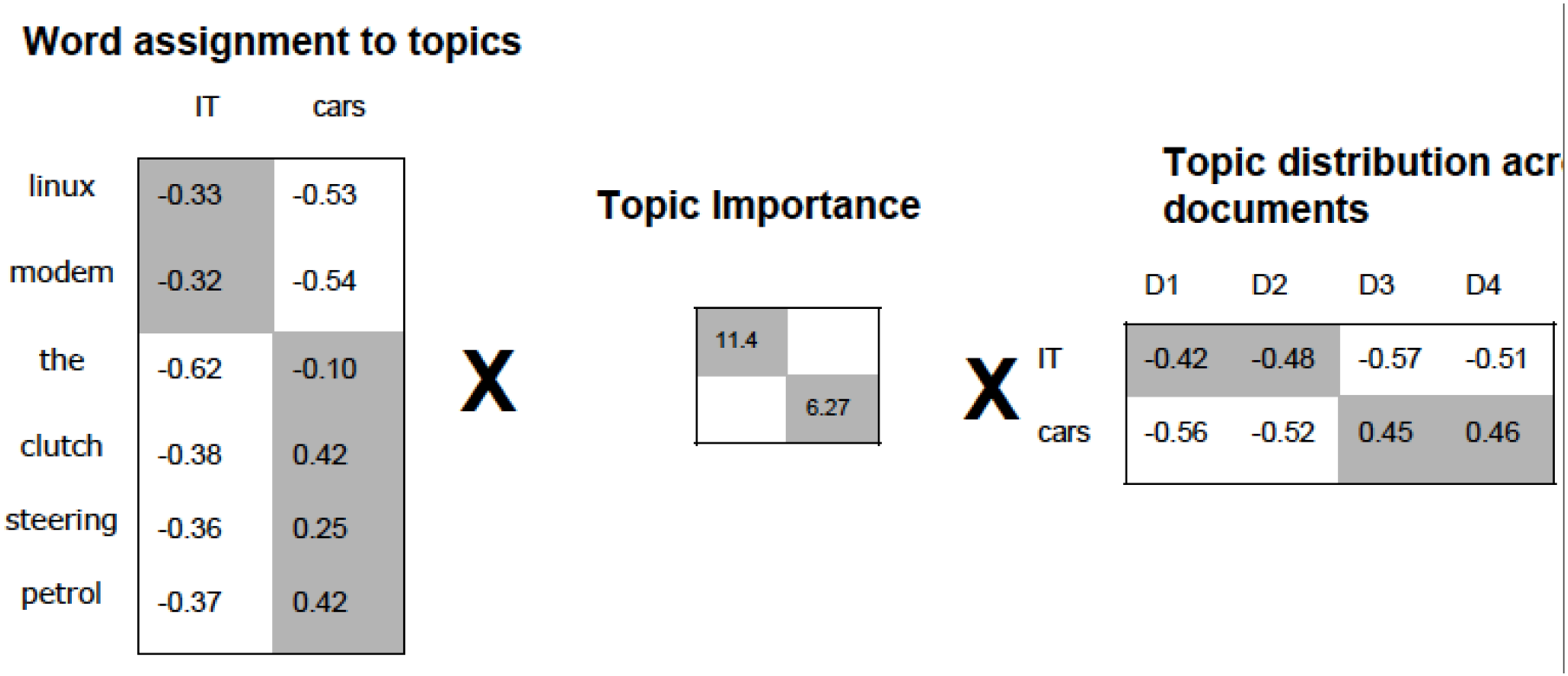
	D1	D2	D3	D4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

Primer preuzet iz https://www.slideshare.net/vitomirkovanovic/topic-modeling-for-learning-analytics-researchers-lak15-tutorial

LATENT SEMANTIC ANALYSIS (LSA)



LATENT SEMANTIC ANALYSIS (LSA)



Eliminisanje koncepata / tema kojima odgovaraju male singularne vrednosti (noise), u cilju detekcije značajnih tema (signal)

PREPOZNAVANJE ENTITETA U TEKSTU

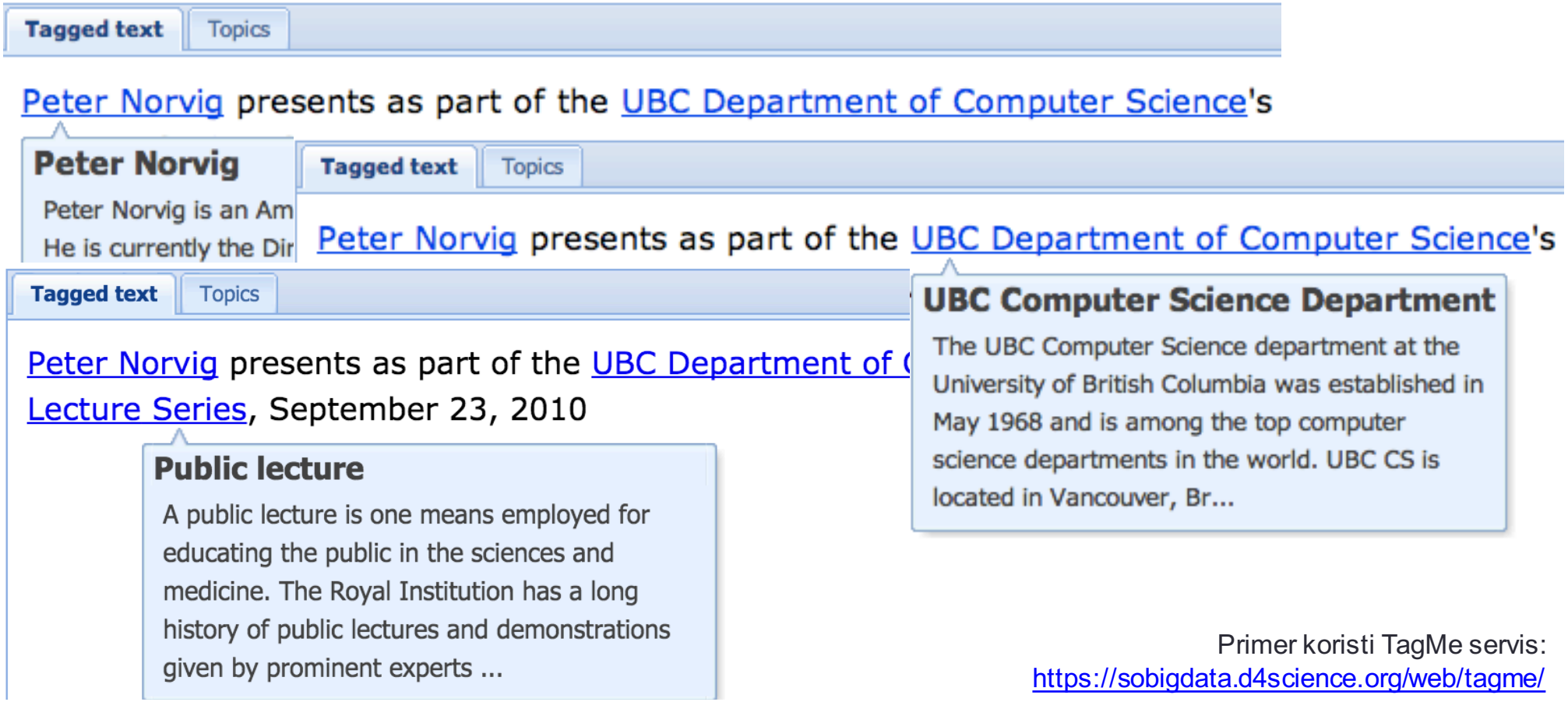
- *Named Entity Recognition* (*NER*)
- Entiteti mogu biti različitog tipa: osoba, organizacija, lokacija, datum, valuta sl.
- Primer:

> Peter Norvig presents as part of the UBC Department of Computer Science's Distinguished Lecture Series, September 23, 2010.

> Peter Norvig [PER] presents as part of the UBC Department of Computer Science's [ORG] Distinguished Lecture Series, September 23, 2010 [DATE].

ENTITY LINKING

- *Entity linking = NER + Disambiguation*
- *Disambiguation* = jedinstveno identifikovanje prepoznatog entiteta, kroz linkovanje termina iz teksta i odgovarajućeg koncepta iz baze znanja



Primer koristi TagMe servis:
https://sobigdata.d4science.org/web/tagme/

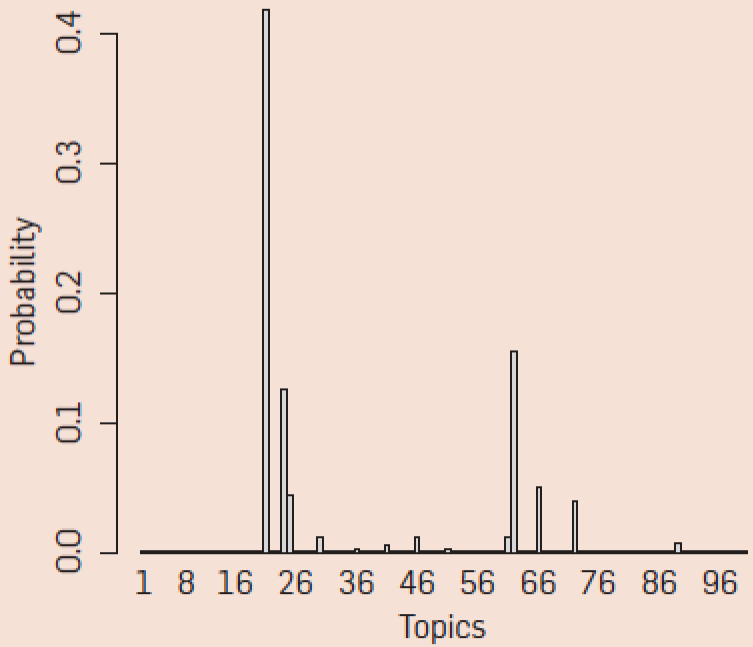
PRIMENA ENTITY LINKING-A ZA KREIRANJE VSM-A

- Identifikovani entiteti formiraju ”rečnik” korpusa, i time definišu vektore kojima se predstavljaju pojedinačni dokumenti
- Tekst iz prethodnog primera bi dodao sledeće entitete u rečnik:
  - wikipedia:Peter_Norvig
  - wikipedia:Department_of_Computer_Science
  - wikipedia:Public_lecture

TOPIC MODELS

- Statističke metode za određivanje “tema” dokumenata u okviru datog korpusa
- Tema se definiše kao raspodela verovatnoća nad rečima iz korpusa
  - za svaku reč definiše verovatnoću pojavljivanja u datoj temi
- Broj tema se unapred zadaje
  - ulazni argument za topic modeling algoritam
- Za svaki dokument se određuje raspodela verovatnoća tema
  - svaki dokument pripada svakoj temi sa određenom verovatnoćom

TOPIC MODELLING – ILUSTRACIJA REZULTATA

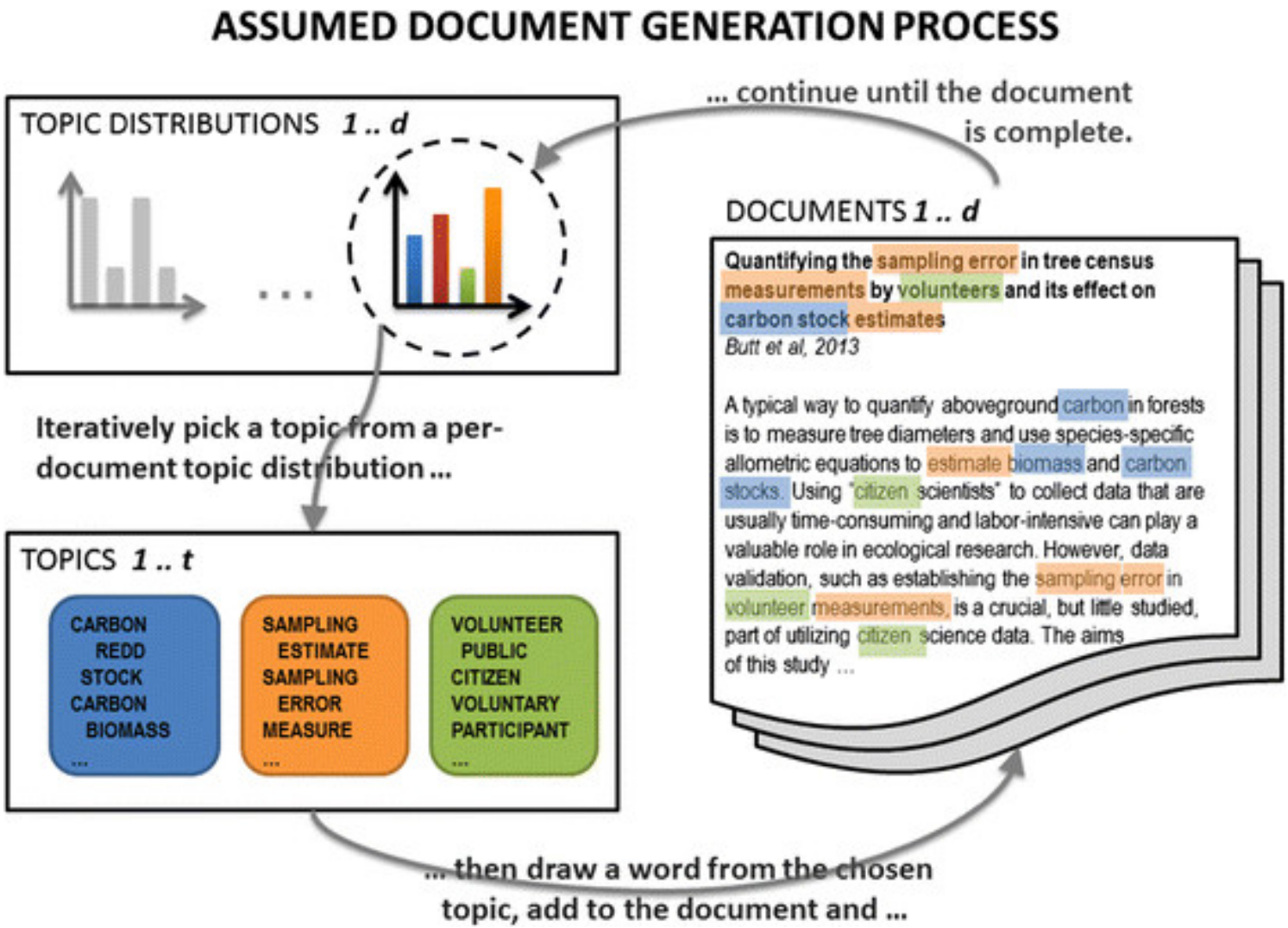


"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Rezultati dobijeni pri kreiranju modela sa 100 tema za korpus od 17,000 radova iz časopisa Science, primenom LDA topic modelling metode

Izvor: Blei, David. 2012. Probabilistic topic models. Communications of the ACM 55(4):77–84.

TOPIC MODELING – ILUSTRACIJA GENERATIVNOG MODELA



Izvor slike:
https://www.researchgate.net/publication/264656298_Assessing_citizen_science_opportunities_in_forest_monitoring_using_probabilistic_topic_modelling

PRIMENA TOPIC MODELA ZA KREIRANJE VSM-A

- Detektovane teme čine “rečnik” korpusa, i predstavljaju osnovu za kreiranje vektora kojima se predstavljaju dokumenti
- Elementi vektora su verovatnoće da dati dokument pripada svakoj od tema iz skupa tema za dati korpus

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 znakova interpunkcije
- pozicija reči u rečenici
- vrsta reči u okruženju
- …

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.

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

Interesantan način kombinovanja leksičkih i semantičkih (topic models) atributa za klasifikaciju kratkih tekstova dat u radu (Yang et al., 2013)

Još jedan interesantan primer kombinovanja različitih kategorija atributa (izvedenih iz teksta) za potrebe detekcije cyberbullying-a u porukama razmenjenim na društvenim mrežama (Hee et al., 2018)

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., … Hoste, V. (2018). Automatic Detection of Cyberbullying in Social Media Text. arXiv:1801.05617 [cs]. Retrieved from http://arxiv.org/abs/1801.05617

Yang, L., Li, C., Ding, Q., & Li, L. (2013). Combining Lexical and Semantic Features for Short Text Classification. Procedia Computer Science, 22(Supplement C), 78–86. https://doi.org/10.1016/j.procs.2013.09.083

KOMBINOVANJE RAZLIČITIH VRSTA ATRIBUTA

Problem: I have two kinds of features that I want to use to build a classification model:

- text data that has been represented as a sparse bag of words,
- more traditional dense features
  - e.g., in the case of spam filtering, that can be an indicator if the poster/sender is anonymous, a numeric variable representing the reputation of the sender,...

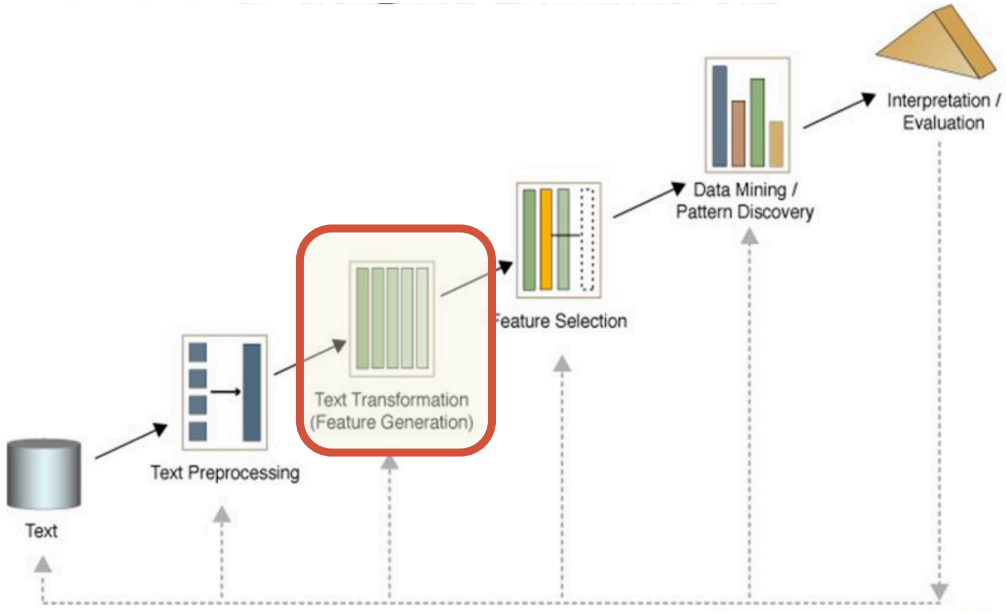
Zasnovano na pitanju postavljenom u okviru Data Science foruma:
https://datascience.stackexchange.com/questions/987/text-categorization-combining-different-kind-of-features

KOMBINOVANJE RAZLIČITIH VRSTA ATRIBUTA

Potential solutions:

- Perform dimensionality reduction (such as SVD) on the sparse data to make it denser, and combine the features into a single dense matrix to train your model(s).
- Add the few dense features to your sparse features matrix so that all the features are in a single sparse matrix, which is then used to train your model(s).
- Create a model using only your sparse text data and then combine its predictions (probabilities if it's classification) as a dense feature with your other dense features to create a (stacked) model. Note: use only CV predictions as features to train your model otherwise you'll risk overfitting

Zasnovano na pitanju postavljenom u okviru Data Science foruma:
https://datascience.stackexchange.com/questions/987/text-categorization-combining-different-kind-of-features



FEATURE SELECTION

SELEKCIJA ATRIBUTA

- Brojne su metrike za selekciju atributa
  - Primenjuju se primarno za selekciju termina (*ngrams*) iz teksta
- Neke od najčešće korišćenih metrika:
  - *Document Frequency (DF)*
  - *Information Gain (IG)*
    - *IG*(*t*) meri smanjenje entropije (*E*) tj. neizvesnosti predikcije koje se postiže ukoliko se za termin *t* zna da li je prisutan u datom dokumentu ili ne

$$IG(t) = E(c) - (P_r(t)E(c|t) + P_r(\bar{t})E(c|\bar{t}))$$

SELEKCIJA ATRIBUTA

- Neke od najčešće korišćenih metrika (nastavak):
  - *Mutual Information (MI)*

$$MI(t,c) = log\frac{P_r(t \wedge c)}{P_r(t) \times P_r(c)}$$

Ili

$$MI(t,c) = logP_r(t|c) - logP_r(t)$$

Nedostatak: pristrasnost prema rečima sa malom učestanošću pojavljivanja u datom korpusu

SELEKCIJA ATRIBUTA

- Neke od najčešće korišćenih metrika (nastavak):
  - *Chi Square (CHI)*
    - meri stepen nezavisnosti termina ***t*** i klase ***c***
    - računa se na osnovu matrice kontigencije termina ***t*** i klase ***c***:

	c	$\bar{c}$
t	A	B
$\bar{t}$	C	D

$$CHI(t,c) = \frac{N \times (A \times D - C \times B)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

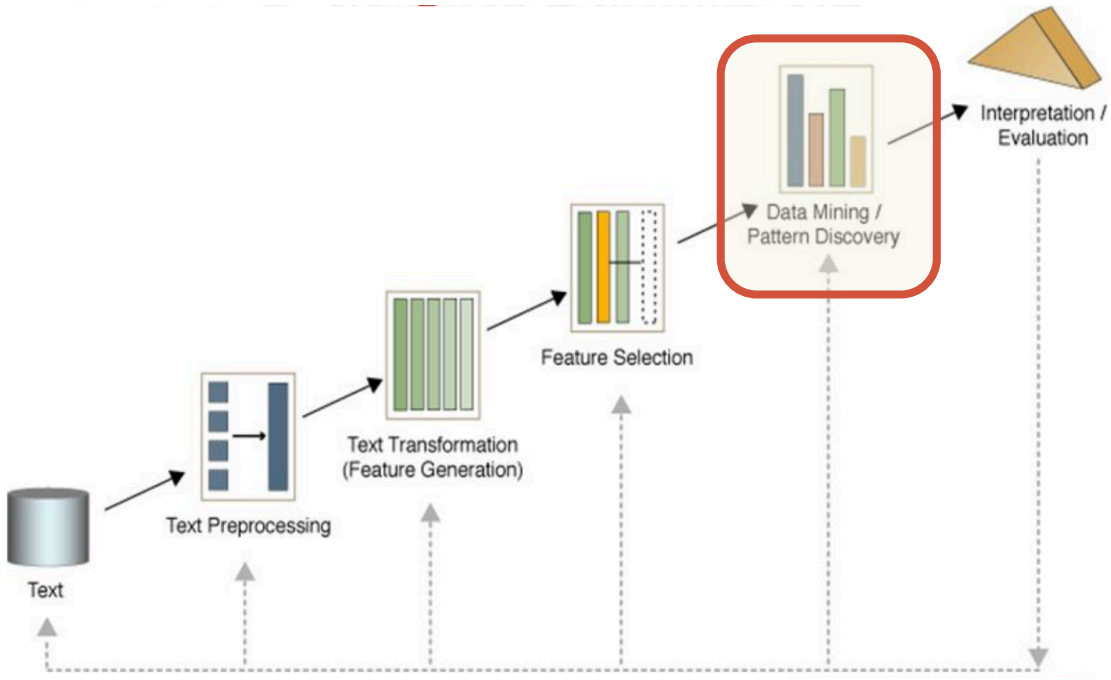
Nedostatak: nepouzdana za termine sa veoma malom učestanošću pojavljivanja (ćelije matrice slabo popunjene)

SELEKCIJA ATRIBUTA

Empirijsko poređenje navedenih metrika u kontekstu zadatka klasifikacije teksta dato u radu (Yang & Pedersen, 1997)

"This paper is a comparative study of feature selection methods in statistical learning of text categorization. The focus is on aggressive dimensionality reduction… We found IG and CHI most effective in our experiments... We found strong correlations between IG, CHI, and DF values of a term. This suggests that DF ... can be reliably used instead of IG or CHI..."

Yang, Y. & Pedersen, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization. In Proc. of the 14th Int'l Conf. on Machine Learning (ICML '97), 412-420.



DATA MINING / PATTERN DISCOVERY

DATA MINING, PATTERN DETECTION

Tekst predstavljen u strukturiranom formatu, tipično u formi TDM matrice, predstavlja ulaz za različite algoritme maš. 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; labele (klase) dodeljenje na osnovu ocene filma (1-4: neg, 7-10: poz)
  - features: unigrami, bigrami, trigrami, i njihove kombinacije
  - korišćena metoda: logistička regresija; dodeljuje svakoj observaciji 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)

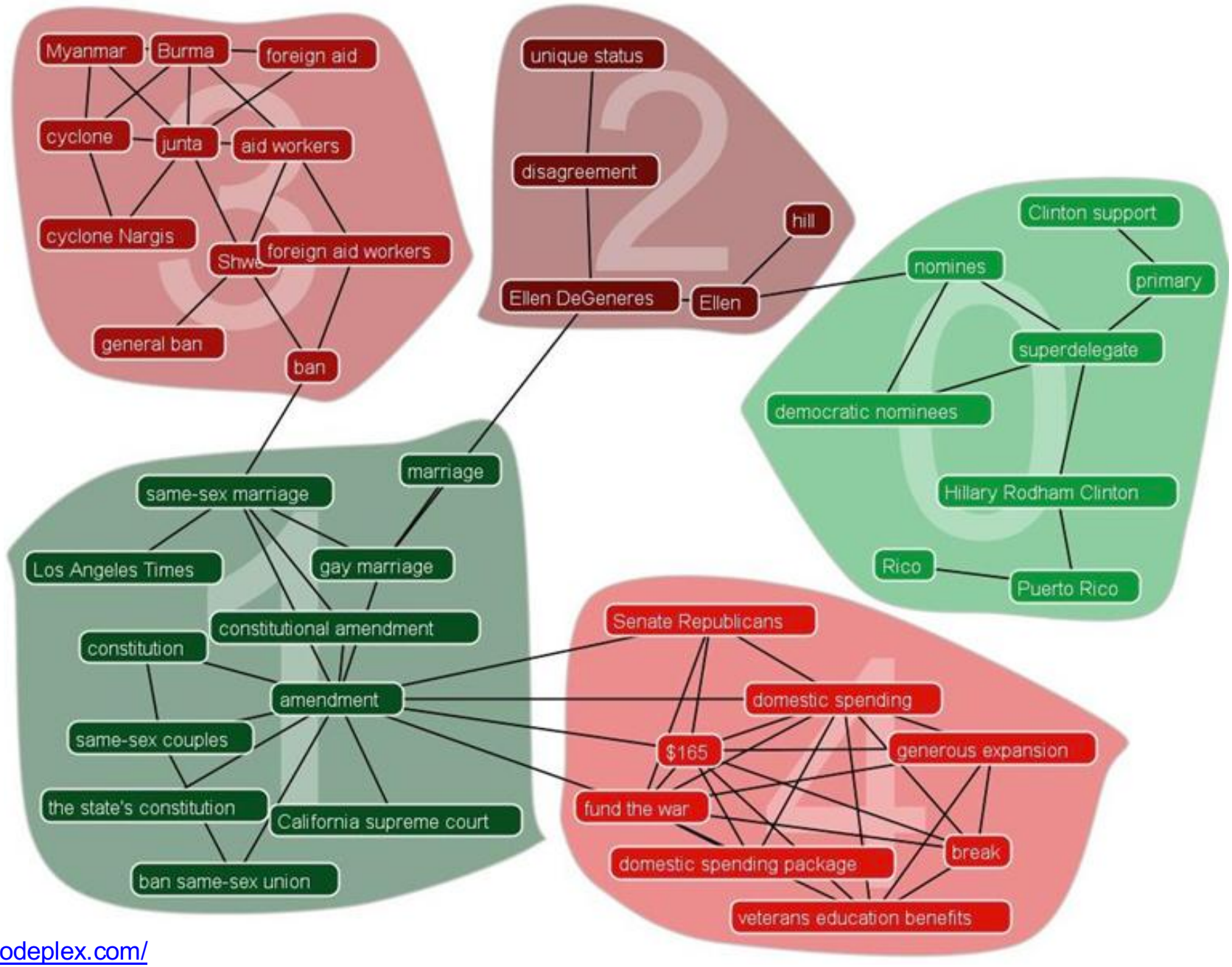
Izlaz klasifikatora (log. regresija): **0.0111**

PRIMER KLASTERIZACIJE DOKUMENATA

- Detekcija tema i grupisanje dokumenata prema identifikovanim temama
- KeyGraph metoda
  - Transformiše korpus dokumenata u graf susedstva ključnih reči iz korpusa
  - Koristi postojeće algoritme za detekciju grupa (communities) u grafu (tj za klasterovanje), u cilju identifikacije ključnih reči koje se mogu grupisati u tematske celine
  - Identifikovane grupe ključnih reči se pokazuju kao dobar proxy za teme dokumenata u korpusu i tematsko grupisanje dokumenata

H. Sayyadi, L. Raschid. "A Graph Analytical Approach for Topic Detection", ACM Transactions on Internet Technology (TOIT), 2013

ILUSTRACIJA REZULTATA KEYGRAPH ALGORITMA



Izvor: http://keygraph.codeplex.com/

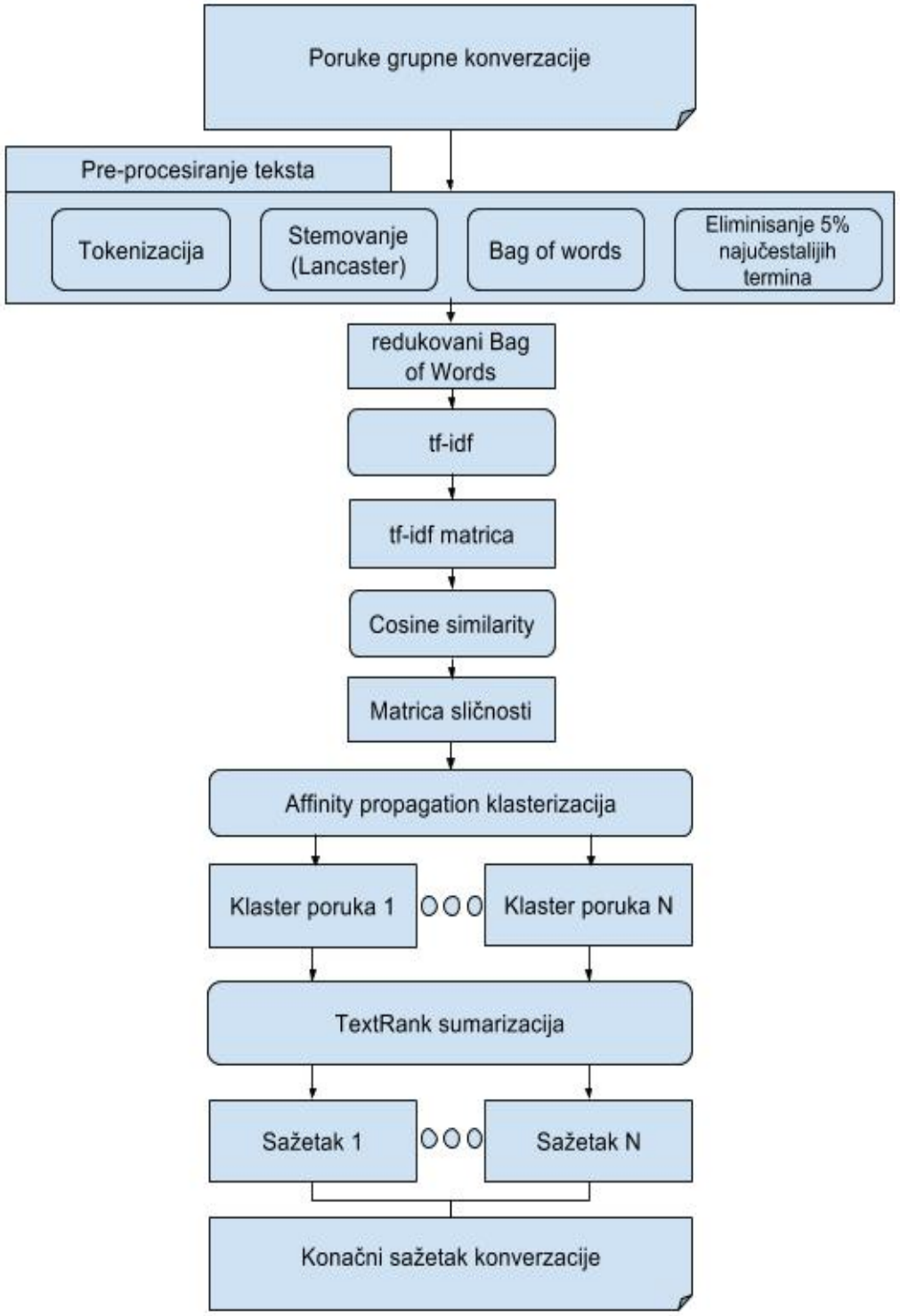
PRIMER KOMBINACIJE VIŠE RAZLIČITIH METODA EKSTRAKCIJE INFORMACIJA IZ TEKSTA

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 (TextRank algoritam)

*Master rad Stefana Stanojevića, FON, Nov 2017

Sumarizacija poruka online grupnih chat diskusija



RELEVANTNI LINKOVI

SOFTVERSKI PAKETI ZA TM

- R paketi:
  - quanteda
  - tm
  - tidytext
  - kompletna lista
- Java frameworks:
  - Stanford CoreNLP
  - Apache OpenNLP
  - LingPIPE
- Python
  - NLTK

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)
- S. Bird, E. Klein, & E. Loper. Natural Language Processing with Python. O'Reilly, 2009. E-verzija raspoloživa: http://www.nltk.org/book/

PREPORUKE

Interesantni članci, projekti:

- NELL - Never Ending Language Learner (website) (NYT article) (video lecture)
- SentiStrength - automatic sentiment analysis of social web texts
- Text analysis of Trump's tweets (blog post with the analysis) (podcast)
- Grimmer, J., & Stewart, B. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis, 21*(3), 267-297. doi:10.1093/pan/mps028

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
- Coarse Discourse dataset – "for researchers trying to characterize the nature of online discussions"
- Dataset of personal attacks in Wikipedia 'talk' pages (discussions around article content)

OSNOVE TEXT MINING-A

Jelena Jovanovic

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net