Prepoznavanje entiteta u tekstu i Semantičko indeksiranje

Jelena Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

PREGLED PREDAVANJA

- Osnovni pojmovi
  - Prepoznavanje entiteta u tekstu (*Named Entity Recognition*)
  - Semantičko indeksiranje (*Semantic Indexing, Entity Linking*)
  - Prepoznavanje tema/koncepata u tekstu (*Topic/concept detection*)
  - *Cognition as a Service*
- Primeri primene
- Prepoznavanje entiteta u tekstu primenom m. učenja
- Semantičko indeksiranje primenom m. učenja i baza znanja

OSNOVNI POJMOVI

PREPOZNAVANJE ENTITETA U TEKSTU

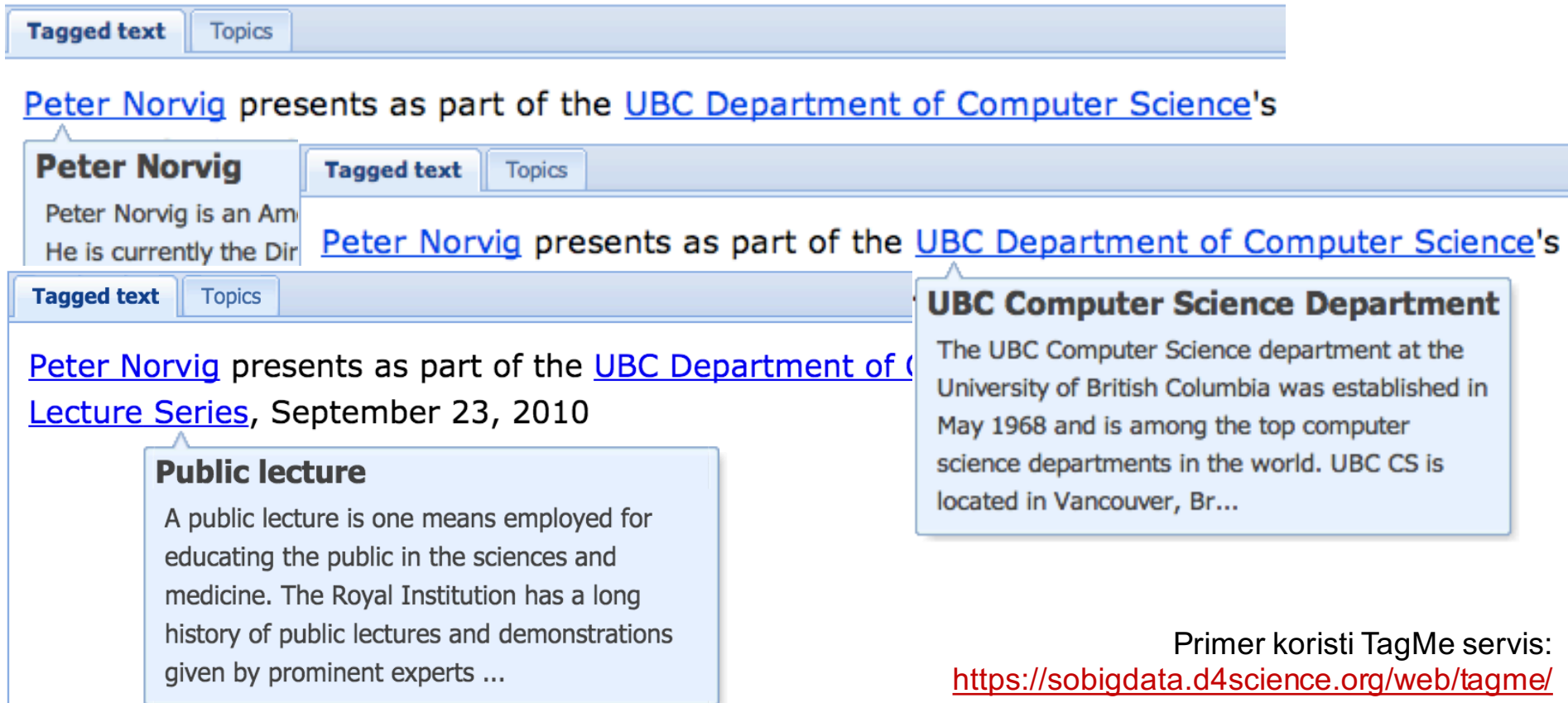
- *Named Entity Recognition* (*NER*)
- Entiteti mogu biti različitog tipa: osoba, organizacija, lokacija, datum, valuta sl.
- Primer:

> 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].

Semantičko indeksiranje

- *Semantic Indexing, Entity Linking*
- *Semantic indexing = NER + Disambiguation*
- *Disambiguation* = jedinstveno identifikovanje prepoznatog entiteta



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

Identifikacija ključnih tema/koncepata

- *Topic detection*
- Slično semantičkom indeksiranju, koristi koncepte iz baze znanja za pridruživanje semantike tekstu
- Razlike u odnosu na semantičko indeksiranje:
  - Teme/koncepti se ne vezuju za pojedinačne reči i/ili fraze u tekstu, već za ceo tekst
  - Nakon identifikacije tema, vrši se njihovo rangiranje po značajnosti

Identifikacija ključnih tema/koncepata

Primer:

"Angela Merkel, Chancellor of Germany, now faces what nearly all here are calling the toughest passage of her 11 years in power, after a terrorist attack on Monday in Berlin left 12 people dead. The Islamic State has claimed responsibility, and the authorities are searching for a Tunisian man with Salafist ties."

Identifikovani koncepti (korišćenjem TextRazor servisa):

- **International security** (1.00)
- **Islam and politics** (1.00)
- **Terrorism** (1.00)
- **Islam-related controversies** (0.97)
- **Politics** (0.93)
- **Conflicts** (0.87)
- **National security** (0.85)
- …

Vrednost u zagradi predstavlja tzv. topic score koji ukazuje na značajnost topic-a za dati tekst

COGNITION AS A SERVICE

- Sve veći broj servisa/alata koji omogućuju ekstrakciju entiteta i tema, i semantičko indeksiranje
- Zajednička karakteristika svih ovih servisa/alata:
  - Kombinuju ‘mašinsku inteligenciju’ i prikupljeno ljudsko znanje, odnosno tehnike mašinskog učenja i ogromne baze znanja (Wikipedia, DBpedia, WikiData,…)

COGNITION AS A SERVICE

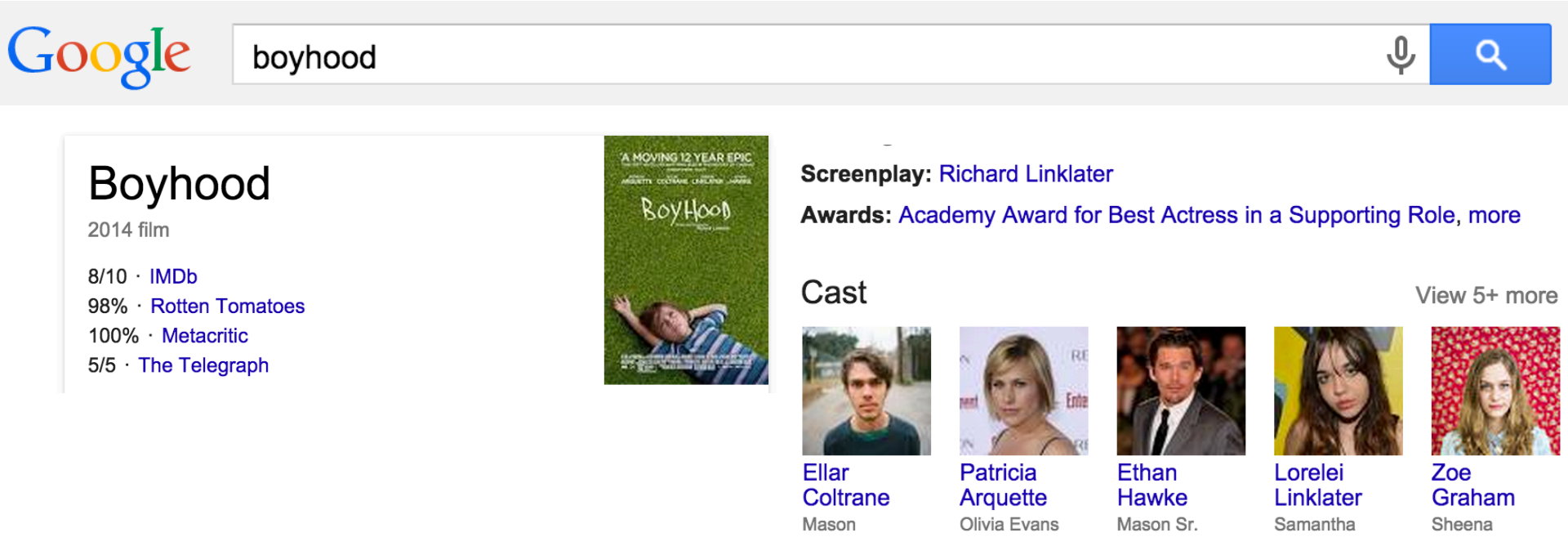
Primeri:

- Alchemy API Language Services (http://www.alchemyapi.com/products/alchemylanguage)
- TextRazor (http://www.textrazor.com/)
- OpenCalais (http://www.opencalais.com/)
- Dandelion API (https://dandelion.eu/)
- TagMe (https://sobigdata.d4science.org/web/tagme/)
- Dbpedia Spotlight (https://github.com/dbpedia-spotlight/dbpedia-spotlight)
- …

PRIMERI PRIMENE

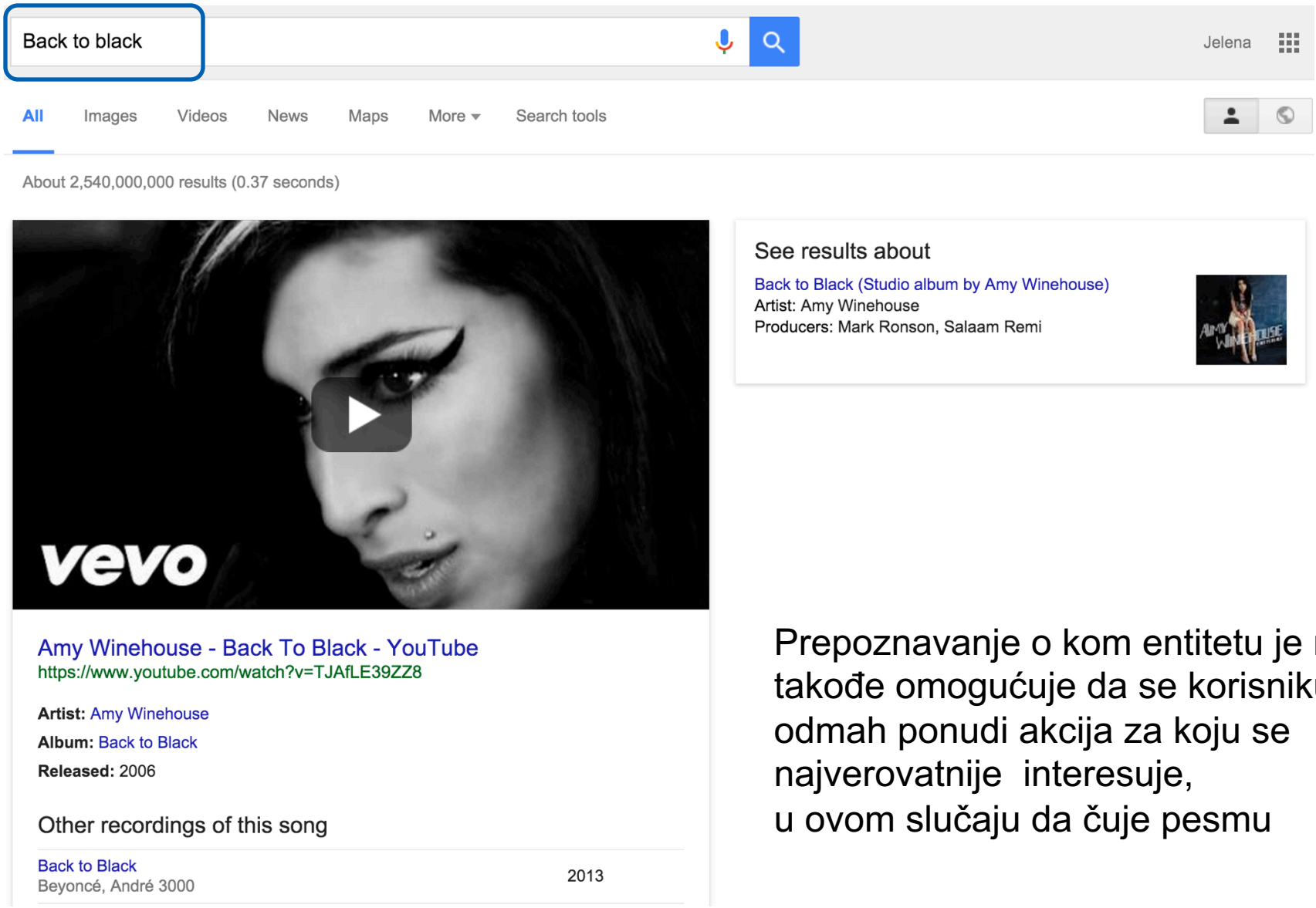
Naprednije pretraživanje

- Preko 50% upita na Web-u odnosi se na neki entitet (osobu, film, muz. numeru/grupu, grad,…)*
- Prepoznavanje o kom je entitetu reč omogućuje preporuku srodnih entiteta za koje bi korisnik mogao biti zainteresovan



* Pound et al. 2010. Ad-hoc object retrieval in the web of data. In: Proceedings of the 19th International Conference on World Wide Web. pp. 771-780. ACM, New York, NY, USA

Naprednije pretraživanje



Prepoznavanje o kom entitetu je reč takođe omogućuje da se korisniku odmah ponudi akcija za koju se najverovatnije interesuje, u ovom slučaju da čuje pesmu

POSLOVNE ANALITIKE

Primer: RavenPack News Analytic

- http://www.ravenpack.com/
- Ekstrakcija entiteta iz novinskih članaka: kompanije, brendovi, proizvodi,…
- Ekstrakcija geo-političkih i makro-ekonomskih događaja, kao i događaja relevantnih za pojedine kompanije i brendove
- Estrahovane informacije predstavljaju ulaz za različite vrste business intelligence sistema

SOCIAL MEDIA MONITORING

Reputation management

- Analiza tekstualnih sadržaja društvenih medija i mreža radi
  - identifikacije relevantnih entiteta: osoba, kompanija, brendova, proizvoda, …
  - detekcije sentimenta o identifikovanim entitetima
- Cilj: upravljanje reputacijom nekog pojedinca ili organizacije
- Primeri:
  - Reputation.com (http://reputation.com/)
  - Rankur (https://rankur.com/)
  - Trackur (http://www.trackur.com/)

ONLINE REKLAMIRANJE

Primer: ADmantX (http://www.admantx.com/)

- Analiza sadržaja Web stranice radi ekstrakcije informacija potrebnih za preporuku reklama za datu stranicu
- Ekstrahuje sledeće informacije:
  - entitete (osobe, lokacije, kompanije, brendove,…),
  - teme o kojima tekst govori,
  - emocije sadržane u tekstu

Prepoznavanje entiteta u tekstu primenom m. učenja

PREPOZNAVANJE ENTITETA U TEKSTU

- Najčešće zasnovano na primeni *nadgledanog* m. učenja, odnosno klasifikaciji
- Osnovna ideja:
  - Program uči karakteristike/osobine koje odlikuju entitete određenog tipa
  - Osobine entiteta se određuju na osnovu termina kojima su entiteti predstavljeni u tekstu, kao i termina koji čine njihov kontekst/okruženje
- Preduslov:
  - Postojanje dovoljno velikog skupa podataka za trening tj. korpusa anotiranih/obeleženih dokumenata

Primena nadgledanog m. učenja

Razmotrićemo osnovne elemente procesa nadgledanog m. učenja pri prepoznavanju entiteta u tekstu:

- kreiranje skupa (obeleženih/anotiranih) podataka za treniranje modela
- definisanje skupa atributa (*features*) koji će se koristiti za formiranje modela
- procesiranje teksta u cilju kreiranja/izdvajanja definisanog skupa atributa
- selekcija algoritma m. učenja
- evaluacija modela

PODACI ZA TRENING

Primer teksta anotiranog za potrebe "obuke" algoritma nadgledanog učenja:

Unlike <PERSON>Robert</PERSON>, <PERSON>John Briggs Jr</PERSON> contacted <ORGANIZATION>Wonderful Stockbrockers Inc </ORGANIZATION> in <LOCATION>New York </LOCATION> and instructed them to sell his <NUMBER> 100</NUMBER> shares in <ORGANIZATION> Acme </ORGANIZATION>

Očigledno, priprema podataka za obučavanje algoritma je prilično zahtevna…

PODACI ZA TRENING

- Dobra vest je da su izvesne organizacije, (istraživačke) groupe i pojedinci publikovali podatke (tj. anotirane dokumente) koji se mogu koristiti za trening
- Primeri:
  - Linguistic Data Consortium (http://www.ldc.upenn.edu) održava katalog lingvističkih datasets (http://www.ldc.upenn.edu/Catalog/)
  - Kolekcija anotiranih datasets za zadatak prepoznavanja entiteta u tekstu: http://www.cs.technion.ac.il/~gabr/resources/data/ne_datasets.html
  - Twitter NER (https://github.com/aritter/twitter_nlp) dataset korišćen za treniranje modela koji prepoznaje entitete u Twitter porukama

DEFINISANJE SKUPA ATRIBUTA

Širok spektar atributa koji se mogu koristiti

- Atributi koji se odnose na pojedinačne reči:
  - dužina reči;
  - prisutnost velikih slova;
  - vrsta reči;
  - učestanost pojavljivanja reči u dok. za trening;
  - prisutnost znakova interpunkcije;
  - pozicija reči u rečenici,…
- Atributi koji se odnose na okruženje reči:
  - opseg okruženja;
  - vrsta reči u okruženju i sl.

DEFINISANJE SKUPA ATRIBUTA

Izbor atributa

- Zavisi od vrste teksta koji je predmet analize (npr. tweet poruke vs. novinski članci vs. stručni tekstovi)
- Pristup: krenuti od tipično korišćenog skupa atributa (prethodni slajd), a zatim prilagoditi taj skup specifičnostima datog teksta
- Potrebno je posvetiti dosta pažnje ovom koraku jer ima veliki uticaj na performanse modela

PROCESIRANJE TEKSTA

Osnovni oblici procesiranja teksta radi kreiranja skupa atrubuta:

- Tokenizacija – podela teksta na tokene (= elementarne jedinice sadržaja)
- Eliminisanje stop-words
- Part-of-Speech (POS) tagging
  - oblik analize teksta u kome se svakoj reči pridružuje tag koji opisuje ulogu te reči (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 je raspoloživa ovde

PROCESIRANJE TEKSTA

Zavisno od definisanog skupa atributa, neki drugi oblici procesiranja teksta mogu biti potrebni; npr:

- Identifikacija i/ili eliminisanje znakova interpunkcije,
- Lematizacija
  - svođenje reči na zajednički osnovni oblik (lema)
  - Na primer: am, are, is => be

    car, cars, car's, cars' => car

KREIRANJE ATRIBUTA

Izabrani skup atributa se koristi za predstavljanje pojedinačnih reči i/ili izraza od kojih je tekst sačinjen

Jednostavan primer

Pretpostavimo da smo izabrali sledeće atribute:

-Boolean atribut koji ukazuje da li reč počinje velikim slovom

-Numerički atribut koji predstavlja dužinu reči

-Nominalni atribut koji predstavlja reč napisanu malim slovima

Rečenica: “The apple sign makes Apple laptops easily recognizable” će imati sledeću reprezentaciju:

<true, 3, "the">, <false, 5, "apple">, <false, 4, "sign">, <false, 5, ”makes">, <true, 5, ”apple">, … , <false, 12, ”recognizable">

Izbor metode m. učenja

Najčešće korišćene metode/algoritmi za prepoznavanje entiteta u tekstu*

- Random forests
- Hidden Markov Models (HMM)
- Maximum Entropy classification
- Support Vector Machines (SVM)
- Conditional Random Fields (CRF)

*samo informativno, ovo su dosta složeniji modeli od onih koje smo radili

EVALUACIJA MODELA

- Evaluacija je zasnovana na metrikama tipičnim za zadatak klasifikacije:
  - Preciznost (Precision), Odziv (Recall), F mera (F measure)

- Softverski okviri (frameworks) za poređenje različitih alata/servisa za prepoznavanje entiteta (tzv. benchmarking):
  - NERD (Named Entity Recognition and Disambiguation): http://nerd.eurecom.fr/
  - GERBIL (General Entity Annotator Benchmark): http://gerbil.aksw.org/gerbil/

Performanse današnjih alata za ekstrakciju entiteta (tip teksta: novinski članci)

		AlchemyAPI	DBpedia Spotlight	Cicero	Lupedia	Opencalais	Saplo	Stanford	Textrazor	Wikimeta	Yahoo	NERD-ML NB	NERD-ML k-NN	NERD-ML SVM	Theoretical limit
PER	p	78.38	66.79	58.13	93.25	**96.57**	58.46	88.96	60.09	78.53	84.16	91.12	91.41	90.83	99.94
	r	56.28	22.14	38.03	27.33	40.01	11.75	90.72	73.65	54.73	5.26	91.40	**92.15**	91.84	98.02
	f	65.51	33.26	45.98	42.28	56.58	19.57	89.83	66.19	64.50	9.90	91.26	**91.78**	91.33	98.97
LOC	p	75.44	67.52	70.21	70.21	65.50	52.47	88.26	50.15	74.09	90.70	87.17	**89.27**	88.33	99.75
	r	69.24	50.48	56.65	60.91	26.98	10.19	89.21	20.62	60.01	2.34	89.99	89.81	**90.29**	97.24
	f	72.21	57.77	62.71	65.23	38.22	17.07	88.73	29.23	66.31	4.56	88.55	**89.54**	89.30	98.48
ORG	p	67.24	66.78	72.91	70.99	62.71	57.97	82.36	47.65	38.59	85.29	72.25	81.15	**82.60**	99.43
	r	16.44	36.91	26.25	11.20	11.14	4.82	79.29	26.91	27.27	3.49	**82.30**	81.64	79.71	93.86
	f	26.42	47.54	38.60	19.34	18.92	8.89	80.80	34.40	31.96	6.71	76.95	**81.39**	81.13	96.56
MISC	p	48.05	4.00	10.95	17.73	2.71	0.00	81.59	11.21	4.65	0.00	62.72	77.70	**81.83**	97.55
	r	5.27	2.14	23.65	3.56	1.85	0.00	77.64	22.65	7.26	0.00	76.92	75.93	**77.64**	85.19
	f	9.50	2.79	14.97	5.93	2.20	0.00	79.56	15.00	5.67	0.00	69.10	76.80	**79.68**	90.95
Overall	p	74.80	59.43	47.85	71.82	60.74	55.91	85.99	42.62	50.33	85.85	80.14	86.09	**86.65**	99.46
	r	42.05	32.37	38.28	29.55	22.93	7.79	85.29	37.91	42.32	3.22	**86.51**	86.35	86.05	94.97
	f	53.84	41.91	42.53	41.87	33.29	13.68	85.64	40.12	45.97	6.21	83.20	86.22	**86.35**	97.17

Izvor: Rizzo et al. 2014. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web, 9th Edition of the Language Resources and Evaluation Conference (LREC'14)

Alternativni oblici m. učenja

- Problem: priprema dovoljno velikog skupa anotiranih dokumenata (korpusa) potrebnog za trening, je prilično zahtevan zadatak
- Usled toga, polu-nadgledano i nenadgledano m. učenje se često nameću kao alternative
  - ovi pristupi ne zahtevaju anotirani skup dokumenata
  - tradicionalno su imali slabije performanse u odnosu na pristupe nadgledanog m. učenja, ali su nova rešenja sve bolja

Polu-nadgledano m. učenje

- *Bootstrapping* je popularna tehnika polu-nadgledanog m. učenja
  - Podrazumeva mali stepen “nadgledanja”, tipično u formi inicijalno zadatog skupa primera, potrebnog za pokretanje procesa učenja
- Ilustracije radi, razmotrimo primer sistema namenjenog prepoznavanju proizvoda koji se pominju u tekstu
  - inicijalno, korisnik zadaje mali broj primera tj. naziva različitih proizvoda;
  - sistem analizira tekst i identifikuje elemente koji karakterišu kontekst zadatih primera; zatim, identifikuje druga pojavljivanja proizvoda na osnovu identifikovanih karakteristika konteksta;
  - proces učenja se ponovo primenjuje polazeći od *novo-otkrivenih instanci* (proizvoda), što vodi otkrivanju novih relevantnih konteksta;
  - ponavljajući ovaj proces, veliki broj proizvoda i konteksta u kojima se oni pojavljuju će biti otkriven.

Polu-nadgledano m. učenje

Preporuka:

Predavanje Tom Mitchell-a pod nazivom

Semisupervised Learning Approaches

održano u okviru

Autumn School 2006: Machine Learning over Text and Images

URL: http://videolectures.net/mlas06_mitchell_sla/

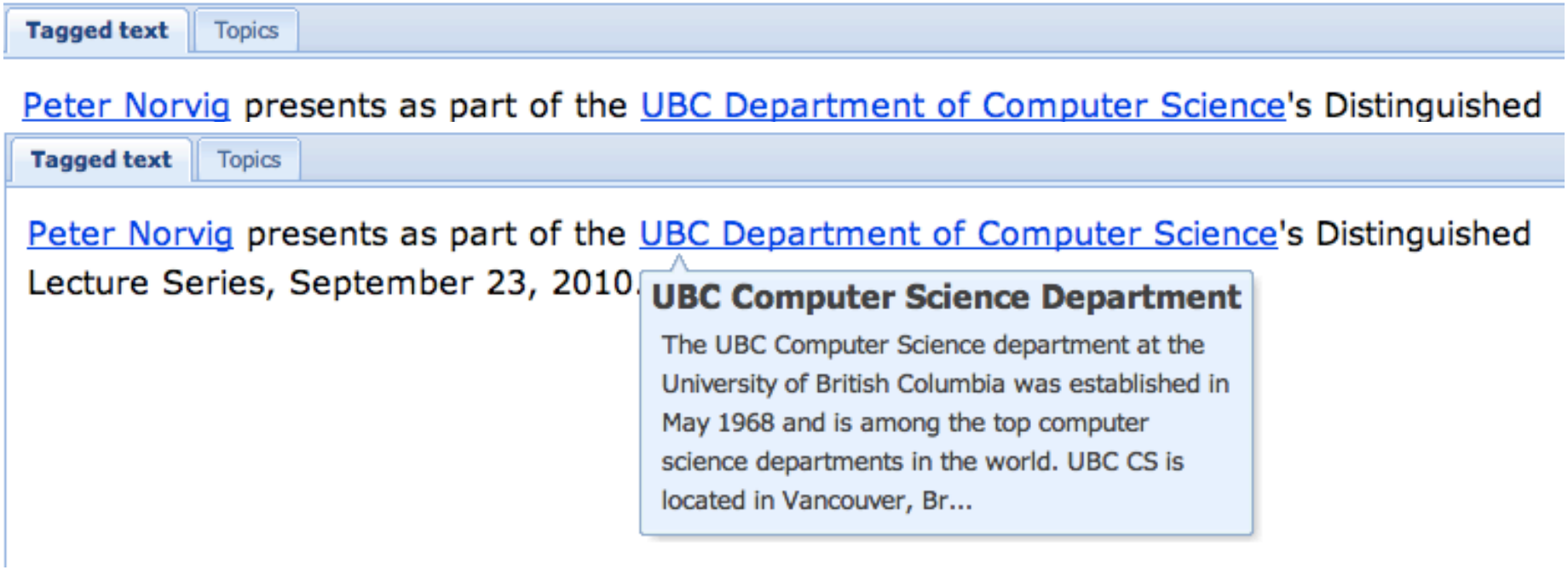
Semantičko indeksiranje kombinovanjem m. učenja i baza znanja

Semantičko indeksiranje (podsećanje)

- Klasično prepoznavanje entiteta u tekstu:

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

- Semantičko indeksiranje:



Semantičko indeksiranje

- Kombinacija nadgledanog m. učenja (klasifikacija) i znanja sadržanog u bazama znanja na Web-u
- Najčešće korišćene baze znanja: Wikipedia, DBpedia, WikiData
- Dodatna pogodnost ovih pristupa je jednostavnije kreiranje skupa podataka za obučavanje algoritma m. učenja

Poteškoće vezane za kreiranje korpusa za trening modela m. učenja

Jedan od glavnih izazova vezanih za prepoznavanje entiteta i semantičko indeksiranje primenom nadgledanog m. učenja, odnosi se na kreiranje dovoljno velikog skupa anotiranih dokumenata za trening modela

Primer teksta anotiranog za potrebe "obuke" modela nadgledanog m. učenja za prepoznavanje entiteta u tekstu:

Unlike <PERSON>Robert</PERSON>, <PERSON>John Briggs Jr </PERSON> contacted <ORGANIZATION>Wonderful Stockbrockers Inc </ORGANIZATION> in <LOCATION>New York</LOCATION> and instructed them to sell his <QUANTITY>100</QUANTITY> shares in <ORGANIZATION>Acme</ORGANIZATION>

Očigledno, priprema skupa podataka (korpusa) za trening je vrlo zahtevan zadatak…

JEDNOSTAVNIJE KREIRANJE TRENING SETA

- Na primer, u slučaju Wikipedia-e:
  - Svaki termin koji predstavlja interni link u Wikipedia-i – zvaćemo ga *anchor* – tretira se kao potencijalni entitet
  - Svaki *anchor* obezbeđuje nekoliko trening instanci:
    - jedan pozitivan primer: destinacija linka (Wikipedia stranica), odnosno "pravo" značenje datog anchor termina u datom kontekstu
    - više negativnih primera: sve ostale moguće destinacije linka, odnosno ostala moguća značenja datog anchor termina

Kreiranje dataset-a za obuku algoritma korišćenjem internih Wikipedia linkova – ilustracija pristupa

Za termin (anchor) *tree* postoji 26 mogućih destinacija (tj. značenja), što daje 1 poz. primer i 25 neg. primera za trening algoritma

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

Ovim pristupom se, npr., od svega 500 Wikipedia članaka, može dobiti dataset od > 50,000 instanci

OSNOVNI KORACI U PROCESU SEM. INDEKSIRANJA

1) *Entity spotting* & *candidate selection* – identifikacija termina koji bi mogli označavati entitete (*entity-mentions*) i selekcija mogućih entiteta iz baze znanja za svaki *entity-mention*
2) *Disambiguation* – izbor "najboljeg" entiteta za svaki *entity-mention,* tj. izbor entiteta koji najbolje odražava semantiku datog termina u datom kontekstu
3) *Filtering* – filtriranje rezultata u cilju eliminacije irelevantnih entiteta

ENTITY SPOTTING & CANDIDATE SELECTION

- Ciljevi prve faza procesa prepoznavanja entiteta su:
  - identifikovati tzv. *entity-mentions* u ulaznom tekstu, tj, delove teksta (pojedinačne reči i izraze) koji označavaju entitete;
  - identifikovati u bazi znanja (npr., Wikipedia ili DBpedia) skup mogućih entiteta za svaki *entity-mention*

ENTITY SPOTTING & CANDIDATE SELECTION

- Primer

"They performed Kashmir, written by Page and Plant.
Page played unusual chords on his Gibson."

Kandidati:
dbpedia:Kashmir – a valley between Pakistan, India and Ladakh
dbpedia:Kashmir_(band) – a Danish rock band
dbpedia:Kashmir_(song) – 1975 song by rock band Led Zeppelin
dbpedia:Kashmir,_Iran – a village in Iran
…

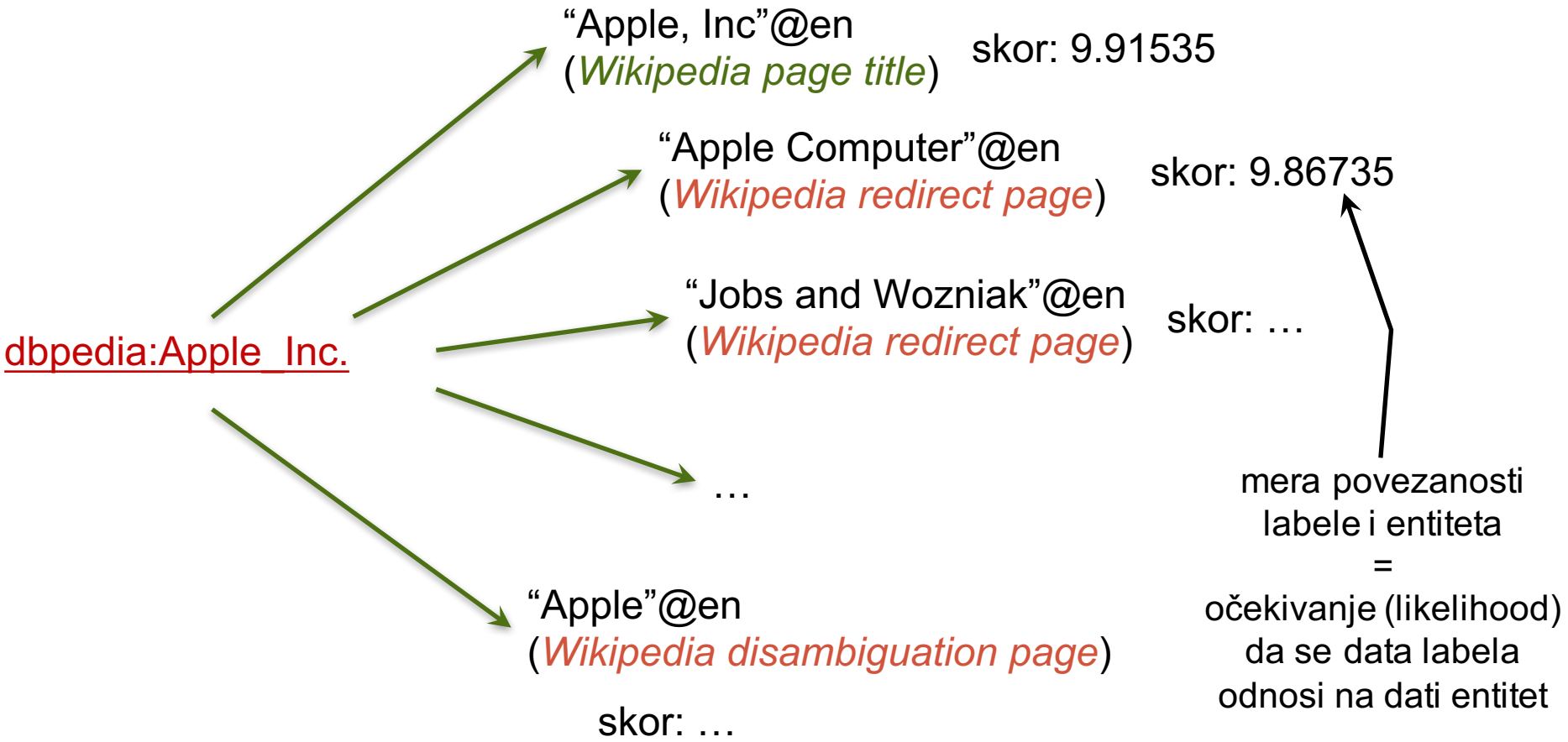
Izvor kandidata: Wikipedia/DBpedia disambiguation stranica za pojam Kashmir

Primer preuzet iz: Hoffart J. et al. (2011). Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), 782-792.

ENTITY SPOTTING & CANDIDATE SELECTION

- Ova faza se tipično realizuje kao *dictionary look-up task*
  - Formira se rečnik putem ekstrakcije labela i opisa svih entiteta sadržanih u izabranoj bazi znanja
  - Wikipedia i DBpedia se najčešće koriste kao baze znanja, odnosno kao izvori iz kojih se estrahuju labele i opisi entiteta
  - Rečnik može sadržati, za svaki entitet, i različite statistike
    - npr. relevantnost određene labele za određeni entitet

Primer: DBpedia Lexicalization dataset

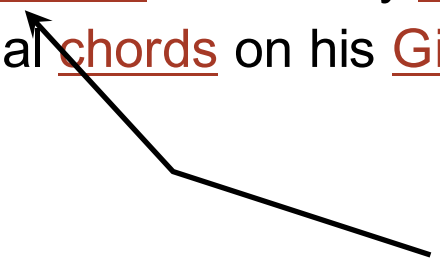


Dataset je raspoloživ na adresi: http://wiki.dbpedia.org/lexicalizations#

DISAMBIGUATION

- Cilj ove faze: za svaki *entity-mention*, selektovati jedan ili više entiteta koji mu po svom značenju (semantici) najviše odgovaraju
  - selekcija se radi iz, obično povećeg, skupa kandidata identifikovanih u prethodnoj fazi procesa
- Nastavljajući sa istim primerom:

"They performed Kashmir, written by Page and Plant.
Page played unusual chords on his Gibson."



dbpedia:Kashmir – a valley between Pakistan, India and Ladakh
dbpedia:Kashmir_(band) – a Danish rock band
dbpedia:Kashmir_(song) – 1975 song by rock band Led Zeppelin
dbpedia:Kashmir,_Iran – a village in Iran
…

DISAMBIGUATION

Postoji više različitih pristupa za realizaciju ove faze; neki od najčešće primenjivanih:

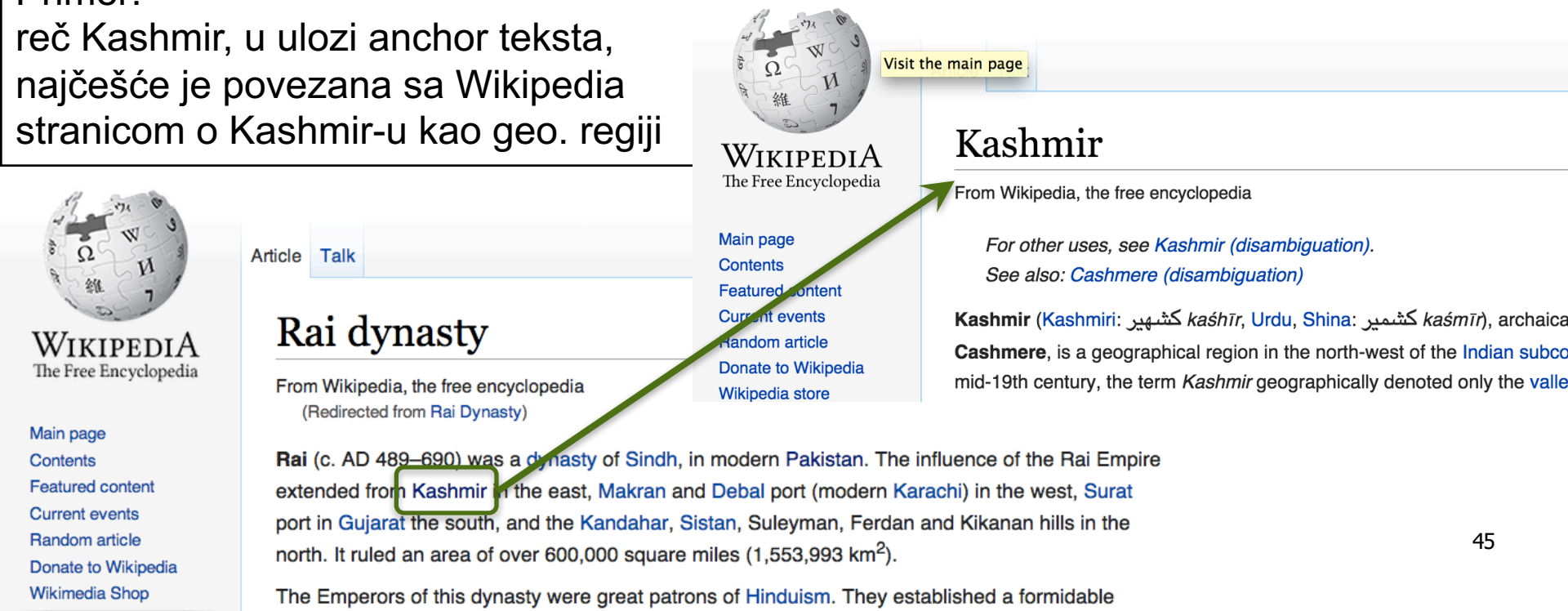
- *Popularity-based (mention-entity) prior*
- *Context-based approach*
- *Collective disambiguation*

DISAMBIGUATION: POPULARITY-BASED (MENTION-ENTITY) PRIOR

Ovaj pristup se sastoji u izboru najistaknutijeg entiteta za datu reč / frazu (*entity mention*)

- Npr., entitet sa kojim je data reč / izraz najčešće povezana kad se (ta reč/izraz) pojavljuje kao *anchor* tekst u Wikipedia-i

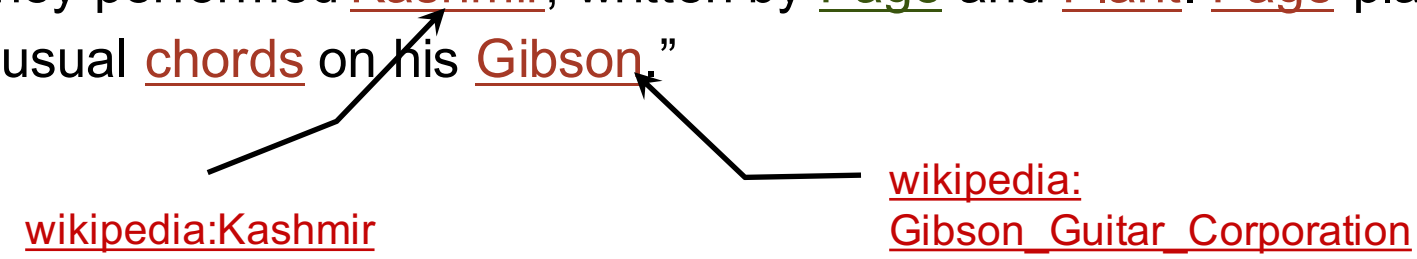
Primer:
reč Kashmir, u ulozi anchor teksta, najčešće je povezana sa Wikipedia stranicom o Kashmir-u kao geo. regiji



DISAMBIGUATION: POPULARITY-BASED (MENTION-ENTITY) PRIOR

Nastavljajući sa istim primerom:

"They performed Kashmir, written by Page and Plant. Page played unusual chords on his Gibson."



U Wikipedia-i,

- "Gibson" je primarno povezan sa Gibson guitar corporation entitetom, dok je samo marginalno povezan sa 24 preostala moguća značenja ovog termina
- "Kashmir" je primarno povezan sa Kashmir region entitetom (90.91% svih pojavljivanja ovog termina), dok se samo retko odnosi na pesmu grupe Led Zeppelin (5.45%)

DISAMBIGUATION: POPULARITY-BASED (MENTION-ENTITY) PRIOR

- Ovo je jednostavan pristup, ali često podložan greškama; zbog toga se koristi u kombinaciji sa drugim pristupima
- Greške se javljaju usled toga što se ne pridaje pažnja
  - kontekstu u kome se reč/izraz (mention) pojavljuje
  - generalnoj temi teksta

Ilustracija greške koja se obično javlja ukoliko se samo popularnost tj učestanost mention-entity konekcije uzima u razmatranje

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

DISAMBIGUATION: PRISTUP ZASNOVAN NA KONTEKSTU

- Jedan od često korišćenih pristupa za realizaciju ove faze
- Zasniva se na poređenju konteksta određenog *entity-mention*-a i konteksta svih entiteta koji su selektovani kao kandidati za taj *entity-mention*
- Za kontekst *entity-mention*-a se tipično uzima rečenica u kojoj se pojavljuje, dok se za kontekst entiteta uzima njegov opis iz baze znanja

DISAMBIGUATION: PRISTUP ZASNOVAN NA KONTEKSTU

- Kontekst se obično predstavlja kao prost skup reči tj. koristi se bag-of-words pristup za predstavljanje teksta
- Poređenje konteksta se vrši primenom neke od metrika za računanje sličnosti vektora
- Često korišćene metrike:
  - Cosine similarity,
  - (weighted) Jaccard coefficient,
  - Wikipedia links-based measure*

* Witten, I.H. & Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence, Chicago, USA, July, 2008. (pp. 25-30).

DISAMBIGUATION: PRISTUP ZASNOVAN NA KONTEKSTU

"They performed Kashmir, written by Page and Plant. Page played unusual chords on his Gibson."

bag-of-words →

perform
Kashmir
write
Page
Plant
play
chord
…

http://en.wikipedia.org/wiki/Kashmir_(song)
…was written by Jimmy Page and Robert Plant…
…performed by the band at almost every concert…

bag-of-words →

write
Jimmy
Page
Robert
Plant
perform
band
concert
…

similarity

similarity

+ 15 more candidate entities

http://en.wikipedia.org/wiki/Kashmir
…northwestern region of the Indian subcontinent.…
…became an important center of Hinduism and later of Buddhism…

bag-of-words →

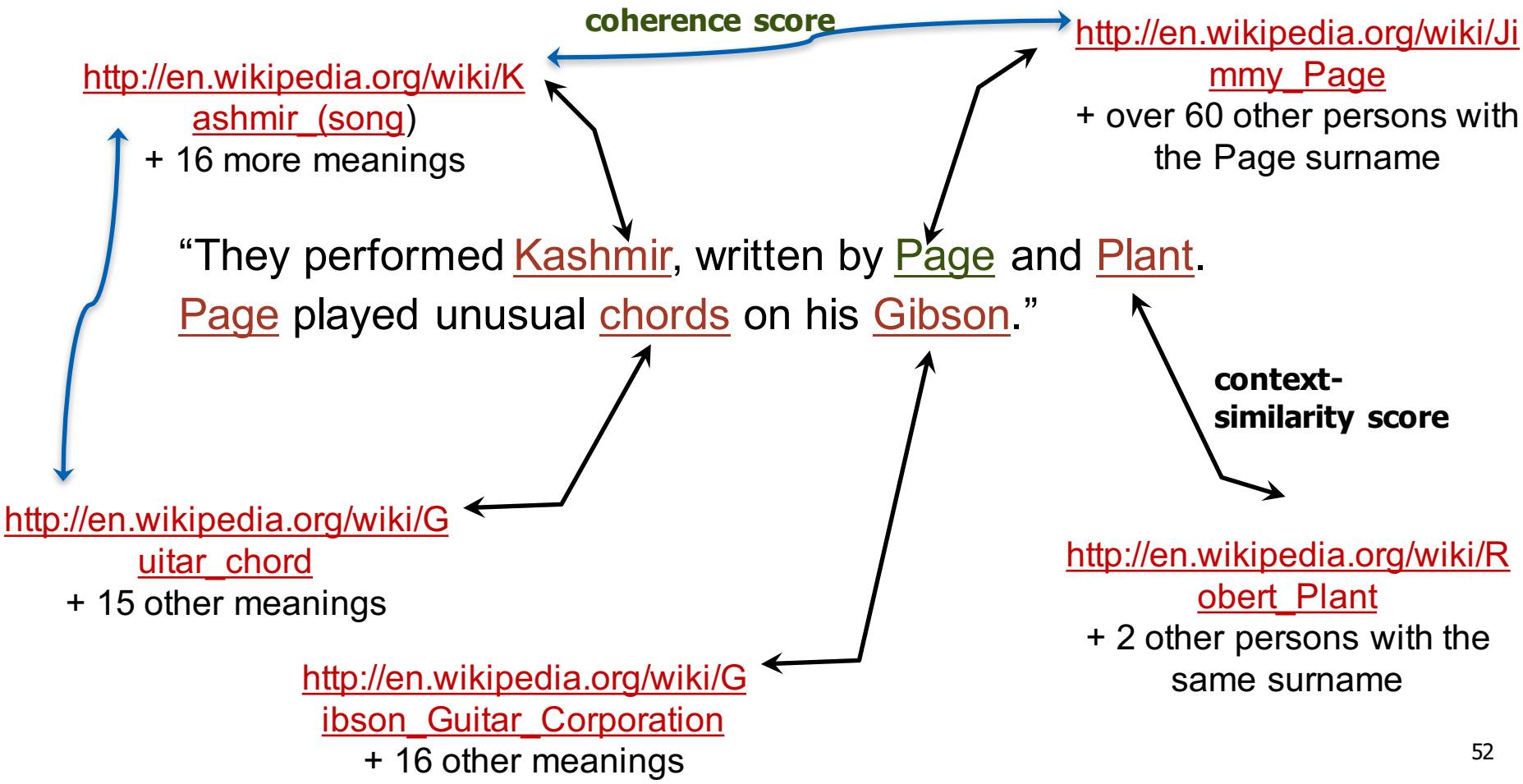
northwest
region
India
subcontinent
center
Hinduism
Buddhism
…

DISAMBIGUATION: COLLECTIVE DISAMBIGUATION APPROACH

- Sastoji se u istovremenom određivanju značenja (*disambiguation*) više reči/izraza (*mentions*) u tekstu koji je predmet analize
- Predstavlja proširenje pristupa zasnovanog na kontekstu:
  - pored računanja skora za kontekstualnu sličnost *(context similarity)* svakog *mention-entity* para, računa se takođe i koherentnost (*coherence score*) za sve parove entiteta
  - koherentnost je u ovom slučaju definisana kao semantička povezanost (*semantic relatedness*) razmatranih entiteta

DISAMBIGUATION: COLLECTIVE DISAMBIGUATION APPROACH

Nastavljajući sa istim primerom:



DISAMBIGUATION: COLLECTIVE DISAMBIGUATION APPROACH

- Ovaj pristup daje dobre rezultate ukoliko
  - postoji dovoljno veliki broj entiteta pomenutih u tekstu, i
  - pomenuti entiteti čine tematski homogen skup
- Greške se najčešće javljaju u slučaju da
  - tekst razmatra više nepovezanih ili slabo povezanih tema
  - entiteti sa kojima reči/izrazi (*mentions*) iz teksta mogu biti povezane, mogu formirati više tematski koherentnih grupa; na primer:

"Real Madrid and Barcelona edge out Manchester and Chelsea to secure trials for Argentine wonder-kid"

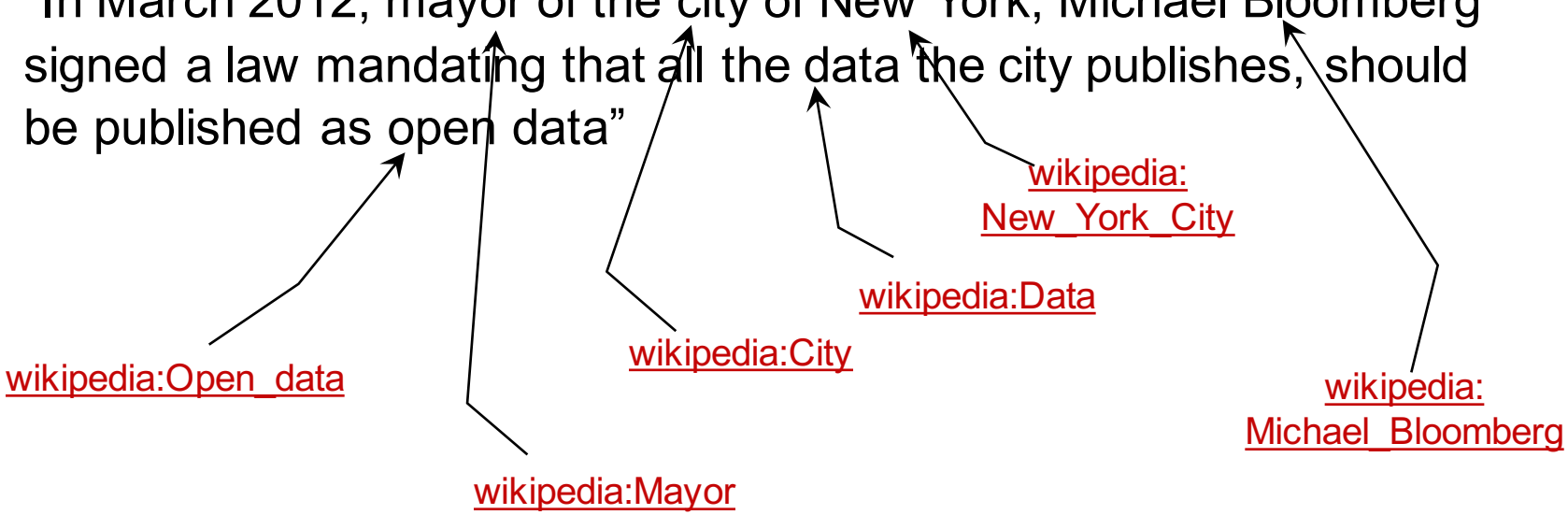
Ovde imamo potencijalno dve tematski koherentne grupe entiteta: lokacije (gradovi) i fudbalski klubovi

FILTERING

- Cilj ove faze je da se iz skupa rezultata uklone oni entiteti koji najverovatnije ne bi bili relevatni korisniku
  - npr., entiteti koji se odnose na neke opšte koncepte ili oni koji su samo marginalno povezani sa glavnom temom teksta

- Primer

"In March 2012, mayor of the city of New York, Michael Bloomberg signed a law mandating that all the data the city publishes, should be published as open data"



Performanse današnjih alata za semantičko indeksiranje

Tip teksta: novinski članci

	AlchemyAPI	dataTXT	DBpedia	Lupedia	Textrazor	THD	Yahoo!	Zemanta
p	**70.63**	39.20	26.93	57.98	49.21	32.50	61.24	35.58
r	14.05	**54.93**	42.21	29.90	51.66	40.10	9.65	7.78
f	23.43	45.75	32.88	39.45	**50.41**	35.90	16.68	12.77

p – precision; r – recall; f – F1 measure

(napomena: dataTXT je prerastao u komercijalni servis DandelionAPI)

Izvor: Rizzo et al. 2014. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web, 9th Edition of the Language Resources and Evaluation Conference (LREC'14)

Performanse današnjih alata za semantičko indeksiranje

Tip teksta: poruke sa Twitter-a

	AlchemyAPI	dataTXT	DBpedia	Lupedia	Textrazor	THD	Yahoo!	Zemanta
p	**72.22**	22.11	13.99	37.37	30.69	23.54	60.68	35.54
r	3.91	34.74	29.70	11.13	**34.89**	23.98	10.68	10.08
f	7.42	27.02	19.02	17.15	**32.65**	23.76	18.16	15.70

p – precision; r – recall; f – F1 measure

Izvor: Rizzo et al. 2014. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web, 9th Edition of the Language Resources and Evaluation Conference (LREC'14)

WIKILINKS CORPUS

- Najveći javno dostupan dataset za obuku algoritama nadgledanog m. učenja za semantičko indeksiranje, konkretno, prepoznavanje Wikipedia entiteta u tekstu
- URL: http://www.iesl.cs.umass.edu/data/wiki-links
- Osnovni podaci o dataset-u:
  - 10 miliona Web stranica
  - 3 miliona Wikipedia entiteta
  - 40 miliona jedinstveno identifikovanih pominjanja entiteta
  - publikovan 08.03.2013. od strane Google Research-a
- Više informacija u članku: Learning from Big Data: 40 Million Entities in Context

FREEBASE ANNOTATIONS OF SOCIAL MEDIA CONTENT

- Google Freebase Annotations of TREC KBA 2014 Stream Corpus
  - TREC – Text Retrieval Conference
  - KBA – Knowledge Base Acceleration
- URL: http://trec-kba.org/data/fakba1/index.shtml
- Osnovni podaci o korpusu:
  - 394M dokumenata sa bar jednim anotiranim Freebase entitetom
  - 9.4 milijarde reči/izraza (*mentions*) povezanih sa Freebase entitetima
  - anotacije su urađene automatski i samim tim nisu savršene
  - na osnovu ručno analiziranog slučajnog uzorka, procenjeno je:
    - ~9% reči/izraza (*mentions*) je povezano sa pogrešnim Freebase entitetima
    - ~8% reči/izraza (*mentions*) koji predstavljaju entitete nisu povezani sa odgovarajućim Freebase entitetima

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

http://goo.gl/cqdp3I