

Language Technology on News Portals

Nikola Ljubešić, research assistant
Department of Information Sciences
Faculty of Humanities and Social Sciences

Information Technology and Journalism 14
Dubrovnik, 25-29 May 2009

Language technology

- language technology - dealing with data encoded in natural language
- two aspects - two areas of interest
 1. searching - morphological normalization (stemming)
 2. browsing - linking news articles by content

Research overview

1. manual exploration of most popular Croatian news portals

- searching and morphological normalization
- browsing through “related news”

2. research on stemming for information retrieval

3. example of content analysis - event detection

Stemming

- process for reducing inflected words to their stem, method of morphological normalization
- first stemmer for english in 1968
- morphological normalization of Croatian mostly through morphological lexicons (Kržak, Boras 1985; Tadić 1994; Agić, Tadić 2006)

Content analysis






- vector space model
- statistical measures for weighting features (TF-IDF)
- measures of vector similarity (cosine)
- clustering data points into clusters

I. Croatian news portals

Methodology

- investigating manually most popular news portals
 1. stemming - “reverse engineering” approach
 2. “related news” methods analyzed

Results

	search	stemming	“related news”
	internal, fair	no	recent news in same category
	internal, fair	no	related news through manual tags
	internal, fair	no	none
	internal, good	no	previous articles, manually assigned
	internal, fair	no	recent news in same category

Google: keyword site:domain.hr?

2. Stemming

Methodology

- using a morphosyntactically tagged corpus - Vjesnik on-line newspaper, cca. 100Mw (tagged by Institute for Linguistics at the Faculty of Humanities and Social Sciences, Zagreb)
- basic resource for empirical evaluation of simple stemmers

Example of tagged corpus

ministara		ministar	Ncmpg
Heraka	Herak	Npmsg	
i	i	Ccs	
Porgesa	Porges	Npmsg	
'			
održanoj		održan	Afpfsl-
prije	prije	Spsg	
dva	dva	Mc-p-l	
dana	dan1	Ncmpg	
na	na	Spsl	
brodu	brod	Ncmsl	
»			
Marko	Marko	Npmsn	
Polo			
«			
između			
Korčule	Korčula	Npfsg	
i	i	Ccs	
Dubrovnika	Dubrovnik	Npmsg	

Methodology 2

- data excerpted from tagged corpus
`{u'dan1': {u'danima': 2542, u'dan': 21871, u'danom': 1294, u'dani': 1661, u'dane': 2331, u'danu': 1286, u'dana': 88883}, u'dan2', {u'dano': 26, u'danome': 2, u'danoj': 37, u'danom': 167, u'danog': 46, u'danoga': 4, u'danih': 139, u'danim': 130}, ...}`
- no stemmer, query “dan” finds 21.871 of 119.868 occurrences
- stripping off vocals, 116.058 found, 26 wrong

Methodology 3

- 10.000 random noun queries, nominative singular, recording difference in precision, recall and F_1
- based on research from 2007 at query sample from <http://www.hr>, 83.98% nouns, 8.99% adjectives, 4.69% english terms, 1.48% acronyms (72.46% nominative singular, 16.66% nominative plural)

Three stemmer implementations

1. stripping off vocals
2. stripping off 20 suffixes chosen by a greedy optimization algorithm
3. 266 most common suffixes - very often method

Results

	precision	recall	F ₁
no stemmer	88,19%	31,91%	46,86%
stemmer 1	81,05%	82,64%	81,84%
stemmer 2	85,44%	89,66%	87,49%
stemmer 3	11,76%	93,99%	20,91%

3. Content analysis

Documents example

	d1	d2
title	Kina i tibetska vlada u egzilu dogovorili nove razgovore	Tibetanska vlada u egzilu dogovorila pregovore
text	Zatvoren za javnost, održan u nedjelju u južnokineskom Shenzhen blizu Hong Konga, sastanak je bio prvi od izbivanja protukineskih prosvjeda u Lhasi i obližnjim pokrajinama, koje su kineske vlasti gušile uhićenjima, javlja agencija...	Izaslanici kineske i tibetske vlade u egzilu dogovorili su, na prvome izravnom sastanku, nastaviti razgovore o Tibetu koji je, zbog vala prosvjeda protiv kineske okupacije te pokrajine postao svjetska tema uoči Olimpijskih igara u Pekingu...
source	dnevnik.hr	javno.com

Document vectors

feature	d1	d2
lhasi	0,0317	0,0193
egzilu	0,0317	0,0387
kine	0,0239	0,0262
vlada	0,0096	0,0088
gradu	0,0062	0,0068
je	0,0005	0,0004
...		

cosine similarity 0,8941

Similarity matrix

	d1	d2	d3	...
d1	1,0	0,894	0,016	
d2	0,894	1,0	0,019	
d3	0,016	0,019	1,0	
...				

... clustering...

Cluster I

Bolivija: prošao referendum za veću autonomiju Santa Cruza, totalportal.hr

Bogati napuštaju Moralesa, glas-slavonije.hr

Najbogatija bolivijska pokrajina izglasala veću autonomiju, business.hr

Prošao referendum za veću autonomiju Santa Cruza, tportal.hr

Bolivija: Prošao referendum za veću autonomiju Santa Cruza, vecernji.hr

Bolivijska najbogatija pokrajina želi autonomiju, javno.com

Hrvat vodi "pobunu" protiv Moralesa, jutarnji.hr

Najbogatija bolivijska provincija proglasila autonomiju, rtl.hr

Morales: referendum je promašena separatistička mjera, seebiz.eu

Bolivija: Prošao referendum za veću autonomiju Santa Cruza, dnevnik.hr

Najveća pokrajina Bolivije na referendumu izabrala autonomiju, index.hr

Cluster 2

Obama vratio veću potporu među demokratskim biračima, jutarnji.hr

Ankete: Obama vratio veću potporu među demokratskih biračima, vecernji.hr

Obama hita prema sve većoj potpori demokrata, javno.com

Obama vratio veću potporu među demokratskih biračima, tportal.hr

Obama povećao prednost pred Hillary Clinton, totalportal.hr

Obama vratio veću potporu među demokratskih biračima, seebiz.eu

Obama ponovo u anketama pobjeđuje Clinton, business.hr

Obama vratio veću potporu među demokratskih biračima - ankete, dnevnik.hr

Obama vratio veću potporu među demokratskim biračima, nacional.hr

Obama vratio potporu glasača i povećao prednost pred Clinton, index.hr

Cluster 3

Tadiću prijete ubojstvom zbog izdaje, javno.com

Prijetnje smrću srbijanskom predsjedniku Borisu Tadiću, totalportal.hr

Borisu Tadiću prijete 'metkom u čelo', dnevnik.hr

Borisu Tadiću prijete smrću, nacional.hr

Predsjedniku Srbije Tadiću prijete "metkom u čelo", business.hr

Borisu Tadiću prijete metkom u čelo, tportal.hr

Borisu Tadiću prijete smrću, rtl.hr

Borisu Tadiću prijete smrću, mojportal.hr

Borisu Tadiću prijete metkom u čelo, jutarnji.hr

Cluster I I

Kovačević razočaran, vijesti.hrt.hr

Kovačević razočaran jer nema sučeljavanja pred SDP-ovim članstvom, mojportal.hr

Dragan Kovačević protiv sučeljavanja u Otvorenom, index.hr

Kovačević razočaran jer je Milanović izbjegao sučeljavanje pred članstvom SDP-a, business.hr

Kovačević: Gdje je nestao čovjek?, javno.com

Kovačević: Gdje je nestao čovjek?, seebiz.eu

Kovačević: "U SDP-u nije zaživjela demokratska praksa javnog sučeljavanja", vecernji.hr

Večeras u Otvorenom "javno sučeljavanje" kandidata SDP-a, nacional.hr

Evaluation results

purity	0,9638
NMI	0,9752
rand index	0,9987
precision	0,8693
recall	0,6427
F_1	0,7390
$F_{0.5}$	0,8120

Thank you