





# Uvod u obradu prirodnog jezika

## 7.1. Što je analiza sentimenta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

# Pozitivna ili negativna kritika filma?

- nevjerojatno razočarenje 
- pun otkačenih likova i bogato primijenjena satira, s nekim velikim zapletima radnje 
- ovo je najveća ekscentrična komedija ikad snimljena 
- ovo je jadno, najgori dio je definitivno scena boksa 

# Google Product Search



**HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner**

**\$89 online, \$100 nearby** ★★★★★ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

## Reviews

**Summary** - Based on 377 reviews



What people are saying

ease of use	<div><div></div><div></div></div>	"This was very easy to setup to four computers."
value	<div><div></div><div></div></div>	"Appreciate good quality at a fair price."
setup	<div><div></div><div></div></div>	"Overall pretty easy setup."
customer service	<div><div></div><div></div></div>	"I DO like honest tech support people."
size	<div><div></div><div></div></div>	"Pretty Paper weight."
mode	<div><div></div><div></div></div>	"Photos were fair on the high quality mode."
colors	<div><div></div><div></div></div>	"Full color prints came out with great quality."

# HP Officejet 6500A E710N Multifunction Printer

[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



**\$121.53 - \$242.39** (14 stores)

☐ Compare

Average rating ★★★★★ (144)

★★★★★	<div></div>	(55)
★★★★★	<div></div>	(54)
★★★★★	<div></div>	(10)
★★★★★	<div></div>	(6)
★★★★★	<div></div>	(23)
★★★★★	<div></div>	(0)

Most mentioned

Performance	<div></div>	(57)
Ease of Use	<div></div>	(43)
Print Speed	<div></div>	(39)
Connectivity	<div></div>	(31)
More ▼		

Show reviews by source

- Best Buy (140)
- CNET (5)
- Amazon.com (3)

# Zašto analizirati sentiment

- Film: je li kritika pozitivna ili negativna?
- Produkti: što ljudi misle o novom štampaču?
- Javni sentiment: koliko je povjerenje potrošača?
- Politika: što ljudi misle o kandidatu?
- Predikcija: predviđanje rezultata izbora ili trendova na tržištu

# Konotacija

- Riječi osim značenja imaju i konotaciju.
- Konotacija je sentiment koju riječ ili fraza posjeduje
- Možemo li izgraditi leksički resurs kojim se reprezentiraju konotacije?
- i iskoristiti ga u obradi prirodnog jezika?

# Shrererova tipologija afektivnih stanja

- **Emocija:** evaluacija glavnih događaja
  - ljut, tužan, veseo, strah, sram, ponosan, ushićen
- **Raspoloženje:** razlikovanje ne izazvanih, dugotrajnih promjena subjektivnog osjeta
  - veseo, tmuran, razdražljiv, bezvoljan, depresivan, poletan
- **Međuljudski odnosi:** afektivni stavovi prema drugoj osobi
  - prijateljski, koketirajući, hladan, topao, podržavajući, prijezirni
- **Stavovi:** postojeća, afektivno obojena vjerovanja, dispozicije prema predmetima ili osobama
  - naklonost, ljubav, mržnja, vrijeđanje, želja
- **Osobne crte:** stabilna stanja osobe i tipično ponašanje
  - nervozan, tjeskoban, bezobziran, mrzovoljan, neprijateljski, ljubomorani

# Analiza sentimenta

- Najjednostavniji zadatak
  - je li stav teksta pozitivan ili negativan
- Složeniji zadatak
  - Ocijeni stav u tekstu od 1 do 5
- Napredni zadatak
  - odredi cilj, izvor ili kompleksne vrste stavova



# Uvod u obradu prirodnog jezika

## 7.2. Osnovni algoritam

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

# Klasifikacija sentimenta kritike filma

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

- Detekcija polariteta
  - je li IMDB kritika filma pozitivna ili negativna
- Podaci: *Polarity Data 2.0*
  - <http://www.cs.cornell.edu/people/pabo/movie-review-data>

# IMDB podaci u Pang i Lee bazi podataka



when \_star wars\_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...]  
when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .  
cool .  
\_october sky\_ offers a much simpler image—that of a single white dot , traveling horizontally across the night sky . [ . . . ]



"snake eyes" is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .  
it's not just because this is a brian depalma film , and since he's a great director and one who's films are always greeted with at least some fanfare .  
and it's not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

# Osnovni algoritam (adaptacija od Pang i Lee)

- Tokenizacija
- Ekstrakcija osobina
- Klasifikacija korištenjem različitih klasifikatora
  - Naivni Bayes
  - MaxEnt
  - SVN

# Problemi tokenizacije sentimenta

- Problemi HTML i XML oznaka
- Twitter oznake (imena, hash oznake)
- Kapitalizacija (sačuvaj riječi koje su pisane u velikim znakovima)
- Brojevi telefona, datumi
- Emotikon

[<>]?	# kapa, obrve
[;=8]	# oči
[\-o\*\']?	# nos
[\\)\]\(\[dDpP/\:}\{\@\\ \\]	# usta
	#### obrnuti smjer
[\\)\]\(\[dDpP/\:}\{\@\\ \\]	# usta
[\-o\*\']?	# nos
[;=8]	# oči
[<>]?	# kapa, obrve

Christopher Potts tokenizator sentimenta:

<http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py>

Brendan O'Connor twitter tokenizator:

<https://github.com/brendano/tweetmotif>

# Ekstrakcija osobina za klasifikaciju sentimenta

- Kako se odnositi s negacijom?
  - Ne sviđa mi se ovaj film
  - vs.
  - Zaista mi se sviđa ovaj film
- Koje riječi koristiti?
  - samo pridjeve
  - sve riječi
    - korištenje svih riječi se pokazuje boljim

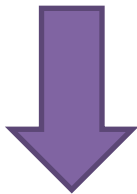
# Negacija

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

- Dodaj NE\_ svakoj riječi između negacije i sljedeće interpunkcije:

Ne sviđa mi se ovaj film, ali ...



Ne NE\_sviđa NE\_mi NE\_se NE\_ovaj NE\_film, ali ...

# Naivni Bayes: podsjetnik

$$c_{NB} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(c) \prod_{i \in \text{pozicije}} P(w_i | c)$$

$$\hat{P}(w_i | c) = \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} (\text{broj}(w, c) + 1)}$$



# Binarni naivni Bayes

- Ideja:
  - za sentiment (i vjerojatno za druge domene klasifikacije teksta)
    - **pojavljivanje riječi** može značiti više od frekvencije riječi
      - pojavljivanje riječi "fantastično" govori nam mnogo
      - činjenica da se riječ "fantastično" pojavljuje 5 puta ne govori nam ništa više
  - Binarni (Booleov) naivni Bayes
    - sažima broj riječi u svakom dokumentu na 1

# Binarni multinominalni naivni Bayes: učenje

**UČENJE**( $D, C$ )

**za svaku** klasu  $c \in C$

$N_{doc} \leftarrow$  broj dokumenata iz  $D$

$N_c \leftarrow$  broj dokumenata iz  $D$  klase  $c$

$prior[c] \leftarrow \frac{N_c}{N_{doc}}$

$V \leftarrow$  riječnik dokumenata iz  $D$

$megadoc[c]$  proširi **s jedinstvenim riječima** iz  $d \in D$   
koji su klase  $c$

**za svaku** riječ  $w \in V$

$broj[w, c] \leftarrow$  broj pojavljivanja od  $w$  u  $megadoc(c)$

$izvjesnost[w, c] \leftarrow \frac{broj[w, c] + 1}{\sum_{w' \in V} (broj[w', c] + 1)}$

**vrați**  $prior, izvjesnost, V$

## Binarni naivni Bayes na testnom dokumentu $d$

- Izbaci sve duplikate riječi iz  $d$
- Izračunaj NB koristeći istu formulu

$$c_{NB} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(c) \prod_{i \in \text{pozicije}} P(w_i | c)$$

# Normalni vs. Binarni multinominalni naivni Bayes

- Normalni

	Dokument	Riječi	Klasa
Treniranje	d <sub>1</sub>	Italija Rim Italija	IT
	d <sub>2</sub>	Italija Italija Firenca	IT
	d <sub>3</sub>	Italija Ankona	IT
	d <sub>4</sub>	Pariz Francuska Italija	FR
Test	d <sub>5</sub>	Italija Italija Italija Pariz Francuska	?

- Binarni

	Dokument	Riječi	Klasa
Treniranje	d <sub>1</sub>	Italija Rim	IT
	d <sub>2</sub>	Italija Firenca	IT
	d <sub>3</sub>	Italija Ankona	IT
	d <sub>4</sub>	Pariz Francuska Italija	FR
Test	d <sub>5</sub>	Italija Pariz Francuska	?

# Binarni naivni Bayes

- B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
- V. Metsis, I. Androutsopoulos, G. Paliouras. 2006. Spam Filtering with Naive Bayes – Which Naive Bayes? CEAS 2006 - Third Conference on Email and Anti-Spam.
- K.-M. Schneider. 2004. On word frequency information and negative evidence in Naive Bayes text classification. ICANLP, 474-485.
- JD Rennie, L Shih, J Teevan. 2003. Tackling the poor assumptions of naive bayes text classifiers. ICML 2003

- Binarni klasifikator se pokazuje boljim
  - On se razlikuje od multivarijabilnog Bernoullijevog naivnog Bayesa (MBNB)
    - MNBN ne radi dobro kod analize sentimenta i drugih zadataka obrade teksta
- Druga mogućnost
  - $\log(freq(w))$
- MaxEnt i SVN teže boljem učinku od naivnog Bayesa

# Problemi: što čini kritikom teškom za obradu?

- Suptilnost
  - Kritika parfema

"Ako ovo čitate, jer se radi o vašem dragom mirisu, molimo vas da ga nosite isključivo kod kuće, a prozore zalijepite trakama"
  - Kritika glumice

"Ona iznosi sve emocije na skali A do B"

# Oprečna očekivanja i efekt redoslijeda

- Ovaj film bi trebao biti **briljantan**. Izgleda da ima **sjajnu** radnju, glumci su **prvoklasni**, kao i sporedne uloge. Stallone pokušava ostvariti **dobru** izvedbu. Međutim, **ne može je održati**.
- Kao i obično Keanu Reeves nije ništa specijalan, ali iznenađujuće je što ni **vrlo talentirani** Laurence Fishbourne **nije toliko dobar**. Iznenađen sam.

# Uvod u obradu prirodnog jezika

## 7.3. Leksikon sentimenta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning



# General Inquirer

- Stranica:
  - <http://www.wjh.harvard.edu/~inquirer>
- Lista kategorija:
  - <http://www.wjh.harvard.edu/~inquirer/homecat.htm>
- Tablica:
  - <http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls>
- Kategorije:
  - Pozitivna (1915 riječi) i Negativna (2291 riječi)
  - Snažno - Slabo, Aktivno - Pasivno, Naglašeno - Nenametljivo
  - Uгода, Bol, Vrlina, Porok, Motivacija, Kognitivna orijentacija, itd.
- Slobodan za istraživačke svrhe

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

# General Inquirer

## Positive

admire, amazing, assure, celebration, charm, eager, enthusiastic, excellent, fancy, fan- tastic, frolic, graceful, happy, joy, luck, majesty, mercy, nice, patience, perfect, proud, rejoice, relief, respect, satisfactorily, sensational, super, terrific, thank, vivid, wise, won- derful, zest

## Negative

abominable, anger, anxious, bad, catastrophe, cheap, complaint, condescending, deceit, defective, disappointment, embarrass, fake, fear, filthy, fool, guilt, hate, idiot, inflict, lazy, miserable, mourn, nervous, objection, pest, plot, reject, scream, silly, terrible, unfriendly, vile, wicked

# LIWC (Linguistic Inquiry and Word Count)

- Stranica:
  - <http://www.liwc.net/>
- 2300 riječi, preko 70 klasa
- **Afektivni procesi**
  - Negativne emocije (loš, čudan, mrzi, problem, naporan)
  - Pozitivne emocije (ljubav, lijepo, slatko)
- **Kognitivni procesi**
  - Probni (možda, valjda, pretpostavljam)
  - Inhibicija (blokirati, ograničiti)
- **Zamjenice, Negacija** (ne, nikad), **Kvantifikatori** (neki, mnogi)
- \$30 do \$90

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

# LIWC (Linguistic Inquiry and Word Count)

Positive emotion	Negative emotion	Insight	Inhibition	Family	Negate
appreciat*	anger*	aware*	avoid*	brother*	aren't
comfort*	bore*	believe	careful*	cousin*	cannot
great	cry	decid*	hesitat*	daughter*	didn't
happy	despair*	feel	limit*	family	neither
interest	fail*	figur*	oppos*	father*	never
joy*	fear	know	prevent*	grandf*	no
perfect*	griev*	knew	reluctan*	grandm*	nobod*
please*	hate*	means	safe*	husband	none
safe*	panic*	notice*	stop	mom	nor
terrific	suffers	recogni*	stubborn*	mother	nothing
value	terrify	sense	wait	niece*	nowhere
wow*	violent*	think	wary	wife	without

## Primjer 5 od 73 leksičke kategorije u LIWC

\* prethodne riječi su prefiksi i sve riječi s ovim prefiksom su također uključene

# MPQA Subjectivity Cues Lexicon

- Stranica:
  - [http://mpqa.cs.pitt.edu/lexicons/subj\\_lexicon/](http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/)
- 6885 riječi od 8221 lema
  - 2718 pozitivnih
  - 4912 negativnih
- Svaka riječ ima oznaku intenziteta (jak, slab)
- GNU GPL

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

# Hu and Liu Opinion Lexicon

- Bing Liu stranica za Opinion Mining
  - <http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>
- 6786 riječi
  - 2006 pozitivne
  - 4783 negativne

- Stranica:
  - <http://sentiwordnet.isti.cnr.it/>
- Svi WordNet-ovi skupovi sinonima imaju oznake stupnja pozitivnosti, negativnosti i neutralnosti (objektivnosti)
  - [procijenjen(J,3)] "može biti izračunat ili procijenjen"
  - Poz 0 Neg 0 Obj 1
  
  - [smatrati(J,1)] "smatra se dobrim učenikom"
  - Poz 0.75 Neg 0 Obj 0.25

# Nesuglasice među polaritetima u leksikonima

Christopher Potts, Sentiment Tutorial, 2011

<http://sentiment.christopherpotts.net/lexicons.html>

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
General Inquirer			520/2306 (23%)	1/204 (0.5%)
SentiWordNet				174/694 (25%)
LIWC				



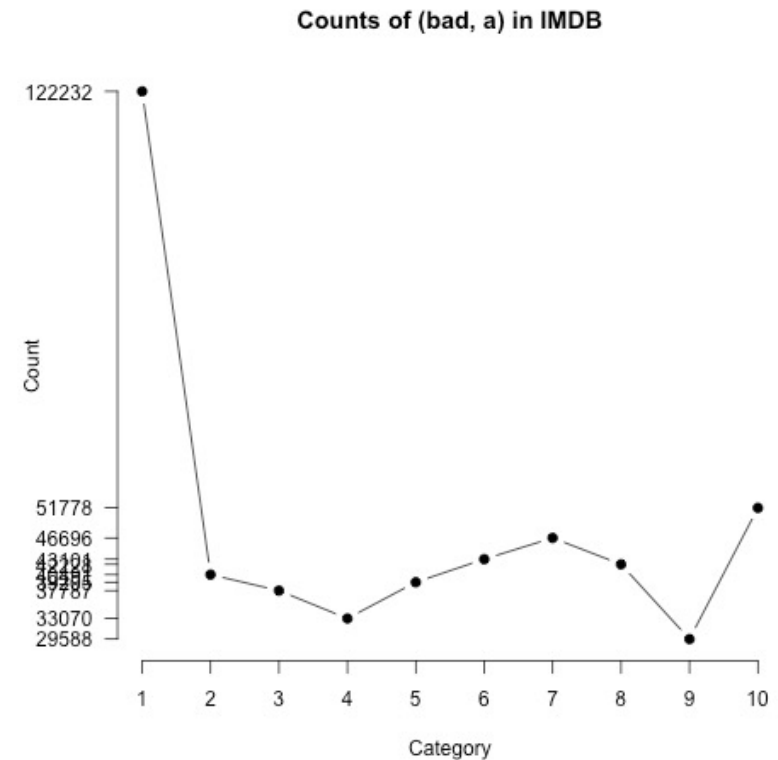
# Analiza polariteta svake riječi u IMDB kritici

- Koliko vjerojatno će se svaka riječ pojaviti u svakoj klasi sentimenta?
  - Prebroji "loš" ("bad") u kritici od 1-zvijezde, 2-zvijezde, 3-zvijezde, itd.
- Ali ne može se koristiti sirovo prebrojavanje:
  - Umjesto **izglednosti**

$$P(w|c) = \frac{f(w, c)}{\sum_{w' \in C} f(w', c)}$$

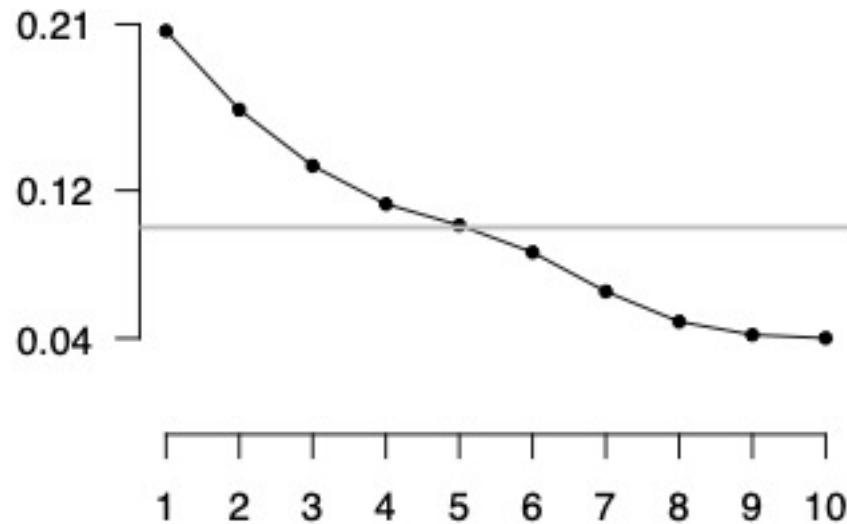
- Potrebno je riječi učiniti međusobno usporedivim
  - **Skalirana izglednost**

$$Pott(w|c) = \frac{P(w|c)}{\sum_c P(w|c)}$$



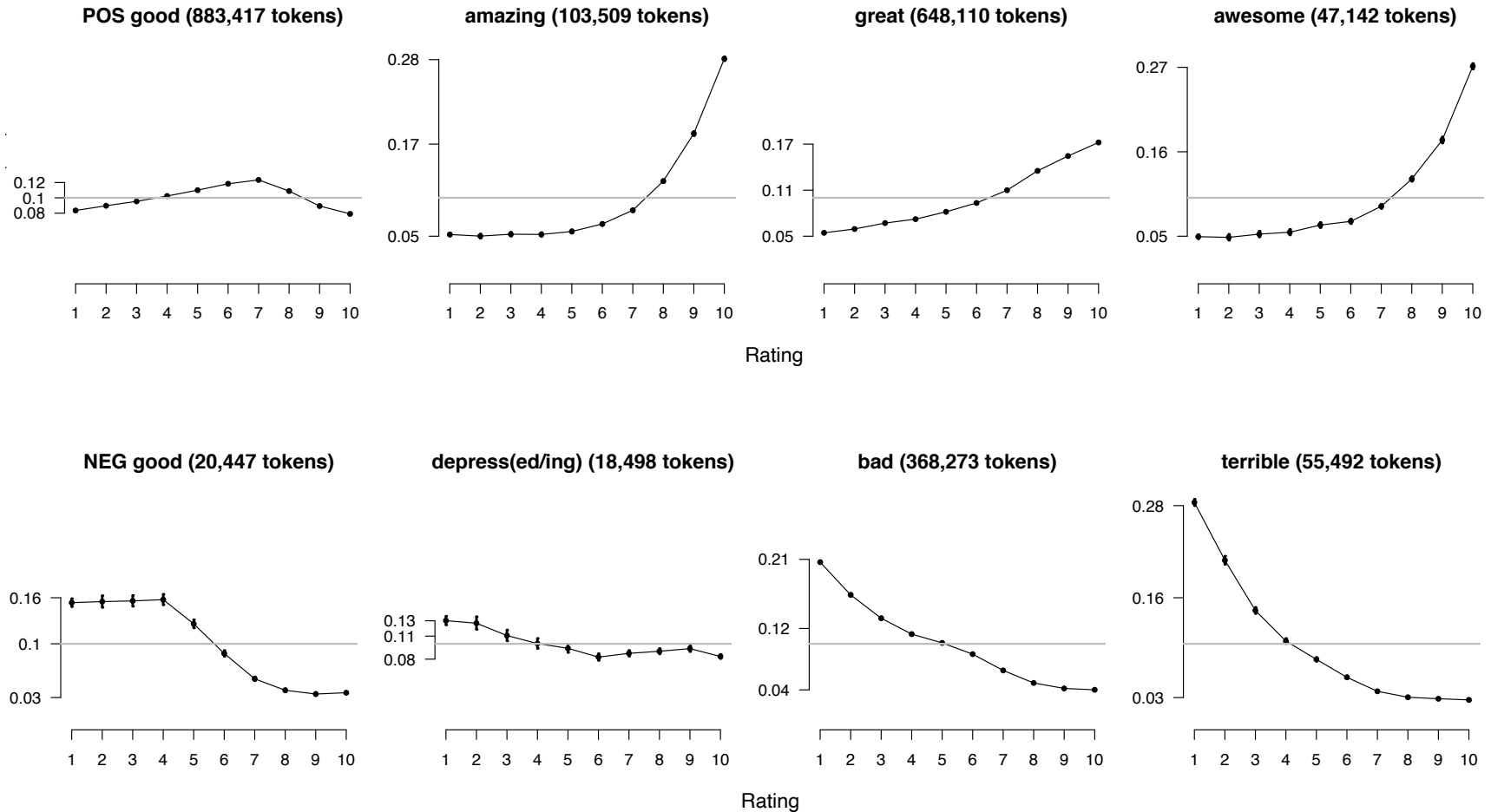
# Analiza polariteta svake riječi u IMDB kritici

- IMDB kritika ima 10 klasa, stoga svakoj riječi se dodjeljuje vektor.
- Vektor od "bad" je  
[0.21 0.14 0.13 0.11 0.10 0.09 0.07 0.06 0.05 0.04]



# Analiza polariteta svake riječi u IMDB kritici

Skalirana vjerodostojnost  $\frac{P(W|C)}{P(w)}$



Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

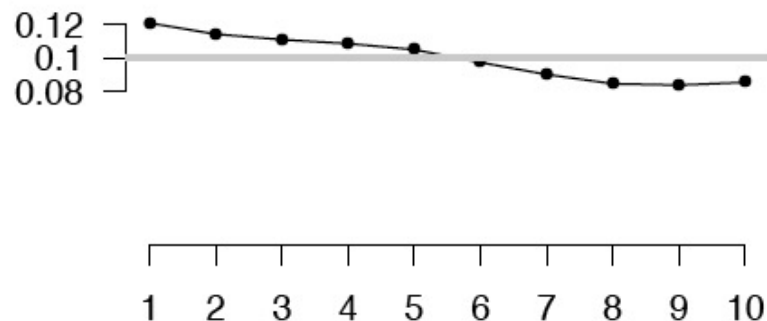
# Ostale osobine sentimenta

- Je li logička negacija (ne, nije) povezana s negativnim sentimentom?
- Pottsov eksperiment:
  - Izbroji negacije (ne, nije, nikad) u kritikama
  - Regresivno usporedi s ocjenom kritike

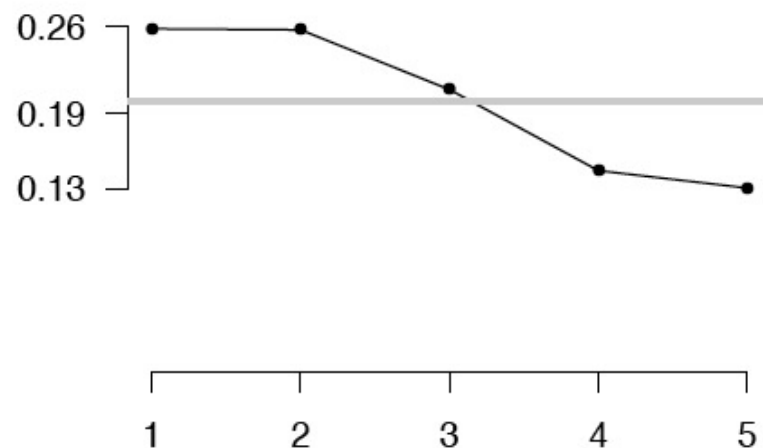
# Potts 2011 rezultati

više negacije u negativnom sentimentu

IMDB (4,073,228 tokens)



Five-star reviews (846,444 tokens)



# Uvod u obradu prirodnog jezika

## 7.4. Učenje leksikona sentimenta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

# Polunadzirano učenje leksikona

- Koristiti malu količinu informacija
  - Nekoliko označenih primjera
  - Nekoliko ručno izrađenih uzoraka
- Podizanje (bootstrap) leksikona

- Pridjevi spojeni s "i" imaju isti polaritet
  - Pošten i odan
  - Pokvaren i svirep
- Pridjevi spojeni s "ali" nemaju isti polaritet
  - Pošten ali strog

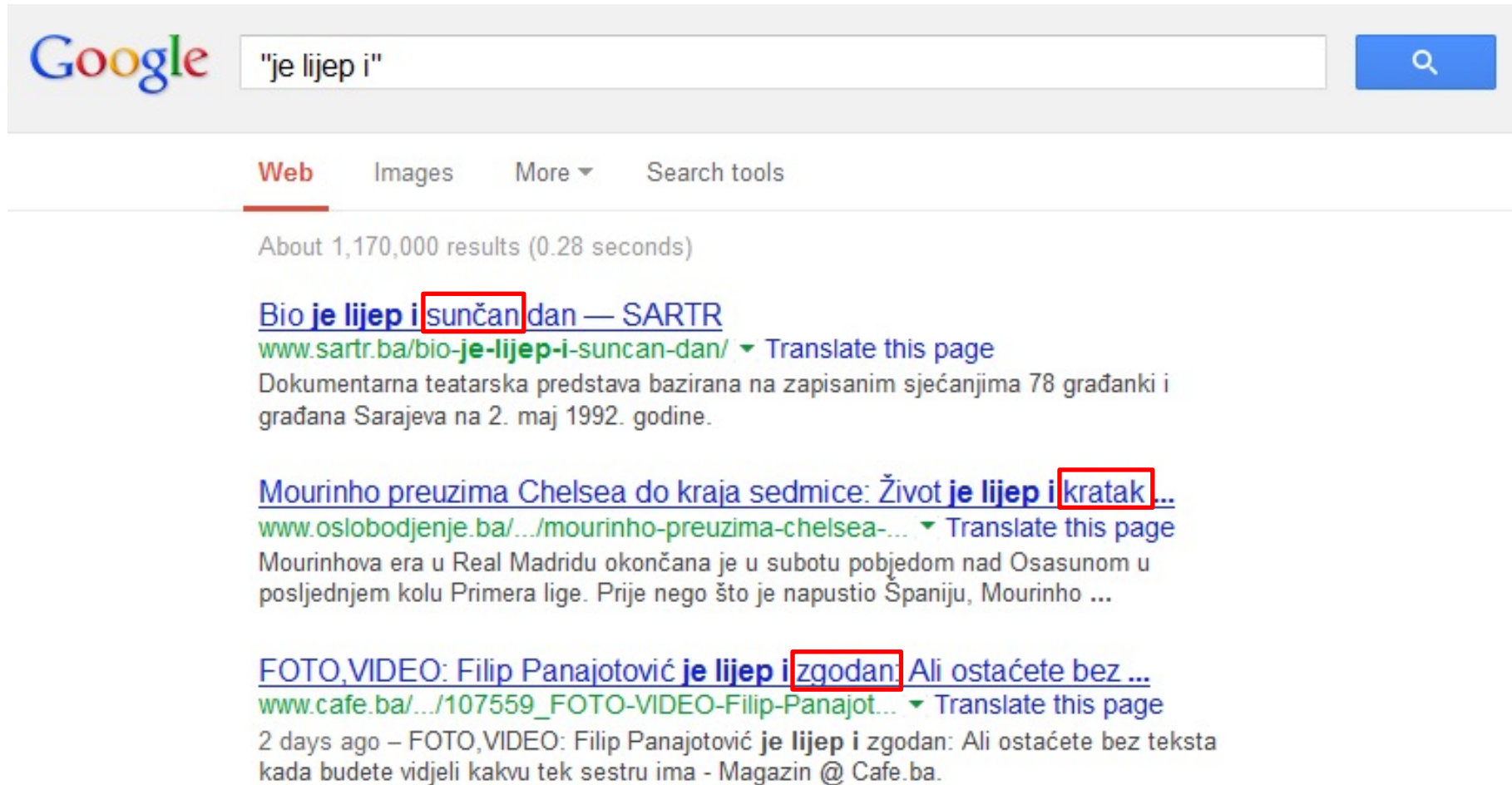


# Hatzivassiloglou i McKeown 1997: 1. Korak

- Označi početni skup od 1336 pridjeva (iz WSJ korpusa od 21 miliona riječi)
  - 657 pozitivnih
    - odgovarajuće, centralno, pametno, poznato, inteligentno, izvanredno, znano, osjetljivo, vitko, uspješno...
  - 679 negativnih
    - zarazno, pijano, neznano, koščato, bezvoljno, primitivno, uznemirujuće, neriješeno, zlobno...

# Hatzivassiloglou i McKeown 1997: 2. Korak

- Proširi početni skup na spojene pridjeve



The screenshot shows a Google search interface. The search bar contains the text "je lijep i". Below the search bar, the "Web" tab is selected. The search results show approximately 1,170,000 results found in 0.28 seconds. Three search results are visible, each with a title, a URL, and a brief description. In each title, a phrase is highlighted with a red box: "sunčan" in the first result, "kratak" in the second, and "zgodan" in the third.

Google

"je lijep i"

Web Images More Search tools

About 1,170,000 results (0.28 seconds)

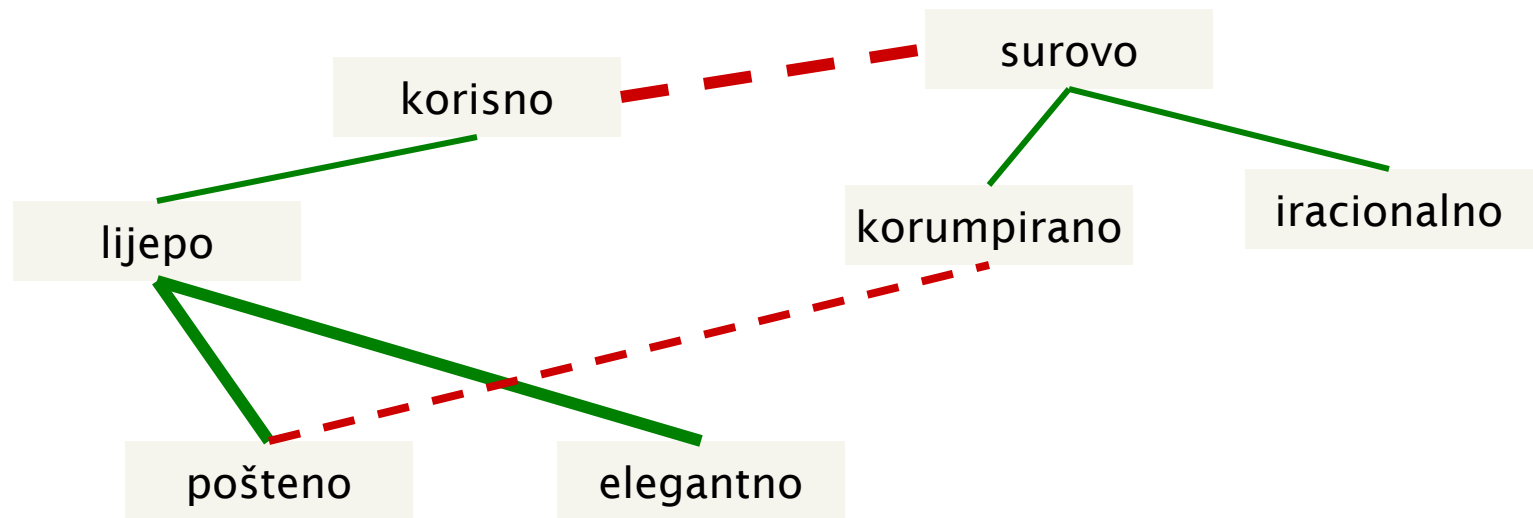
[Bio je lijep i sunčan dan — SARTR](#)  
[www.sartr.ba/bio-je-lijep-i-suncan-dan/](http://www.sartr.ba/bio-je-lijep-i-suncan-dan/) Translate this page  
Dokumentarna teatarska predstava bazirana na zapisanim sjećanjima 78 građanki i građana Sarajeva na 2. maj 1992. godine.

[Mourinho preuzima Chelsea do kraja sedmice: Život je lijep i kratak ...](#)  
[www.oslobodjenje.ba/.../mourinho-preuzima-chelsea-...](http://www.oslobodjenje.ba/.../mourinho-preuzima-chelsea-...) Translate this page  
Mourinhova era u Real Madridu okončana je u subotu pobjedom nad Osasunom u posljednjem kolu Primera lige. Prije nego što je napustio Španiju, Mourinho ...

[FOTO,VIDEO: Filip Panajotović je lijep i zgodan: Ali ostaćete bez ...](#)  
[www.cafe.ba/.../107559\\_FOTO-VIDEO-Filip-Panajot...](http://www.cafe.ba/.../107559_FOTO-VIDEO-Filip-Panajot...) Translate this page  
2 days ago – FOTO,VIDEO: Filip Panajotović je lijep i zgodan: Ali ostaćete bez teksta kada budete vidjeli kakvu tek sestru ima - Magazin @ Cafe.ba.

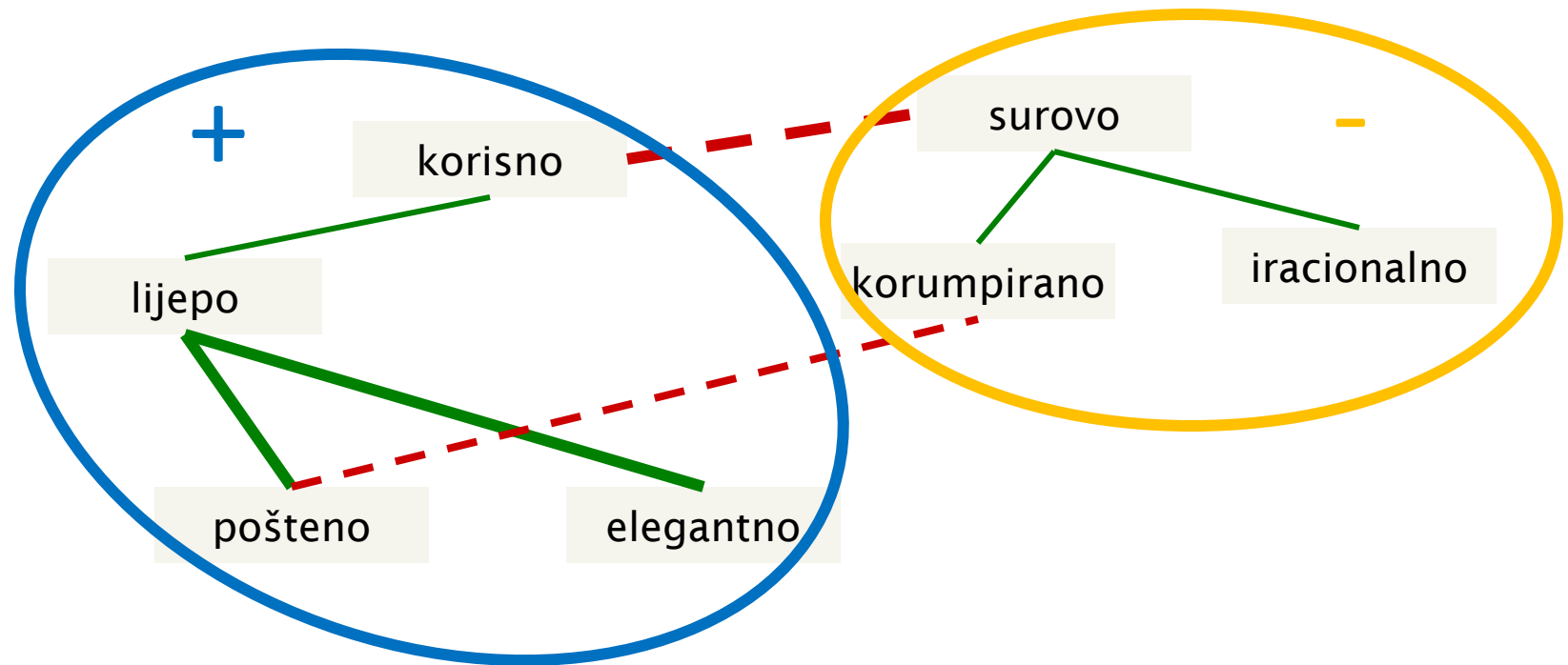
# Hatzivassiloglou i McKeown 1997: 3. Korak

- Nadzirani klasifikator pridružuje "sličnost polariteta" svakom paru riječi



# Hatzivassiloglou i McKeown 1997: 4. Korak

- Grupiranje grafa



# Izlaz iz leksikona polarnosti

- Pozitivno
  - hrabar odlučujuća uznemirujuće velikodušni dobri poštení važna velika zreła strpljivi mirni pozitivno ponosni zvuk poticanje jednostavan neobično snažno talentirana duhovita ...
- Negativno
  - dvosmislena oprezni cinični izbjegavajući štetna licemjerno neučinkovit nesigurno iracionalno neodgovorno maloljetnika gorljivi ugodno osvrće rizično sebični zamorno nepotkrijepljene ranjiva rastrošna ...

# Izlaz iz leksikona polarnosti

- Pozitivno

- hrabar odlučujuća **uznemirujuće** velikodušni dobri  
pošteni važna velika zrela strpljivi mirni pozitivno  
ponosni zvuk poticanje jednostavan neobično **stran**  
talentirana duhovita ...

- Negativno

- dvosmislena **oprezni** cinični izbjegavajući štetna  
licemjerno neučinkovit nesigurno iracionalno  
neodgovorno maloljetnika gorljivi **ugodno** osvrće rizično  
sebični zamorno nepotkrijepljene ranjiva rastrošna ...

# Turney-ov algoritam

1. Izvuci leksikon fraza iz kritika
2. Nauči polaritet za svaku frazu
3. Ocjeni kritiku temeljem srednje vrijednosti polarnosti njenih fraza

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

# Izvlačenje fraze od dvije riječi koje imaju pridjev

Prva riječ	Druga riječ	Treća riječ (nije izvučena)
JJ	NN ili NNS	bilo što
RB, RBR, RBS	JJ	nije NN ni NNS
JJ	JJ	nije NN ni NNS
NN ili NNS	JJ	nije NN ni NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	bilo što

J – pridjev

N – imenica

R – prilog

V - glagol



# Kako izmjeriti polaritet fraza

- Pozitivne fraze
  - pozitivne fraze se češće ko-javljaju s "dobro"
  - negativne fraze se češće ko-javljaju s "loše"
- ali kako izmjeriti ko-javljanje?

# Srednji uzajamni sadržaj informacije

- Uzajamni sadržaj informacije dviju slučajnih varijabli  $X$  i  $Y$
- mjeri količinu informacija koju imamo o pojavljivanju jedne riječi, ako imamo informacije o pojavljivanju druge riječi

$$I(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Srednji uzajamni sadržaj informacije  
(Pointwise mutual information)
- koliko često se događaji  $X$  i  $Y$  ko-javljaju ako su nezavisni?

$$PMI(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

# Srednji uzajamni sadržaj informacije

- Srednji uzajamni sadržaj informacije između dvije riječi
  - koliko često se riječi  $w_1$  i  $w_2$  ko-javljaju ako su nezavisni?

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

# Kako procijeniti srednji uzajamni sadržaj informacije?

- Corpus Query System (<http://filip.ffzg.hr/bonito2/>)
  - $P(w)$  procjenjuje se kao  $\frac{broj(w)}{|V|}$
  - $P(w_1, w_2)$  procjenjuje se kao broj  $\frac{broj(w_1 \text{ POKRAJ } w_2)}{|V|^2}$

$$PMI(w_1, w_2) = \log_2 \frac{broj(w_1 \text{ POKRAJ } w_2)}{broj(w_1)broj(w_2)}$$

# Da li se fraza pojavljuje više uz "izvrstan" ili "jadan"?

$$\begin{aligned} \text{Polarnost(fraza)} &= \text{PMI(fraza, dobro)} - \text{PMI(fraza, loše)} \\ &= \log_2 \left( \frac{\text{broj(fraza POKRAJ dobro)}}{\text{broj(fraza) broj(dobro)}} \right) - \log_2 \left( \frac{\text{broj(fraza POKRAJ loše)}}{\text{broj(fraza) broj(loše)}} \right) \\ &= \log_2 \left( \frac{\text{broj(fraza POKRAJ dobro)}}{\text{broj(fraza) broj(dobro)}} \right) - \log_2 \left( \frac{\text{broj(fraza POKRAJ loše)}}{\text{broj(fraza) broj(loše)}} \right) \\ &= \log_2 \left( \frac{\text{broj(fraza POKRAJ dobro)} \text{ broj(fraza) broj(loše)}}{\text{broj(fraza) broj(dobro) broj(fraza POKRAJ loše)}} \right) \\ &= \log_2 \left( \frac{\text{broj(fraza POKRAJ dobro) broj(loše)}}{\text{broj(fraza POKRAJ loše) broj(dobro)}} \right) \end{aligned}$$

# Fraze iz pozitivne kritike

Fraza	POS	Polarnost
online usluga	JJ NN	2.8
online iskustvo	JJ NN	2.3
izravna uplata	JJ NN	1.3
lokalna grana	JJ NN	0.42
...		
niske pristojbe	JJ NNS	0.33
prava usluga	JJ NN	-0.73
druga banka	JJ NN	-0.85
nezgodno nalazi	JJ NN	-1.5
<b>Prosjek</b>		<b>0.32</b>

# Fraze iz negativne kritike

Fraza	POS	Polarnost
izravna uplata	JJ NNS	5.8
online web	JJ NN	1.9
veoma praktično	RB JJ	1.4
...		
virtualni monopol	JJ NN	-2.0
manje zlo	RBR JJ	-2.3
drugi problemi	JJ NNS	-2.8
niska sredstva	JJ NNS	-6.8
neetične prakse	JJ NNS	-8.5
<b>Prosjek</b>		<b>-1.2</b>

# Rezultati Turneyovog algoritma

- 410 kritika iz [www.epinions.com](http://www.epinions.com)
    - 170 (41% negativnih)
    - 210 (59% pozitivnih)
  - Većinska klasa: 59%
  - Turney algoritam: 74%
- 
- Koristiti fraze rađe nego same riječi
  - Učiti nad područnim informacijama



# Korištenje WordNet-a za određivanje polarnosti

- WordNet: online leksikon sinonima (thesaurus)
- Ideja
  - Stvoriti pozitivni skup ("dobar") i negativni skup ("loš") riječi
  - Pronađi sinonime i antonime
  - Pozitivni skup: Dodaj sinonime pozitivne riječi ("sjajan") i antonime negativne riječi ("užasan")
  - Negativni skup: Dodaj sinonime negativne riječi ("grozan") i antonime pozitivne riječi ("odličan")
  - Ponovi, koristeći lanac sinonima
  - Filtriraj

# Zaključak

- Prednosti:
  - namijenjen za specifično područje
  - može biti robusniji (više riječi)
- Ideja
  - Započeti s inicijalnim skupom riječi ("dobar", "loš")
  - Pronaći ostale riječi koje imaju sličan polaritet:
    - koristeći "i" i "ili"
    - koristeći riječi koje se ko-javljaju u istom dokumenti
    - koristeći sinonime i antonime