Defining Significant Terms

J. Savoy
Université de Neuchâtel

C. Müller : *Principes et méthodes de statistique lexicale*. Champion. Paris.
F. Smadja: *Retrieving Collocations form Text: Xtract*. Computational Linguistics, 19(1), 1993, 143-177.
J. Savoy: *Lexical Analysis of US Political Speeches*. Journal of Quatitative Linguistics, 17(2), 123-141, 2010.

Discriminating Features

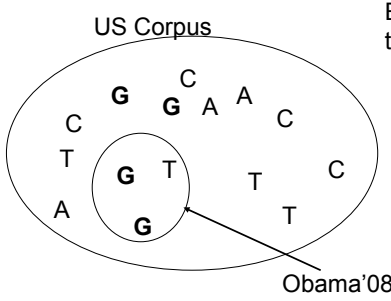
- Various methods have been proposed to define / weight the importance of each word / term in describing the semantic content of a document
- Usually related to Information Retrieval (IR)
- Here we will focus on a *comparative* basis
- How can we characterize a corpus (or a document or a set of documents) in comparison with another?
  Compare two works of two different authors
  Compare two works of the same author
  Compare a web site with another

Discriminating Features

- To define whether a given feature (e.g., word, bigram, POS, etc.) is used significantly more often in a given corpus, we may subdivide the whole corpus (C) into two (or more) disjoint parts
- Example: US electoral speeches

Our US Corpus

US: all speeches given by B. Obama & J. McCain during the years 2007 & 2008



Example with 15 tokens and 4 types

Contingency Table

- We can resume all needed information into a contingency table (one per word / feature)
- A large corpus **C** is subdivided into two (disjoint) parts **S** and **C-** (with **C** = **S** U **C-**)

	S	**C-**	
ω	*a*	*b*	*a* + *b*
not ω	*c*	*d*	*c* + *d*
	a + *c*	*b* + *d*	*n* = *a* + *b* + *c* + *d*

5

Bernoulli Process

- Example
  The word "IT" in Obama's speeches in 2008 (**S**) vs. all other US electoral Speeches (**C-**)

	Obama'08	**C-**	
"IT"	1	0	1
not "IT"	294,552	334,541	629,093
	294,553	334,541	629,094

- Prob[ω] = Prob["IT" in C] = (*a*+*b*)/*n* = 1/629,024 = 0.0000016.
- *n'* = *a* + *c* = 294,553

6

Bernoulli Process

- We can view the distribution of ω as follows.
- We draw a (biased) coin (Bernoulli process).
  For each "head" (success) we generate the word ω.
  For each "tail" (failure), another word.
- The probability of obtaining "head" is small (e.g., Prob[ω] = 0.0000016).
- We repeat this process *n'* times (e.g., *n'* = 294,553)
- We may expect finding *n'* · Prob[ω] heads (or successes or word ω in a document composed of 294,553 word tokens)
  In our example, we obtain 0.468.
  This value is the mean of the underlying Bernoulli process

7

Bernoulli Process

- Another example
- We draw a (biased) coin.
  The probability of obtaining "head" (success) is *p* = 0.4
  The probability of "tail" (failure), 1 - *p* = 0.6.
- We repeat this process *n'* times (*n'* = 10)
- We may expect finding *n'* · *p* heads.
  In our example, we have 10 · 0.4 = 4.

8

Bernoulli Process

- We can then compare the expected number of occurrence ($n' \cdot$ Prob[ω]) of the word ω with *a* (the observed number of occurrence).
- In our case, we obtain 0.468 and *a* = 1.
- The difference must be analyzed with respect to the underlying (normal) variability. This is measured by the standard deviation (denoted σ) defined as:

$$\sigma = \sqrt{n' \cdot Prob[\omega] \cdot (1 - Prob[\omega])}$$

If σ is large, we may expect a larger (but normal) difference between ($n' \cdot$ Prob[ω]) and *a*

9

The Z Score

- As a general measure to take account for the difference between:
  - an observed value (x), a random variable
  - its mean (μ)
  - its standard deviation (σ) (or its variance σ^2)

we may compute its Z score (standardized score) as

$$Z\ score = \frac{x - \mu}{\sigma} = \frac{x - \mu}{\sqrt{\sigma^2}}$$

on our case,

$$Z\ score = \frac{x - \mu}{\sigma} = \frac{a - (n' \cdot Prob[\omega])}{\sqrt{n' \cdot Prob[\omega] \cdot (1 - Prob[\omega])}}$$

10

The Z Score

- In our example (word "IT"), we have

$$Z\ score = \frac{1 - (294,553 \cdot 1/629,094)}{\sqrt{294,553 \cdot 1/629,094 \cdot (1 - 1/629,094)}} = 0.777$$

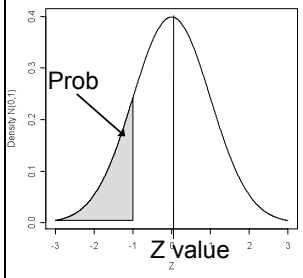
is this value significantly large?

- To have a complete answer, we need to compare it with "normal" values. Is this possible? Yes, because it is known that the Z score follows a Normal distribution N(μ=0,σ^2=1) or in short, N(0,1).

11

The Z Score

The interesting values of a N(0,1) distribution are …



Probability	Z value
0.01	-2.33
0.025	-1.96
0.05	-1.64
0.1	-1.28
0.5	0.0
0.9	1.28
0.95	1.64
0.975	1.96
0.99	2.33

12

Characteristics Terms

- Back to our example
  The word “IT” in Obama’s speeches in 2008 (**S**) vs. all other US electoral Speeches (**C-**)

	Obama’08	C-	
“IT”	1	0	1
not “IT”	294,552	334,541	629,093
	294,553	334,541	629,094

$$Z\ score = \frac{x-\mu}{\sigma} = \frac{1-(294,553 \cdot 1/629,094)}{\sqrt{294,553 \cdot 1/629,094 \cdot (1-1/629,094)}} = 0.777$$

13

Characteristics Terms

- In our example, we have Z score = 0.777
  This value is not really an exception and thus the corresponding term (“IT” or “astronaut”) occurring only once cannot be qualify as “significant” for Obama 2008.
- We can consider another word type / subset.

	McCain’08	C-	
“Bush”	26	398	424
not “Bush”	154,339	474,331	628,670
	154,365	474,729	629,094

14

Characteristics Terms

- For the word “Bush” in McCain’s speeches in 2008 we compute the Z score as

$$Z\ score = \frac{x-\mu}{\sigma} = \frac{26-(154,365 \cdot (424/629,094))}{\sqrt{154,365 \cdot (424/629,094) \cdot (1-(424/629,094))}} = -7.654$$

The resulting value is -7.654 (very small). The probability of having a Z score value lower than -2.33 is around 0.01.

Clearly the word “Bush” is underused in McCain’s speeches (in 2008) compared to the rest of the US corpus.

15

Other (Related) Questions

- Do we use all word types or remove some (not useful) types (e.g., “the”, “of”)?
- Do we use the surface (inflected) form or the lemma (e.g., “is”, “was” or “be”)?
- Do we apply a deeper morphological analysis to conflate related word types under the same stem (e.g., “American” and “America”)?
- Do we use only a subset of all possible POS tags (e.g., only nouns, adjectives, adverbs and verbs)?
- What is the difference between the frequency and the Z score?

16

Most Frequent Words

	McCain 2008		Obama 2008
Freq.	Word	Freq.	Word
2345	*I*	6203	*we*
2160	*we*	4216	*I*
1602	*our*	3276	*our*
1540	*will*	3164	*will*
821	my	2389	*you*
775	*you*	1566	*American*
775	*American*	1444	*they*
709	*they*	1313	can
640	he	1107	*America*
540	country	1081	year
530	tax	1047	need
485	*America*	958	tax

17

Most Significant Words

Z	McCain 2008	Z	Obama 2008
14.5	Obama	17.8	McCain
9.8	government	11.1	John
9.6	my	9.9	we
8.6	Canada	8.7	Bush
8.1	federal	7.7	jobs
7.9	among	7.5	Washington
7.8	small	7.4	up
7.7	judicial	7.3	relief
7.4	Arizona	7.2	working
7.4	court	7.1	why
7.3	very	7.1	street
7.1	such	7.0	family
7.0	business	7.0	because

18

Using Filter?

- We want to study the most significant bigrams (sequence of two words)
- Looking at the most frequent ones we obtain
  - of/IN the/DT
  - in/IN the/DT
  - i/PRP be/VB
  - to/TO the/DT
- Not really helpful
- Adding constraints?

19

Example of Filters

- We admit the following POS sequences
  - JJ NN white house
  - NN NN mortgage rate
- And for trigrams
  - NN NN NN stem cell research
  - JJ JJ NN next big idea
  - JJ NN NN clean energy economy
  - NN IN NN academy of science
- Difference between the frequency and the Z score (both with POS constraints)

20

Most Frequent Bigrams

	McCain 2008		Obama 2008
Freq.	**Bigram**	**Freq.**	**Bigram**
326	Senator Obama	479	*health care*
158	*health care*	384	Senator McCain
131	small business	322	*United States*
123	*United States*	300	*Wall Street*
111	*American people*	289	John McCain
48	*Wall Street*	284	*American people*
40	next street	245	*middle class*
40	new president	214	tax cut
38	tax increase	148	George Bush
35	health insurance	132	insurance company
35	government spending	131	tax break
34	*middle class*	129	new job

21

Most Significant Bigrams

Z	McCain 2008	Z	Obama 2008
28.5	Senator Obama	20.0	Senator McCain
8.4	small business	17.2	John McCain
8.1	government spending	13.9	Wall Street
6.7	tax increase	11.9	middle class
6.6	bad economy	11.4	tax cut
6.3	higher tax	11.0	Main Street
6.2	business tax	9.6	tax break
6.2	flex fuel	9.1	insurance company
6.1	law enforcement	8.5	George Bush
5.9	more job	8.4	more year
5.9	energy security	7.9	oil company
5.6	great country	7.6	rescue plan
5.6	tax rate	7.5	21st century

22

Most Frequent Trigrams

Freq.	McCain 2008	Freq.	Obama 2008
50	President I will	69	President United States
28	I elected President	67	President I will
25	you thank you	57	United States America
22	thank you thank	42	I running President
21	I believe we	40	we can afford
21	health care system	38	million new jobs
20	dependence foreign oil	35	we can choose
18	small business owner	34	we will make
17	I thank you	34	I President we
16	thank you I	33	President we will
16	I will work	33	I will make
15	I will make	32	will make sure
12	our country I	26	change we need

23

Most Significant Trigrams

Z	McCain 2008	Z	Obama 2008
5.0	hybrid flex fuel	8.2	State of America
4.6	nuclear power plant	5.6	common sense regulation
4.6	cost of energy	5.5	last eight years
4.5	strong have courage	5.3	middle class family
4.5	stronger better country	5.2	capital gain tax
4.5	selfishness in Washington	4.8	source of energy
4.5	mess of corruption	4.6	world class education
4.4	percent of American	4.6	month in Iraq
4.3	manufacture of hybrid	4.4	time for change
4.3	excess of Wall	4.2	jobs of tomorrow
4.0	worse keep tax	4.1	mountain of debt
4.0	tax increase spending	4.0	uncertainty for America
4.0	single government program	4.0	early childhood education

24

Most Frequent Terms (2007)

PS		PDC		PRD		UDC	
Freq.	Type	*Freq.*	Type	*Freq*	Type	*Freq*	Type
237	*nous*	643	*nous*	178	*être*	864	*suisse*
198	*politique*	347	*suisse*	176	*suisse*	456	*pas*
192	*doit*	261	*pas*	166	*doit*	445	*politique*
190	*pas*	245	*être*	143	*politique*	384	*ne*
178	*ne*	230	notre	138	*nous*	323	*être*
150	*être*	222	*ne*	108	sécurité	321	état
133	*suisse*	177	*politique*	108	*ne*	320	AI
132	culture	174	PDC	91	*pas*	295	droit
106	culturelle	156	*doit*	90	doivent	286	UDC
104	sociale	144	formation	88	armée	248	étranger

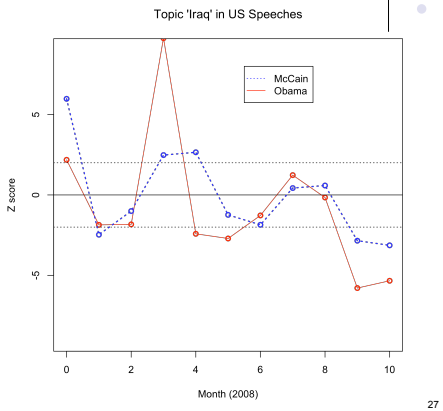
Most Significant Terms (2007)

PS		PDC		PRD		UDC	
Z	Type	*Z*	Type	*Z*	Type	*Z*	Type
15.2	état	21.8	nous	18.9	PRD	14.6	AI
14.0	II	18.9	PDC	16.0	radical	13.2	UDC
13.0	culture	11.8	demandons	12.2	mission	11.3	neutralité
11.9	culturelle	10.4	énergie	12.0	armée	10.0	gauche
11.7	artiste	10.1	internet	11.7	défense	9.6	naturalisation
10.3	encouragement	9.1	enfant	11.3	sécurité	9.0	rente
10.1	art	9.1	notre	9.6	militaire	8.8	état
10.0	autogestion	8.9	énergétique	9.6	*easy*	8.7	nationalité
10.0	CO2	8.2	PDC	9.5	imposition	8.0	milliard
9.5	pro	8.1	formation	9.2	*tax*	7.4	étranger

Dynamic Evaluation

Topic “Iraq”

Month by month in 2008

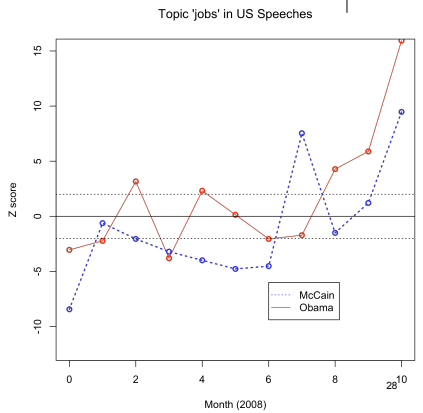


27

Dynamic Evaluation

Topic “jobs”

Month by month in 2008

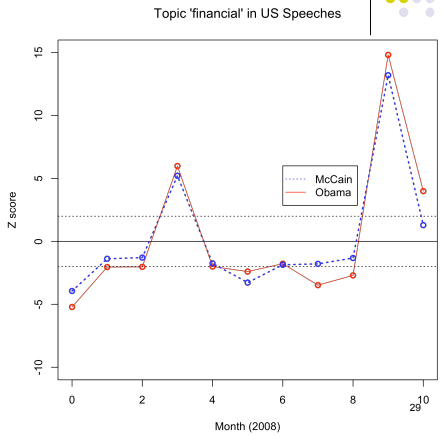


28

Dynamic Evaluation

Topic "financial"

Month by month in 2008



The Context of a Term

	Obama 2008
6	Washington we can
6	failure politician Washington
5	Washington player expect
5	status quo Washington
5	know happen Washington
5	dime Washington lobbyist
5	broken system Washington
4	Washington twenty six
4	Washington think long
4	Washington game Washington
4	they back Washington
4	politician Washington think
4	George Bush Washington

30

And for the President Obama?

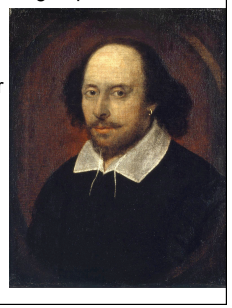
Terms overused by the President

budget	thank
Chrysler	Turkey
department	secretary
recovery plan	recovery act
new foundation	economic recovery
American recovery	new investment
reinvestment act	mutual interest
auto loan	mutual respect
higher education	
health care reform	kind of energy
clean energy economy	long term deficit

31

Authorship Attribution

- Did Shakespeare write all his plays?
  - Various authors including Bacon and Marlowe are said to have written parts or all of several plays
  - "Shakespeare" may even be a nom-de-plume for a group of writers?
- Plays written by more than one author
  - *Edward III* – Shakespeare? & Kyd?
  - *Two Noble Kinsmen* – Shakespeare & Fletcher
  - *Timon of Athens* – Shakespeare & Middleton?
  - *Henry VIII* – Shakespeare & Fletcher?

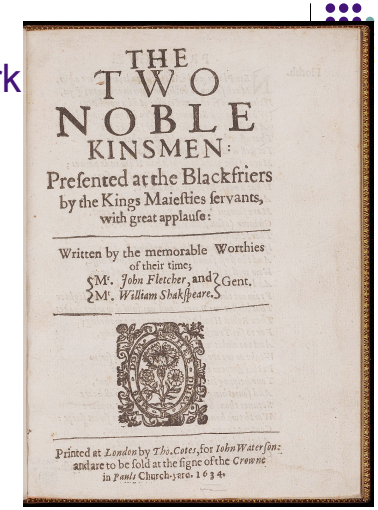


Craig, H. & Kinney A.F. (Eds): *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge Univ. Press, 2009

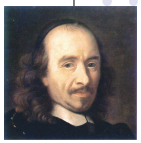
A Common Work

Two Noble Kinsmen

Shakespeare &
Fletcher


THE
TWO
NOBLE
KINSMEN:
Presented at the Blackfriers
by the Kings Maiesties servants,
with great applause:
Written by the memorable Worthies
of their time;
M^r. *John Fletcher*, and } Gent.
M^r. *William Shakspeare*.
Printed at *London* by *Tho. Cotes*, for *John Waterson*:
and are to be sold at the signe of the *Crowne*
in *Paul's* Church-yard. 1634.

Some Classical Examples



- The debate *Molière* vs. *Corneille*?
  Jean Baptiste Poquelin (1622-1673)
  Pierre Corneille (1606-1684)
- *Psyché* (1671), both are authors
- Plays (comedies) from 1658
- Corneille needs money, well-known for his dramas (but cannot write comedies, and inferior genre)
- Pierre Louys (1919) (and Voltaire) indicates that Corneille was the real author based on the rhythmus, versification.

Labbé, D. (2009). Si deux et deux font quatre, Molière n'a pas écrit Dom Juan. Paris, Max Milo.

Stylometry

- How?
  - A single measurement
  - Multivariate analysis
  - Text Categorization
    (larger set of the vocabulary)
  - Others (syntax, layout, …)

35

Single Measurement

- Letter counts
- "What disturb me in Shakespeare's plays is the over-used of the letter "o". I can live with a lot of "e" or "I", but not a lot of "o". So, yes clearly, I prefer reading Marlowe."

Letter Counts

- T. Merriam reports
  "of counting the letters in the 43 plays was the implausible discovery that the letter 'o' differentiates Marlowe and Shakespeare plays to an extent well in excess of chance" (used also letter 'a')
- Frequency less than 0.0078, 6 plays of Marlowe
  Frequency greater than 0.0078, 36 plays of Shakespeare

T. Merriam: Letter Frequency as a Discriminator of Authors. *Notes & Queries*, 239, 1994, p. 467-469.
T. Merriam: Heterogeneous Authorship in Early Shakespeare and the Problem of *Henry V*. *Literary and Linguistic Computing*, 13, 1998, p. 15-28.

37

French Corpus

Author	Title 1	Title 2
Marivaux	*La Vie de Marianne*	*Le Paysan parvenu*
Voltaire	*Zadig*	*Candide*
Rousseau	*La nouvelle Héloïse*	*Emile*
Chateaubriand	*Atala*	*Vie de Rancé*
Balzac	*Les Chouans*	*Le cousin Pons*
Sand	*Indiana*	*La Mare au Diable*
Flaubert	*Madame Bovary*	*Bouvard et Pécuchet*
Maupassant	*Une Vie*	*Pierre et Jean*
Zola	*Thérèse Raquin*	*La Bête humaine*
Verne	*De la Terre à la Lune*	*Le Secret de Wilhelm Storitz*
Proust	*Du côté de chez Swann*	*Le Temps retrouvé*

Z Score in French Literature

The Z score values for some very frequent German lemmas
between -2 and 2, normal usage
negative value → under-used, positive value → over-used

Lemma	Balzac	Chateaub.	Flaubert	Proust	Verne	Zola
le	**5.90**	**6.67**	**8.54**	-1.66	2.42	-1.98
.	-0.18	0.93	4.09	**-6.25**	1.52	**4.79**
il	**-4.83**	**-3.16**	0.58	-1.73	**-5.24**	**8.94**
être	-2.52	-0.31	**-4.96**	1.17	0.34	**-4.17**
que	**-5.24**	-1.46	**-7.30**	**6.42**	-2.59	**-3.62**
je	**-9.50**	-0.13	**-11.77**	**3.42**	-1.23	**-3.82**
de	**3.97**	**8.79**	2.50	**3.20**	1.30	**-1.21**

German Corpus

Author	Title 1	Title 2	Title 3
Goethe	*Die Wahlverwandschaften*	*Die Leiden des jungen Werther*	*Wilhelm Meisters Wanderjahre*
Heyse	*L'Arrabbiata*	*Beatrice*	*Der Weinhüter von Meran*
Fontane	*Unterm Birnbaum*		
Nietzsche	*Also Sprach Zarathustra*	*Ecce Homo*	
Hauptmann	*Bahnwärter Thiel*	*Bahnwärter Thiel*	
Falke	*Der Mann im Nebel*		
H. Mann	*Flöten und Dolche*	*Der Vater*	
T. Mann	*Der Tod in Venedig*	*Tonio Kroeger*	*Tristan*
Kafka	*Die Verwandlung*	*In der Strafkolonie*	
Wassermann	*Caspar Hauser*	*Der Mann von vierzig Jahren*	*Mein Weg als Deutsche und Jude*
Hesse	*Knulp*	*Siddhartha*	
Graf	*Zur Freundlichen Erinnerung*		

Z Score in German Literature

The Z score values for some very frequent German lemmas
between -2 and 2, normal usage
negative value → under-used, positive value → over-used

Lemma	Goethe	Kafka	Nietsche	Hesse	T. Mann
d	**-3.66**	**3.39**	-0.75	**-5.80**	**3.31**
.	**-4.20**	**-2.76**	**-4.66**	0.54	-0.44
und	**-2.79**	**-5.51**	0.57	**2.42**	**4.91**
sein	-1.13	-0.01	0.72	**4.14**	1.58
ich	**4.76**	**-4.66**	**7.51**	1.55	**-8.07**
nicht	0.67	**3.60**	0.40	1.23	**-2.60**

Z Score

- To compare two texts (one with known author, the second with disputed authorship)
- When comparing two texts, considering all Z scores from a set (*m* in this case) of terms (lemmas, word types, etc.)

$$Dist(D_j, D_k) = \frac{1}{m} \sum_i^m (Zscore(t_{ij}) - Zscore(t_{ik}))^2$$

- The smallest the distance, the highest the chance that both texts were written by the same author
- Instead of using all texts, we can concatenate all texts written by a given author to form an author profile.

42

Evaluation

English Corpus, 52 text excerpts (~10 000 tokens), 9 authors

French Corpus, 44 texts excepts (~10 000 tokens), 11 authors

German Corpus, 59 texts excepts (~10 000 tokens), 15 authors

	English	French	German
Z score	**100%**	**100%**	**84.7%**
Delta, 150 word types	96.2%	90.9%	**84.7%**
PCA, 5 axes, 100 lemmas	92.3%	70.4%	66.1%

J. Savoy: Authorship Attribution Based on Specific Vocabulary. *Journal of Quantitative Linguistics*. 19, 2012, to appear

43

Conclusion

- Various methods have been proposed to define / weight the importance of each word / term in describing the semantic content of a document
- The Z score is relatively effective to discriminate between terms used by both speakers and terms overused by one of them
- Adding POS constraints is useful (but we need a POS tagger)
- Chi-square requires at least 5 observations in each cell
- Mutual Information (MI) does not have a clear decision rule

44

Other Association Measures

- We can resume all needed information into a contingency table (one per word / feature)
- A large corpus **C** is subdivided into two (disjoint) parts **S** and **C-** (with **C** = **S** U **C-**)

	S	**C-**	
ω	*a*	*b*	*a + b*
not ω	*c*	*d*	*c + d*
	a + c	*b + d*	*n = a + b + c + d*

45

Mutual Information

- Basic Idea: Comparing two models (Church & Hanks, 1990)
- Under independence

$$Prob[S \cap \omega] = Prob[S] \cdot Prob[\omega] = \frac{a+c}{n} \cdot \frac{a+b}{n}$$

- Estimation (MLE)

$$Prob[S \cap \omega] = \frac{a}{n}$$

- How to measure the deviation between the two models?
- Mutual information (MI) for the word ω in the subset S

$$I(S;\omega) = \log_2 \left[\frac{Prob[S \cap \omega]}{Prob[S] \cdot Prob[\omega]} \right]$$

Mutual Information

I(S;ω) ≈ 0 Independence (random)
I(S;ω) > 0 Positive association
I(S;ω) < 0 Negative association

Example IM("IT";Obama'08) = 1.09
No clear decision rule

	Obama'08	**US-**	
"IT"	1	0	1
not "IT"	294 552	334 541	629 093
	294 553	334 541	629 094

Chi-square

$$\chi^2 = \sum_{i,j=0,1} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Compute the statistics followings a chi-square distribution
Example word = "Bush", S = McCain'08: χ^2 = **78.13**
Limit values: 6,63 α = 0,01 (1 dof)
10,83 α = 0,001

	McCain'08	**US-**	
"Bush"	26	398	424
not "Bush"	154 339	474 331	628 670
	154 365	474 729	629 094

48