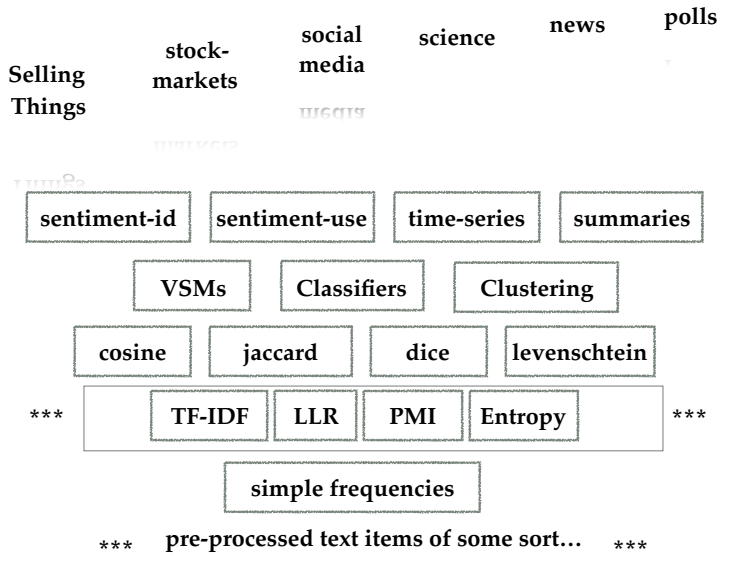
Beyond (Simple) Frequencies:

Weighting and So On…

Lecture 4: Text Analytics for Discovering Meaning
Mark Keane, Insight/CSI, UCD



Introduction

- Simple frequencies can be informative; but, are really too simple (e.g., normalisation)
- However, frequencies can be weighted differently, to make them more informative
- Or, you can use probabilistic approaches to make them convey more (is this 50 = that 50)

The Overview

- Beyond Frequency we look at:
  - Weighting Words: TF-IDF
  - Finding Discriminating Words: LLRs
- Co-occurrence & Collocation
  - (Pointwise) Mutual Information (PMI)
- Redundancy & Interestingness
  - Entropy (normalised and un-normalised)

Beyond Frequencies

Weighting Words:
TF-IDF

Imagine #1

- Sometimes *frequency* of a word in a **text-item*** tells you what that text-item is about
- In a collection of my personal emails, those from my Mammy will mention "tea" a lot
- So, Mammy-emails, may have high frequencies of the word-term "tea"

* Generic term for doc/tweet/text-snippet…

Imagine #2

- Other times *infrequency* can be informative
- If all of my e-mails except for two mention "japan" then this tells us something
- The two Japan-emails stick out 'cos of the rarity of the "japan" word-term in all emails

Intuitions

- Intuitions about the common occurrence of "tea" and the rarity of "japan" leads to most-commonly-used method for weighting word frequencies
- Namely, TF-IDF or term-frequency X inverse-document-frequency

Term Frequency

- **Term-Frequency (TF)** refers to the count of a *term* (word) in a given *document* (text-item)
- So, text-item (email) may be about Topic-X (visiting-mammy) if it has high-counts of certain terms (words like "tea")

Formally, *tf(t, d)* is…

- usually, the count of a term, *t*, in a text-item, *d:*

  $$\mathrm{tf}(t,d) = \mathrm{f}(t,d)$$

- Boolean "frequencies": $\mathrm{tf}(t,d) = 1$ if t occurs in d, else 0
- Log-scaled frequency: $\mathrm{tf}(t,d) = \log(\mathrm{f}(t,d) + 1)$
- Augmented frequency*: $\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$

where bias created by longer text-items; raw frequency divided by max raw-frequency of any term in the text-item (like normalisation)

Term Frequency: Crudities

- **Term-Frequency (TF)** is still fairly crude, even using different options
- The cleverness in TF-IDF is that the IDF bit is used to weight the frequency using rarity
- A term that is really frequent in a given document, but really *rare* in every other document is not-equal-to a term that is really frequent in every document (cf. "japan")

Document Frequency

- **Document-Frequency (DF)** refers to the count of a *term's* (word's) occurrence in a set of *documents* (text-items)
- Basic intuition: If only one of my emails mentions "japan" then it may be about Japan

Inverse Document Frequency

- **Inverse Document Frequency (IDF)** uses
  - the size of the document set, and the
  - the frequency of the *term's* (word's) occurrence in set of *documents* (text-items)

Formally, *df(t,D)* is…

- Number of times the term, *t*, is found at least once in a text-item, *d*, in the set of text-items, *D*

$$|\{d \in D : t \in d\}|$$

- or, a count of the number of text-items, *d*s, containing the term, *t*, in a set of text-items, *D*
- *D* is called the **corpus**

Formally, i*df(t, D)* is…

- The inverse of the df; the number text-items in the corpus, *N*, over the count of text-items containing the term, *t:* *N/df(t, D)*
- Typically, this is "logged" to smoothen the value; log to the base 10
- And to avoid dividing by zero, we may use:

$$1 + |\{d \in D : t \in d\}|$$

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

What IDF does to DF

Text-Item	1	2	3	4	5	6	7	8	9	10
N = 10	BC	BCD	XYBZC	BDSC	BFXC	DFGC	ABC	AXYC	BCD	CREP

	DF	IDF-no-log	IDF-log
A	2	5	0.70
B	7	1.43	0.15
C	10	1	0
…			

Significance of IDF in TF-IDF

- The cleverness in TF-IDF is that the IDF bit is used to weight the frequency by rarity
- A term that is really frequent in a given document, but really *rare* in every other document is not-equal-to a term that is really frequent in every document
- A term that is in every text-item has no weight

Formally, *tf-idf(t,d,D)* is…

- the term-frequency scaled by the inverse-document frequency

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

- It modifies the raw frequency by the overall rarity of term in the corpus of documents

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28(1),* 11-21.

What IDF does in TF-IDF…

(1) "mammy, great to see you again, and thanks for the **tea** and **scones**, last week"

(2) "thanks for sending up the parcel of **tea** and **scones**, for mary, she loves **scones**!"

(3) "**tea** 'nd **scones**, **tea** 'nd **scones**, **tea** 'nd **scones**, that's what I chant all day"

(4) "Nagonshu, I will be in **Japan** in September, for great **Japan tea**".

(5) "I can't make the meeting on tuesday, will wed do, we might have **tea**?"

	(1)	(2)	(3)	(4)	(5)
tea	1	1	3	1	1
scone	1	2	3	0	0
japan	0	0	0	2	0

What IDF does in TF-IDF…

(1) "mammy, great to see you again, and thanks for the **tea** and **scones**, last week"

(2) "thanks for sending up the parcel of **tea** and **scones**, for mary, she loves **scones**!"

(3) "**tea** 'nd **scones**, **tea** 'nd **scones**, **tea** 'nd **scones**, that's what I chant all day"

(4) "Nagonshu, I will be in **Japan** in September, for great **Japan tea**".

(5) "I can't make the meeting on tuesday, will wed do, we might have **tea**?"

	DF	IDF-log
tea	5	0
scone	3	0.22
japan	1	0.7

What IDF does in TF-IDF…

(1) "mammy, great to see you again, and thanks for the **tea** and **scones**, last week"

(2) "thanks for sending up the parcel of **tea** and **scones**, for mary, she loves **scones**!"

(3) "**tea** 'nd **scones**, **tea** 'nd **scones**, **tea** 'nd **scones**, that's what I chant all day"

(4) "Nagonshu, I will be in **Japan** in September, for great **Japan tea**".

(5) "I can't make the meeting on tuesday, will wed do, we might have **tea**?"

	(1)	(2)	(3)	(4)	(5)
tea	0.00	0.00	0.00	0.00	0.00
scone	0.22	0.24	0.67	0.00	0.00
japan	0.00	0.00	0.00	1.40	0.00

What IDF does in TF-IDF

	(1)	(2)	(3)	(4)	(5)
tea	1	1	3	1	1
scone	1	2	3	0	0
japan	0	0	0	2	0

tea looks important in TF, but is zeroed by IDF in TF-IDF; whereas **japan** is boosted in TF-IDF

	DF	IDF
tea	5	0
scone	3	0.22
japan	1	0.7

	(1)	(2)	(3)	(4)	(5)
tea	0.00	0.00	0.00	0.00	0.00
scone	0.22	0.44	0.67	0.00	0.00
japan	0.00	0.00	0.00	1.40	0.00

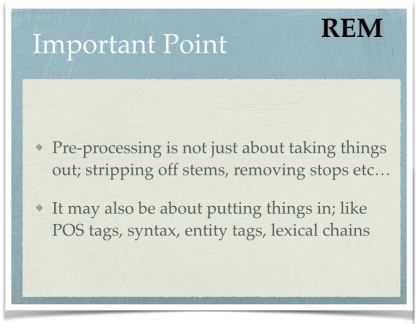
TF-IDF: Issues

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

- TF-IDF's penalising of repeated terms in text-items of a collection is a potential drawback
- Does not distinguish repeated-but-important (**tea**) from repeated-but-not-important terms (**the, and**); hence, stop-word removal can become important
- If full document-set is not known; may need approximations
- NB issue of defining the text-item; all docs in a week, each line of every document; the actual document article…

http://www.r-bloggers.com/the-tf-idf-statistic-for-keyword-extraction/

Remember Lect2, Lect3…

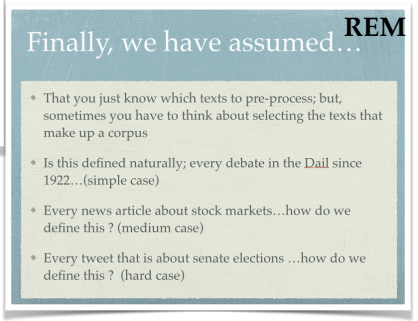


Pre-processing Non-Trivial

Normalisation Non-Trivial

Corpus Selection Non-Trivial

Item Selection Non-Trivial



TF-IDF: Solo?

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

- TF-IDF is rarely used on its own, though you will see it used occasionally
- Fundamental part of Vector Space Models (VSMs), comparing whole sets of text-items on their similarity to one another; basically by comparing the word-vectors for the whole text-item (i.e., word tf-idf scores)

Vector Space Model…

- After pre-processing you have a word/term by doc matrix or doc by word/term matrix

$$\begin{array}{c|cccc} & T_1 & T_2 & \ldots & T_t \\ D_1 & w_{11} & w_{21} & \ldots & w_{t1} \\ D_2 & w_{12} & w_{22} & \ldots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \ldots & w_{tn} \end{array}$$

	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Eg of TF-IDF Solo

- Kaptein et al. (2010) looked at using TF versus TF-IDF in tag clouds along with other modifications

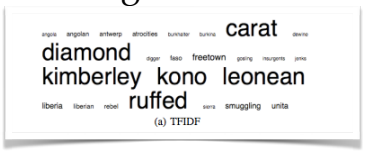


Fig. 1. Word cloud from 10 results for the topic "diamond smuggling"

Kaptein, R., Hiemstra, D., & Kamps, J. (2010). How different are language models and word clouds? In Advances in information retrieval (pp. 556-568). Springer Berlin Heidelberg.

Adieu…adieu…mon copain

- In lectures on similarity, classification and clustering, we will see extensive use of *tf-idf* and VSMs
- So, while we will say goodbye for now…you will always remain in my heart…

Beyond Frequency

Finding Discriminating Words: Log Likelihood Ratios

What Sticks Out…

- Sometimes you are looking for exceptional words…words that stick out…that are rare but indicate something about a text-item
- We could use simple frequencies for this; look for low-frequency words, or words that are lower in frequency than some average, word-frequency
- We could look for frequency peaks in one text-item versus other text-items in a corpus?

When Frequencies are no good…

- But, one of the problems with looking at raw frequencies is determining whether a given count really "means" anything
- The high/low frequency count could, essentially, be the same but the likelihood of the occurrence of that frequency could be different (e.g., because they are different sizes of corpus/text-items)

The Intuition #1

- We can solve this by computing log-likelihood ratios (LLRs) for the terms and checking whether there are significant differences
- If term frequency in Play-X differs from its frequency in all Shakespeare's plays (in some normalised sense); then there is something different about Play-X

LLR used for many tasks…

- Could concern the frequency of Word-X in Corpus-A versus its frequency in Corpus-B
- Frequency of Word-X in Item-A, versus frequency of Word-X in Corpus-B (set of Items; set of docs, as in Authorship Tests)
- Frequency of Word-X in Time-period-Y, versus its frequency in Time-period-Z (a New Event)

Formulae

$$\text{LLR} = -2\log\lambda$$

LLR is minus 2 times log of lambda

$$\lambda = \frac{L(H_0)}{L(H_1)}$$

Lambda is the likelihood of null hypothesis (H_0) over alternative (H_1)

$$H_0 : \ \theta = \theta_0,$$
$$H_1 : \ \theta = \theta_1.$$

$$L(n,k,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

L is computed relative to the binomial distribution

Likelihood-ratio test

In statistics, a **likelihood ratio test** is a statistical test used to compare the fit of two models, one of which (the *null model*) is a special case of the other (the *alternative model*). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model in favour of the alternative model. When the logarithm of the likelihood ratio is used, the statistic is known as a **log-likelihood ratio statistic**, and the probability distribution of this test statistic, assuming that the null model is true, can be approximated using **Wilks's theorem**.

Definition [edit]

A **likelihood ratio test** can be used to make a decision about two competing hypotheses or models: a null hypothesis H_0 and an alternative hypothesis H_1. The likelihood function is defined as the probability of observing x given the hypothesis. The likelihood function is defined as $f(x|H_0)$ for the null hypothesis and $f(x|H_1)$ for the alternative. The likelihood of the null hypothesis over the alternate is

$$\Lambda(x) = \frac{f(x|H_0)}{f(x|H_1)} = \frac{L(H_0|x)}{L(H_1|x)}.$$

To decide whether to reject the null hypothesis, the likelihood is compared to a threshold c:

$$\text{Do not reject } H_0 \text{ if } \Lambda(x) > c$$
$$\text{Reject } H_0 \text{ if } \Lambda(x) \leq c$$

Usually the likelihood is determined by a set of parameters θ that are different under each hypotheses. For simple hypotheses, the parameters take fixed values and do not need to be estimated; in composite hypotheses, the parameters may take a range of values.

Likelihood-ratio test

From Wikipedia, the free encyclopedia
(Redirected from Log-likelihood ratio)

Wilkes Theorem is that all the foregoing can be computed using the χ^2_c or Chi² statistic…

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

A **chi square** (X^2) **statistic** is used to investigate whether distributions of categorical variables differ from one another. Basically categorical variable yield data in the categories and numerical variables yield data in numerical form.

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

LLR Example $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

- 20 People of different backgrounds asked yes/no? So, 6 students out of 20 say NO.

	Business owner	School student	Adult male resident	Adult female resident	Senior citizen	Total
o	4	6	14	10	16	50
e	10	10	10	10	10	50
o - e	-6	-4	4	0	6	--------
$(o-e)^2$	36	16	16	0	36	--------
$\frac{(o-e)^2}{e}$	3.6	1.6	1.6	0	3.6	--------
x^2	3.6	1.6	1.6	0	3.6	10.4

In this example, the expected data (e) is simply taken as being the mean negative frequency of response. It is calculated by adding up all of the observed data (o) and then dividing by the number of categories, i.e. 5. This gives an expected frequency of 10 for each category.

LLR Example $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

- Do people fall into expected categories…

Clip slide

Interpreting your Chi-Squared Value

- Negative responses = 10.4
- Calculate degrees of freedom:
  - **df = n-1** (where **n** is the no. of categories)
  - In this case **df = 5 - 1 = 4**
- Use a *critical values* table to work out the *significance* of your result.
- Significance tells us *how confidently* we can disprove the null hypothesis

TABLE IV

Chi-Square (χ^2) Distribution

Area to the Right of Critical Value

Degrees of Freedom	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

LLR Example $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

- Do two corpora differ in the words used…

Table 1 Contigency table for word frequencies

	CORPUS ONE	CORPUS TWO	TOTAL
Freq of word	a	b	a+b
Freq of other words	c-a	d-b	c+d-a-b
TOTAL	c	d	c+d

Rayson, P., & Garside, R. (2000, October). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora (pp. 1-6). Association for Computational Linguistics.

LLR Example $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

Table 1 Contigency table for word frequencies

	CORPUS ONE	CORPUS TWO	TOTAL
Freq of word	a 70 52.5	b 140 157.5	a+b 210
Freq of other words	c-a 30 47.5	d-b 160 142.5	c+d-a-b 190
TOTAL	c 100	d 300	c+d 400

$f_e = \frac{(N_r)(N_c)}{N}$, where N_r is the total number of cases in the respective row, N_c is the total number in the respective column, and N is the number in the full sample.[1]

LLR = Chi^2

χ_c^2 is written as Chi^2

$Chi^{2\text{-Observed}} = 16.37$

$Chi^{2\text{-Critical}}(1) = 3.84, p < 0.05$

$p = 0.0005$

Rayson, P., & Garside, R. (2000, October). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1-6). Association for Computational Linguistics.

LLR used for many tasks… REM

- Could concern the frequency of Word-X in Corpus-A versus its frequency in Corpus-B
- Frequency of Word-X in Item-A, versus frequency of Word-X in Corpus-B (set of Items; set of docs, as in Authorship Tests)
- Frequency of Word-X in Time-period-Y, versus its frequency in Time-period-Z (a New Event)

Corpus-A to Corpus-B

- Demographically-sampled spoken English part of the British National Corpus (4.5M words)
- Gender; using Corpus-A (words used by males) and Corpus-B (words used by females)
- Social Class; using Corpus-A (A/B/C1) and Corpus-B (C1/D/E)

Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of English vocabulary. *International Journal of Corpus Linguistics, 2(1)*, 133-152.

TABLE 2: Words most characteristic of male speech[10]

WORD	MALES	M %	FEMALES	F %	χ^2
fucking	1401	0.08	325	0.01	1233.1
er	9589	0.56	9307	0.36	945.4
the	44617	2.60	57128	2.20	698.0
yeah	22050	1.29	28485	1.10	310.3
aye	1214	0.07	876	0.03	291.8
right	6163	0.36	6945	0.27	276.0
hundred	1488	0.09	1234	0.05	251.1
fuck	335	0.02	107	0.00	239.0
is	13608	0.79	17283	0.67	233.3
of	13907	0.81	17907	0.69	203.6

TABLE 3: Words most characteristic of female speech

WORD	MALES	M %	FEMALES	F %	χ^2
she	7134	0.42	22623	0.87	3109.7
her	2333	0.14	7275	0.28	965.4
said	4965	0.29	12280	0.47	872.0
n't	24653	1.44	44087	1.70	443.9
I	55516	3.24	92945	3.58	357.9
and	29677	1.73	50342	1.94	245.3
to	23467	1.37	39861	1.54	198.6
cos	3369	0.20	6829	0.26	194.6
oh	13378	0.78	23310	0.90	170.2
Christmas	288	0.02	1001	0.04	163.9

Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of English vocabulary. *International Journal of Corpus Linguistics, 2(1)*, 133-152.

TABLE 11 Words used more by social classes A/B/C1

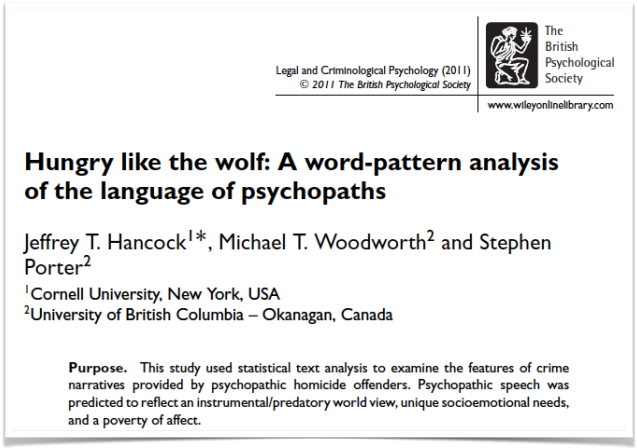
WORD	A/B/C1	%	C2/D/E	%	χ^2
yes	6485	0.49	5089	0.28	876.9
really	2897	0.22	2833	0.16	155.1
okay	1073	0.08	858	0.05	136.5
are	5150	0.39	5622	0.31	127.8
actually	1159	0.09	983	0.05	119.7
just	5924	0.44	6707	0.37	103.5
good	2622	0.20	2748	0.15	90.0
you	38616	2.89	49263	2.72	82.6
erm	4874	0.36	5551	0.31	79.9
right	4468	0.33	5158	0.28	62.7

TABLE 12 Words used more by social classes C2/D/E

WORD	A/B/C1	%	C2/D/E	%	χ^2
he	11308	0.85	19707	1.09	452.1
says	731	0.05	2332	0.13	432.0
said	4168	0.31	8178	0.45	379.7
fucking	235	0.02	1006	0.06	280.3
ain't	312	0.02	1031	0.06	202.6
yeah	14017	1.05	22132	1.22	197.3
its	122	0.01	571	0.03	174.8
them	3748	0.28	6550	0.36	153.4
aye	373	0.03	1031	0.06	144.6
she	8284	0.62	13249	0.73	137.9

Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of English vocabulary. *International Journal of Corpus Linguistics, 2(1)*, 133-152.

Corpus-A (Psychopaths) V Corpus-B (Controls)



Legal and Criminological Psychology (2011)
© 2011 The British Psychological Society
The British Psychological Society
www.wileyonlinelibrary.com

Hungry like the wolf: A word-pattern analysis of the language of psychopaths

Jeffrey T. Hancock[1]*, Michael T. Woodworth[2] and Stephen Porter[2]
[1]Cornell University, New York, USA
[2]University of British Columbia – Okanagan, Canada

Purpose. This study used statistical text analysis to examine the features of crime narratives provided by psychopathic homicide offenders. Psychopathic speech was predicted to reflect an instrumental/predatory world view, unique socioemotional needs, and a poverty of affect.

Corpus-A to Norm-Corpus

- 103 pages of conversations with Air Traffic Controllers and ethnographer reports; systems analysis of main concepts in domain
- Used LLR to compare words in ATC-corpus and Normative-Corpus (BNC subset of 2.3M)
- Specialist vocabulary of ATCs have high LLR scores (plane, flight, airport, rack, strip)

Sawyer, P., Rayson, P., & Garside, R. (2002). REVERE: support for requirements synthesis from documents. Information Systems Frontiers, 4(3), 343-353.

Presence / Absence in Same Corpus

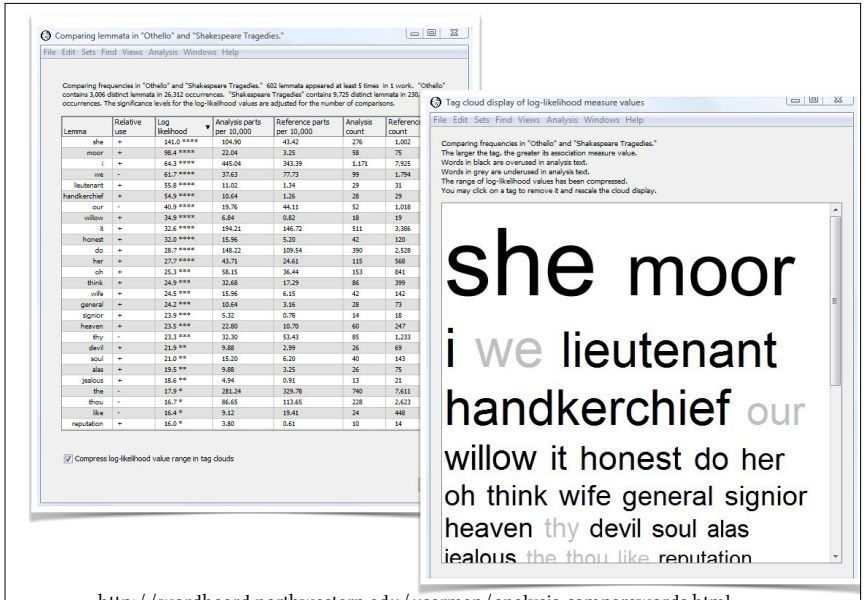
- Automated Machine Translation (of manuals) often uses word / phrase / sentence alignment; "cat"->"chat"
- Phrase translations are normally first extracted from word-aligned bilingual text segments; generative models use HMMs, but hard to add new features
- Compare when Both-Words occurs in Aligned-Set-of-Sentences (Corpus A) versus Neither-Word occurs in Aligned-Set-of-Sentences (Corpus-A)

Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In *Proceedings Conf on Human Language Technology and Empirical Methods in NLP* (pp. 81-88). ACL.

Doc to Corpus

- Comparing count in a Doc versus count in a Doc-Corpus: what words stick out in Play-X versus other Shakesheare plays
- Will show you the words that are "overused" in that play versus the other plays
- Compare word in Othello (Doc) versus word in all other Shakespeare Tragedies (Corpus-A)

http://wordhoard.northwestern.edu/userman/analysis-comparewords.html



http://wordhoard.northwestern.edu/userman/analysis-comparewords.html

Keyword Extraction…

- How do I select a set of features from some document set to index them…or to use as a set of features for a classifier or ML
- Where you want to reduce the set of features used…these methods will signal features that stick out?

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *ICML* (Vol. 97, pp. 412-420).

LLR:Issues

- Sidak correction for multiple comparisons; should consider reducing number of comparisons in statistical testing
- Selection of comparison corpora is critical

The Sidak correction computes an adjusted significance level "alpha" as follows.

$$\text{adjusted alpha} = 1 - (1 - \text{alpha})^{1/k}$$

where "alpha" is the nominal significance level and "k" the number of comparisons.

For example, let's say there are k=1,000 comparisons and alpha=0.01 . The nominal breakpoint for achieving a 0.01 level of significance is 6.63 . The Sidak adjusted level is $1 - (1 - 0.05)^{1/1000}$ or 0.000051 . This is roughly the same as dividing the nominal significance level by the number of comparisons. The corresponding adjusted breakpoint for G^2 is 16.4 .

Beyond Frequencies

Co-Occurrences & Collocation:

Pointwise Mutual Information

Counting Things Together

- We have looked at just counting individual items (words, word-pairs, emoticons), but often want to find out if certain words co-occur together
- "the" and "of" are very frequent and you might find them together a lot "of the"; but that does not mean they are a meaningful pair (unlike "lame duck")
- Intuition: "of" and "the" are found with many other words but "lame" "duck" usually only occur together

Collocation Used for…

- Habitual ways of putting words together… the "co" "location" of words can reveal a lot:
  - Named entities: New York, Big Fella
  - Idioms: lame duck, white elephant
  - Words & POS tags: fish-noun, fish-verb
  - Words in parallel texts…for translation

Intuition

- Association measure of two items: considering how likely are they to be found together or apart in some larger set of items (words, etc)
- Tells us how informative the occurrence of one word is about the occurrence of another
- Words that are highly informative about one another form a collocation

Keller, F. (2006). *Formal Modelling in Cognitive Science (Lect 27)*. University of Edinburgh

Idea: MI and PMI

Mutual Information (MI) and Pointwise Mutual Information (PMI): use observed frequency of the pair and an expected frequency of the pair under the assumption that the parts are independent

Church, K. & Hanks., P. (1989). Word association norms, mutual information, & lexicography. *ACL*, 76- 83.

Formula: PMI

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

log base 2 of probability of w1 and w2 over the probability of w1-alone by w2-alone

$$\frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \frac{\frac{C(w_1, w_2)}{N}}{\frac{C(w_1)}{N}\frac{C(w_2)}{N}} =$$

$$\frac{C(w_1, w_2)}{N} \times \frac{N^2}{C(w_1)C(w_2)} = \frac{C(w_1, w_2)N}{C(w_1)C(w_2)}$$

C() is a counting function; N is sample size, no of items in corpus or list of pairs examined

$$PMI(w_1, w_2) = \log_2(C(w_1, w_2)) + \log_2(N) - \log_2(C(w_1)) - \log_2(C(w_2))$$

For ease of computation…

Formula

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

log base 2 of probability of w1 and w2 over the probability of w1-alone by w2-alone

$$\frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \frac{\frac{C(w_1, w_2)}{N}}{\frac{C(w_1)}{N}\frac{C(w_2)}{N}} =$$

$$\frac{C(w_1, w_2)}{N} \times \frac{N^2}{C(w_1)C(w_2)} = \log_2 \frac{C(w_1, w_2)N}{C(w_1)C(w_2)}$$

C() is a counting function; *N* is sample size, no of items in corpus or list of pairs examined

$$PMI(w_1, w_2) = \log_2(C(w_1, w_2)) + \log_2(N) - \log_2(C(w_1)) - \log_2(C(w_2))$$

For ease of computation…

PMI Example

$$PMI(w_1, w_2) = \log_2(C(w_1, w_2)) + \log_2(N) - \log_2(C(w_1)) - \log_2(C(w_2))$$

$PMI(w_1; w_2)$	$c(w_1)$	$c(w_2)$	$c(w_1, w_2)$	w_1	w_2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13487	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

PMI Example

$$PMI(w_1, w_2) = \log_2(C(w_1, w_2)) + \log_2(N) - \log_2(C(w_1)) - \log_2(C(w_2))$$

Example

Take an example from the table:

$$PMI(x; y) = \log \frac{f(x, y)}{f(x)f(y)} = \log \frac{\frac{c(x,y)}{N}}{\frac{c(x)}{N}\frac{c(y)}{N}}$$

$$PMI(\text{unsalted}; \text{butter}) = \log \frac{\frac{20}{14307668}}{\frac{24}{14307668}\frac{320}{14307668}} = 15.19$$

This means: the amount of information we have about *unsalted* at position i increases by 15.19 bits if we are told that *butter* is at position $i + 1$ (i.e., uncertainty is reduced by 15.19 bits).

PMI :Issues

Random selection from 734 V+N pairs with highest PMI in BNC

V	N	C(VN)	C(V)	C(N)	PMI
Asalam	alekum	1	1	1	6.4719
Astynax	mexicana	1	1	1	6.4719
cholyglycine	hydrolase	1	1	1	6.4719
choose{	gth	1	1	1	6.4719
christopher	Columbus	1	1	1	6.4719
ek	badmash	1	1	1	6.4719
elk	n'a	1	1	1	6.4719
perswade	yong	1	1	1	6.4719
royall	maiesty	1	1	1	6.4719
sont	superbe	1	1	1	6.4719

- Seriously over-estimates low frequency events because of how it treats counts
- Solution: to have a minimal frequency cut-off; only consider words with frequency >100

Beroni, M. (2014). *Text Processing*. University of Trento, Italy.

Beyond Frequencies

Finding Redundancy & Interest:

Information Gain & Entropy

Intuition

- Beyond word-frequencies we can examine the distribution of those frequencies to see if a text-item (or corpus of text-items) is redundant or interesting
- If a text-item has many words with the same frequencies (flat distribution) then it is likely to be about a single topic and quite repetitive
- If a text-item has a few words with high peaks and others with very low dips (peaky distribution) then it is likely to be much more diverse and, possibly, interesting

Prior Usage…

- Concept used in thermodynamics: to measure of disorder in a changing system, where energy is transformed from an ordered state to a disordered state; higher entropy = higher disorder
- Shannon Entropy, Information theory; average amount of information in a message received, where *message* is an event/character/sample drawn from a distribution or data-stream (so, the less likely an event, the more info it provides)
- In text analytics, used to assess degree of order/repetitiveness/redundancy (low entropy) or disorder/diversity/interestingness (high entropy) in a text-item

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27 (3):* 379–423.

Idea: Entropy

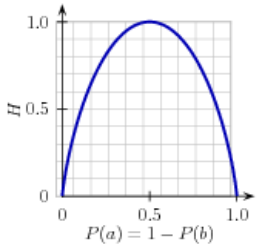
Entropy is defined as the sum of the probability of each label times the log probability of that same label

Shannon, Claude E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27 (3):* 379–423.

Formula: Entropy

$$H(A) = -\sum_{a \in A} p_a \log_2 p_a$$

A is some distribution* (word freq in a doc) and *a* is item (word), sum of the probability of *a* by log-base-2 of probability of *a*



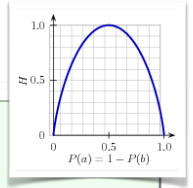
Entropy values as a function of percentage of a word-item (“male”) in a set of words where there are just two options (“male”, “female”)

http://www.nltk.org/book/ch06.html#fig-entropy

*from a discrete random variable

Entropy Eg

$$H(A) = -\sum_{a \in A} p_a \log_2 p_a$$




```
import math
def entropy(labels):
    freqdist = nltk.FreqDist(labels)
    probs = [freqdist.freq(l) for l in freqdist]
    return -sum(p * math.log(p,2) for p in probs)
```

```
>>> print(entropy(['male', 'male', 'male', 'male']))
0.0
>>> print(entropy(['male', 'female', 'male', 'male']))
0.811...
>>> print(entropy(['female', 'male', 'female', 'male']))
1.0
>>> print(entropy(['female', 'female', 'male', 'female']))
0.811...
>>> print(entropy(['female', 'female', 'female', 'female']))
0.0
```

Example 4.3 (code_entropy.py): **Figure 4.3**: Calculating the Entropy of a List of Labels

http://www.nltk.org/book/ch06.html#fig-entropy

Issues

Sometimes you should use normalised entropy, which divides it by the length of the signal (eg doc length); if e.g. you are comparing entropy scores for same items in different documents

Beroni, M. (2014). Text Processing. University of Trento, Italy.

Tweet Finding

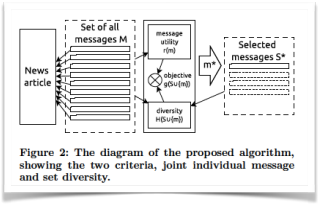


Figure 2: The diagram of the proposed algorithm, showing the two criteria, joint individual message and set diversity.

- Problem: to select an maximally interesting (diverse) set of tweets commenting on a news article
- Find all tweet with links to news-articles on event-x; take this set and identify a whole bunch of features about them
- Find the set, S*, of tweets of maximal interest on these features (i.e., set with the highest entropy)

Štajner, T., Thomee, B., Popescu, A. M., Pennacchiotti, M., & Jaimes, A. (2013). Automatic selection of social media responses to news. 19th ACM SIGKDD Conf. on Knowledge discovery and data mining (pp. 50-58). ACM.

"Filter Bubbles" on Web

- With query personalisation, we may only get to web-pages that confirm our views on shocking news (e.g. mass killings)
- Analysed web-browser logs before & after SandyHook, looking for change in type of search (61k users on 297k sites)
- Got raters to classify webpages within a domain and then checked diversity of labels using normalised entropy (as diff domains have diff web-page-nos)

Koutra, D., Bennett, P., & Horvitz, E. (2014). Events and Controversies: Influences of a Shocking News Event on Information Seeking (No. arXiv: 1405.1486).

"Filter Bubbles" on Web

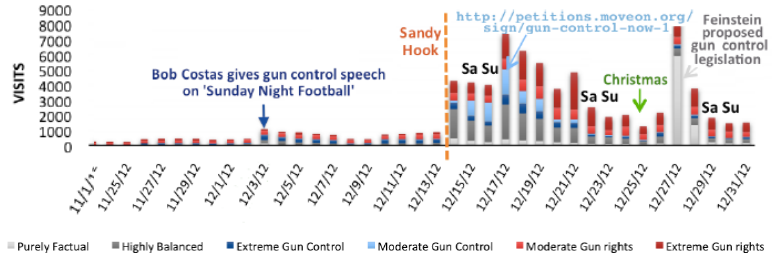
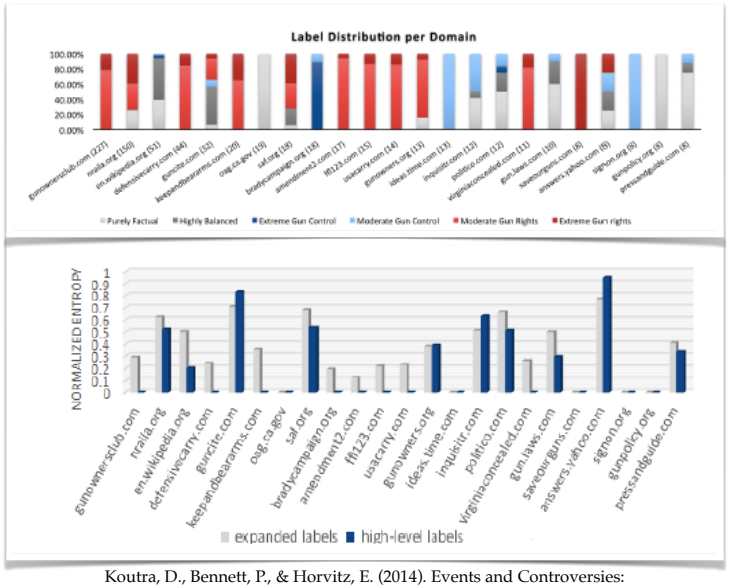


Figure 1: Number of visits to gun control/rights related webpages over time (November-December 2012). The colors correspond to webpages categories: gray for factual and balanced pages; blue for pages supporting gun control; and red for pages supporting gun rights. The categories and the labeling process are described in the Appendix and Sec. 3.2 respectively.

Koutra, D., Bennett, P., & Horvitz, E. (2014). Events and Controversies: Influences of a Shocking News Event on Information Seeking (No. arXiv: 1405.1486).



Koutra, D., Bennett, P., & Horvitz, E. (2014). Events and Controversies: Influences of a Shocking News Event on Information Seeking (No. arXiv: 1405.1486).

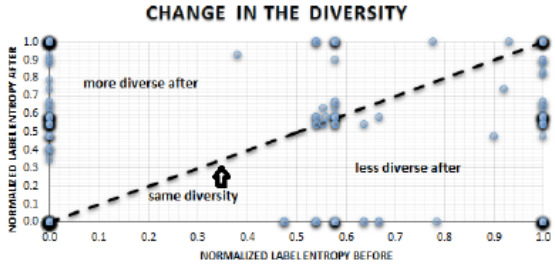


Figure 5: Change in the user diversity after S.H. for users who visited at least two different domains both before and after the event. Every point in the plot corresponds to a user.

With Feinstein's website. Figure 5 shows the change in the user diversity after the Sandy Hook shootings. On average, the normalized entropy increased by 8.14% after Sandy Hook. However, the results vary across users: the entropy remained the same for 36% of the users (and a majority of those, about 70%, are highly polarized before and after the event, visiting domains in one category only); for 36.2% of the users the diversity increased by 66.53%; and for 27.8% decreased by 57.3%.

Without Feinstein's website. By repeating the analysis described above after removing the website which was heavily visited when the gun ban list was announced, we find that, on average, the normalized entropy of the users decreased by only 0.59%. For 43.75% of the users, the diversity remained the same; for 27.6% of the users, the entropy increased by 62%, and for 28.6% it decreased by the same percent.

The change in user diversity, both with and without the outlier webpage, suggests that people peek outside of their bubble when events have potential for individual impact on the user – such as the proposed legislation on banning some types of guns –, but remain in an "echo chamber" otherwise.

Tweet Finding

Table 1: The features used by our methods. Superscripts s and d respectively indicate features used in r and H_0, respectively.

N-grams[d]	All unigrams, bigrams, and trigrams taken over the words in the tweet.
Tf-idf score[s]	Average tf-idf score of all words in the tweet, emphasizing rarely-used and penalizing out-of-vocabulary words.
Log-likelihood[s]	Likelihood of the tweet, based on a bigram language model constructed from all of the tweets of the article.
Number of words[s]	A higher number of words may indicate a tweet with more useful content.
Me-information[s]	Presence or absence of a first-person pronoun, indicating if the tweet is mainly about the author himself [24].
Question[s]	Presence of a question mark in the content.
Quote sharing[s]	Presence of a quotation from the news article in the tweet.
Quality[s]	An measure of the tweet quality, indicating how well-written the tweet is. We use a supervised approach using on a lexicon of low-quality terms, an English dictionary, and the proportion of words, hashtags, capitalized characters, and repetitions.
Sentiment[s]	% of positive and negative words, % of subjective words, and a mixed sentiment score, reflecting the presence of highly emotional or opinionated expressions in a tweet [29].
Intensity[s]	A measure indicating the strength of the user's reaction [30].
Explicit controversy[s]	A measure reflecting the mention of common controversial issues [29].
Location[d]	The geographic location derived from the tweet, either from a mentioned place or the location of the user.
Retweet and reply[s,d]	A flag indicating whether the tweet is a retweet or a reply.
Followers, friends[s]	Number of followers and friends.
Follower-friend ratio[s]	Ratio between number of followers and friends.
Number of retweets[s]	The total number of retweets of the user's posts over the time of data gathering, signaling user authority [34, 7].
Number of users retweeting[s]	The total number of other users that retweeted this user at least once.
Tweet-retweet ratio[s]	Ratio between the number of tweets a user posts and the number of retweets received.
User verified[s]	A flag indicating if the user is verified by Twitter, which may increase the credibility of the user's posts.
User spam[s]	A flag indicating if the user is a spammer, based on a large spam dictionary trained on web page spam.
User authority score[s]	A global authority score, computed to have a topic-independent estimate of the user's overall importance.
User topic authority[s]	We consider a user u authoritative for the article's topic, when previous tweets on the same topic were often retweeted. We extract the three most relevant named entities from the article using our "aboutness" system [28] and consider these entities as a synthetic summary of the topic of the article. For each of these entities we then extract the set of tweets of the user that mention the entity and look up the number of times any tweet in in this set was retweeted by other users. The topic authority score is then computed as a combination of the relevance of the entities for the article and the number of retweets.

Conclusions

Conclusions I

- We started in Lect4 talking about frequency counts and their use in text analysis
- Here we have examined how more statistical considerations of frequency (using probability models) can inform analyses

Conclusions II

- TF-IDF weighted word frequencies in a corpus
- LLR helped us find exceptional terms/words
- PMI allows us to find words that "go together"
- Entropy tells us about redundancy or randomness/interestingness in a text-item

Next…

In the next lecture we look at how if you take a vector of (weighted) frequencies for a text-item, you can compare it to other text-items to determine its similarity