Text Mining for the Social Sciences
Lecture 3: Count-based analysis

Stephen Hansen

Introduction

Last time we discussed selecting terms for the document-term matrix X, which recall is high dimensional and often very sparse.

Now we want to begin to extract information from X.

Count-based methods use the elements of X to describe text without an explicit probability model.

The sparsity of X makes intelligent storage a practical concern for all the methods we'll see.

Outline

1. Boolean search
2. Dictionary methods
3. Term weighting
4. Vector space model

Boolean Methods

Boolean search provide a binary representation of each document based on whether it includes certain terms or not.

Define an *incidence matrix* X^I where $X^I_{dv} = \mathbb{1}(X_{dv} > 0)$.

One can now think of each document as a bit vector corresponding to a row in X^I.

We can define Boolean expressions involving AND, OR, and NOT on the columns of X^I.

Examples

The simplest Boolean expression is just “term v in document d”, equivalent to the vth column in X^I.

A more complex expression is “term v1 in document d” AND “term v2 in document d”, equivalent to multiplying $v1$th and $v2$th columns in X^I.

“term v1 in document d” AND NOT “term v2 in document d”, equivalent to multiplying $v1$th column and $1 - v2$th column.

“term v1 in document d” OR “term v2 in document d”, equivalent to $\mathbb{1}(v1\text{th column} + v2\text{th column} > 0)$.

Advantages

Boolean search is important for many document-retrieval systems, and has been built into many search engines.

An advantage is that the hard work has been done for you if the documents of interest have already been indexed.

Moreover, if you only care about the number of documents satisfying a Boolean query, no need to collect the data for yourself.

For example, Google returns the number of web pages that satisfy any particular search.

Economics Application

The recent work of Baker, Bloom, and Davis (http://www.policyuncertainty.com/) is largely based on Boolean search.

BBD are interested in measuring economic policy uncertainty, and create an index based in large part on Boolean searches of newspaper articles from major US and European newspapers.

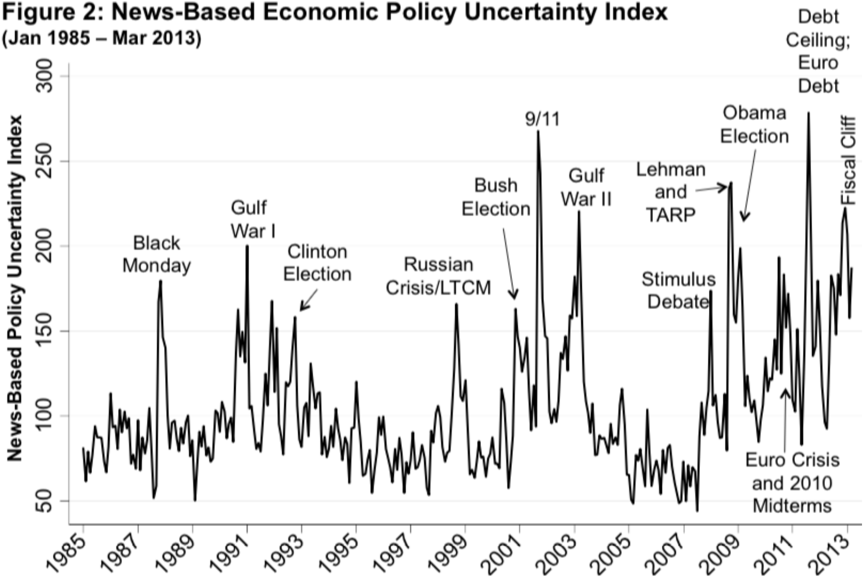
For each paper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Take resulting article counts, and normalize by total newspaper articles that month.

Results

Figure 2: News-Based Economic Policy Uncertainty Index
(Jan 1985 – Mar 2013)



Why Text?

VIX is an asset-based measure of uncertainty: implied S&P 500 volatility at 30-day horizon using option prices.

So what does text add to this?

1. Focus on broader type of uncertainty besides equity prices.
2. Much richer historical time series.
3. Cross-country measures.

Dictionary Methods

Dictionary methods operate on the document-term matrix X rather than the incidence matrix X^I.

They involve two steps:

1. Define a list of key words that captures content of interest.
2. Represent each document in terms of the (normalized) frequency of words in the dictionary.

For example, let the dictionary be $\mathfrak{D} = \{\text{labor}, \text{wage}, \text{employ}\}$.

One could then represent each document d as

$$s_d = \frac{\#\text{ labor occurrences} + \#\text{ wage occurrences} + \#\text{ employ occurrences}}{\text{total words in document d}}$$

Boolean Search versus Dictionary Methods

Boolean search considers the presence of a word to be informative of its content.

Dictionary methods consider the intensity of word use to be informative.

Consider a Boolean search for the presence of “information”. Equivalence between

- The textbook *Introduction to Information Retrieval*
- Newspaper article on a new drug about which the FDA has insufficient information to approve.

The textbook is clearly more about “information” than the article, and dictionary methods measure this.

Application

A large literature in finance uses dictionary methods to measure sentiment of text (newspaper columns, 10-K filings, press releases, etc.)

In first lecture, we identified sentiment analysis as example of supervised learning. This literature does not use labels to inform its estimate of content; the labels are instead incorporated in second-stage regressions.

Tetlock (2007)

Tetlock (2007) is a highly cited paper that applies dictionary methods to the Wall Street Journal's "Abreast of the Market" column.

Uses Harvard IV-4 dictionaries `http://www.wjh.harvard.edu/~inquirer`.

Large number of categories: positive, negative, pain, pleasure, rituals, natural processes, etc. 77 in all.

Count number of words in each dictionary in each column from 1984-1999.

Principal components analysis shows most variation on dimensions that reflect pessimism: negative, weak, fail, fall.

Tetlock (2007)

Tetlock (2007) is a highly cited paper that applies dictionary methods to the Wall Street Journal's "Abreast of the Market" column.

Uses Harvard IV-4 dictionaries `http://www.wjh.harvard.edu/~inquirer`.

Large number of categories: positive, negative, pain, pleasure, rituals, natural processes, etc. 77 in all.

Count number of words in each dictionary in each column from 1984-1999.

Principal components analysis shows most variation on dimensions that reflect pessimism: negative, weak, fail, fall.

Main result: pessimism predicts low short-term returns (measured with the Dow Jones index) followed by reversion.

Loughran and McDonald (2011)

Following Tetlock (2007), popular to use just negative word dictionary from Harvard IV-4.

This includes words like ‘tax’, ‘cost’, ‘capital’, ‘liability’, and ‘vice’.

Unclear that these are appropriate for describing negative content in financial context.

Loughran and McDonald (2011) use 10-K filings to define their own finance-specific word lists, available from `http://www3.nd.edu/~mcdonald/Word_Lists.html`.

Negative list includes words like ‘restated’, ‘litigation’, ‘termination’, ‘unpaid’, ‘investigation’, etc.

Loughran and McDonald (2011)

Following Tetlock (2007), popular to use just negative word dictionary from Harvard IV-4.

This includes words like ‘tax’, ‘cost’, ‘capital’, ‘liability’, and ‘vice’.

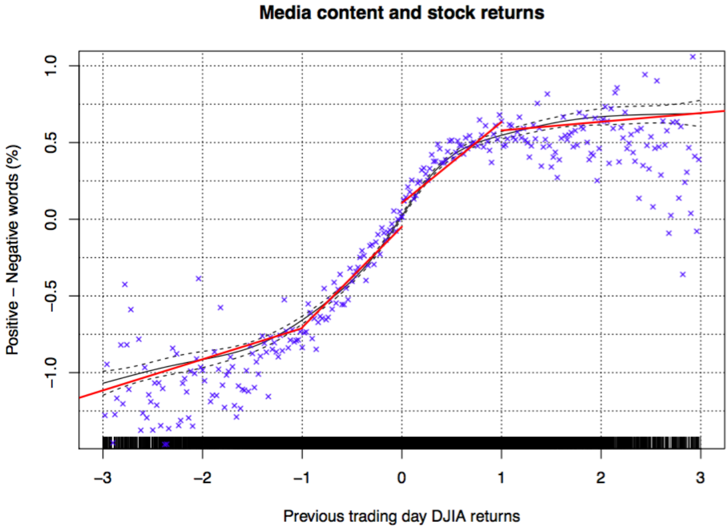
Unclear that these are appropriate for describing negative content in financial context.

Loughran and McDonald (2011) use 10-K filings to define their own finance-specific word lists, available from `http://www3.nd.edu/~mcdonald/Word_Lists.html`.

Negative list includes words like ‘restated’, ‘litigation’, ‘termination’, ‘unpaid’, ‘investigation’, etc.

Main result: the context-specific list has greater predictive power for return regressions than the generic one.

Garcia (2015)


Media content and stock returns
Positive – Negative words (%)
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
Previous trading day DJIA returns

Dictionary Methods and Psychology

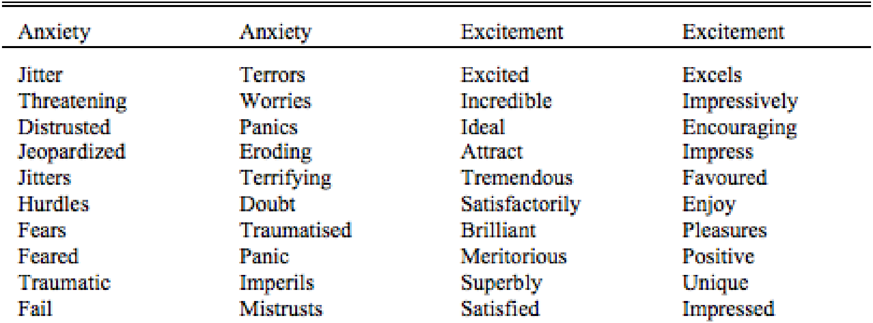
Established idea in economics is that agents may face ambiguity, or uncertainty about the probability distribution from which payoffs are drawn.

Conviction Narrative Theory is framework from psychology that informs ambiguous choices; behavior motivated by anxiety and excitement.

Tuckett et. al. (2015) use dictionary methods and describe documents as the number of excitement words net of the number of anxiety words (custom dictionaries).

When applied to Reuters News Archive (1996-2014), this index positively predicts future values of GDP growth, even after controlling for asset prices.

Illustration of Words



Anxiety	Anxiety	Excitement	Excitement
Jitter	Terrors	Excited	Excels
Threatening	Worries	Incredible	Impressively
Distrusted	Panics	Ideal	Encouraging
Jeopardized	Eroding	Attract	Impress
Jitters	Terrifying	Tremendous	Favoured
Hurdles	Doubt	Satisfactorily	Enjoy
Fears	Traumatised	Brilliant	Pleasures
Feared	Panic	Meritorious	Positive
Traumatic	Imperils	Superbly	Unique
Fail	Mistrusts	Satisfied	Impressed

Term Weighting

Dictionary methods are based on raw counts of words.

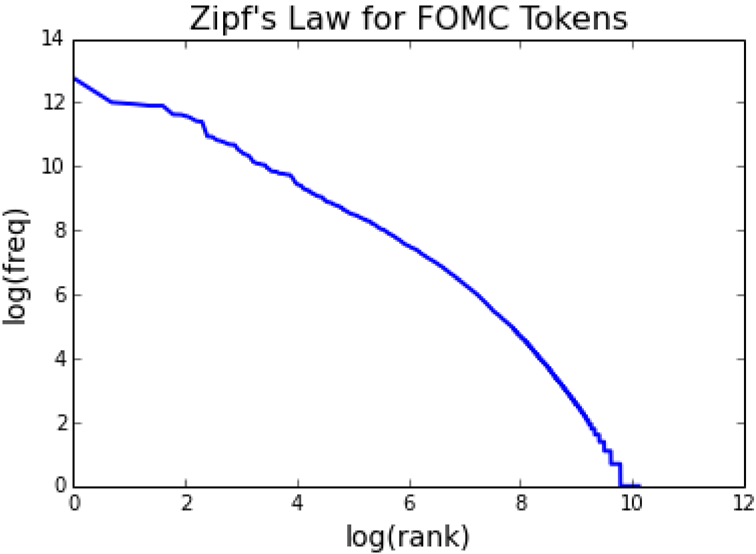
But the particular frequency of words in natural language makes this rather distorted.

Zipf's Law is an empirical regularity for most natural languages that maintains that the frequency of a particular term is inversely proportional to its rank.

Means that a few terms will have very large counts, many terms have small counts.

Example of a *power law*.

Zipf's Law in FOMC Transcript Data



Rescaling Counts

Let $x_{d,v}$ be the count of the vth term in document d.

To dampen the power-law effect can express counts as $\log(1 + x_{d,v})$.

Thought Experiment

Consider a two-term dictionary $\mathfrak{D} = \{v', v''\}$.

Suppose two documents d' and d'' are such that:

$$x_{d',v'} > x_{d'',v'} \text{ and } x_{d',v''} < x_{d'',v''}.$$

Now suppose that no other document uses term v' but every other document uses term v''.

Which document is "more about" the theme the dictionary captures?

Inverse Document Frequency

Let df_v be the number of documents that contain the term v.

The *inverse document frequency* is

$$\mathrm{idf}_v = \log\left(\frac{D}{df_v}\right),$$

where D is the number of documents.

Properties:

1. Higher weight for words in fewer documents.
2. Log dampens effect of weighting.

TF-IDF Weighting

Combining the two observations from above allows us to express the *term frequency - inverse document frequency* of term v in document d as

$$\text{tf-idf}_{d,v} = \overbrace{\log\left(1 + x_{d,v}\right)}^{tf_{d,v}} \times \overbrace{\log\left(\frac{D}{df_v}\right)}^{idf_v}.$$

Gives prominence to words that occur many times in few documents.

Can now score each document as $s_d = \sum_{v \in \mathfrak{D}} \text{tf-idf}_{d,v}$ and then compare.

In practice, this provides better results than simple counts.

Note that divergence between generic and specific dictionaries in Loughran and McDonald (2011) is greatly reduced after tf-idf correction.

Data-Driven Stopwords

Stopword lists are useful for generic language, but there are also context-specific frequently used words.

For example, in a corpus of court proceedings, words like ‘lawyer’, ‘law’, ‘justice’ will show up a lot.

Can also use collection frequency in place of $x_{d,v}$, i.e. $x_v = \sum_d x_{d,v}$, to endogenously choose stopwords.

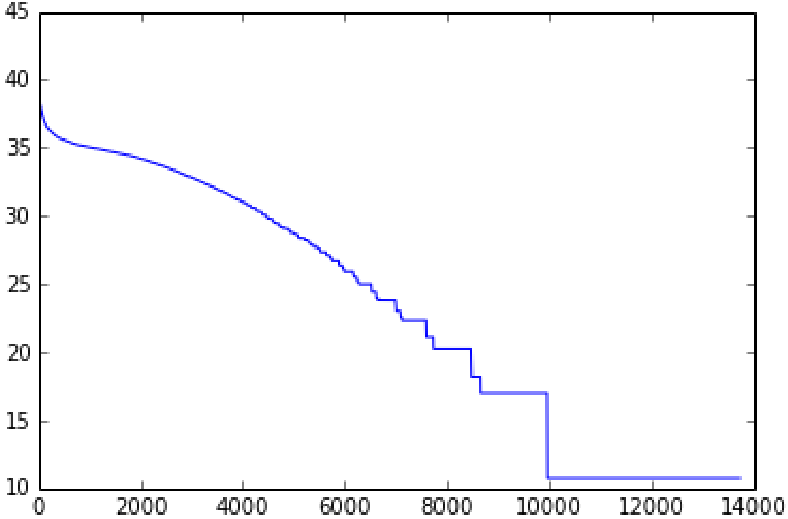
Stem Rankings in FOMC Transcript Data

R1 = collection frequency ranking

R2 = tf-idf-weighted ranking

Rank	1	2	3	4	5	6	7	8	9
R1	rate	think	year	will	market	growth	inflat	price	percent
R2	panel	katrina	graph	fedex	wal	mart	mbs	mfp	euro

Ranking of All FOMC Stems



Vector Space Model

Rather than focus on a particular set of meaningful words, we may wish to compare documents across all dimensions of variation.

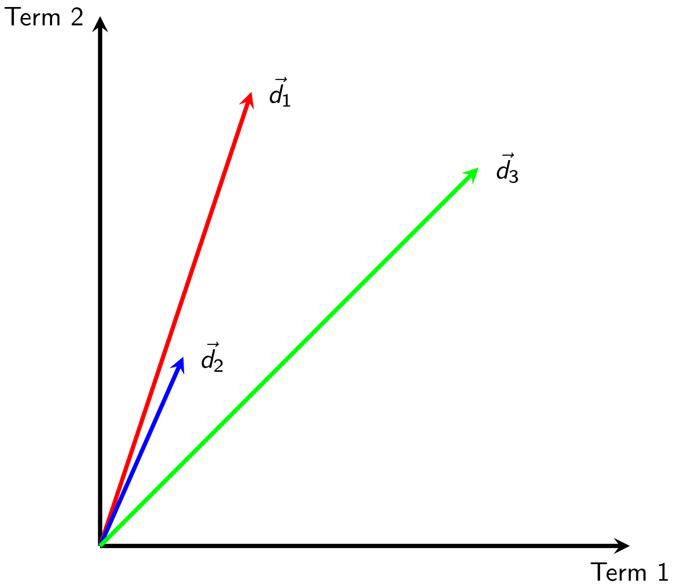
Can view rows of document-term matrix as vectors lying in a V-dimensional space, and represent document i as $\vec{d}_i$.

Tf-idf weighting usually used, but not necessary.

The question of interest is how to measure the similarity of two documents in the vector space.

Initial instinct might be to use Euclidean distance $\sqrt{\sum_v \left(x_{i,v} - x_{j,v}\right)^2}$.

Three Documents



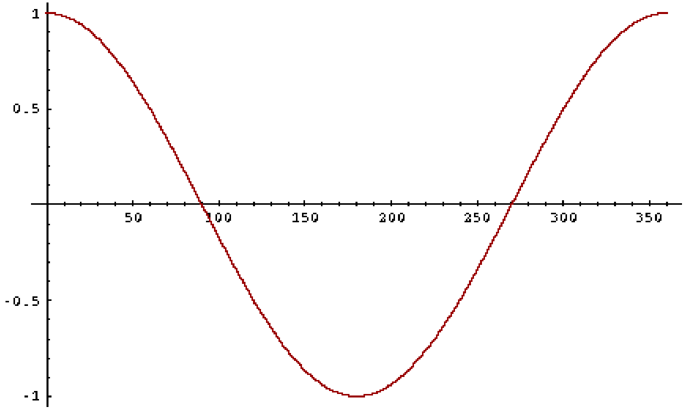
Problem with Euclidean Distance

Semantically speaking, documents 1 and 2 are very close, and document 3 is an outlier.

But the Euclidean distance between 1 and 2 is high due to differences in document length.

What we really care about is whether vectors point in same direction.

Cosine



Cosine Similarity

Define the cosine similarity between documents *i* and *j* as

$$CS(i,j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\left\|\vec{d}_i\right\| \left\|\vec{d}_j\right\|}$$

1. So long as document vectors have no negative elements, we have that $CS(i,j) \in [0,1]$.
2. $\vec{d}_i / \left\|\vec{d}_i\right\|$ is unit-length, correction for different distances.
3. Can use vector space model for clustering and classification (e.g. kNN).

Information Retrieval Application

An important task in information retrieval is to compute the similarity of an out-of-sample document with respect to within-sample documents.

For example, users of an information system might submit search terms, often called a *query*, and the system should return the most relevant document(s).

One can treat the query as a vector in the same vector space as the documents, and return to the user an ordered list of documents according to cosine similarity.

Example of ranked retrieval, much more informative than Boolean search.

Only caveat: some terms in query may not be in the set of terms used in documents.

Economics Application

An important theoretical concept in industrial organization is location on a product space.

Industry classification measures are quite crude proxies of this.

Hoberg and Phillips (2010) take product descriptions from 49,408 10-K filings and use the vector space model (with bit vectors defined by dictionaries) to compute similarity between firms.

Data available from `http://alex2.umd.edu/industrydata/`.

Conclusion

In this lecture, we have discussed how to use the document-term matrix X to describe content.

All the approaches we have seen have widespread application in information retrieval.

We have ignored computationally efficient implementation of these—see MRS for details.