Text Mining for Economics and Finance
Dictionary Methods / Vector Space Model

Stephen Hansen

Introduction

Last time we introduced the document-term matrix to quantify documents.

In general, it is both high dimensional (on the column dimension) and sparse.

A first challenge in using it for empirical work is therefore to reduce its dimensionality.

In this lecture, we look at ways the current economics and finance literature has by and large addressed this problem.

Boolean Methods

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

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

One can represent each document as a bit vector corresponding to a row in $\mathbf{X}^I$.

We can define Boolean expressions involving AND, OR, and NOT on the columns of $\mathbf{X}^I$.

Examples

The simplest Boolean expression is just "term v in document d", equivalent to the vth column in $\mathbf{X}'$.

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 $\mathbf{X}'$.

"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.

E.g. Google returns the number of web pages that satisfy Boolean searches.

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.

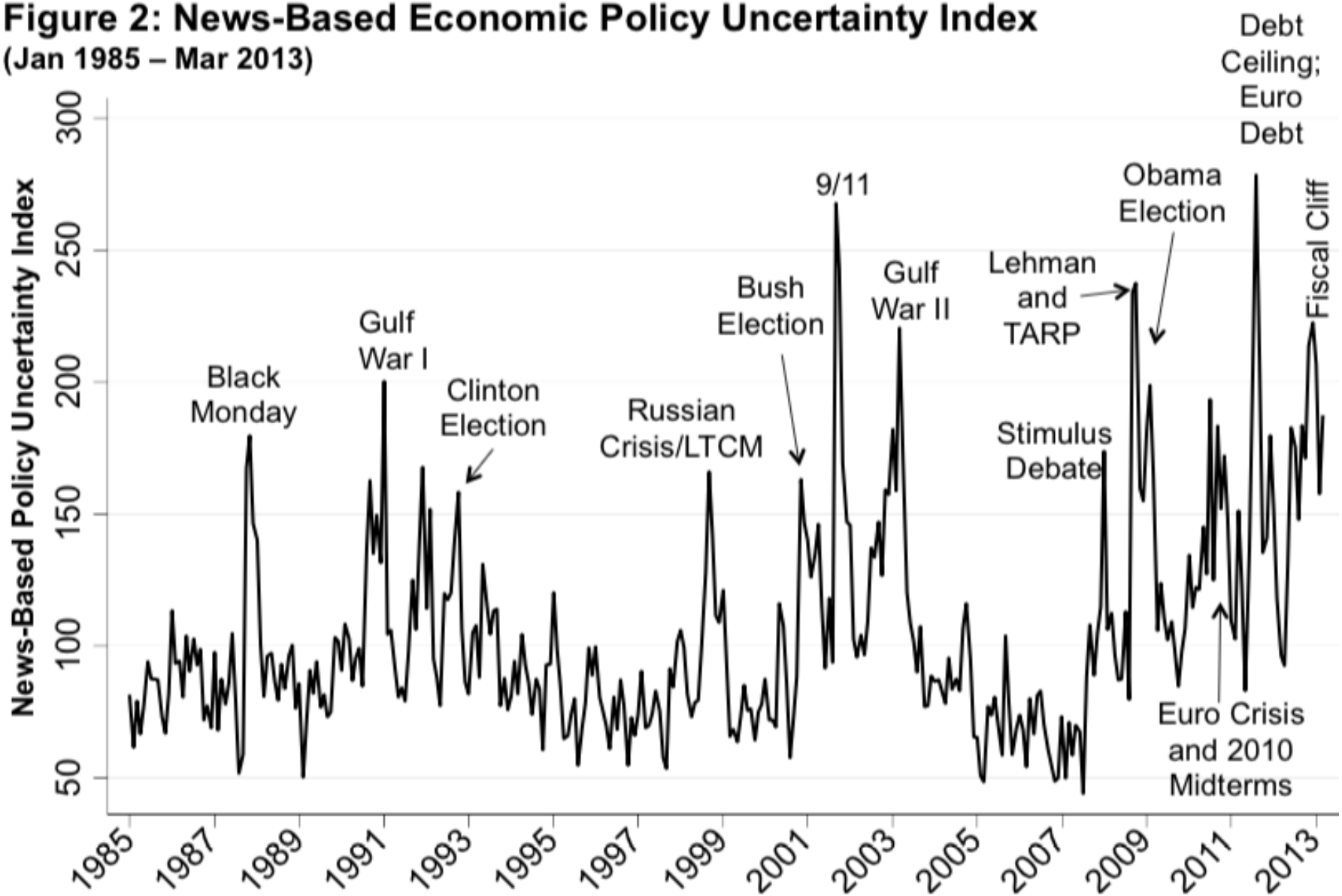
For each paper on each day since 1985, submit the following query:

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

Let $\mathfrak{D}$ be the set of keywords of interest.

We can then represent each document d as $x_d = \sum_{v \in \mathfrak{D}} x_{d,v}$ or perhaps normalize, e.g. $s_d = \sum_{v \in \mathfrak{D}} x_{d,v} / \sum_v x_{d,v}$.

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

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.

Term Weighting

Dictionary methods are based on raw counts of words.

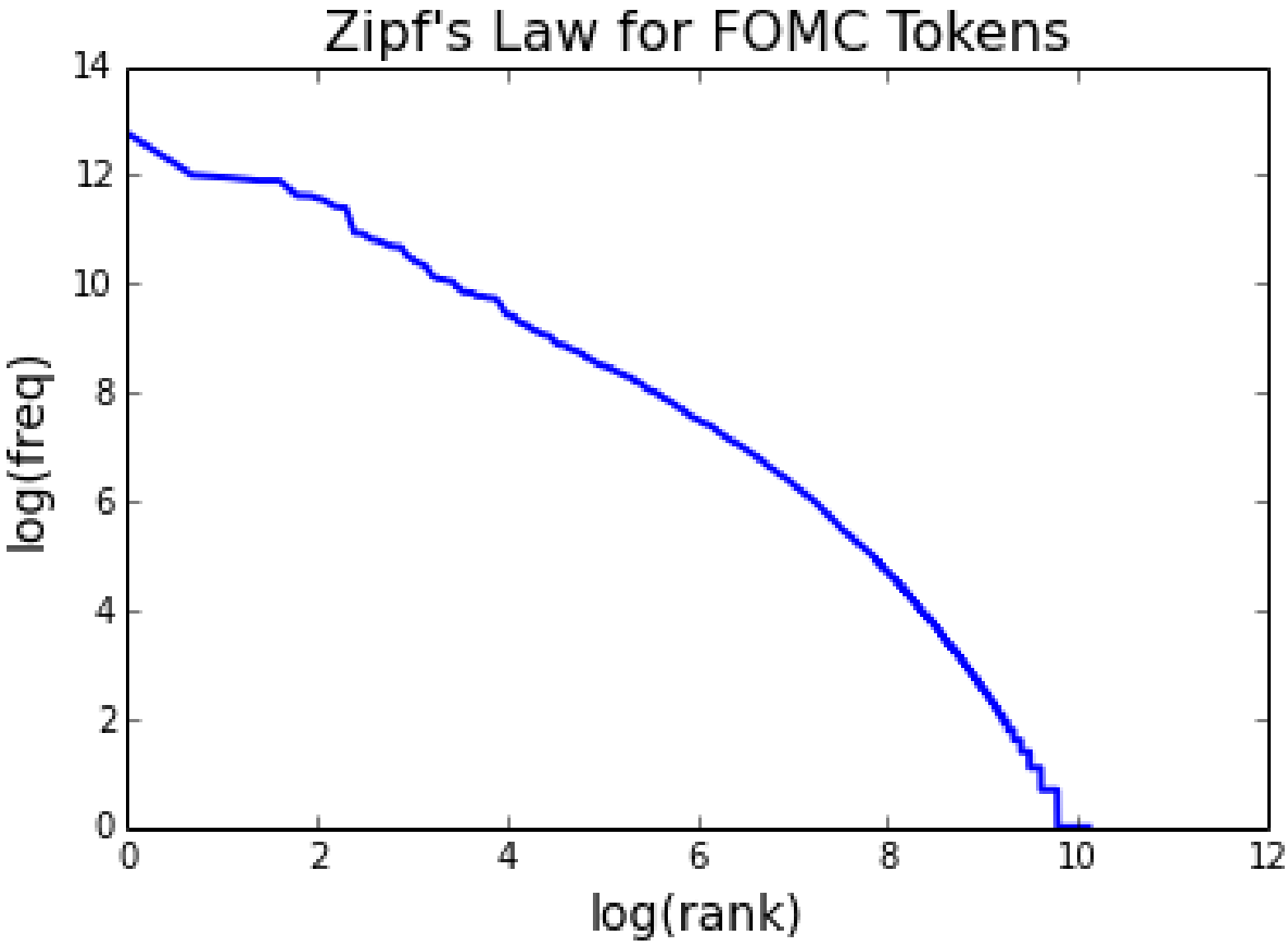
But the frequency of words in natural language can distort raw counts.

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

$$tf_{d,v} = \begin{cases} 0 & \text{if } x_{d,v} = 0 \\ 1 + \log(x_{d,v}) & \text{otherwise} \end{cases}$$

which is the *term frequency* of v in d.

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

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

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

$$\text{tf-idf}_{d,v} = tf_{d,v} \times idf_v.$$

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 define term-frequency across entire corpus as

$$tf_v = 1 + \log\left(\sum_d x_{d,v}\right).$$

One can then rank each term in the corpus according to $tf_v \times idf_v$, and choose a threshold below which to drop terms.

This provides a means for data-driven stopword selection.

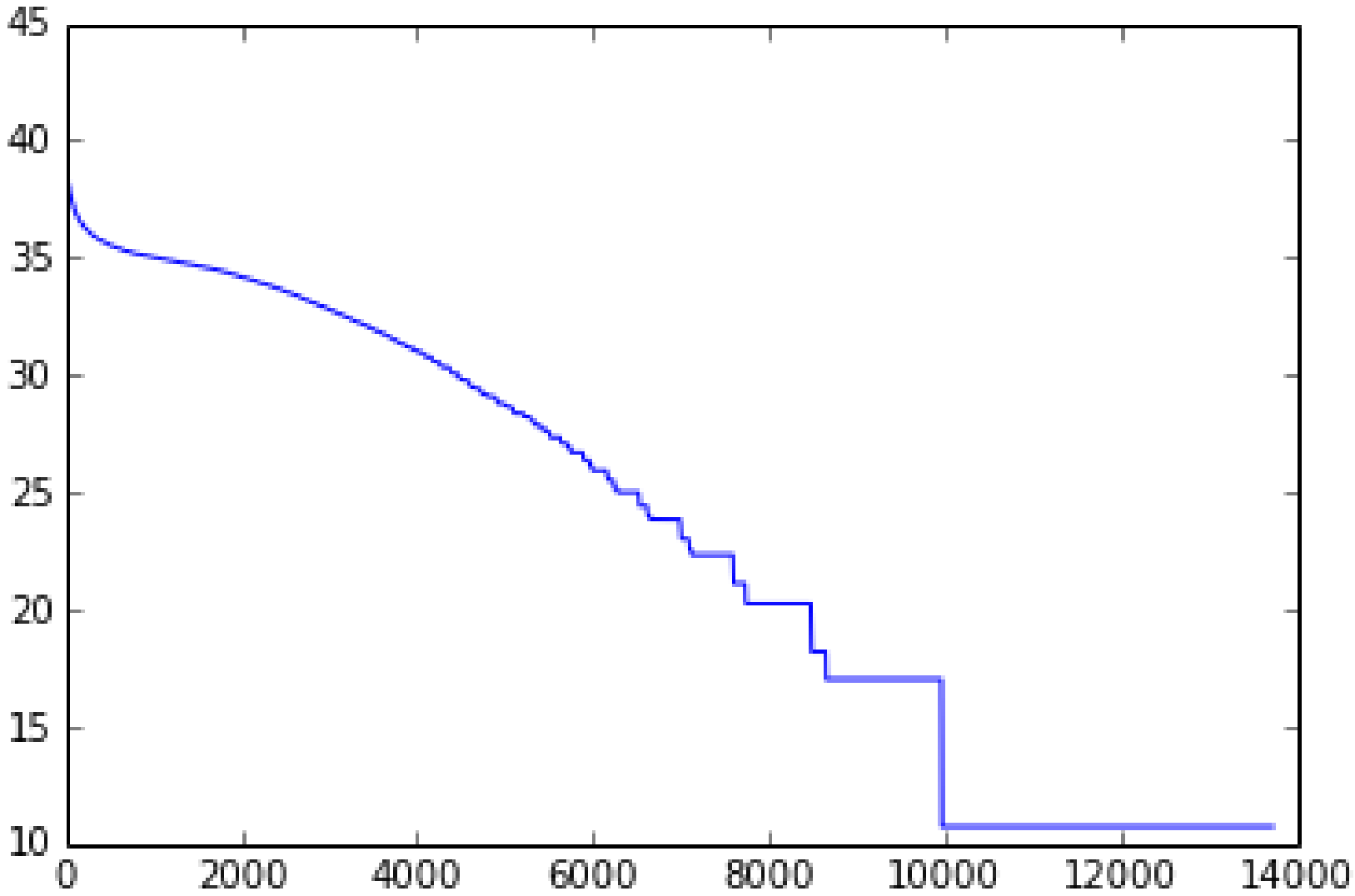
Stem Rankings in FOMC Transcript Data

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

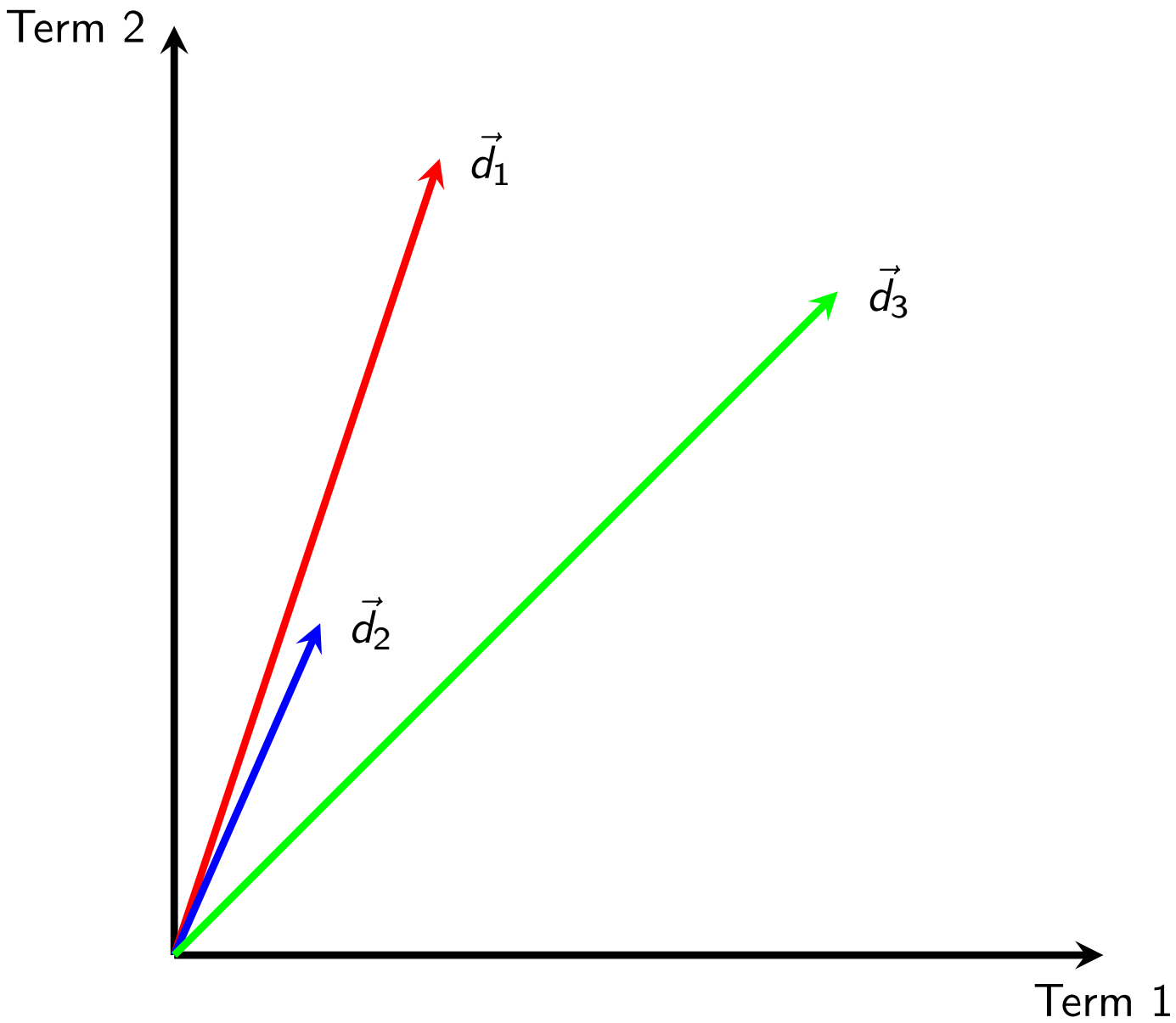
One can also view rows of document-term matrix as vectors lying in a V-dimensional space, and represent document i as $\vec{d_i}$.

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 $\sqrt{\sum_v (x_{i,v} - x_{j,v})^2}$.

Three Documents



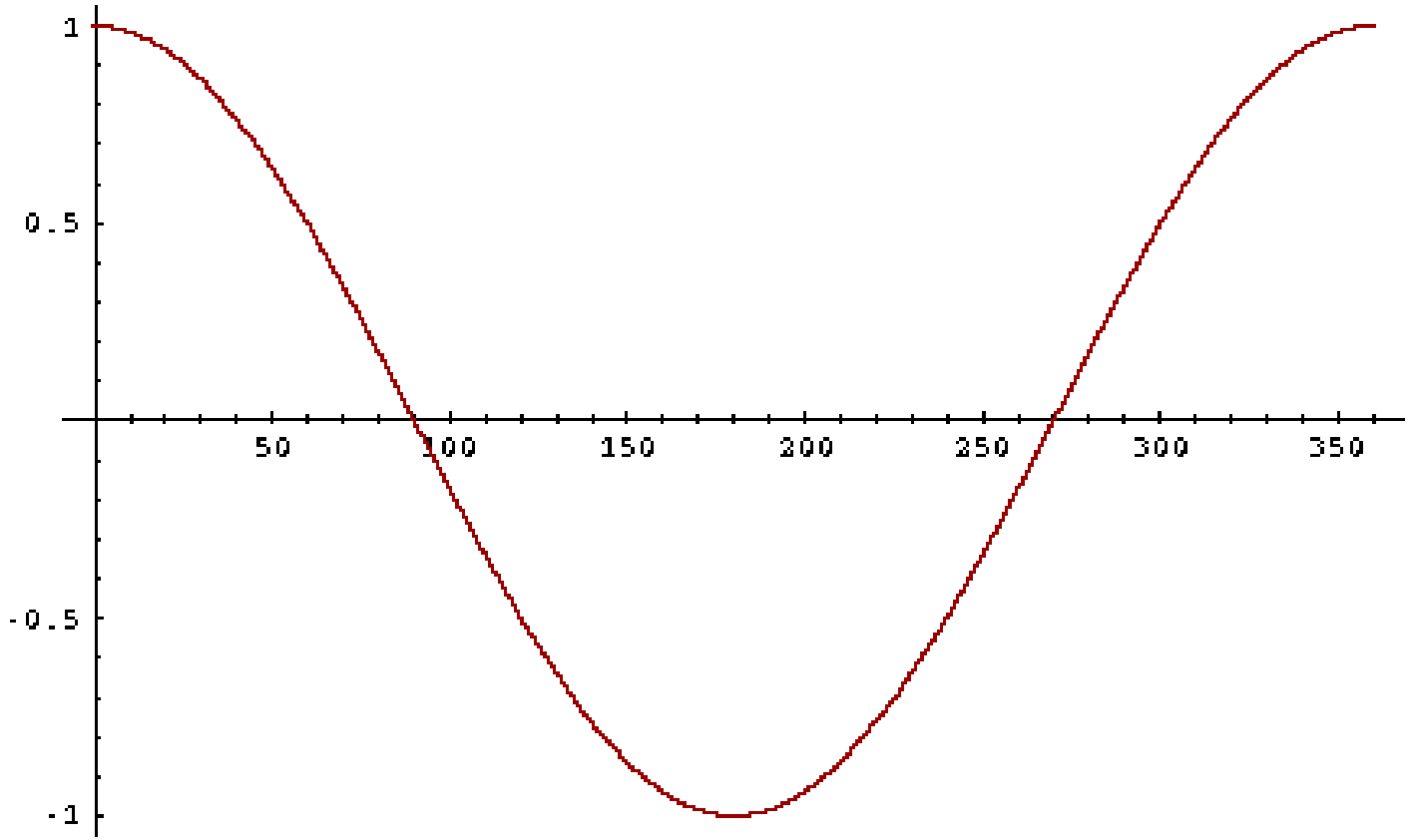
Problem with Euclidean Distance

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. Since document vectors have no negative elements $CS(i,j) \in [0,1]$.
2. $\vec{d_i} / \left\|\vec{d_i}\right\|$ is unit-length, correction for different distances.

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/`.

Towards Machine Learning

Dictionary methods focus on variation across observations along a limited number of dimensions and ignore the rest.

Ideally we would use variation across *all* dimensions to describe documents.

This obviously provides a richer description of the data, but a deeper point relevant for many high-dimensional datasets is that economic theory does not tell us which dimensions are important.

At the same time, incorporating thousands of independent dimensions of variation in empirical work is difficult.

Machine learning approaches exploit all dimensions of variation to estimate a lower-dimensional set of types for each documents.