

# A Primer on Text Mining for Economists

Michael McMahon (and Stephen Hansen)

# Introduction

Traditional focus in (monetary) economics on quantitative information

# Introduction

Traditional focus in (monetary) economics on quantitative information

We also have access to lots of text data

- Papers, reports, speeches, statements, questionnaires, emails, etc.

Typically this data has been ignored, or analyzed in a qualitative way.

# Introduction

Traditional focus in (monetary) economics on quantitative information

We also have access to lots of text data

- Papers, reports, speeches, statements, questionnaires, emails, etc.

Typically this data has been ignored, or analyzed in a qualitative way.

This presentation focuses on ways of analyzing text **quantitatively**.

# Text as Data

Broadly speaking, text mining is the study of the quantitative representation of text.

Text data is a sequence of characters called *documents*.

The set of documents is the *corpus*.

Text data is *unstructured*: the information we want is mixed together with (lots of) information we don't. **How to separate the two?**

All text mining algorithms will throw away some information  
⇒ key is knowing what is the valuable information to retain.

## Example

A representation of text we will consider a lot today is the *bag-of-words* model, which represents text as a frequency count of words.

Doc1 = ['text', 'mining', 'is', 'more', 'fun', 'than', 'coal', 'mining']

## Example

A representation of text we will consider a lot today is the *bag-of-words* model, which represents text as a frequency count of words.

Doc1 = ['text', 'mining', 'is', 'more', 'fun', 'than', 'coal', 'mining']

Bag of words:

| text | mining | is | more | fun | than | coal |
|------|--------|----|------|-----|------|------|
| 1    | 2      | 1  | 1    | 1   | 1    | 1    |

## Example

A representation of text we will consider a lot today is the *bag-of-words* model, which represents text as a frequency count of words.

Doc1 = ['text', 'mining', 'is', 'more', 'fun', 'than', 'coal', 'mining']

Bag of words:

| text | mining | is | more | fun | than | coal |
|------|--------|----|------|-----|------|------|
| 1    | 2      | 1  | 1    | 1   | 1    | 1    |

Note that the bag of words representation is the same as for the document

Doc2 = ['coal', 'mining', 'is', 'more', 'fun', 'than', 'text', 'mining']

Does this matter?



# Unigrams and N-grams

The bag-of-words model is sometimes called the unigram model because we model each word in isolation from its neighbors.

This makes sense when we care about what a document is about (such as in information retrieval applications).

# Unigrams and N-grams

The bag-of-words model is sometimes called the unigram model because we model each word in isolation from its neighbors.

This makes sense when we care about what a document is about (such as in information retrieval applications).

- 'I like dogs',
- 'My kids likes dogs',
- 'My wife thinks dogs are dirty'

All clearly different sentences, but all concern dogs.

# Unigrams and N-grams

The bag-of-words model is sometimes called the unigram model because we model each word in isolation from its neighbors.

This makes sense when we care about what a document is about (such as in information retrieval applications).

- 'I like dogs',
- 'My kids likes dogs',
- 'My wife thinks dogs are dirty'

All clearly different sentences, but all concern dogs.

*N*-gram models consider sequences of words in sentences and are used for tasks like speech recognition, machine translation, and spelling correction.

## Our Way into Text Data

We have long-standing interest in behavior of monetary policymakers.

Initial papers on voting behavior in committees. Votes are easy to quantify.

Most time in committees spent deliberating, but little research on this.

# Our Way into Text Data

We have long-standing interest in behavior of monetary policymakers.

Initial papers on voting behavior in committees. Votes are easy to quantify.

Most time in committees spent deliberating, but little research on this.

Started research in this area in 2012 (with Andrea Prat):

1. “Transparency and Communication on the FOMC: A Computational Linguistics Approach”
2. “Shocking Language: Understanding the macroeconomic effects of central bank communication”

# The Data

A large amount of text data available from [http://www.federalreserve.gov/monetarypolicy/fomc\\_historical.htm](http://www.federalreserve.gov/monetarypolicy/fomc_historical.htm).

We first focused on FOMC transcripts from the era of Alan Greenspan.  
149 meetings from August 1987 through January 2006.

A document is a single statement by a speaker in a meeting.

There are 46,502 such statements.

Associated metadata: speaker biographical information, macroeconomic conditions, etc.

No memory problems: data is around 40 MB file.

# The Challenge

6,249,776 total alphanumeric tokens in the data: 26,030 unique tokens.

Consider the bag of words representation in terms of a *document-term matrix*  $\mathbf{X}$ , whose  $(d, v)$ th element is the count of token  $v$  in document  $d$ .

Matrix is high-dimensional and sparse, challenge is how to reduce dimensionality while preserving the important variation across documents.

# Outline

1. Pre-processing
2. Dictionary methods
3. Bag of Words and the Vector Space Model
4. Topic Models
5. Learning Models



## Pre-processing Steps

To generate a term-document matrix from raw text, we must process strings. Typical steps are:

1. *Tokenize*. Break up strings into constituent parts (e.g. words, numbers, and punctuation).
2. *Case-fold*. Convert all strings to lower case.
3. *Stopword removal*. Drop extremely common words like 'a', 'the', 'it', and so on.
4. *Equivalence class*. Bring words into linguistic root through stemming or lemmatizing.
5. *Further token removal*. Drop rare words, along with punctuation and/or numbers depending on the context.

# Equivalence Classing

Option #1: Lemmatizing (first-best). Requires POS tagging first.

Converts 'ran' to 'run' if tagged as verb.

Option #2: Stemming (dirty but fast). Deterministic algorithm to remove ends from words. 'prefer', 'prefers', and 'preference' all become 'prefer'.

Not necessarily English word. 'inflation' becomes 'inflat'.

## Effect of Processing on Dimensionality

The following table represents the effect of pre-processing steps on the dimensionality of the FOMC transcript data.

|              | All terms | Alpha terms | No stopwords | Stems   |
|--------------|-----------|-------------|--------------|---------|
| # terms      | 6249776   | 5519606     | 2505261      | 2505261 |
| Unique terms | 26030     | 24801       | 24611        | 13734   |

Substantial reductions, but still very high-dimensional space.

# Dictionary Methods

Dictionary methods operate on the document-term matrix  $\mathbf{X}$ .

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  $\mathcal{D} = \{\text{labor, wage, employ}\}$ .

One could then represent each document  $d$  as

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

# Measuring Uncertainty

Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.

(See <http://www.policyuncertainty.com/>).

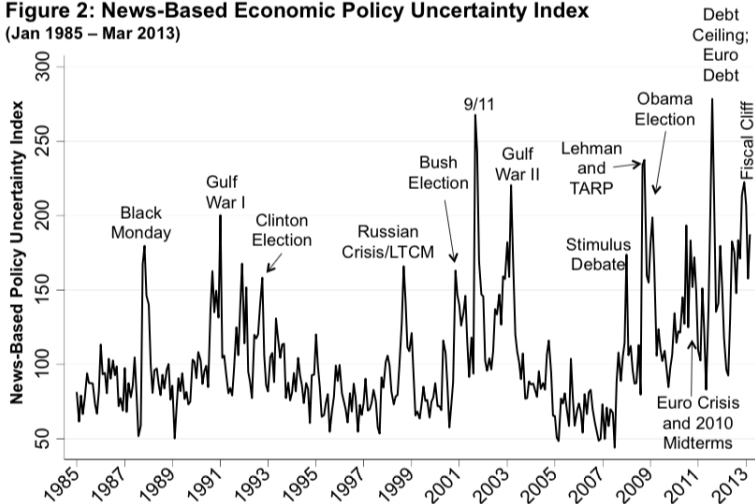
For each paper on each day since 1985, submit the following query:

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”

Normalize resulting article counts by total newspaper articles that month.

## Results

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



# Why Text?

So what does text add to a measure like VIX?

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

## 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.

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](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 particular frequency of words in natural language makes this rather distorted.

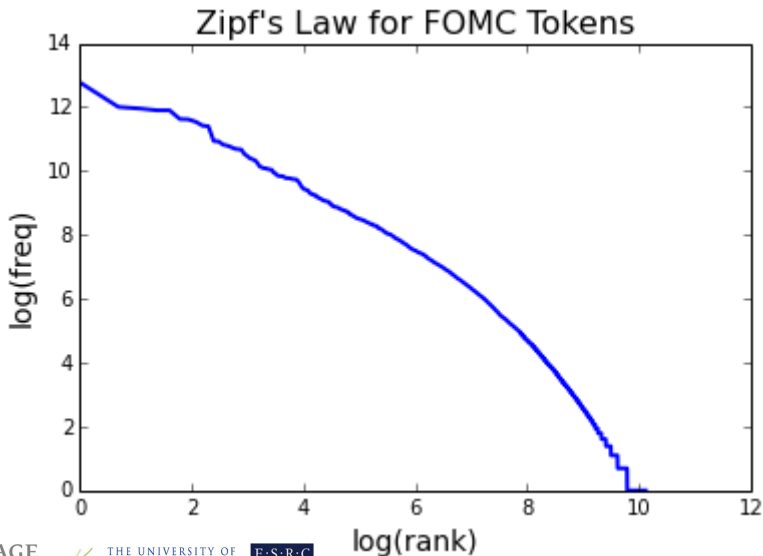
Most natural languages are an example of a *power law*

Zipf's Law

⇒ the frequency of a particular term is inversely proportional to its rank.

A few terms will have very large counts, many terms have small counts.

# Zipf's Law in FOMC Transcript Data



## Rescaling Counts: TF-IDF Weighting

We want to

- dampen power law effect we use logs of counts
- give higher weight for words in fewer documents

*term frequency - inverse document frequency* of term  $v$  in document  $d$  as

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

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

In practice, this provides better results than simple counts.

# Vector Space Model

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

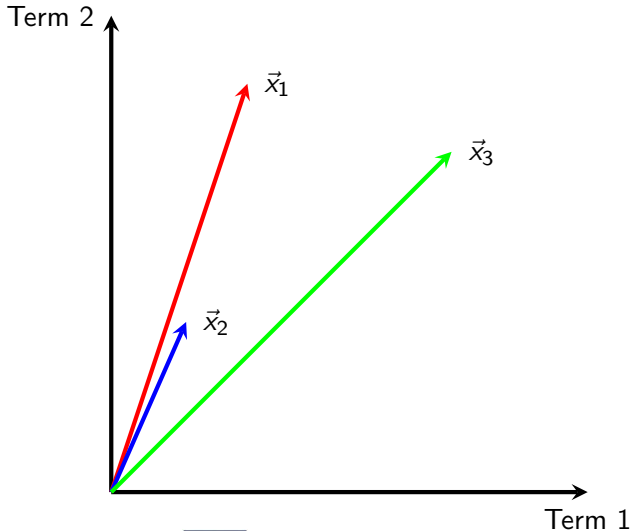
Can view rows of document-term matrix as vectors lying in a  $V$ -dimensional space, and represent document  $d$  as  $\vec{x}_d$ .

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  $\sqrt{\sum_v (x_{d,v} - x_{d',v})^2}$ .

# 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 Similarity

Define the cosine similarity between documents  $d$  and  $d'$  as

$$CS(i, j) = \frac{\vec{x}_d \cdot \vec{x}_{d'}}{\|\vec{x}_d\| \|\vec{x}_{d'}\|}$$

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



# Beyond Word Counts

The vector space model treats each term in the vocabulary as an independent source of variation.

Properties of language make this assumption quite restrictive:

1. synonymy
2. polysemy

## Synonymy

The same underlying concept can be described by many different words.

Words associated with the theme of education are 'school', 'university', 'college', 'teacher', 'professor', etc.

Consider the following two documents

|        |            |         |         |           |
|--------|------------|---------|---------|-----------|
| school | university | college | teacher | professor |
| 0      | 5          | 5       | 0       | 2         |
| school | university | college | teacher | professor |
| 10     | 0          | 0       | 4       | 0         |

# Synonymy

The same underlying concept can be described by many different words.

Words associated with the theme of education are 'school', 'university', 'college', 'teacher', 'professor', etc.

Consider the following two documents

| school | university | college | teacher | professor |
|--------|------------|---------|---------|-----------|
| 0      | 5          | 5       | 0       | 2         |
| school | university | college | teacher | professor |
| 10     | 0          | 0       | 4       | 0         |

How big is their cosine similarity?

# Polysemy

Polysemy refers to the same word having multiple meanings in different contexts.

Consider the following two documents

| tank | seal | frog | animal | navy | war |
|------|------|------|--------|------|-----|
| 5    | 5    | 3    | 2      | 0    | 0   |
| tank | seal | frog | animal | navy | war |
| 5    | 5    | 0    | 0      | 4    | 3   |

# Polysemy

Polysemy refers to the same word having multiple meanings in different contexts.

Consider the following two documents

| tank | seal | frog | animal | navy | war |
|------|------|------|--------|------|-----|
| 5    | 5    | 3    | 2      | 0    | 0   |
| tank | seal | frog | animal | navy | war |
| 5    | 5    | 0    | 0      | 4    | 3   |

How related are these documents? How large is their cosine similarity?

# Latent Variables

The implicit assumption of the dictionary  $\mathfrak{D} = \{\text{labor, wage, employ}\}$  is that each word maps back into an underlying topic “labor markets”.

We cannot observe the topics in text, only observe the words that those topics tend to generate.

A natural way forward is to model topics with latent variables.

# Features of Latent Variable Models

Latent variable models generally share the following features:

1. Associate each word in the vocabulary to any given latent variable.
2. Allow each word to have associations with multiple topics.
3. Associate each document with topics.

Give documents a representation in a latent, more accurate, semantic space rather than the raw vocabulary space:

- Latent Semantic Analysis - SVD
- Latent Dirichlet Allocation

Allow algorithm to find best association between words and latent variables without pre-defined word lists or labels.

# The Latent Dirichlet Allocation (LDA) model

- Blei, Ng and Jordan (2003) cited 11,500+ times
  - Hansen, McMahon and Prat (2014)
- LDA (and its extensions) estimates what fraction of each document in a collection is devoted to each of several “topics.”
- LDA is an unsupervised learning approach - we don't set probabilities



# The Latent Dirichlet Allocation (LDA) model

- Blei, Ng and Jordan (2003) cited 11,500+ times
  - Hansen, McMahon and Prat (2014)
- LDA (and its extensions) estimates what fraction of each document in a collection is devoted to each of several “topics.”
- LDA is an unsupervised learning approach - we don't set probabilities

## 1. Start with words in statements

# The Latent Dirichlet Allocation (LDA) model

- Blei, Ng and Jordan (2003) cited 11,500+ times
    - Hansen, McMahon and Prat (2014)
  - LDA (and its extensions) estimates what fraction of each document in a collection is devoted to each of several “topics.”
  - LDA is an unsupervised learning approach - we don't set probabilities
1. Start with words in statements
  2. Tell the model how many topics there should be
    - Perplexity scores

# The Latent Dirichlet Allocation (LDA) model

- Blei, Ng and Jordan (2003) cited 11,500+ times
    - Hansen, McMahon and Prat (2014)
  - LDA (and its extensions) estimates what fraction of each document in a collection is devoted to each of several “topics.”
  - LDA is an unsupervised learning approach - we don't set probabilities
1. Start with words in statements
  2. Tell the model how many topics there should be
    - Perplexity scores
  3. Model will generate  $\beta_K$  **topic distributions**
    - the distribution over words for each topic

# The Latent Dirichlet Allocation (LDA) model

- Blei, Ng and Jordan (2003) cited 11,500+ times
    - Hansen, McMahon and Prat (2014)
  - LDA (and its extensions) estimates what fraction of each document in a collection is devoted to each of several “topics.”
  - LDA is an unsupervised learning approach - we don't set probabilities
1. Start with words in statements
  2. Tell the model how many topics there should be
    - Perplexity scores
  3. Model will generate  $\beta_K$  **topic distributions**
    - the distribution over words for each topic
  4. Model also generates  $\theta_d$  **document distributions**

## Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

noticed change relationship between core CPI  
 chained core CPI suggested maybe something going  
 relating substitution bias upper level index focused  
 nonmarket component PCE wondered something  
 unusual happening core CPI relative measures

# Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → **Stemming** → Multi-word tokens = Bag of Words

chain      notic      chang      relationship between      core CPI  
 relat      core CPI      suggest      mayb      someth      go  
 nonmarket compon      upper level      index      focus  
 unusu      happen      PCE      wonder      someth  
                                  core CPI rel      measur

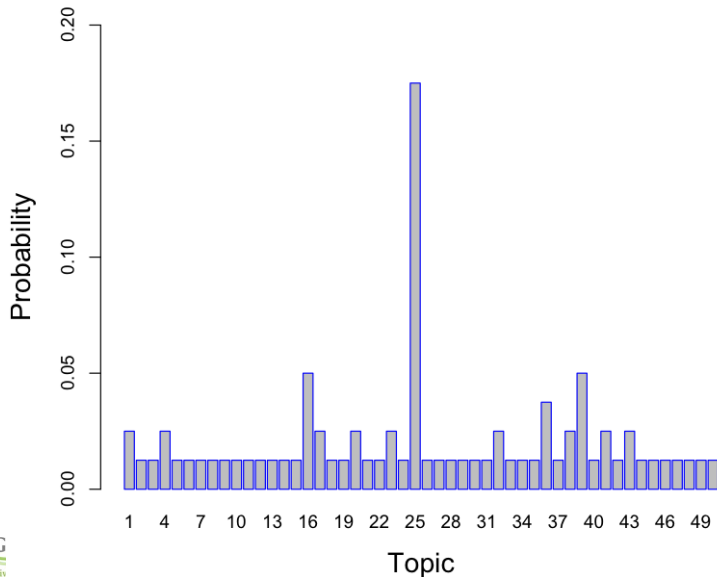
# Example statement: Yellen, March 2006, #51

## Allocation

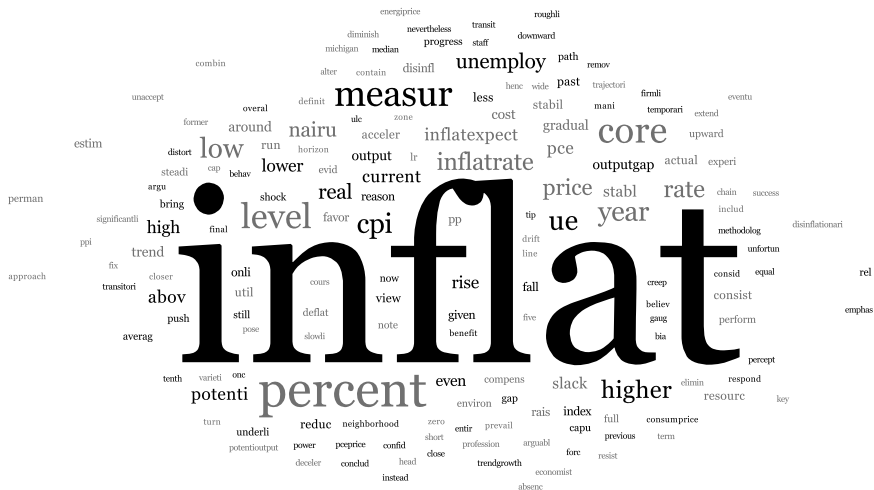
|    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|
|    | 17 |    | 39 |    | 39 |    | 1  |    | 25 | 25 |    |
| 41 | 25 | 25 |    | 25 |    |    | 36 | 36 |    |    | 38 |
| 43 |    | 25 |    | 20 | 25 | 39 |    | 16 |    |    | 23 |
|    | 25 |    | 25 |    | 25 |    | 32 |    | 38 |    |    |
| 16 |    |    | 4  |    | 25 | 25 | 16 |    |    | 25 |    |



# Distribution of Attention



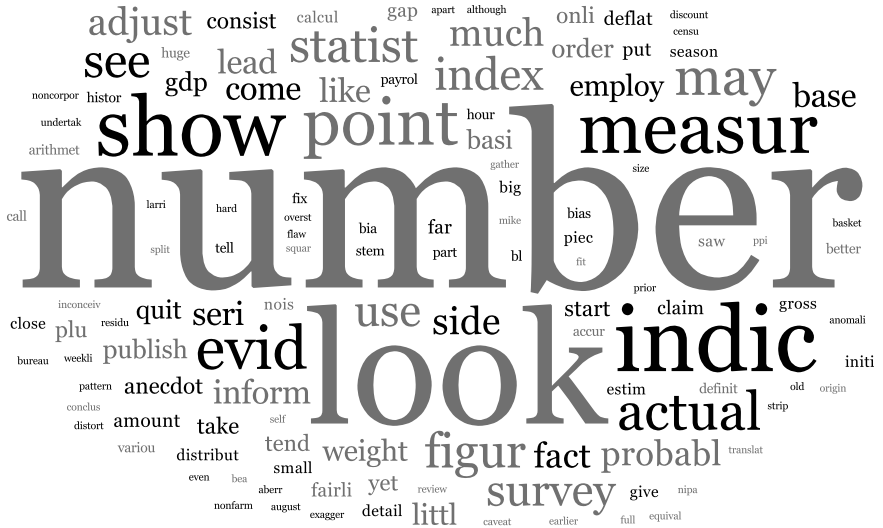
# Topic 25



# Advantage of Flexibility

'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11

## Topic 11



## Advantage of Flexibility

'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11.

It gets assigned to 25 in this statement consistently due to the presence of other topic 25 words.

In statements containing words on evidence and numbers, it consistently gets assigned to 11.

Sampling algorithm can help place words in their appropriate context.

# Topics Used to Create our Dependent Variables

Statement  $\xrightarrow{LDA}$   $K$ -dimensional vector  $\xrightarrow{Function}$  dependent variable.

# Topics Used to Create our Dependent Variables

Statement  $\xrightarrow{LDA}$   $K$ -dimensional vector  $\xrightarrow{Function}$  dependent variable.

1. Percentage of statement about economic topics (preparedness):  
Define conditional topic distribution  $\chi_{i,t,s}$
2. Herfindahl concentration index (breadth of discussion).
3. Percentage of time on quantitative topics (information acquisition).
4. Proximity to Chairman in FOMC2

# Topics Used to Create our Dependent Variables

Statement  $\xrightarrow{LDA}$   $K$ -dimensional vector  $\xrightarrow{Function}$  dependent variable.

1. Percentage of statement about economic topics (preparedness):  
Define conditional topic distribution  $\chi_{i,t,s}$
2. Herfindahl concentration index (breadth of discussion).
3. Percentage of time on quantitative topics (information acquisition).
4. Proximity to Chairman in FOMC2

Dependent variables are then used in standard econometric applications



# Combining dictionary methods and topic models

Hansen & McMahon (2015):

- Propose a simple way of combining these two approaches
  - measure topic-level tone
  - deals, somewhat, with the weakness of dictionary methods.
- Identify the paragraphs in which topic  $k$  makes up at least  $\alpha = 0.5$  fraction of attention as measured by  $\phi_{p,k,d}$  allocation.
- Compute the tone measures within that subset of paragraphs

# Active Learning Models

- Advantages of automated techniques:
  - scalability with consistency
  - scalability to larger corpora
  - Reduces the biases that might creep in
  - Might pick up some nuance (while also missing other nuance)

# Active Learning Models

- Advantages of automated techniques:
  - scalability with consistency
  - scalability to larger corpora
  - Reduces the biases that might creep in
  - Might pick up some nuance (while also missing other nuance)
- Sometimes distinguishing the nuance is key - 'narrative approach'
- Active learning models can bridge the gap

# Conclusion

We have looked at basic ideas for quantifying text.

# Conclusion

We have looked at basic ideas for quantifying text.

These and related tools can help unlock the information embedded in text corpora to allow us to address new questions.

# Conclusion

We have looked at basic ideas for quantifying text.

These and related tools can help unlock the information embedded in text corpora to allow us to address new questions.

Rather than being an end in themselves, Stephen and I think of these representations as producing outputs that will serve as the inputs into some statistical model for testing an economic hypothesis.

## Further Material

<https://github.com/sekhansen/text-mining-tutorial>

<https://github.com/sekhansen/text-mining-course>

CCBS Handbook “Text Mining for Central Banks” (Bholat, Hansen, Santos, and Schonhardt-Bailey)

# Product Space

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 to compute similarity between firms.

Data available from <http://alex2.umd.edu/industrydata/>.



# Transparency

How transparent should a public organization be?

Benefit of transparency: accountability.

Costs of transparency:

1. Direct costs
2. Privacy
3. Security
4. Worse behavior → “chilling effect”

# Transparency and Monetary Policy

**Mario Draghi (2013):** “It would be wise to have a richer communication about the rationale behind the decisions that the governing council takes.”

Table: Disclosure Policies as of 2014

|              | Fed | BoE | ECB |
|--------------|-----|-----|-----|
| Minutes?     | ✓   | ✓   | X   |
| Transcripts? | ✓   | X   | X   |

# Natural Experiment

FOMC meetings were recorded and transcribed from at least the mid-1970's in order to assist with the preparation of the minutes.

Committee members unaware that transcripts were stored prior to October 1993.

Greenspan then acknowledged the transcripts' existence to the Senate Banking Committee, and the Fed agreed:

1. To begin publishing them with a five-year lag.
2. To publish the back data.

## Difference-in-Differences

Hansen, McMahon and Prat (2014) use LDA output to study behavioral response to increased transparency.

Career concerns literature predicts that effect of transparency should decline with labor market experience.

$$y_{its} = \alpha_i + \delta_t + \beta D(Trans) + \eta FedExp_{it} + \phi D(Trans) \times FedExp_{it} + \epsilon_{it}$$

Two main meetings sections:

1. FOMC1. Discussion of economic conditions.
2. FOMC2. Discussion of policy preferences. Greenspan speaks first.



## seven

### Table: Evidence for career concerns

| Discipline                            | Conformity                         |
|---------------------------------------|------------------------------------|
| ↑ use of numbers in FOMC1             | ↓ statements in FOMC2              |
| ↑ topic breadth in FOMC1              | ↓ questions in FOMC2               |
| ↑ references to data topics in FOMC1  | ↓ distance from Greenspan in FOMC2 |
|                                       | ↓ topic breadth in FOMC2           |
| ↑ economics topic percentage in FOMC2 |                                    |

# Rescaling Counts

Let  $x_{d,v}$  be the count of the  $v$ th 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  $\mathcal{D} = \{v', v''\}$ .

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

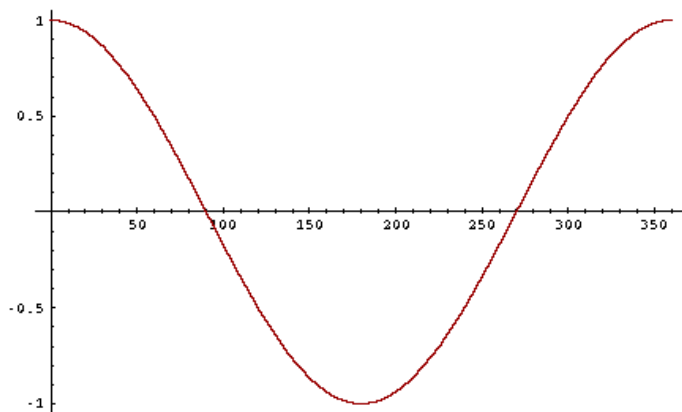
$$\text{idf}_v = \log \left( \frac{D}{df_v} \right),$$

where  $D$  is the number of documents.

Properties:

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

# Cosine



# Latent Semantic Analysis

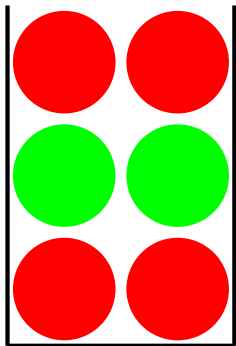
One of the earliest topic models applied a singular value decomposition to the term-document matrix (Deerwester et. al. 1990).

Called Latent Semantic Analysis, sometimes Latent Semantic Indexing.

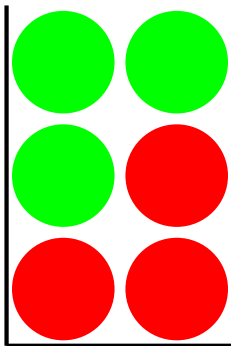
Influential approach, but principal components are difficult to interpret and LSA is a fundamentally linear algebra approach rather than a statistical one.

Correspondence Analysis is an extension of LSA, see work of Schonhardt-Bailey.

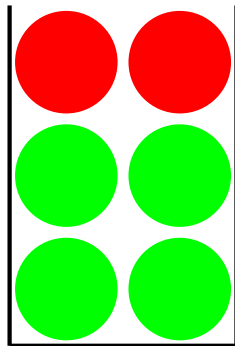
# Topics as Urns



Topic 1



Topic 2



Topic 3

# Modeling Documents

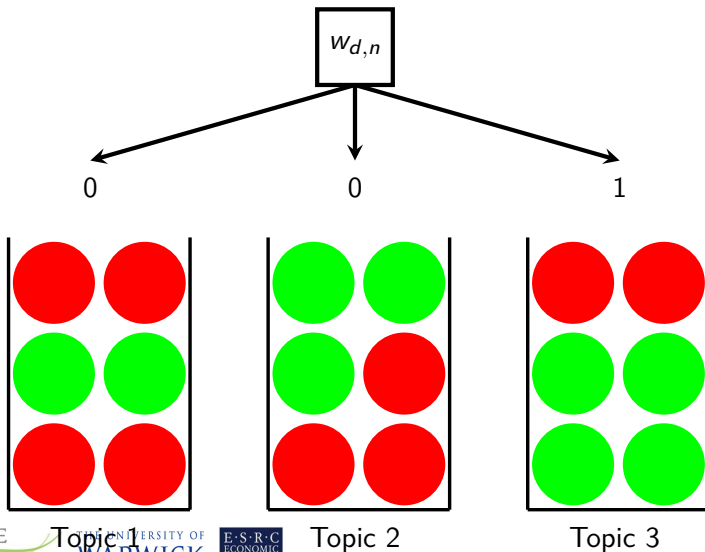
An important idea in topic modeling is that documents are composed of latent topics, and the words we observe are random draws from those topics.

A simple topic model is one in which each document  $d$  in the corpus is assigned a single topic  $z_d$ .

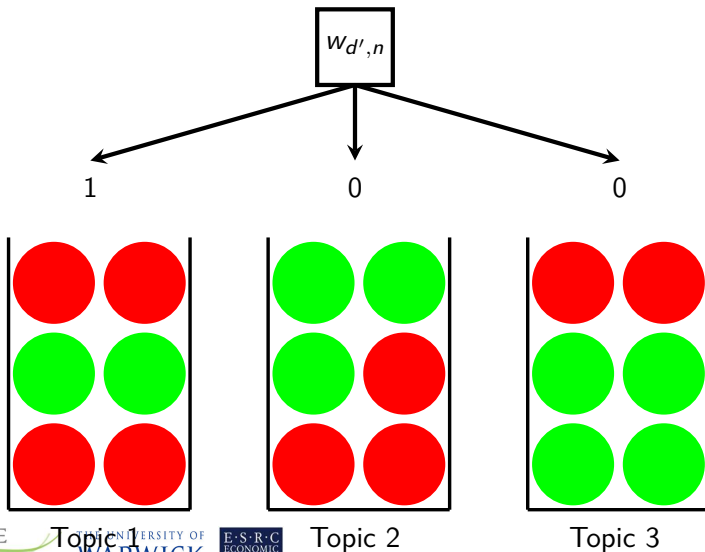
Each element word  $n$  in document  $d$  is then drawn from  $\beta_{z_d}$ .

This defines a multinomial mixture model.

# Mixture Model



# Mixture Model





# Mixed-Membership Model

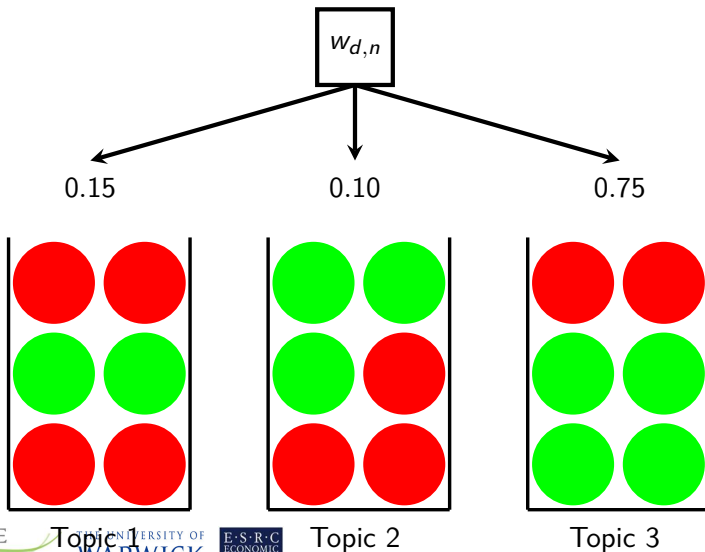
A feature of the mixture model is that every word in a document is forced to be drawn from the same distribution.

An alternative to this model is that documents can cover multiple topics (but may be inclined to cover some topics more than others).

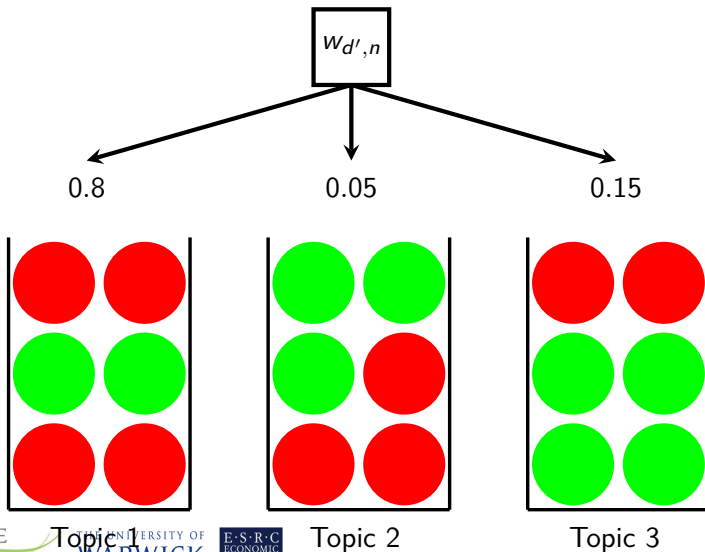
This is called a *mixed-membership model* because the same document can belong to multiple topics.

However, each word  $w_{d,n}$  in a document belongs to a single topic  $z_{d,n}$ .

# Mixed-Membership Model



# Mixed-Membership Model



# Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003) is a hugely influential mixed-membership topic model.

The data generating process of LDA is the following:

1. Draw  $\beta_k$  independently for  $k = 1, \dots, K$  from  $\text{Dirichlet}(\eta)$ . (Note that original model did not have Dirichlet prior).
2. Draw  $\theta_d$  independently for  $d = 1, \dots, D$  from  $\text{Dirichlet}(\alpha)$ .
3. Each word  $w_{d,n}$  in document  $d$  is generated from a two-step process:
  - 3.1 Draw topic assignment  $z_{d,n}$  from  $\theta_d$ .
  - 3.2 Draw  $w_{d,n}$  from  $\beta_{z_{d,n}}$ .