Big Data for Public Policy

Instructor: Elliott Ash

Introduction to Text Analysis

Checkpoints for a big data project

1. **What is the policy problem or research question?**

Checkpoints for a big data project

1. **What is the policy problem or research question?**
2. **Data:**
    - **obtain, clean, preprocess, and link.**
    - **Produce descriptive visuals and statistics on the text and metadata**
3. Machine learning / analysis
4. Solution / answers / system

Text as Data

Text as Data

- Text data is a sequence of characters called **documents**.
- The set of documents is the **corpus**.

Text as Data

- 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.

Text as Data

- 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.
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information.

Write it Down: How would you use text data for policy?

- Think of a policy problem – maybe something related to your job/research, or just something you are interested in. It needs to involve some text.
    - What is your corpus (text data source)?
    - What would you like to achieve using it?

Outline

1. Dictionary Methods
2. Featurizing Texts
3. Document Distance/Similarity
4. Topic Models
5. Machine Learning with Text
6. Chat GPT

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
- Corpus-specific: counting sets of words or phrases across documents
    - (e.g., number of times a judge says “justice” vs “efficiency”)

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says "justice" vs "efficiency")
- General dictionaries: WordNet, LIWC, MFD, etc.

Sentiment Analysis

Extract a “tone” dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., “good” versus “not good” versus “not very good”

Sentiment Analysis

Extract a “tone” dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., “good” versus “not good” versus “not very good”
- flair’s pre-trained sentiment model uses a context-sensitive neural net

Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"
- flair's pre-trained sentiment model uses a context-sensitive neural net

- Off-the-shelf scores are designed for online writing – may not work for legal text, for example.

Write it Down: Dictionaries

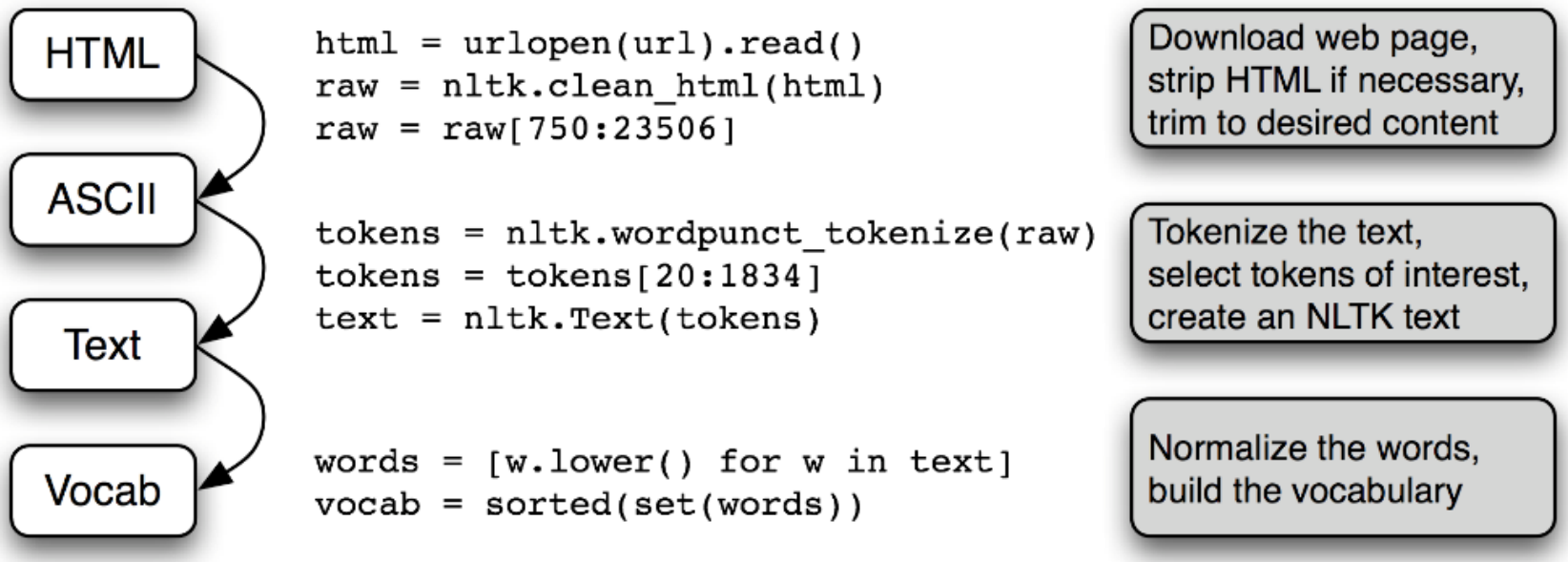
- How could you use a dictionary for your policy problem or project? Think of a list of terms and explain how they would capture the dimension/concept of interest.

Outline

1. Dictionary Methods
2. Featurizing Texts
3. Document Distance/Similarity
4. Topic Models
5. Machine Learning with Text
6. Chat GPT

Goals of Featurization

- The goal: produce features that are
  - **predictive** in the learning task
  - **interpretable** by human investigators
  - **tractable** enough to be easy to work with

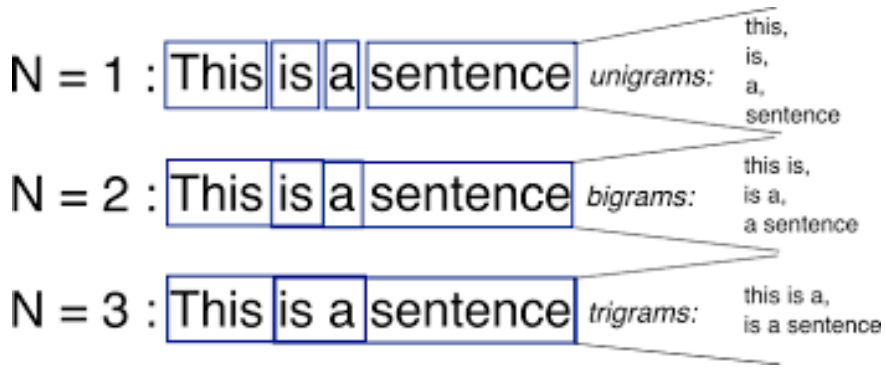


Basics

- Standard pre-processing steps:
  - drop capitalization, punctuation, numbers, stopwords (e.g. "the", "such")
  - remove word stems (e.g., "taxes" and "taxed" become "tax")
- In the "bag-of-words" representation, a document is represented as counts over words in the document.
  - term frequency would divide that by the total number of words in the document

N-grams

- N-grams are phrases, sequences of words up to length N.
  - bigrams, trigrams, quadgrams, etc.



- capture information and familiarity from local word order.
  - e.g. “estate tax” vs “death tax”

Parts of speech

- Parts of speech (POS) tags provide useful word categories corresponding to their functions in sentences:
  - **Content**: noun (NN), verb (VB), adjective (JJ), adverb (RB)
  - **Function**: determinant (DT), preposition (IN), conjunction (CC), pronoun (PR).

Parts of speech

- Parts of speech (POS) tags provide useful word categories corresponding to their functions in sentences:
  - **Content**: noun (NN), verb (VB), adjective (JJ), adverb (RB)
  - **Function**: determinant (DT), preposition (IN), conjunction (CC), pronoun (PR).
- Parts of speech vary in their informativeness for various functions:
  - For categorizing **topics**, nouns are usually most important
  - For **sentiment**, adjectives are usually most important.

Application: What Drives Media Slant?

Gentzkow and Shapiro (2010)

Application: What Drives Media Slant?

Gentzkow and Shapiro (2010)

- Corpora:
  - news text from large sample of US daily newspapers.
  - congressional text is 2005 Congressional Record.

Application: What Drives Media Slant?

Gentzkow and Shapiro (2010)

- Corpora:
  - news text from large sample of US daily newspapers.
  - congressional text is 2005 Congressional Record.
- Pre-process text, stripping away prepositions, conjunctions, pronouns, and common words
  - get bigrams and trigrams

Application: What Drives Media Slant?

Gentzkow and Shapiro (2010)

- Corpora:
  - news text from large sample of US daily newspapers.
  - congressional text is 2005 Congressional Record.
- Pre-process text, stripping away prepositions, conjunctions, pronouns, and common words
  - get bigrams and trigrams
- Identify polarizing phrases using χ^2 statistical test.
  - in sklearn, it is feature_selection.chi2

TABLE I

MOST PARTISAN PHRASES FROM THE 2005 *CONGRESSIONAL RECORD*[a]

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases		
private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife
Three-Word Phrases		
veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

TABLE I—*Continued*

Panel B: Phrases Used More Often by Republicans

Two-Word Phrases		
stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program
Three-Word Phrases		
embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

[a]The top 60 Democratic and Republican phrases, respectively, are shown ranked by χ^2_{pl}. The phrases are classified as two or three word after dropping common "stopwords" such as "for" and "the." See Section 3 for details and see Appendix B (online) for a more extensive phrase list.

Consumers drive media slant (GS 2010)

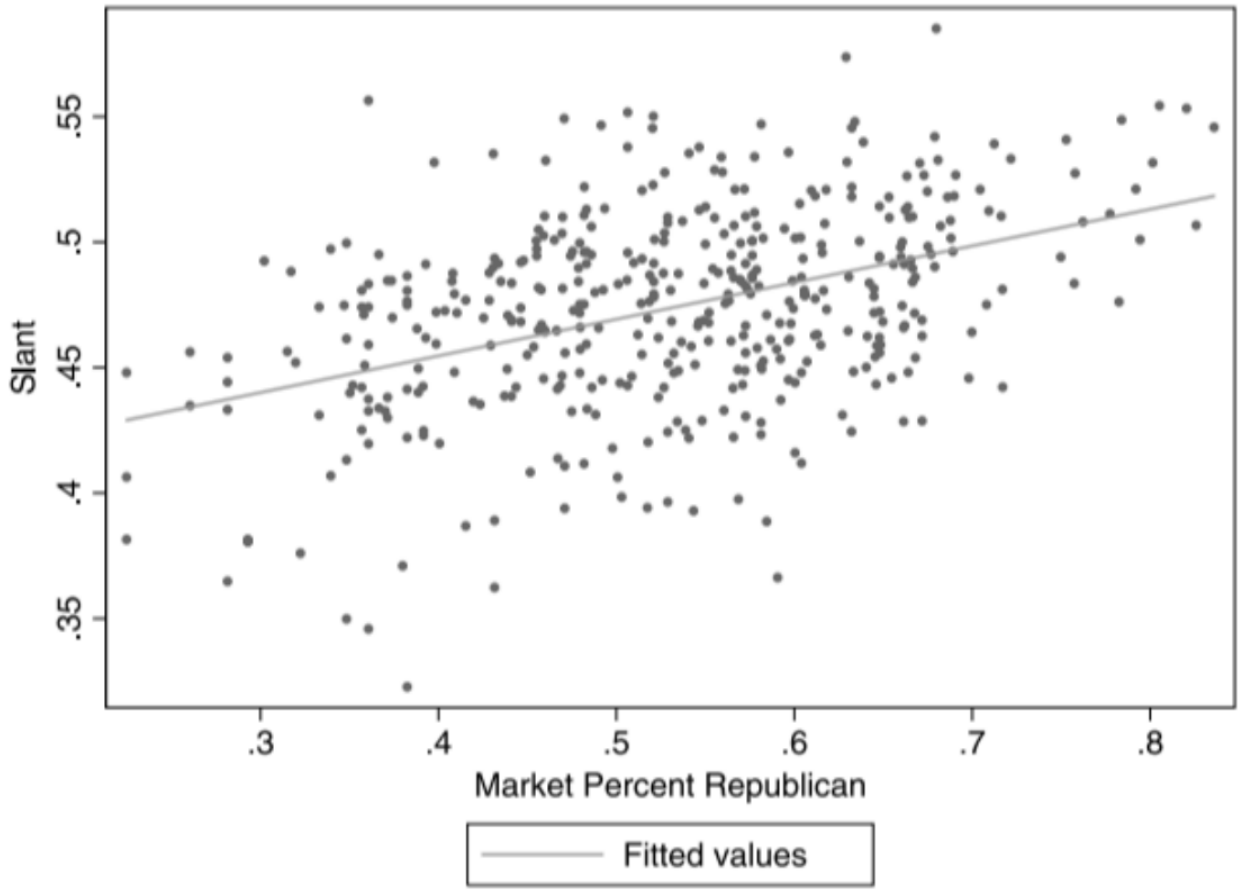


FIGURE 4.—Newspaper slant and consumer ideology. The newspaper slant index against Bush's share of the two-party vote in 2004 in the newspaper's market is shown.

Write it Down: Text Features

- What features are informative in your corpus, or about your problem? Single words? Phrases? Particular parts of speech (eg nouns)? Explain.

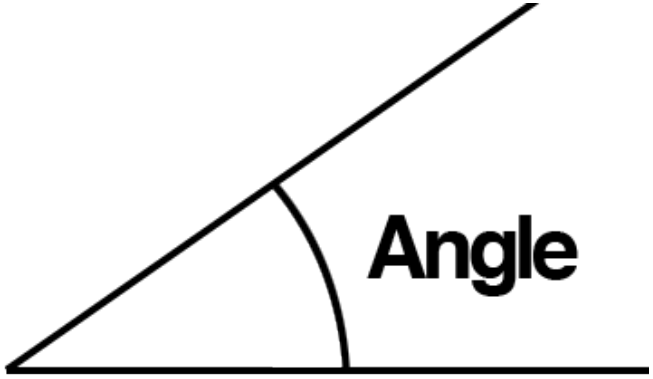
Outline

1. Dictionary Methods
2. Featurizing Texts
3. Document Distance/Similarity
4. Topic Models
5. Machine Learning with Text
6. Chat GPT

Text Re-Use

- Text Re-Use algorithms (like "Smith-Waterman") measure similarity by finding and counting shared sequences in two texts above some minimum length, e.g. 10 words.
    - useful for plagiarism detection, for example.
- precise but slow

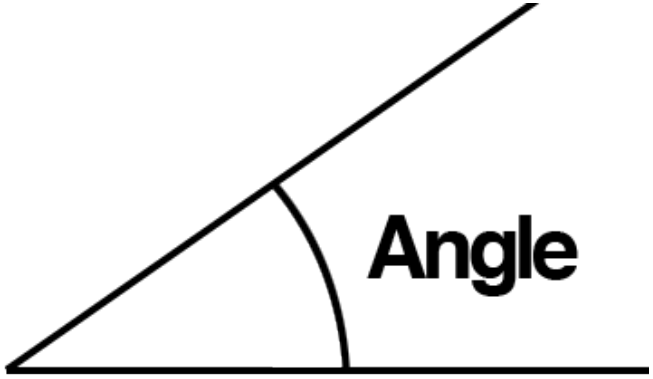
Cosine Similarity



$$\cos\_\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||}$$

where v_1 and v_2 are vectors, representing documents (e.g., IDF-weighted frequencies).

Cosine Similarity

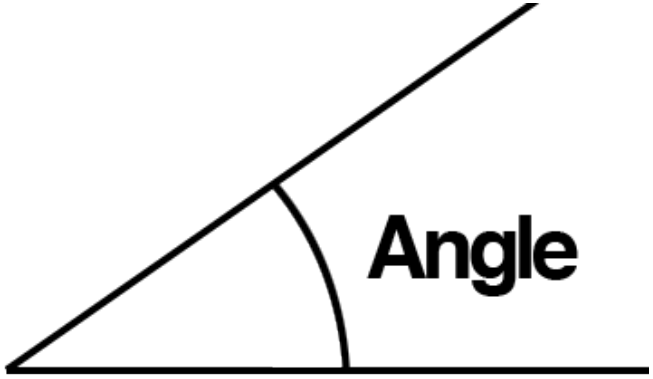


$$\cos\_\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||}$$

where v_1 and v_2 are vectors, representing documents (e.g., IDF-weighted frequencies).

- each document is a non-negative vector in an m-space (m = size of dictionary):
  - closer vectors form smaller angles: $\cos(0) = +1$ means identical documents.
  - furthest vectors are orthogonal: $\cos(\pi/2) = 0$ means no words in common.

Cosine Similarity



$$\cos\_\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||}$$

where v_1 and v_2 are vectors, representing documents (e.g., IDF-weighted frequencies).

- each document is a non-negative vector in an m-space (m = size of dictionary):
  - closer vectors form smaller angles: $\cos(0) = +1$ means identical documents.
  - furthest vectors are orthogonal: $\cos(\pi/2) = 0$ means no words in common.
- For n documents, this gives $n \times (n-1)$ similarities.

Text analysis of patent innovation

Kelly, Papanikolau, Seru, and Taddy (2018)

"Measuring technological innovation over the very long run"

- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - date, inventor, backward citations
  - text (abstract, claims, and description)

Text analysis of patent innovation

Kelly, Papanikolau, Seru, and Taddy (2018)

"Measuring technological innovation over the very long run"

- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - date, inventor, backward citations
  - text (abstract, claims, and description)
- Text pre-processing:
  - drop HTML markup, punctuation, numbers, capitalization, and stopwords.
  - remove terms that appear in less than 20 patents.
  - 1.6 million words in vocabulary.

Measuring Innovation

Kelly, Papanikolau, Seru, and Taddy (2018)

- For each patent i:
  - compute cosine similarity ρ_{ij} to all future patents j

Measuring Innovation

Kelly, Papanikolau, Seru, and Taddy (2018)

- For each patent i:
  - compute cosine similarity ρ_{ij} to all future patents j
- 9m$\times$9m similarity matrix = 30TB of data.
  - enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Novelty, Impact, and Quality

Kelly, Papanikolau, Seru, and Taddy (2018)

- "Novelty" is defined by dissimilarity (negative similarity) to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

Novelty, Impact, and Quality

Kelly, Papanikolau, Seru, and Taddy (2018)

- "Novelty" is defined by dissimilarity (negative similarity) to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).
- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{j \in F(i)} \rho_{ij}$$

  where $F(i)$ is the set of future patents (in, e.g., next 100 years).

Novelty, Impact, and Quality

Kelly, Papanikolau, Seru, and Taddy (2018)

- "Novelty" is defined by dissimilarity (negative similarity) to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).
- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{j \in F(i)} \rho_{ij}$$

  where $F(i)$ is the set of future patents (in, e.g., next 100 years).
- A patent has high **quality** if it is **novel** and **impactful**:

Validation

Kelly, Papanikolau, Seru, and Taddy (2018)

- For pairs with higher ρ_{ij}, patent j more likely to cite patent i.

Validation

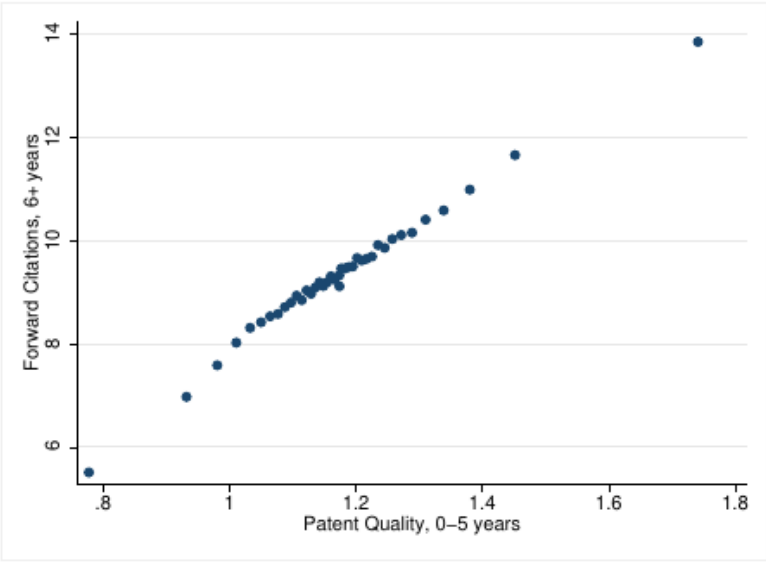
Kelly, Papanikolau, Seru, and Taddy (2018)

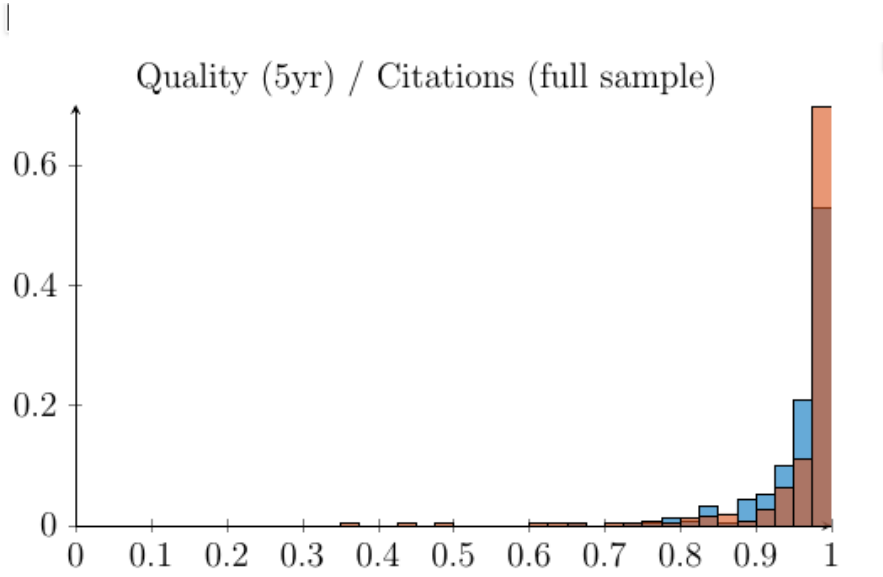
- For pairs with higher ρ_{ij}, patent j more likely to cite patent i.
- Within technology class (assigned by patent office), similarity is higher than across class.

Validation

Kelly, Papanikolau, Seru, and Taddy (2018)

- For pairs with higher ρ_{ij}, patent j more likely to cite patent i.
- Within technology class (assigned by patent office), similarity is higher than across class.
- Higher quality patents get more cites:





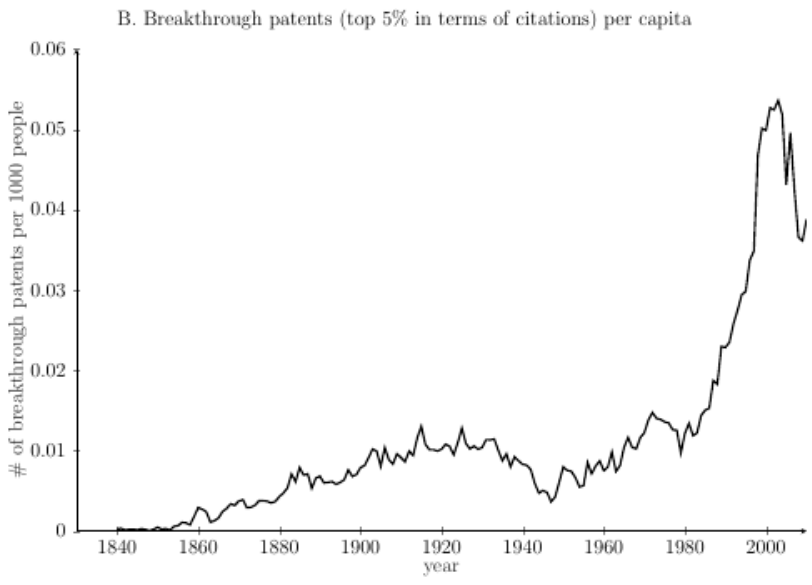
Most Innovative Firms

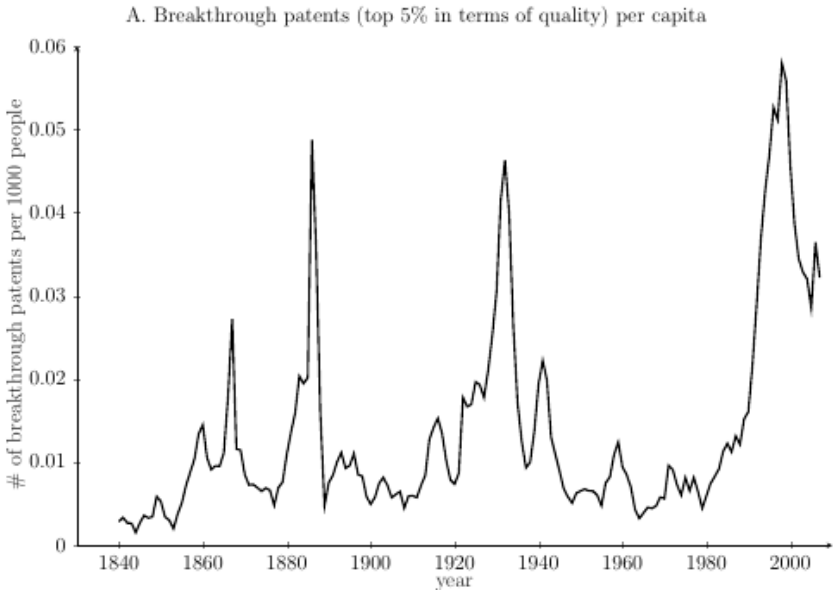
Kelly, Papanikolau, Seru, and Taddy (2018)

Assignee	First Year	# Breakthroughs
General Electric	1872	3,457
Westinghouse Electric Co.	1889	1,762
Eastman Kodak Co.	1890	2,244
Western Electric Co.	1899	1,222
AT&T (includes Bell Labs)	1899	5,645
Standard Oil Co.	1900	1,212
Dow Chemical Co.	1902	1,235
Du Pont	1905	3,353
International Business Machines	1908	14,913
American Cyanamid Co.	1909	690
Universal Oil Products Co.	1919	590
RCA	1920	3,222
Monsanto Company (inc. Monsanto Chemicals)	1921	902
Honeywell International, inc.	1928	872
General Aniline & Film Corp.	1929	1,181
Massachusetts Institute of Technology	1935	504
Philips	1939	1145
Texas Instruments	1960	2,088
Xerox	1961	2,198
Applied Materials	1971	510
Digital Equipment	1971	1,101
Hewlett-Packard Co.	1971	2,661
Intel	1971	2,629
Motorola, inc.	1971	4,129
Regents of the University of California	1971	823
United States Navy	1945	791
NCR	1973	737
Advanced Micro Devices	1974	1,195
Apple Computer	1978	864

Breakthrough patents: citations vs quality

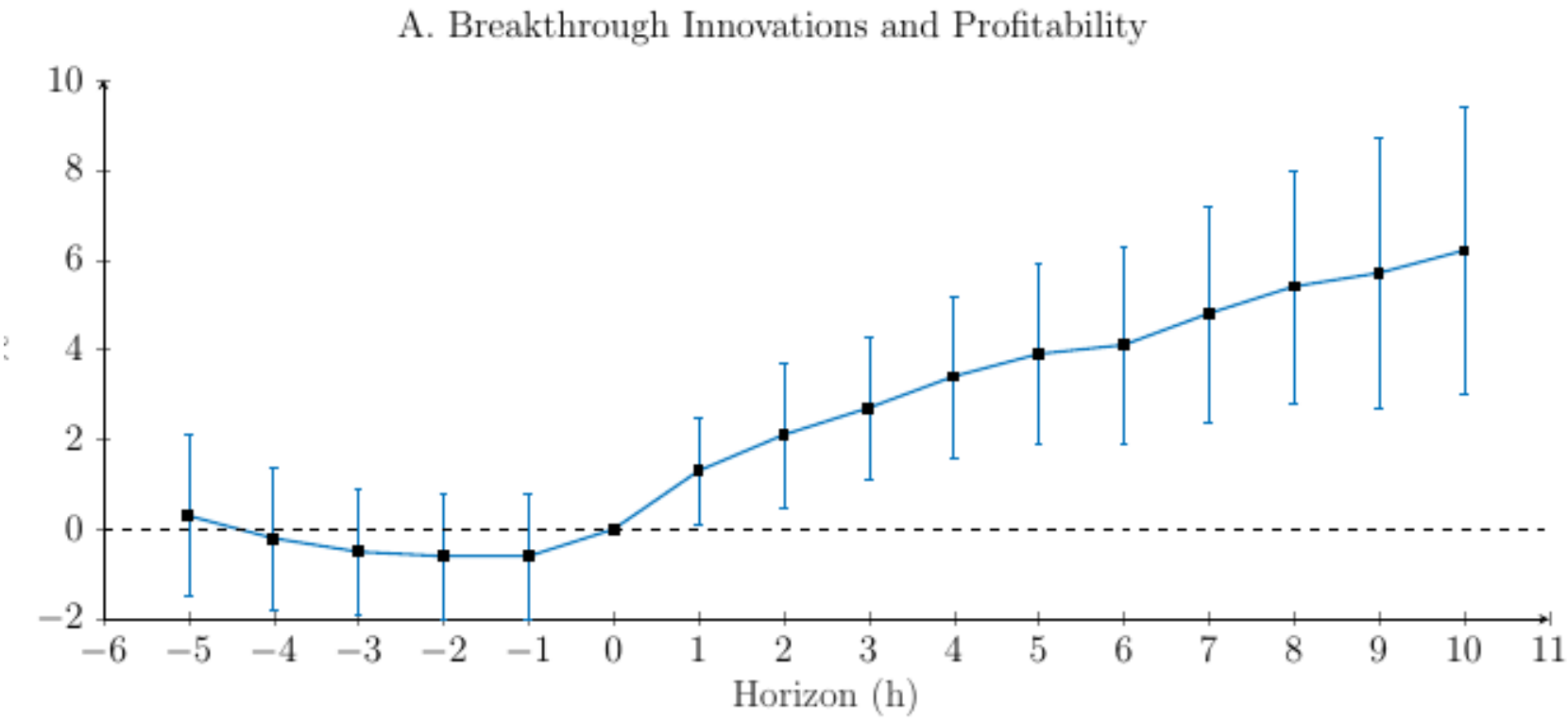
Kelly, Papanikolau, Seru, and Taddy (2018)





Breakthrough patents and firm profits

Kelly, Papanikolau, Seru, and Taddy (2018)



Write it Down – Document Distance

- How could you use document comparisons (distance) to analyze your corpus or solve your problem?

Outline

1. Dictionary Methods
2. Featurizing Texts
3. Document Distance/Similarity
4. Topic Models
5. Machine Learning with Text
6. Chat GPT

Topic Models in Social Science

- Idea: documents exhibit each topic in some proportion.
  - Each **document** is a distribution over **topics**.
  - Each **topic** is a distribution over **words**.

Topic Models in Social Science

- Idea: documents exhibit each topic in some proportion.
  - Each **document** is a distribution over **topics**.
  - Each **topic** is a distribution over **words**.
- Latent Dirichlet Allocation (e.g. Blei 2012) is the most popular topic model because it is easy to use and (usually) provides great results.

Topic Models in Social Science

- Idea: documents exhibit each topic in some proportion.
  - Each **document** is a distribution over **topics**.
  - Each **topic** is a distribution over **words**.
- Latent Dirichlet Allocation (e.g. Blei 2012) is the most popular topic model because it is easy to use and (usually) provides great results.
- Social scientists use topics as a form of measurement
  - how observed covariates drive trends in language
  - **topic models are more interpretable** than other methods, e.g. principal components analysis.
    - can tell a story not just about what, but how and why

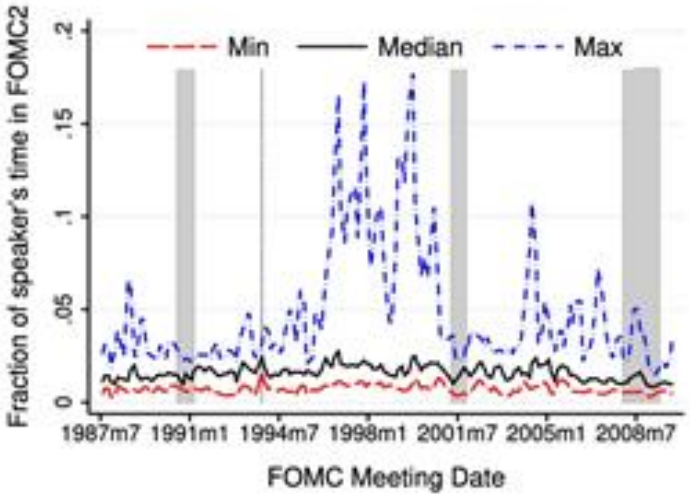
Topic modeling Federal Reserve Bank transcripts

Hansen, McMahon, and Prat (QJE 2017)

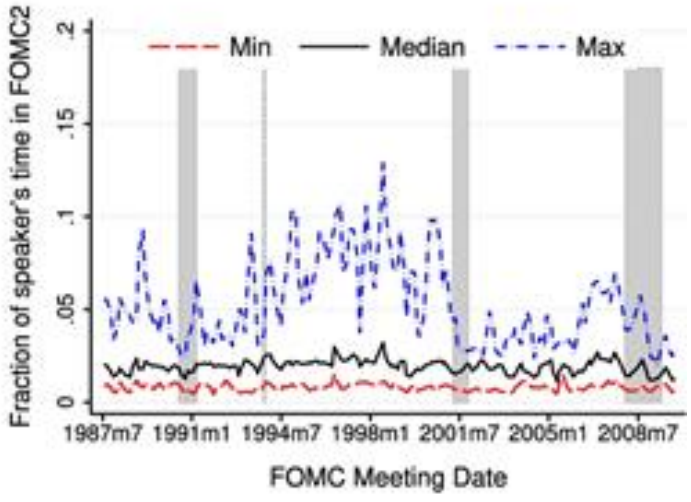
- Use LDA to analyze speech at the FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - transcripts: 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- $K = 40$ topics selected for interpretability / topic coherence.

Pro-Cyclical Topics

Hansen, McMahon, and Prat (QJE 2017)



(A) Topic 0 'Productivity'



(B) Topic 1 'Growth'

Counter-Cyclical Topics

Hansen, McMahon, and Prat (QJE 2017)

(A) Topic 38 'Financial sector'

(B) Topic 39 'Economic weakness'

Write it Down – Topic Models

- How would you use a topic model to provide some information about your corpus and support your work?

Outline

1. Dictionary Methods
2. Featurizing Texts
3. Document Distance/Similarity
4. Topic Models
5. Machine Learning with Text
6. Chat GPT

Machine Learning with Text Data

- Say each document has an associated label y – e.g. legal topic.
- We want a **text classifier** – identify the annotated legal topic from the text features.

Machine Learning with Text Data

- Say each document has an associated label y – e.g. legal topic.
- We want a **text classifier** – identify the annotated legal topic from the text features.

- This is machine learning – tools for predicting or relating text to metadata.
    - ML with text data is the same as with non-text data – the topic of the next lecture segment.

Write it Down – ML with Text

- What metadata do you have associated with your corpus? Which of these would be useful to predict or else relate to your text data?

Outline

1. Dictionary Methods
2. Featurizing Texts
3. Document Distance/Similarity
4. Topic Models
5. Machine Learning with Text
6. Chat GPT

ChatGPT API

https://platform.openai.com/playground

https://colab.research.google.com/drive/1O14qeQc_pJGDv5csYmpcNr51KZLTzhNl?usp=sharing

- How could you use an AI assistant like Chat GPT to solve your problem?

```
# system prompt
'''
You are a star student in ETH Zurich's Big Data for Public Policy Course.  Answer the following
questions to get the best possible participation grade.

'''
# user prompt
'''

1.  Think of a policy problem maybe something related to your job/research, or just something
you are interested in.  It needs to involve some text.  What is your corpus (text data source)?
What would you like to achieve using it?
2.  How could you use a dictionary for your policy problem or project?  Think of a list of terms
and explain how they would capture the dimension/concept of interest.
3.  What features are informative in your corpus, or about your problem?  Single words?
Phrases?  Particular parts of speech (eg nouns)?  Explain.
4.  How could you use document comparisons (distance) to analyze your corpus or solve your
problem?
5.  How would you use a topic model to provide some information about your corpus and support
your work?
6.  What metadata do you have associated with your corpus?  Which of these would be useful to
predict or else relate to your text data?
7.  How could you use an AI assistant like Chat GPT to solve your problem?
'''
```