Machine Learning for Textual and Unstructured Data

Lecture 1: Introduction and Bag-of-Words Model

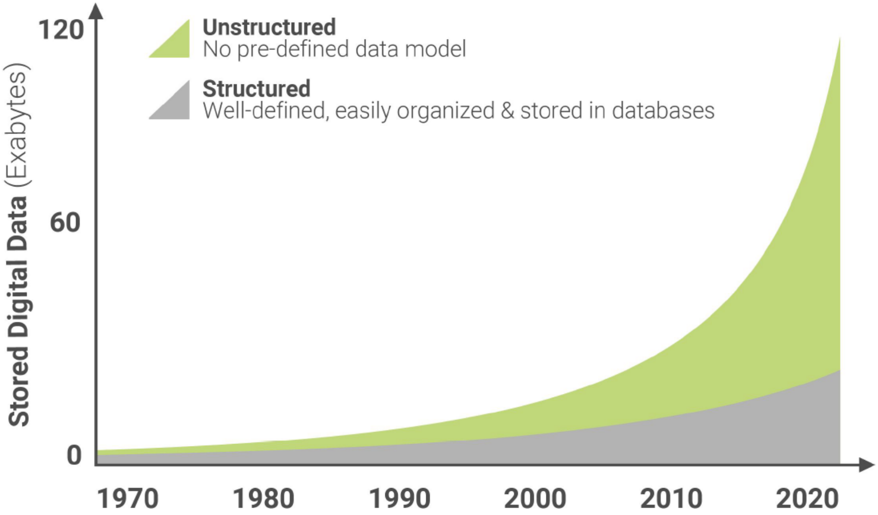
Stephen Hansen
University College London

FUNDACIÓN RAMÓN ARECES

IESE Business School University of Navarra

Center for International Finance

Trends in Data Types



Unstructured Data

Unstructured data does not come organized in a traditional relational database.

Extracting relevant information and separating it from irrelevant information is a primary challenge.

Examples:

1. Text
2. Audio
3. Images
4. Videos

Happenstance Data[1]

Traditional economic data is constructed with a particular measurement in mind, e.g. GDP statistics.

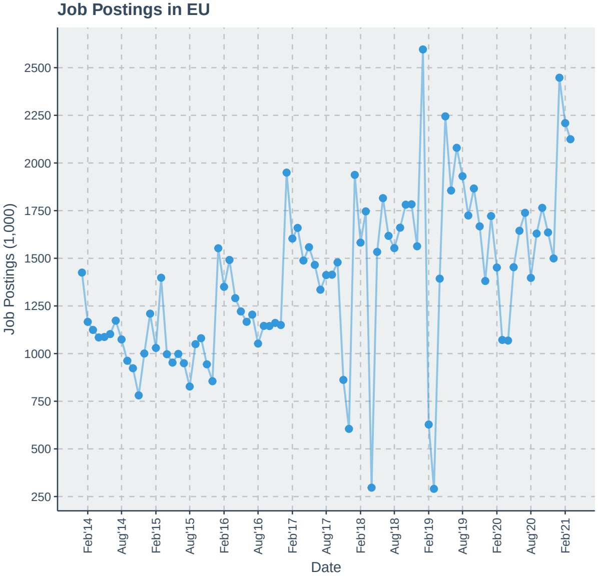
Most data generated in the private sector is happenstance, and arises via the everyday activities of agents ("digital exhaust").

Statistical challenge is that data is not collected with a consistent, representative sample frame.

Organizational challenge is that data access arrangements have yet to be normalized.

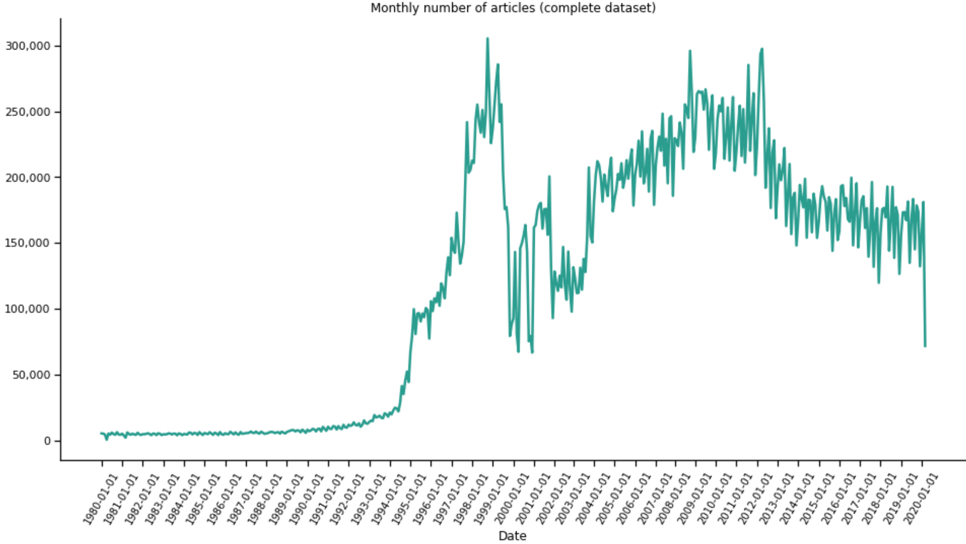
[1] Discussed more fully in `https://rs-delve.github.io/reports/2020/11/24/data-readiness-lessons-from-an-emergency.html`

Monthly Online Job Postings



Source: Web−scraped Job Ads from EU27 Countries, provided by Burning Glass Technologies.

Monthly Newswire Postings



Unstructured vs Happenstance Data

	Administrative	Happenstance
Structured	Traditional Economic Data	Credit Card Transactions Amazon product ratings
Unstructured	10-K Filings FOMC press conferences	Tweets Online Job Postings

What is the Value of Unstructured Data?

The main application of unstructured data in economics and related disciplines has been to measure important phenomena.

Can complement existing measures: e.g. build more granular versions of official data.

Or create entirely new measures: economic policy uncertainty, media slant, central bank communication.

Makes information retrieval methods useful in a wide variety of fields.

This Course

1. Bag-of-words model
2. Factor models for discrete data (aka topic models)
3. Word embeddings
4. Sequence embeddings and large language models
5. Image data (time permitting)

What is Text?

At an abstract level, text is simply a string of characters.

Some of these may be from the Latin alphabet—'a', 'A', 'p' and so on—but there may also be:

1. Decorated Latin letters (e.g. ö)
2. Non-Latin alphabetic characters (e.g. Chinese and Arabic)
3. Punctuation (e.g. '!')
4. White spaces, tabs, newlines
5. Numbers
6. Non-alphanumeric characters (e.g. '@')

Key Question: How can we obtain an informative, quantitative representation of these character strings?

First step is to pre-process strings to convert them into lists of units of meaning, sometimes called tokens.

We delay discussion of pre-processing steps for the first practical class. See also [Denny and Spirling, 2018].

Notation

The corpus is composed of D documents indexed by d.

After pre-processing, each document is a finite, length-N_d list of terms $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,N_d})$ with generic element $w_{d,n}$.

Let $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$ be a list of all terms in the corpus, and let $N \equiv \sum_d N_d$ be the total number of terms in the corpus.

Suppose there are V **unique** terms in $\mathbf{w}$, where $1 \leq V \leq N$, each indexed by v.

We can then map each term in the corpus into this index, so that $w_{d,n} \in \{1, \ldots, V\}$.

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

Example

Consider three documents:

1. 'stephen is nice'
2. 'john is also nice'
3. 'george is mean'

We can consider the set of unique terms as $\{\text{stephen}, \text{is}, \text{nice}, \text{john}, \text{also}, \text{george}, \text{mean}\}$ so that $V = 7$.

Construct the following index:

stephen	is	nice	john	also	george	mean
1	2	3	4	5	6	7

We then have $\mathbf{w}_1 = (1, 2, 3)$; $\mathbf{w}_2 = (4, 2, 5, 3)$; $\mathbf{w}_3 = (6, 2, 7)$.

Moreover $x_{1,1} = 1$, $x_{2,1} = 0$, $x_{3,1} = 0$, etc.

Bag-of-Words Model

Document-Term Matrix

A popular quantitative representation of text is the *document-term matrix* $\mathbf{X}$, which collects the counts $x_{d,v}$ into a $D \times V$ matrix.

In the previous example, we have

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Real-World Example

In “Transparency and Deliberation” we use a corpus of verbatim 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 (46,502).
- Associated metadata: speaker biographical information, macroeconomic conditions, etc.

Executive Time Use Project

Data on each 15-minute block of time for one week of 1,114 CEOs' time classified according to

1. type (e.g. meeting, public event, etc.)
2. duration (15m, 30m, etc.)
3. planning (planned or unplanned)
4. number of participants (one, more than one)
5. functions of participants, divided between employees of the firms or "insiders" (finance, marketing, etc.) and "outsiders" (clients, banks, etc.).

There are 4,253 unique combinations of these five features in the data.

One can summarize the data with a 1114×4253 matrix where the (i, j)th element is the number of 15-minute time blocks that CEO i spends in activities with a particular combination of features j.

Other Examples

Network data can be represented by an adjacency matrix which is typically high dimensional, sparse, and discrete.

Bag-of-visual words model in image processing.

Four Measurement Problems

[Ash and Hansen, 2023] organize measurement problems associated with text into four categories:

1. Distance between documents, e.g. how similar are corporate filings from each other.
2. Whether a concept is present (and degree of presence) in a document, e.g. sentiment.
3. How concepts relate in a document, e.g. sentiment and individual companies.
4. Associating documents to metadata, e.g. mapping newspaper text into recession vs expansion periods.

Document Similarity

Documents as Vectors

We can view the documents that make up the rows of $\mathbf{X}$ as vectors.

Let each vocabulary term v have its own vector $\mathbf{e}_v \in \mathbb{R}^V$ where

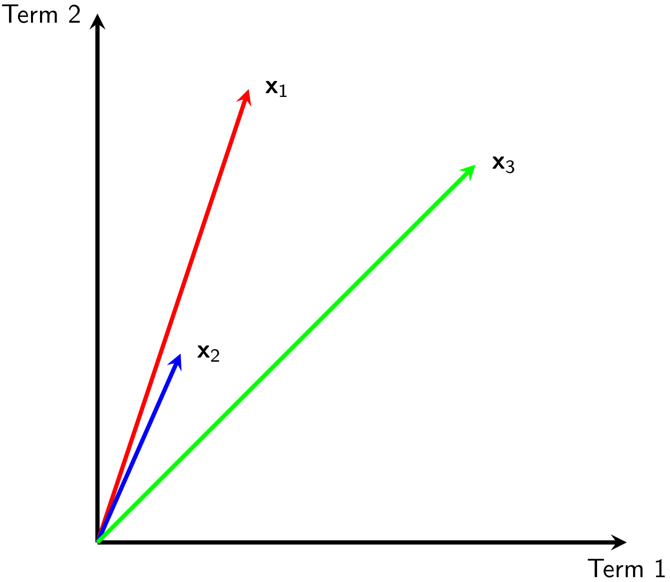
$$e_{v,v'} = \begin{cases} 1 & \text{if } v = v' \\ 0 & \text{otherwise} \end{cases}$$

Note that each term's vector is orthogonal to every other term's vector.

We can express document d as

$$\mathbf{x}_d = x_{d,1}\mathbf{e}_1 + x_{d,2}\mathbf{e}_2 + \ldots + x_{d,V}\mathbf{e}_V$$

Three Documents



Distance in the Vector Space

An initial question of interest is how similar are any two documents in the vector space.

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

What is the problem with Euclidean distance? How can we correct this?

Cosine Similarity

Define the cosine similarity between documents i and j as

$$CS(i, j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}$$

1. Since document vectors have no negative elements $CS(i, j) \in [0, 1]$.
2. $\mathbf{x}_i / \|\mathbf{x}_i\|$ 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] and [Hoberg and Phillips, 2016] 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/`.

Term Weighting

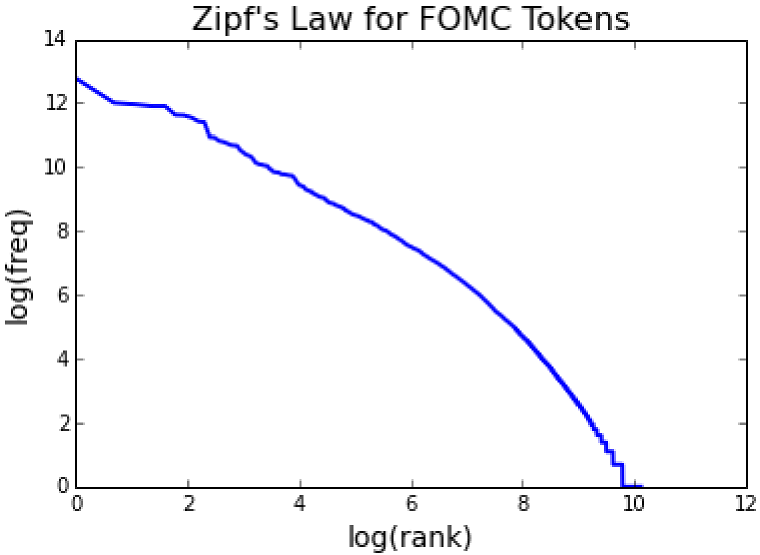
The frequency of words in natural language can distort raw counts.

Zipf's Law is an empirical regularity for natural language: 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.

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

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

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

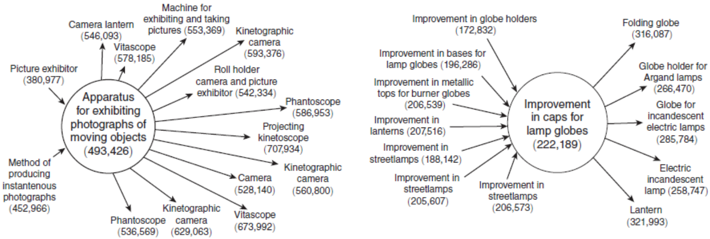
Application

[Kelly et al., 2021] uses the text of US patents to identify radical innovation.

An individual patent is said to be influential when its backward similarity is low and its forward similarity is high.

Measure validated with historically important patents, forward citations, market value.

Similarity Networks



Concept Detection

Dictionary Methods

The most common strategy for concept detection is to define a list of terms that capture the concept of interest, and to express documents as counts over those terms.

Strategy is referred to as *dictionary methods*.

Where do the dictionaries come from?

1. Pre-defined lists
2. Domain expertise
3. Ability to predict objective label

Measuring Sentiment

[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.

Measuring Sentiment

[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.

Dictionaries Using Domain Expertise

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.

Dictionaries Using Domain Expertise

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.

Economics Application

The Economic Policy Uncertainty (EPU) index of [Baker et al., 2016] (`http://www.policyuncertainty.com/`) is based on keyword search applied to newspaper articles from major US and European newspapers.

Search logic is the result of extensive manual audits of newspaper articles.

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

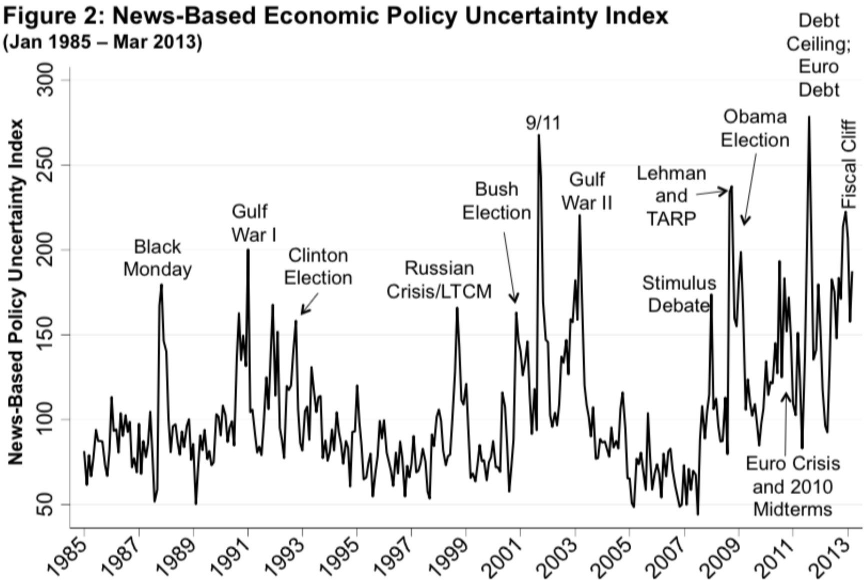


Figure 2: News-Based Economic Policy Uncertainty Index
(Jan 1985 – Mar 2013)
News-Based Policy Uncertainty Index
300
250
200
150
100
50
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Black Monday
Gulf War I
Clinton Election
Russian Crisis/LTCM
Bush Election
9/11
Gulf War II
Lehman and TARP
Stimulus Debate
Obama Election
Euro Crisis and 2010 Midterms
Debt Ceiling; Euro Debt
Fiscal Cliff

Relationship among Concepts

Combining Dictionaries

One common strategy to measure how concepts relate to each other is to:

1. Define separate dictionaries for each concept.
2. Count the instances in which terms from each dictionary co-occur in some local window.

Firm-Level Political Risk

[Hassan et al., 2019] measures firm-level political risk from quarterly earnings calls made by firms traded on US stock markets.

Transcripts from 175,797 conference calls made by 9,478 firms between 2002 and 2016 (downloaded from Thomson Reuters' StreetEvents).

BBD uncertainty measures aggregate risk arising from policymaking, but not firm-specific risks.

Uses a risk/uncertainty dictionary, but the method for associating these to political vs. non-political risks is novel.

Define corpora of canonical political language $\mathbb{P}$ and non-political language $\mathbb{N}$, and compute all bigrams from each.

Sources for these training libraries are undergraduate textbooks or, alternatively, newspaper articles.

Political Risk Measure

$$PRisk_{it} = \frac{\sum_{b}^{B_{it}} \left(1[b \in \mathbb{P} \backslash \mathbb{N}] \times 1[|b-r| < 10] \times \frac{f_{b,\mathbb{P}}}{B_{\mathbb{P}}} \right)}{B_{it}}$$

B_{it} is the total number of bigrams for firm i at time t.

b is an individual bigram.

r is the position of the nearest synonym of risk or uncertainty.

$f_{b,\mathbb{P}}$ is the count of bigram b in the political corpus.

$B_{\mathbb{P}}$ is the total number of bigrams in the political corpus.

Results

Firms with higher levels of political risk have higher volatility in their stock prices.

Firms with higher political risk engage more in lobbying.

Sector membership and time explain little variation in firm-level risk.

Main conclusion is that location in cross-section of risk exposures seems to matter for firms at least as much as time-series variation.

Relating Text to Metadata

Text Regression

Suppose that the text has associated metadata $\mathbf{y}_d$, which might contain speaker ID, timestamp, or any other numeric covariate.

Associating text with metadata involves associating $\mathbf{x}_d$ and $\mathbf{y}_d$.

Most straightforward approach would regress $y_{d,j}$ on $\mathbf{x}_d$ and $\mathbf{y}_{d,-j}$.

Due to strong dependence structure in $\mathbf{x}_d$, strong case for use of non-linear models.

Generative vs Discriminative Models

A generative model estimates the full joint distribution $p(\, y_d , \mathbf{x}_d \,)$ whereas typical regression estimates discriminative model $p\,(\, y_d \mid \mathbf{x}_d \,)$.

[Efron, 1975] shows that discriminative classifiers obtain a lower asymptotic error than generative ones.

Two motivations for nevertheless studying generative models:

1. [Ng and Jordan, 2001] show that generative classifiers can approach their (higher) asymptotic error faster.
2. They can reveal interesting structure, e.g. $p\,(\, \mathbf{x}_d \mid y_d \,)$.

A generative model requires a probability model for $\mathbf{x}_d$.

One example is a Naive Bayes Classifier.

Inverse Regression

Inverse regression models specify a model for $p(\mathbf{x_d} \mid y_d)$.

Well-known example is [Gentzkow and Shapiro, 2010].

Drawing on this paper as motivation, [Taddy, 2013] and [Taddy, 2015] propose fully generative models for inverse regression.

[Gentzkow et al., 2019] uses these models to study political polarization.

Multinomial Inverse Regression

Model takes the form

$$\mathbf{x}_d \sim \mathrm{MN}(\mathbf{q}_d, N_d) \text{ where } q_{d,v} = \frac{\exp(a_v + \mathbf{y}_d^T \mathbf{b}_v)}{\sum_v \exp(a_v + \mathbf{y}_d^T \mathbf{b}_v)}.$$

Generalized linear model with a (multinomial) logistic link function.

MLE estimates of multinomial regression coefficients can be approximated by estimating V separate Poisson regression models of $x_{d,v}$ on $\mathbf{y}_d$.

LASSO prior used to regularize regression parameters.

Application to Congressional Speech

[Gentzkow et al., 2019] use MNIR to model speech data from the *US Congressional Record* from 1873-2016.

Select speeches by Democrats/Republicans (7,732 speakers). Total 36,161 unique speaker-session.

Count two-word phrases (bigrams): 508,351 phrases with count $\geq$ 10 in at least one session.

$\mathbf{y}_d$ includes party, state, chamber, gender.

Democratic Phrases

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

Republican Phrases

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

Polarization

Let $q^D_{t,v}(\mathbf{y}')$ be the probability that a Democrat at time t with observables $\mathbf{y}'$ speaks phrase v. Similarly define $q^R_{t,v}(\mathbf{y}')$.

Given phrase v, posterior probability of the speaker being a Democrat is (assuming uniform prior)

$$\rho_{t,v}(\mathbf{y}') = \frac{q^D_{t,v}(\mathbf{y}')}{q^D_{t,v}(\mathbf{y}') + q^R_{t,v}(\mathbf{y}')}$$

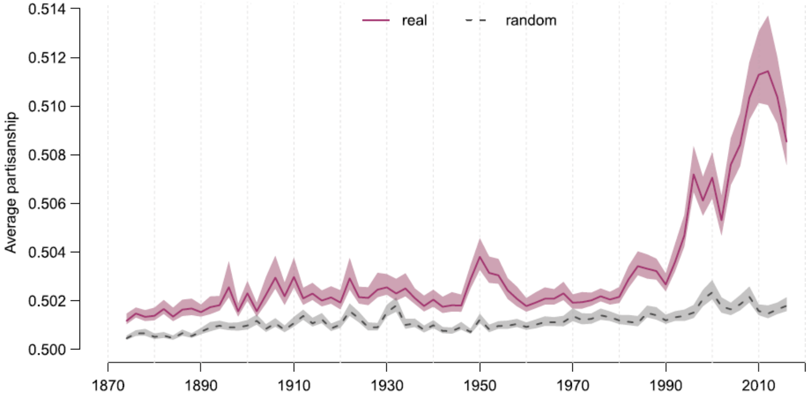
Partisanship is the expected posterior after hearing a single phrase by a speaker with characteristics $\mathbf{y}'$:

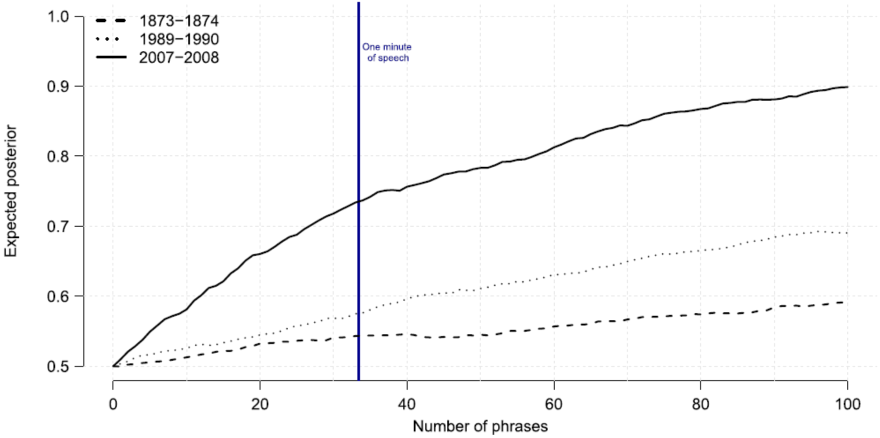
$$\pi_t(\mathbf{y}') = \frac{1}{2}\mathbf{q}^D_t(\mathbf{y}') \cdot \boldsymbol{\rho}_t(\mathbf{y}') + \frac{1}{2}\mathbf{q}^R_t(\mathbf{y}') \cdot (1 - \boldsymbol{\rho}_t(\mathbf{y}'))$$

Let s_t be total speakers in session t. Average partisanship is

$$\overline{\pi}_t = \frac{1}{s_t}\sum_{i=1}^{s_t} \pi_{it}(\mathbf{y}'_{it})$$

Panel B: Partisanship from Preferred Penalized Estimator ($\hat{\pi}_t^*$)




1.0
0.9
0.8
0.7
0.6
0.5
1873−1874
1989−1990
2007−2008
One minute
of speech
Expected posterior
0
20
40
60
80
100
Number of phrases

Sufficient Reduction Projection

There remains the issues of how to use the estimated model for classification.

Let $z_{d,j} = \mathbf{f}_d^T \hat{\mathbf{b}}_j$ be the *sufficient reduction projection* for the jth covariate for document d, where $\mathbf{f}_d = \mathbf{x}_d / N_d$ is a vector of term frequencies.

$z_{d,j}$ is sufficient for predicting $y_{d,j}$ in the sense that

$$y_{d,j} \perp \mathbf{x}_d, N_d \mid z_{d,j}, \mathbf{y}_{d,-j}.$$

All the information contained in the high-dimensional frequency counts relevant for predicting $y_{d,j}$ can be summarized in the SR projection.

Dimensionality reduction targeted at specific covariate.

Classification

For classification, use the SR projections to build a forward regression that models $y_{d,j}$ as some function of $z_{d,j}, \mathbf{y}_{d,-j}$: OLS; logistic; with or without non-linear terms in $z_{d,j}$, etc.

To predict $y_{d',j}$ for an out-of-sample document d':

1. Form $z_{d',j}$ given the estimated $\hat{\mathbf{b}}_j$ coefficients in the training data.
2. Use the estimated forward regression to generate a predicted value for $y_{d',j}$.

Conclusion

The document-term matrix can be used to address each of the four measurement problems relevant for text-as-data in economics and finance.

Term-count analysis has been, and will continue to be, very influential.

Strength is that matrix-structured data is relatively familiar to economists, and analysis is relatively straightforward.

Nevertheless, all sequential information is ignored and much of natural language's meaning depends on context.

References I

Ash, E. and Hansen, S. (2023).
Text Algorithms in Economics.
Annual Review of Economics, forthcoming.

Baker, S. R., Bloom, N., and Davis, S. J. (2016).
Measuring Economic Policy Uncertainty.
The Quarterly Journal of Economics, 131(4):1593–1636.

Denny, M. J. and Spirling, A. (2018).
Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.
Political Analysis, 26(2):168–189.

Efron, B. (1975).
The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis.
Journal of the American Statistical Association, 70(352):892–898.

Gentzkow, M. and Shapiro, J. M. (2010).
What Drives Media Slant? Evidence From U.S. Daily Newspapers.
Econometrica, 78(1):35–71.

References II

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019).
Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech.
Econometrica, 87(4):1307–1340.

Hassan, T. A., Hollander, S., van Lent, L., and Tahoun, A. (2019).
Firm-Level Political Risk: Measurement and Effects.
The Quarterly Journal of Economics, 134(4):2135–2202.

Hoberg, G. and Phillips, G. (2010).
Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis.
The Review of Financial Studies, 23(10):3773–3811.

Hoberg, G. and Phillips, G. (2016).
Text-Based Network Industries and Endogenous Product Differentiation.
Journal of Political Economy, 124(5):1423–1465.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021).
Measuring Technological Innovation over the Long Run.
American Economic Review: Insights, 3(3):303–320.

References III

Loughran, T. and Mcdonald, B. (2011).
When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.
The Journal of Finance, 66(1):35–65.

Ng, A. Y. and Jordan, M. I. (2001).
On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes.
In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01, pages 841–848, Cambridge, MA, USA. MIT Press.

Taddy, M. (2013).
Multinomial Inverse Regression for Text Analysis.
Journal of the American Statistical Association, 108(503):755–770.

Taddy, M. (2015).
Distributed Multinomial Regression.
The Annals of Applied Statistics, 9(3):1394–1414.

Tetlock, P. C. (2007).
Giving Content to Investor Sentiment: The Role of Media in the Stock Market.
The Journal of Finance, 62(3):1139–1168.