Supervised Machine Learning III: Unstructured and Unsupervised ML

Paul Goldsmith-Pinkham

May 3, 2022

Unstructured Data and Unsupervised Learning

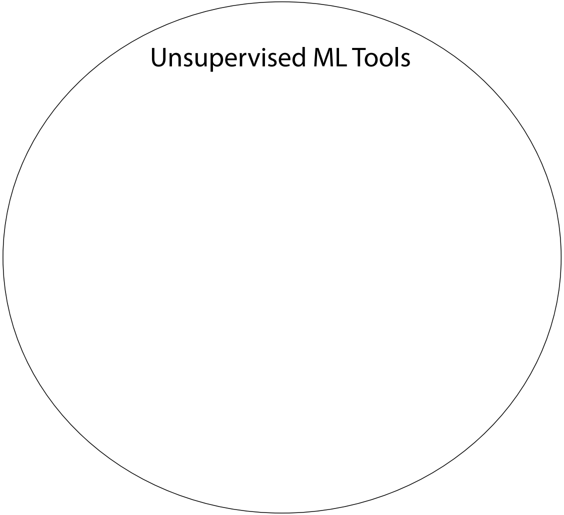
- We have, so far, discussed ML in the context of two assumptions:
  1. Our data was a relatively well-defined (*feature* matrix X and outcome y)
  2. We had an outcome (y) to go with each set of predictors

- Today we'll talk about settings where these are relaxed

- Data that is
  - unstructured (e.g. some data $\mathcal{X}$ that we need to turn into a matrix X)
  - unlabeled (no outcome y defined)

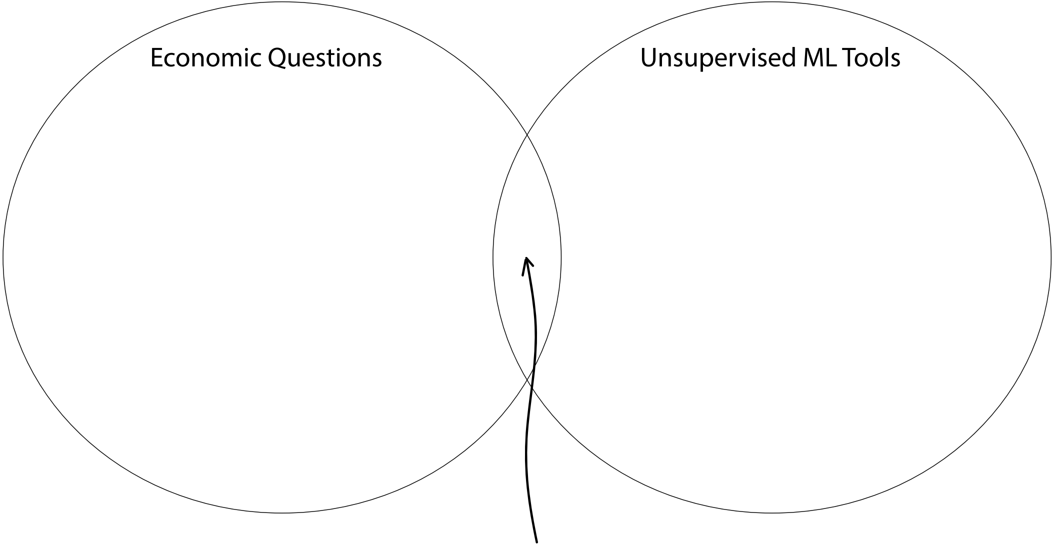
- The key challenge with this literature is keeping eye on the prize:
  - Our goal is to answer *economic* questions

Huge space of tools



Unsupervised ML Tools

Huge space of tools



Economic Questions

Unsupervised ML Tools

How big is this intersection?

Some high level notation

- Consider a data object $\mathcal{X}$ which is complex and challenging to describe
    - A set of firms or products with various characteristics
    - The collection of news articles over time
    - Evaluations of banks' health
    - A set of congressional speeches
    - Etc.

- First step in the process is a mapping, $\psi(\mathcal{X}) \rightarrow X$
    - This typically involves some sort of quantification
    - This also include the construction or addition of a label, y that goes along with the data
        - This will give the data a supervised ML structure
    - This object will likely be very high dimensional! (e.g. $dim(X_i) >$ observations)

- Next step in the process: constructing economic masures or features from X
    - Calculating "interesting" subdimensions of X (summarization)
    - Projecting labels y onto dimensions of X
    - Projecting units into new dimensions based on X (e.g. relative distance metrics)

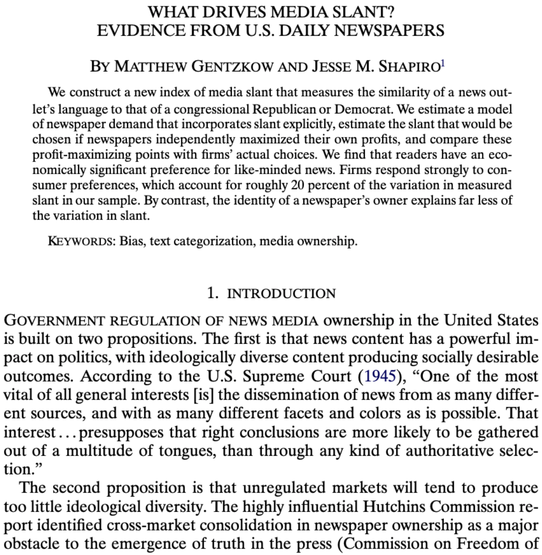
- Will provide examples for each case...

Today’s Class

- A overview of two different examples / applications where unusual unstructured data was used
- An brief dive into one particular unsupervised ML technique, Latent Dirichlet Allocation (LDA)
    - Commonly used in text data (things with counts)
- Goal: highlight that these techniques can be very powerful at unlocking new measures
    - But they require *extremely* judicious selection of applications / approaches
- What I want you to avoid is a common situation (that I have been in):
    - “Amazing data in search of a question” (a real quote from one of my advisors)

Example 1: Gentzkow and Shapiro (2010)

- How do we evaluate the “slant” of a newspaper?
    - Subjective: go through and label yourself (or get others)
- $\mathcal{X}$ is the “newspaper” and “politics”
    - X is now two sets of data:
    - X_1 - text from newspapers
    - X_2 - text from congressional speakers
    - Y_2 - *labels* of political party
- How are X_1 and X_2 constructed?



WHAT DRIVES MEDIA SLANT?
EVIDENCE FROM U.S. DAILY NEWSPAPERS

BY MATTHEW GENTZKOW AND JESSE M. SHAPIRO[1]

We construct a new index of media slant that measures the similarity of a news outlet’s language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms’ actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to consumer preferences, which account for roughly 20 percent of the variation in measured slant in our sample. By contrast, the identity of a newspaper’s owner explains far less of the variation in slant.

KEYWORDS: Bias, text categorization, media ownership.

1. INTRODUCTION

GOVERNMENT REGULATION OF NEWS MEDIA ownership in the United States is built on two propositions. The first is that news content has a powerful impact on politics, with ideologically diverse content producing socially desirable outcomes. According to the U.S. Supreme Court (1945), “One of the most vital of all general interests [is] the dissemination of news from as many different sources, and with as many different facets and colors as is possible. That interest . . . presupposes that right conclusions are more likely to be gathered out of a multitude of tongues, than through any kind of authoritative selection.”

The second proposition is that unregulated markets will tend to produce too little ideological diversity. The highly influential Hutchins Commission report identified cross-market consolidation in newspaper ownership as a major obstacle to the emergence of truth in the press (Commission on Freedom of

Aside on quantifying text data

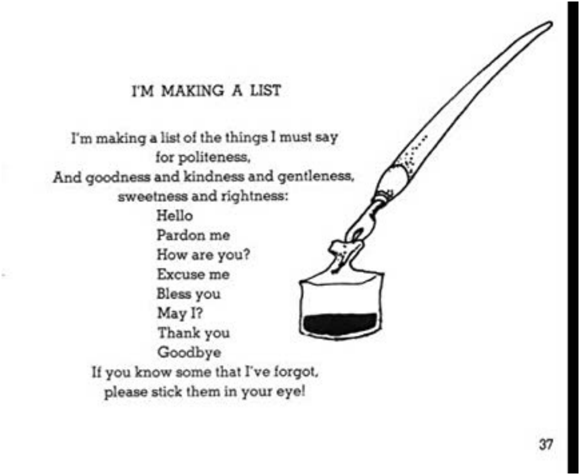
- Given a *corpus* of text, this unstructured data can be made structured in a number of ways
  - Corpus: a collection of written texts

- Simplest: bag of single words
  - E.g. a sentence is converted into counts
  - “the branch of knowledge concerned with the production, consumption, and transfer of wealth.”
  - becomes $[2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1]$ for [”of”,”the”,”and”,”branch”,”concerned”,”consumption” ”knowledge”,”production”, ”transfer”,”wealth”,”with”]
  - We would also have a lot of zeros for all the words we don’t use!
    - *Sparse* matrices

- Can consider bigrams, trigrams, etc.
  - The issue is that dimensionality blows up
  - Why would be do bigrams? More specific meaning

- Note that there is tremendous resolution to the data that is lost by doing this! We lose the structure of the data, etc.

Losing information with bag of words



I'M MAKING A LIST

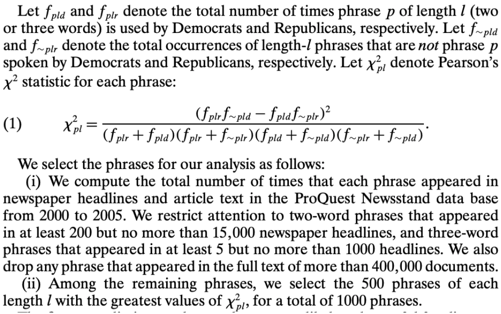
I'm making a list of the things I must say
for politeness,
And goodness and kindness and gentleness,
sweetness and rightness:
Hello
Pardon me
How are you?
Excuse me
Bless you
May I?
Thank you
Goodbye
If you know some that I've forgot,
please stick them in your eye!

37

Example 1: Gentzkow and Shapiro (2010)

- How are X_1 and X_2 constructed?
    - G&S focus on highly split phrases (bigrams and trigrams) in X_2
    - The focus is then on this set of words in X_1 and X_2
    - Note that y_2 is not used to sign things!

- Then, a *supervised* measure is used to construct a mapping: $y_2 = f(X_2)$ and then applied to X_1 to construct $\hat{y}_1$



Let f_{pld} and f_{plr} denote the total number of times phrase p of length l (two or three words) is used by Democrats and Republicans, respectively. Let $f_{\sim pld}$ and $f_{\sim plr}$ denote the total occurrences of length-l phrases that are *not* phrase p spoken by Democrats and Republicans, respectively. Let χ^2_{pl} denote Pearson's χ^2 statistic for each phrase:

$$(1) \qquad \chi^2_{pl} = \frac{(f_{plr}f_{\sim pld} - f_{pld}f_{\sim plr})^2}{(f_{plr}+f_{pld})(f_{plr}+f_{\sim plr})(f_{pld}+f_{\sim pld})(f_{\sim plr}+f_{\sim pld})}.$$

We select the phrases for our analysis as follows:

(i) We compute the total number of times that each phrase appeared in newspaper headlines and article text in the ProQuest Newsstand data base from 2000 to 2005. We restrict attention to two-word phrases that appeared in at least 200 but no more than 15,000 newspaper headlines, and three-word phrases that appeared in at least 5 but no more than 1000 headlines. We also drop any phrase that appeared in the full text of more than 400,000 documents.

(ii) Among the remaining phrases, we select the 500 phrases of each length l with the greatest values of χ^2_{pl}, for a total of 1000 phrases.

Example 1: Gentzkow and Shapiro (2010)

- How are X_1 and X_2 constructed?
    - G&S focus on highly split phrases (bigrams and trigrams) in X_2
    - The focus is then on this set of words in X_1 and X_2
    - Note that y_2 is not used to sign things!

- Then, a *supervised* measure is used to construct a mapping: $y_2 = f(X_2)$ and then applied to X_1 to construct $\hat{y}_1$

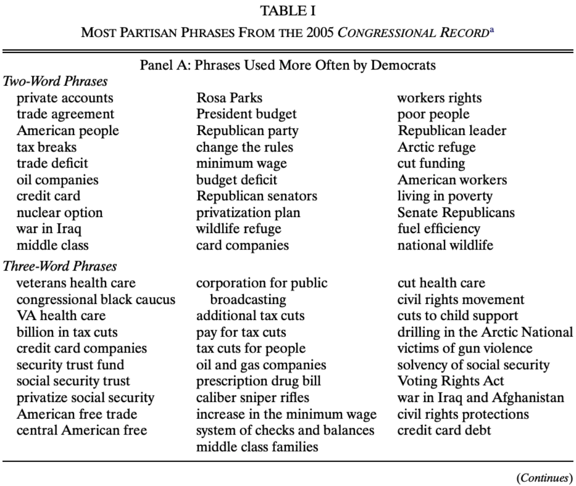


TABLE I

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

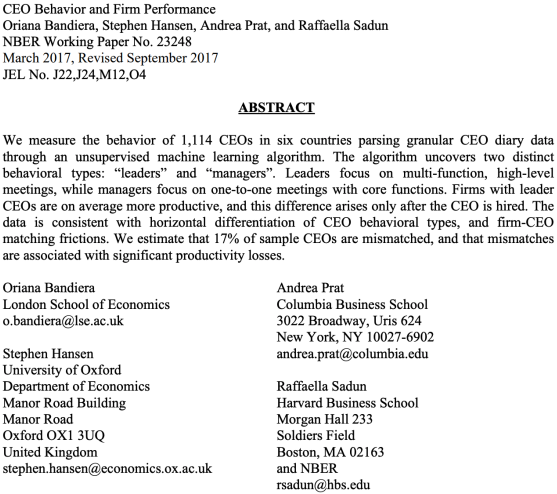
Panel A: Phrases Used More Often by Democrats

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

(*Continues*)

Example 2: Bandiera et al. (2017)

- In this example, $\mathcal{X}$ are CEO behavior at firms
  - Data is converted to X using diaries of activity which are "coded" using surveys
  - "Data on 42,233 activities of different duration, equivalent to 225,721 15-minute blocks, 90% of which cover work activities"
- This high dimensional object is then converted into a lower dimensional θ, which is then correlated with firm outcomes
  - The move to θ is doing dimension reduction!
  - So how do they do it? LDA



CEO Behavior and Firm Performance
Oriana Bandiera, Stephen Hansen, Andrea Prat, and Raffaella Sadun
NBER Working Paper No. 23248
March 2017, Revised September 2017
JEL No. J22,J24,M12,O4

ABSTRACT

We measure the behavior of 1,114 CEOs in six countries parsing granular CEO diary data through an unsupervised machine learning algorithm. The algorithm uncovers two distinct behavioral types: "leaders" and "managers". Leaders focus on multi-function, high-level meetings, while managers focus on one-to-one meetings with core functions. Firms with leader CEOs are on average more productive, and this difference arises only after the CEO is hired. The data is consistent with horizontal differentiation of CEO behavioral types, and firm-CEO matching frictions. We estimate that 17% of sample CEOs are mismatched, and that mismatches are associated with significant productivity losses.

Oriana Bandiera
London School of Economics
o.bandiera@lse.ac.uk

Stephen Hansen
University of Oxford
Department of Economics
Manor Road Building
Manor Road
Oxford OX1 3UQ
United Kingdom
stephen.hansen@economics.ox.ac.uk

Andrea Prat
Columbia Business School
3022 Broadway, Uris 624
New York, NY 10027-6902
andrea.prat@columbia.edu

Raffaella Sadun
Harvard Business School
Morgan Hall 233
Soldiers Field
Boston, MA 02163
and NBER
rsadun@hbs.edu

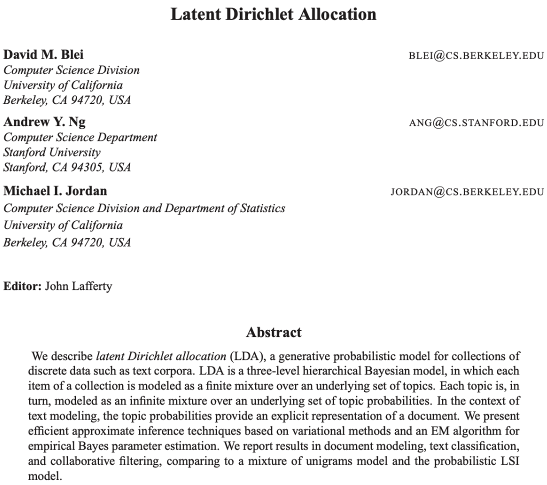
What is LDA? Latent Dirichlet Allocation

- Originally described by Blei, Ng and Jordan (2003), LDA is a generative model of how a matrix of count variables, X, of dimension $n \times p$ is made

- p is the number of potential words (or bigrams), n is the number of documents (e.g. CEO surveys)

- LDA is, in essence, a structured mixture model
  - Uses a Hierarchical Bayesian structure (recall our lecture!)
  - The structure provides a way to inform structure by shrinking across

- Assume an "unobserved" dimensionality



Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

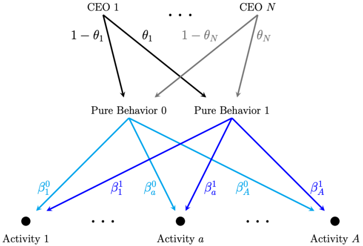
We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

What is LDA? Latent Dirichlet Allocation

- Simple example from Bandiera et al.: there are two types (e.g. unobserved dimension of 2)
    - All CEOs are drawn from one of two types

- Consequentially, LDA model will estimate:
    - For a given CEO, what is the probability that they are type 1 or type 2 (0 or 1)
    - For each type, what is the relative distribution of each activity

Figure 3: Data Generating Process for Activities with Two Pure Behaviors



Notes: This figure provides a graphical representation of the data-generating process for the time-use data. First, CEO i chooses – independently for each individual unit of his time – one of the two pure behaviors according to a Bernoulli distribution with parameter θ_i. The observed activity for a unit of time is then drawn from the distribution over activities that the pure behavior defines.

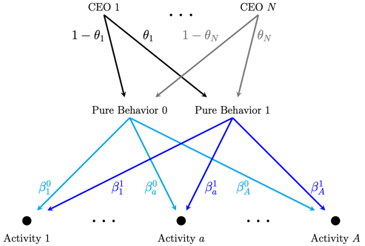
What is LDA? Latent Dirichlet Allocation

- Output of this model gives a number of pieces: for each CEO, we have an measure of how much they are each type

- For each activity, we know how much they reflect each “type”

- For Bandiera et al., they use the type measure (θ), as an index

Figure 3: Data Generating Process for Activities with Two Pure Behaviors



Notes: This figure provides a graphical representation of the data-generating process for the time-use data. First, CEO i chooses – independently for each individual unit of his time – one of the two pure behaviors according to a Bernoulli distribution with parameter θ_i. The observed activity for a unit of time is then drawn from the distribution over activities that the pure behavior defines.

The issues or challenges with LDA

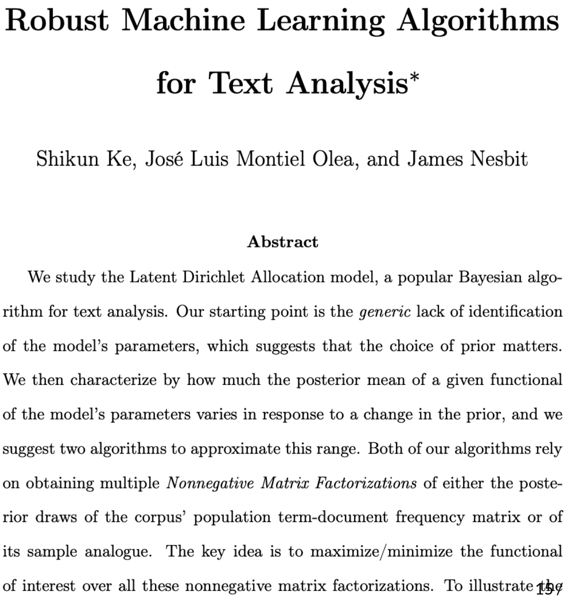
- What do the types even mean?
  - E.g. what is type 1? What is type 2?

- Why is 2 the right number?
  - Consider the analogy to Principal Component Analysis
  - Dimension choice can be done using maximum Bayes Factor (see Bybee et al. (2020))

- There are a number of ways to diagnose the types:
  - Correlate them with some other label from outside the data
  - Subjectively label them by examining the β frequencies for each document
    - E.g. if one type puts a lot on one type of activity, you could construct a name for it
    - This is just correlating using the human mind

The issues or challenges with LDA

- This model is Bayesian, and uses priors to initialize the model
- It turns out that the parameters of the model are unidentified, generically
    - The joint probability of the corpus from model is given by $P = B\Theta$, where B is the matrix of β ($p \times K$), and Θ is $K \times n$
    - Concretely, imagine $p = 1, andK = 2$
- The priors are necessary for estimation!
    - As a result, choice of prior can move your results
    - Many empiricists might feel uncomfortable with this



Robust Machine Learning Algorithms for Text Analysis*

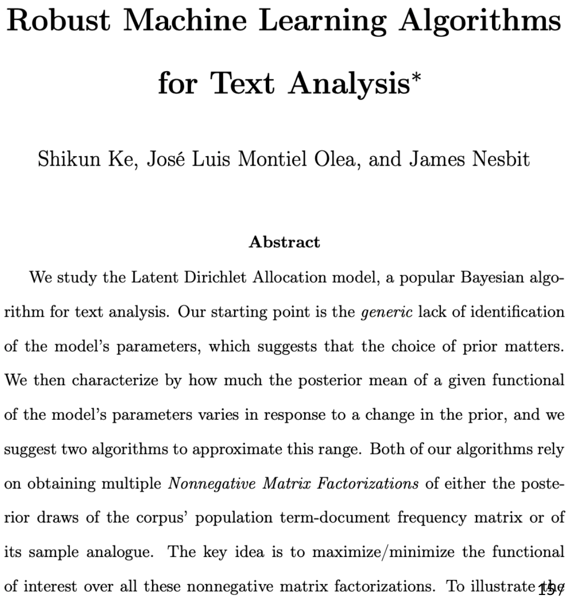
Shikun Ke, José Luis Montiel Olea, and James Nesbit

Abstract

We study the Latent Dirichlet Allocation model, a popular Bayesian algorithm for text analysis. Our starting point is the *generic* lack of identification of the model's parameters, which suggests that the choice of prior matters. We then characterize by how much the posterior mean of a given functional of the model's parameters varies in response to a change in the prior, and we suggest two algorithms to approximate this range. Both of our algorithms rely on obtaining multiple *Nonnegative Matrix Factorizations* of either the posterior draws of the corpus' population term-document frequency matrix or of its sample analogue. The key idea is to maximize/minimize the functional of interest over all these nonnegative matrix factorizations. To illustrate the

The issues or challenges with LDA

- This model is Bayesian, and uses priors to initialize the model
- It turns out that the parameters of the model are unidentified, generically
  - The joint probability of the corpus from model is given by $P = B\Theta$, where B is the matrix of β ($p \times K$), and Θ is $K \times n$
  - Concretely, imagine $p = 1, andK = 2$
- The priors are necessary for estimation!
  - As a result, choice of prior can move your results
  - Many empiricists might feel uncomfortable with this



Robust Machine Learning Algorithms for Text Analysis*

Shikun Ke, José Luis Montiel Olea, and James Nesbit

Abstract

We study the Latent Dirichlet Allocation model, a popular Bayesian algorithm for text analysis. Our starting point is the *generic* lack of identification of the model's parameters, which suggests that the choice of prior matters. We then characterize by how much the posterior mean of a given functional of the model's parameters varies in response to a change in the prior, and we suggest two algorithms to approximate this range. Both of our algorithms rely on obtaining multiple *Nonnegative Matrix Factorizations* of either the posterior draws of the corpus' population term-document frequency matrix or of its sample analogue. The key idea is to maximize/minimize the functional of interest over all these nonnegative matrix factorizations. To illustrate the

Example 3: TFIDF + Cosine Similarity

- Define a concept called TFIDF: term-frequency inverse document frequency

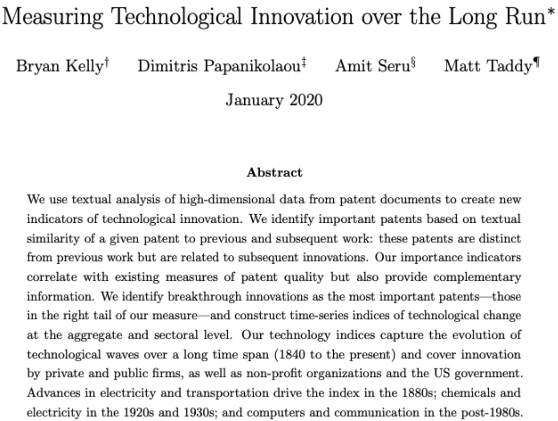
$$TF_{pw} = \frac{c_{pw}}{\sum_k c_{pk}} \tag{1}$$

which is the frequency that word w shows up in document p relative to the other words.

- Define IDF_w as

$$IDF_w = \log\left(\frac{d}{d\text{with word } w}\right) \tag{2}$$

- Then $TFIDF_{pw}$ is the product of those two.



Measuring Technological Innovation over the Long Run[*]

Bryan Kelly[†] Dimitris Papanikolaou[‡] Amit Seru[§] Matt Taddy[¶]

January 2020

Abstract

We use textual analysis of high-dimensional data from patent documents to create new indicators of technological innovation. We identify important patents based on textual similarity of a given patent to previous and subsequent work: these patents are distinct from previous work but are related to subsequent innovations. Our importance indicators correlate with existing measures of patent quality but also provide complementary information. We identify breakthrough innovations as the most important patents—those in the right tail of our measure—and construct time-series indices of technological change at the aggregate and sectoral level. Our technology indices capture the evolution of technological waves over a long time span (1840 to the present) and cover innovation by private and public firms, as well as non-profit organizations and the US government. Advances in electricity and transportation drive the index in the 1880s; chemicals and electricity in the 1920s and 1930s; and computers and communication in the post-1980s.

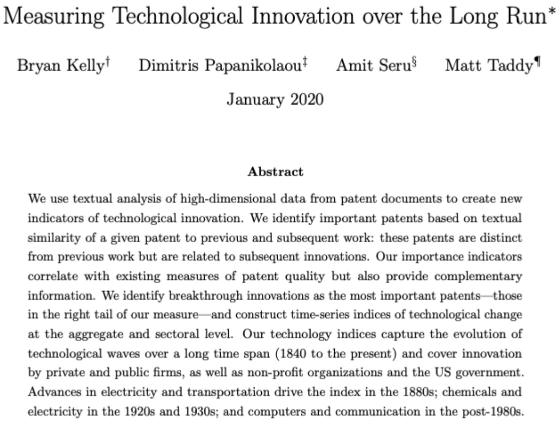
Example 3: TFIDF + Cosine Similarity

- This paper constructs BIDF, which is a backwards looking version of IDF:

$$IDF_{wp} = \log\left(\frac{\text{patents priors to p}}{\text{1+ patents prior to p that i}}\right. \tag{3}$$

- Finally, they look at the cosine distance between these TFBIDF for a given patent.

- They can identify "new" patents using this!



Measuring Technological Innovation over the Long Run*

Bryan Kelly[†] Dimitris Papanikolaou[‡] Amit Seru[§] Matt Taddy[¶]

January 2020

Abstract

We use textual analysis of high-dimensional data from patent documents to create new indicators of technological innovation. We identify important patents based on textual similarity of a given patent to previous and subsequent work: these patents are distinct from previous work but are related to subsequent innovations. Our importance indicators correlate with existing measures of patent quality but also provide complementary information. We identify breakthrough innovations as the most important patents—those in the right tail of our measure—and construct time-series indices of technological change at the aggregate and sectoral level. Our technology indices capture the evolution of technological waves over a long time span (1840 to the present) and cover innovation by private and public firms, as well as non-profit organizations and the US government. Advances in electricity and transportation drive the index in the 1880s; chemicals and electricity in the 1920s and 1930s; and computers and communication in the post-1980s.

My Main takeaway

- This is a really powerful way to take new data and apply to problems
- However, really easy to parse and summarize data without a good economic question in mind
  - Still need exogeneous variation and an economic question!
- Without a research design in mind, it becomes very hard to describe “why” you’re doing something.
  - Personal expereince with my own work
- Thoughts?