Text as Data

Amy Catalinac

New York University

June 18, 2018

Why a Talk on Text Analysis?

Text is the new frontier of . . .

1. **Data**: lots of it (literally petabytes) on the web. . . not to mention archives.
2. **Quantitative methods**: for the most part, it is unstructured. Needs to be harvested and modeled.
3. **Social science**: politicians give speeches, thinkers write articles, nations sign treaties, users post on Facebook, etc.

Today: overview of quantitative 'text-as-data' approaches as strategies to learn more about social scientific phenomena of interest.

Goal of Text Analysis

- In many (most?) applications of text as data in social science, we are trying to make an inference about a **latent variable**.
  - → something which we cannot observe directly but which we can make inferences about from things we can observe.
  - → Examples include ideology, priorities, ambition, narcissism, propensity to vote, etc.
- In **traditional** social science research, we use data that is structured.
  - → For example, surveys *we wrote* to understand ideology, donation decisions to understand which party a voter supports, or roll-call votes to understand where a politician stands on a policy.
- Here, the thing we **can** observe are the words spoken, the passages written, the issues debated, etc.

And . . .

- The latent variable of interest may pertain to the. . .

author ‘what does this Senator prioritize?’, ‘where is this party in ideological space?’

document ‘does this treaty represent a fair deal for American Indians?’, ‘how did the discussion of lasers change over time?’

both ‘how does the way Japanese politicians talk about national security change in response to electoral reform?’

We need to think carefully about . . .

1. The appropriate **population** and **sample**.
   - → "I have so many texts!", but must remain cognizant of **selection biases** in our sample that influence our ability to draw valid inferences about the population.
   - → must incorporate **uncertainty** around any estimates of the latent variable derived from a sample.
2. The **features** we care about.
   - → how to **get at them** with unstructured text, how to **represent** them numerically, and how to **describe** them.
3. How to **model** our data (the texts with their relevant features measured/coded).
   - → entails **statistical** decisions
4. What we can infer about our **latent variable** of interest.
   - → **compare** our estimates across people, **validate** them

In general, we will . . .

Get Texts

```
An expert hospital
consultant has written
to my hon. Friend…
```

```
Order. The Minister
must be allowed to
reply without
interruption.
```

```
I am grateful to my
hon. Friend for her
question. I pay tribute
to her work with the
International Myeloma
Foundation…
```

```
My constituent, Brian
Jago, was fortunate
enough to receive a
course of Velcade, as a
result of which he does
not have to…
```

→ Make Document Term Matrix

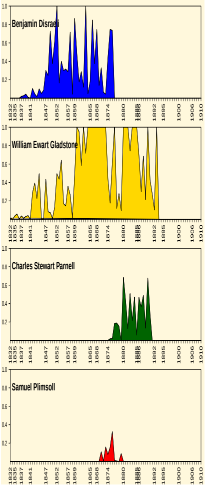
$$\begin{array}{r|cccc} & a & an & \dots & ze \\ MP_{001} & 2 & 0 & \dots & 1 \\ MP_{002} & 0 & 3 & \dots & 0 \\ MP_i & \vdots & \vdots & \dots & \vdots \\ MP_{654} & \vdots & \vdots & \ddots & \vdots \\ MP_{655} & 0 & 0 & \dots & 2 \end{array}$$

→ Model

- (dis)similarity
- diversity
- readability
- scale
- classify
- topic model
- burstiness
- sentiment

. . .

→ Make inferences



I. From Text to (Numeric) Data

Reducing Complexity

Language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, in transforming text to (numeric) data, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

→ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes, and the fit of our models.

NB. Inevitably, the degree to which one simplifies is dependent on the particular task at hand.

→ there is no 'one best way' to go from texts to numeric data.

→ good idea to check sensitivity of one's inferences to these decisions (see Denny & Spirling, 2018).

From Texts to Numeric Data

1. Collect raw text in machine readable/electronic form. Decide what constitutes a document.
2. Strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.
3. Cut document up into useful elementary pieces: tokenization.
4. Add descriptive annotations to the tokens that preserve context: tagging.
5. Map tokens back to common form: lemmatization, stemming.
6. Model.

From Texts to Numeric Data

1. Collect raw text in machine readable/electronic form. Decide what constitutes a document.

“PREPROCESSING”

6. Model.

Exceptions

In some languages, tokenizing is a non-trivial problem because whitespace may not be used:

来年の事を言えば鬼が笑う

We may want to deal directly with multiword expressions in some contexts. There are rules which help us identify them relatively quickly and accurately.

e.g. ‘`White House`’ , ‘`traffic light`’

NB. these words mean something ‘special’ (and slightly opaque) when combined.

We Don't Care about Word Order

Generally, we are willing to **ignore the order of the words** in a document. This **simplifies** things. And we do (almost) as well without that information as when we retain it.

→ we are treating a document as a $\boxed{\text{bag-of-words}}$ (BOW).

but we keep multiplicity—i.e. multiple uses of same token

e.g. "The leading Republican presidential candidate has said Muslims should be banned from entering the US."

→ "lead republican presidenti candid said muslim ban enter us"

= "us lead said candid presidenti ban muslim republican enter"

Could use *n*-grams to take order into account, but typically not worth it.

Vector Space Model

A document can be thought of as a **collection of features** (tokens/words, etc.)

if each feature can be placed on the real line, then the document can be thought of as a point in space.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions: one corresponds to how '`Bob`'-ish it is, one corresponds to how '`goes`'-ish it is, one corresponds to how '`home`'-ish it is.

Features will typically be the *n*-gram (mostly unigram) frequencies of the words/tokens in the document, or some function of those frequencies.

e.g. '`the cat sat on the mat`' becomes `(2,1,1,1,1)` if we define the dimensions as `(the, cat, sat, on, mat)` and use simple counts.

Later, we take the vectors for all of our documents and arrange them as a **term document matrix** or a document term matrix.

II. Supervised Learning

Overview of Supervised Learning

Goal: categorize documents (or sentences) as belonging to a certain class (mutually exclusive? Jointly exhaustive?)

1. **Decide** on the taxonomy
2. **Label** examples of documents belonging to each category
   - e.g. some movie reviews that were positive ($y = 1$), some that were negative ($y = 0$); some statements that were liberal, some that were conservative.
3. **Train** a 'machine' on these documents (e.g. logistic regression), using the features (DTM, other stuff) as the 'independent' variables.
   - e.g. could flag use of word '`fetus`' relative to '`baby`' in discussing abortion as indicative of certain categories.
4. Use the learned relationship —some $f(x)$—to **classify** the outcomes of documents ($y \in \{0, 1\}$, review sentiment) not in the training set.

Classification with Dictionary Methods

In doing this, we learn **which words are associated with which category**. The model we trained tells us this.

We could use this to build a **dictionary** of words indicative of our categories (e.g. liberal, conservative, positive, negative, etc.)

Scholars have already done this. Provide **pre-determined lists of words**, sometimes with weights, indicating **which category they belong to** and sometimes, the extent to which they belong to that category.

We can rely on their dictionaries to classify our documents, too. Although, presents some problems.

Example: Barnes' review of *The Big Short*

Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in The Big Short, wasting an A-list ensemble cast in the process. Steve Carrel, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-drunk America walking towards that cliff's edge, but not enough to save the film.

Retain words in Hu & Liu Dictionary. . .

Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in The Big Short, wasting an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-drunk America walking towards that cliff's edge, but not enough to save the film.

Retain words in Hu & Liu Dictionary...

great savage

crisis wasting

tenuously

killing

superficially swelled

bad

complex

brutal

drunk

enough

Simple math...

negative	11
positive	2
total	13

$$\text{tone} = \frac{2-11}{13} = \frac{-9}{13}$$

Naive Bayes Classification

Is a way of **building our own dictionary**.

Motivation: emails d arrive and must classified as belonging to one of two classes $c \in \{\text{spam}, \text{ham}\}$.

by using the frequencies of the features (words, tokens) the emails contain.

use Naive Bayes, also simple Bayes, or independence Bayes,

is a family of classifiers which apply Bayes's theorem and make 'naive' assumptions about independence between the features of a document.

→ fast, simple, accurate, efficient and therefore popular.

Set up

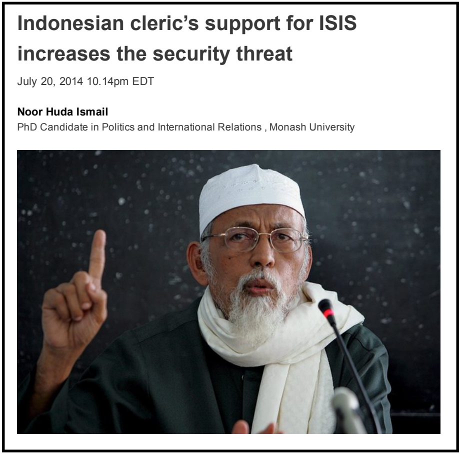
We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

NB we assume (naively) terms basically occur randomly throughout the document/no position effects

When we see certain terms which are disproportionately prevalent in certain classes of document, the probability we are observing a document of that class increases.

e.g. perhaps 99% of spam emails make reference to `dollars`, but only 1% of non-spam emails do. If the machine sees `dollars` it should update that it's likely a spam email.

Social Science Application



Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10.14pm EDT

Noor Huda Ismail

PhD Candidate in Politics and International Relations , Monash University

Nielsen (2012) investigates why certain scholars of Islam become Jihadi: i.e. why they encourage armed struggle (especially against the West)

Requires him to first classify scholars as Jihadi and ¬ Jihadi: has 27,142 texts from 101 clerics, so difficult to do by hand.

Jihadi Clerics

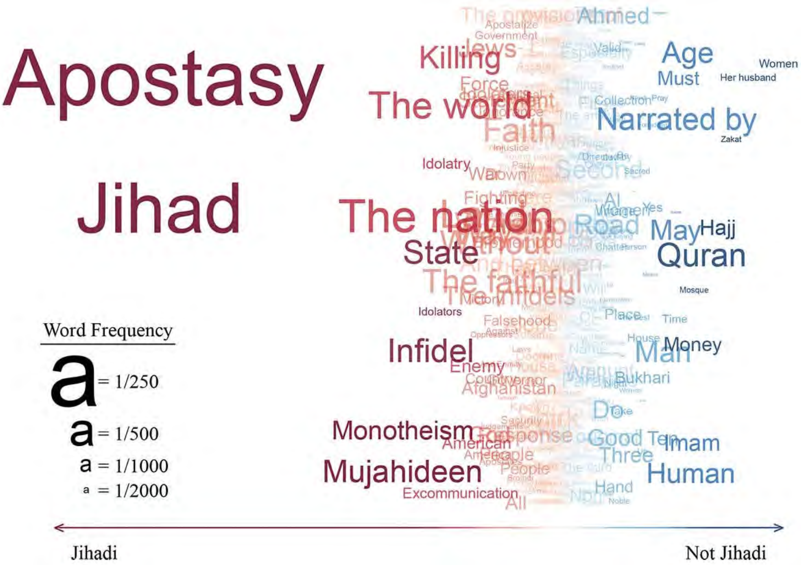
Training set: self-identified Jihadi texts (765), and sample from Islamic website as $\neg$ Jihadi (1951)

Preprocess: drops terms occurring in less than 10%, or more than 40% of documents, and uses 'light' stemmer for Arabic

Assigns a *Jihad Score* to each document: basically the logged likelihood ratio, $\sum_i \log \frac{\Pr(t_k|\text{Jihad})}{\Pr(t_k|\neg \text{ Jihad})}$ (note: doesn't know what 'real world' priors are, so drops them here)

Then for each cleric, he concatenates all works into one and gives this 'document'/cleric a score.

Discriminating Words



III. Unsupervised Learning

Overview of Unsupervised Learning

Don't have to label our data as belonging to categories we come up with, or decide how to recognize those categories.

→ while we know **who** gave a speech, we don't yet know what that speech 'represents' in terms of its latent properties (what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.)

Goal is to take the documents and **uncover their hidden structure and meaning**. By this, we mean that we look for **similarities across documents**.

→ Because we did not label our data, it is then up to **us** to **interpret** why some documents are estimated to be similar to others and not to others (what the groups/dimensions/concepts represent).

- Typically, these techniques put documents **on a scale** or **in a group**.

Scaling

We suspect that the German Greens and Social Democrats have moved steadily rightwards, post-reunification.

→ This is a time series problem, but extant techniques struggle. . .

i.e. hand-coding is expensive (manifestos are long)

and hard to find reference texts for Naive Bayes over time

→ need to assume lexicon is pretty stable, and that you can identify texts that contain all relevant terms.

Slapin and Proksch, 2008

Helpful to have an **unsupervised approach**, which is not dependent on **reference texts**

Suggest `Wordfish` scaling technique ("A Scaling Model for Estimating Time-Series Party Positions from Text")

1. Begin with **naive Bayes assumption**: idea that each word's occurrence is independent of all other words in the text.

→ surely false, but convenient starting point.

2. Need a (parametric) model for **frequencies** of words.

→ Choose *Poisson*: extremely simple because it has only one parameter—λ (which is mean and variance).

The model

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

- α_{it} fixed effect(s) for party *i* in time *t*: some parties have longer manifestos in certain years (which boosts all counts)
- ψ_j word fixed effect: some parties just use certain words more (e.g. their own name)
- β_j word specific weight: importance of this word in discriminating between party positions.
- ω_{it} estimate of party's position in a given year (so, this applies to specific party manifesto)
- → fit via **expectation maximization** (details in paper)

Catalinac, 2017

Question: theoretical work suggests that candidates **adopt different ideological positions** in different electoral systems. Little empirical evidence of this.

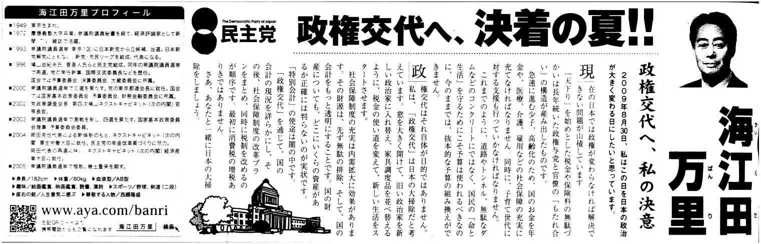
Empirical strategy: did candidates in elections to Japan's Lower House adopt different positions after electoral reform in 1994?

Specifically:

HI Did MMDs → SMDs bring candidates trying to win the district closer together ideologically?

H2 Did intra-party competition → no intra-party competition bring candidates of the same party closer together ideologically?

Texts: Candidate Manifestos



海江田万里

政権交代へ、決着の夏!!

The Democratic Party of Japan
民主党

政権交代へ、私の決意

2009年8月30日、私はこの日を日本の政治が大きく変わる日にしたいと思っています。

現在の日本では政権が変わらなければ解決できない問題が山積しています

「天下り」を始めとした税金や保険料の無駄づかいは長年続いた政権与党と官僚の「もたれ合い」の構造が産み出したものです。

急速に進む少子、高齢化のため、国のお金を年金や、医療、介護、雇用などの社会保障の充実に充てなければなりません。同時に、子育て世代に対する支援も行っていかなければなりません。

これまでのように、道路やトンネル、無駄なダムなどのコンクリートにではなく、国民の「命と生活」を守るためにこそ予算は使われるべきなのに、今のままでは、抜本的な予算の組み換えができません。

政権交代はそれ自体が目的ではありません。私は、「政権交代」は日本の大掃除だと考えています。窓を大きく開けて、旧い政治家を新しい政治家に入れ替え、家具調度品を並べ替えるように、税金の使い道を変えて、新しい生活をスタートさせます。

社会保障制度の充実は内需拡大に効果があります。その財源は、先ず無駄の排除、そして、国の会計をもっと透明にすることです。国の財産についても、どこにいくらの資産があるか正確には判らないのが実状です。「特別会計」の使途は闇の中です。

「政権交代」を通じて、国の会計の現況を詳らかにし、その後、社会保障制度の改革プランをまとめ、同時に税制を改めるのが順序です。最初に消費税の増税ありきではありません。

さあ、あなたと一緒に日本の大掃除をしましょう。

海江田万里プロフィール

■1949 東京生まれ。
■1972 慶應義塾大学卒業。参議院議員秘書を経て、経済評論家として新聞、TV、雑誌で活躍。
■1993 衆議院議員選挙（東京1区）に日本新党から立候補、当選。日本新党解党にともない、新党・市民リーグを結成、代表になる。
■1996 鳩山由紀夫氏、菅直人氏らと民主党結成。同年の衆議院議員選挙で再選。党の筆頭幹事、国際交流委員長などを歴任。国会では予算委員会、決算委員会、大蔵委員会に所属。
■2000 衆議院議員選挙で三選を果たす。党の東京都連会長に就任。国会では国家基本政策委員会、予算委員会、財務金融委員会に所属。
■2002 党政策調査会長、第四次鳩山ネクストキャビネット（次の内閣）官房長官。
■2003 衆議院議員選挙で激戦を制し、四選を果たす。国家基本政策委員会理事、予算委員会委員。
■2004 岡田克也代表による新体制のもと、ネクストキャビネット（次の内閣）厚生労働大臣に就任。民主党の年金改革案づくりに尽力。岡田代表の再選に伴い、ネクストキャビネット（次の内閣）経済産業大臣に就任。
■2005 衆議院議員選挙で惜敗。捲土重来を期す。

▶身長／182cm ▶体重／80kg ▶血液型／AB型
▶趣味／絵画鑑賞、映画鑑賞、読書、漢詩 ▶スポーツ／野球、剣道（二段）
▶座右の銘／人生意気ニ感ズ ▶尊敬する人物／西郷隆盛

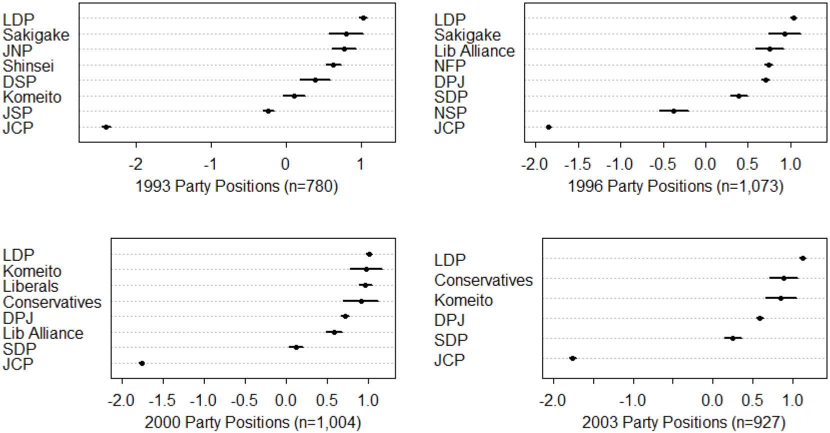
www.aya.com/banri
左記QRコードより
携帯電話からもご覧になれます 海江田万里 検索

→ Likely to be **representative** of a candidate's campaign strategy. Likely to be **taken seriously**. 40% of voters say they saw it (surveys, 1972-2005).

→ Collected and digitized **7,497**, produced by universe of non-frivolous candidates running in 8 HOR elections, 1986-2009.

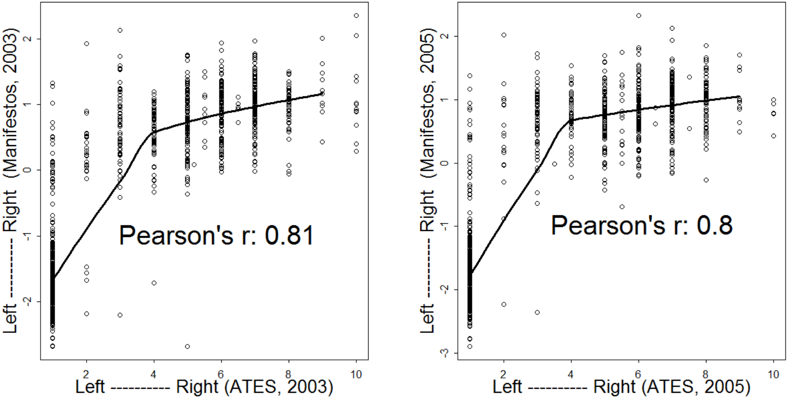
→ Used Wordfish to locate each manifesto on a unidimensional scale.

Validation with Locations of Average Party Members



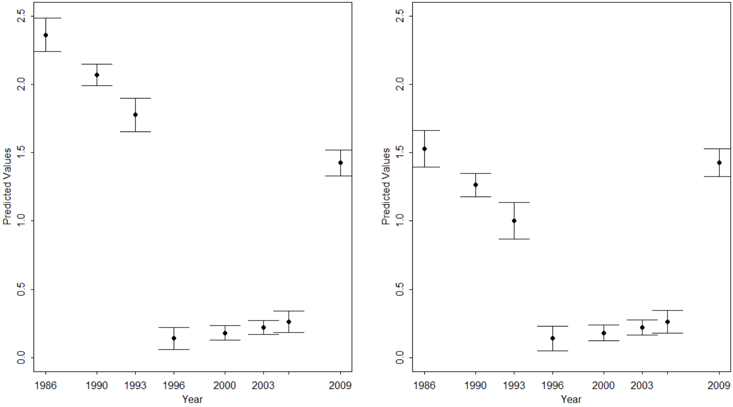
→ In all elections, JCP is to the **left** of all other parties and LDP is to the **right** of the JSP, SDP, NSP, and DPJ. In 1986, LDP and NLC positions are **same**.

Validation with ATES Data



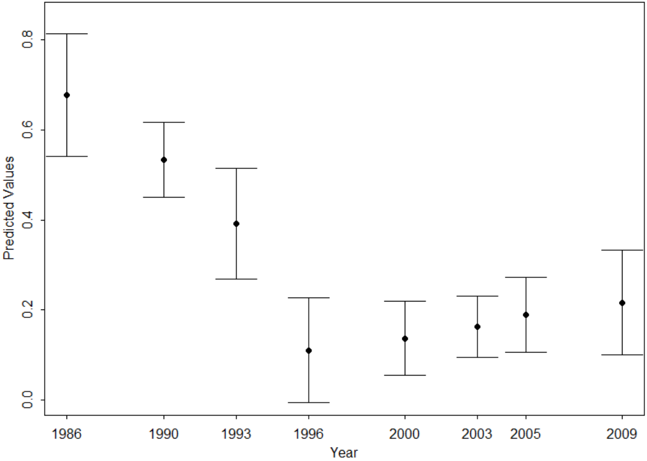
→ 2003 and 2005: candidates asked to **locate themselves** on ideological scale (response rates: 95% and 91%). **Correlate highly** with estimates from Wordfish.

Results: Hypothesis 1



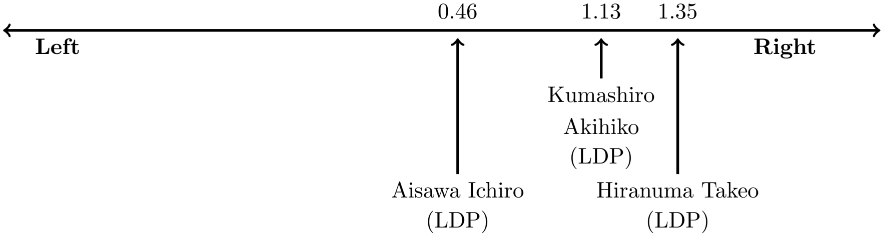
Predicted values of **within-district dispersion** with 95% confidence intervals. **Lower after reform.**

Results: Hypothesis 2



Predicted values of **within-party dispersion** with 95% confidence intervals. **Lower after reform.**

Okayama 1st District, 1993 (m=5)



0.46

1.13

1.35

Left

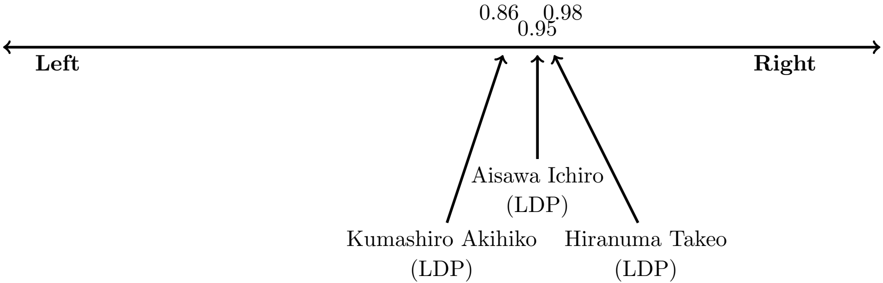
Right

Kumashiro Akihiko (LDP)

Aisawa Ichiro (LDP)

Hiranuma Takeo (LDP)

Okayama 1st, 2nd, 3rd Districts, 1996 (m=1 each)



Topic Models

> *Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.*
>
> Blei, 2012

NB. in social science we often use the outputs from topic models as a measurement strategy:

"who pays more attention to education policy, conservatives or liberals?"

Clustering

Document 1

Document 2

Document 3

⋮

Document N

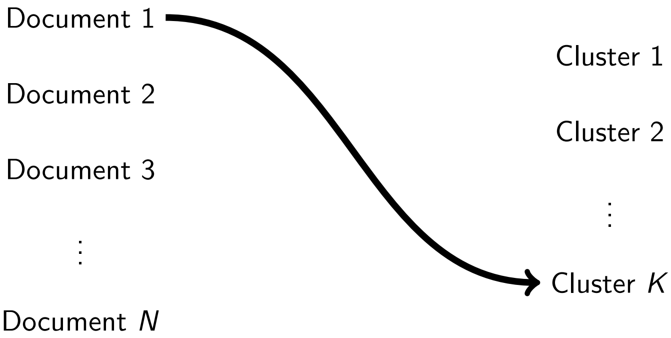
Cluster 1

Cluster 2

⋮

Cluster K

Clustering



Document 1

Document 2

Document 3

⋮

Document N

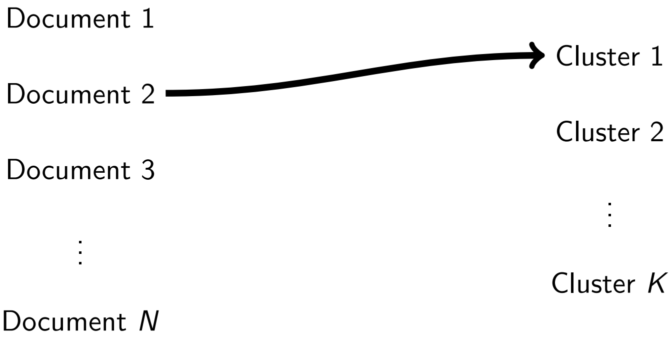
Cluster 1

Cluster 2

⋮

Cluster K

Clustering



Document 1

Document 2

Document 3

⋮

Document N

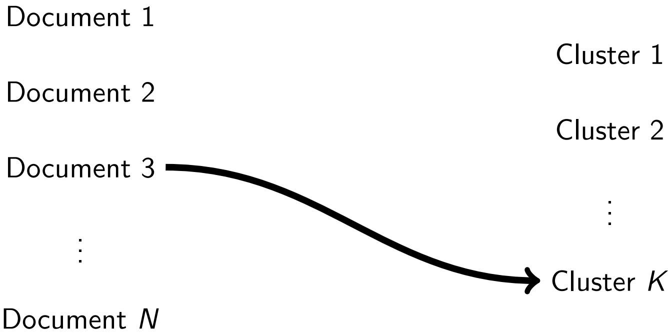
Cluster 1

Cluster 2

⋮

Cluster K

Clustering



Document 1

Document 2

Document 3

⋮

Document N

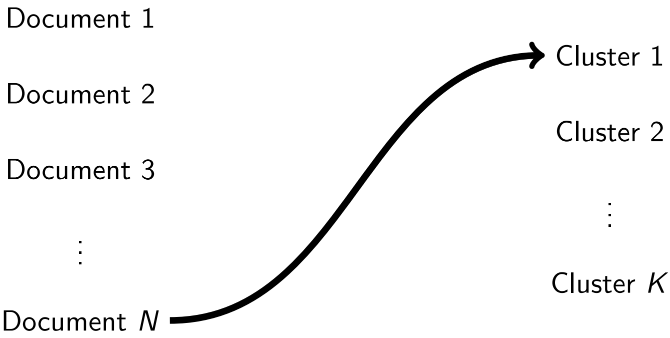
Cluster 1

Cluster 2

⋮

Cluster K

Clustering



Document 1

Document 2

Document 3

⋮

Document N

Cluster 1

Cluster 2

⋮

Cluster K

Clustering

Document 1

Document 2

Document 3

⋮

Document N

Cluster 1

Cluster 2

⋮

Cluster K

Topic Modeling

Document 1

Document 2

Document 3

⋮

Document N

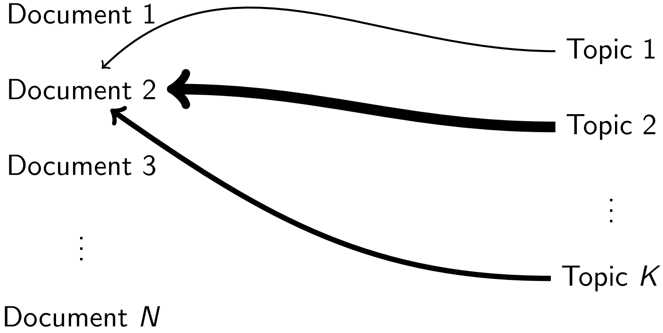
Topic 1

Topic 2

⋮

Topic K

Topic Modeling



Document 1

Document 2

Document 3

⋮

Document N

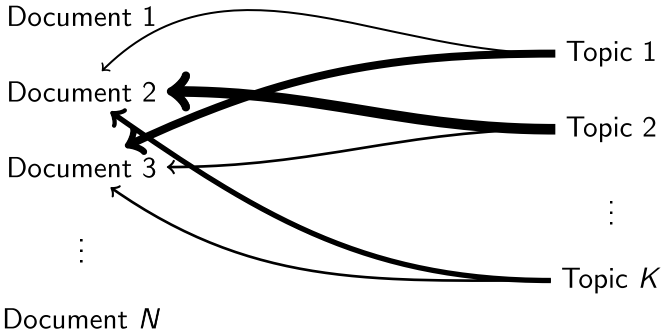
Topic 1

Topic 2

⋮

Topic K

Topic Modeling



Intuition behind data-generating process

Documents exhibit same topics, but in different proportions.

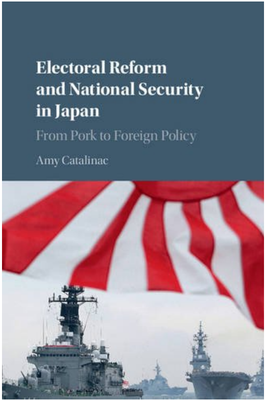
e.g. a speech by the Finance minister might be 50% drawn from the trade topic, 40% from the spending topic, 9.9% from the taxation topic, 0.1% from the health topic.

Think of a topic as a distribution over a fixed vocabulary.

e.g. the trade topic will have words like import and tariff with a high probability.

Technically we assume the topics are generated first, and the documents are generated second (from those topics).

Catalinac, 2016


Electoral Reform and National Security in Japan
From Pork to Foreign Policy
Amy Catalinac

Japan is a curious case for IR: wealthy post-war yet not very interested in foreign policy. Recent times have seen more interest in this area. Why?

1. Rise of China? Need to focus on security.

vs.

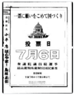
2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of a Westminster-style political system.

To decide, need a data source that covers all Lower House legislators where they set out their policy priorities over time. See if/when they shift priorities.

Manifestos

00008.tif

00009.tif


00010.tif

00011.tif

00012.tif

00013.tif

00014.tif

00015.tif

00016.tif

00017.tif

00018.tif

00019.tif

00020.tif

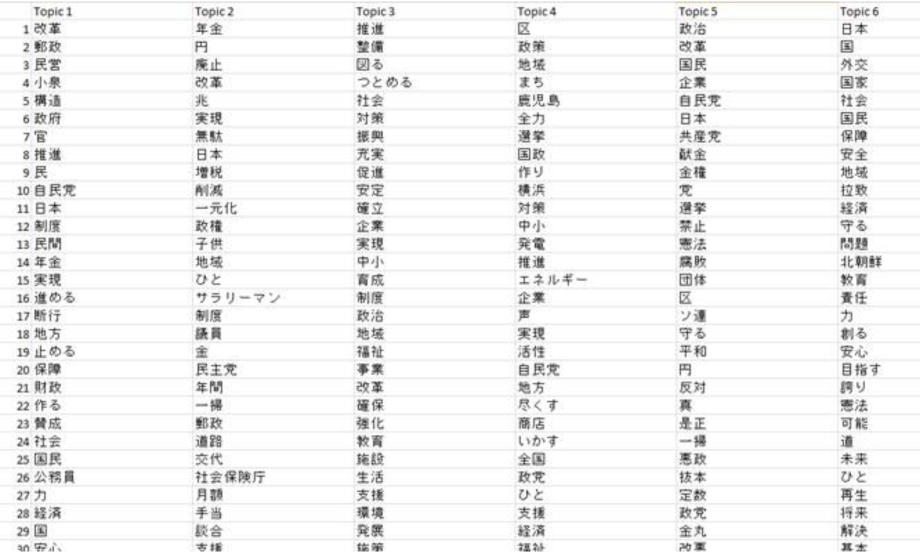
00021.tif

- Collected universe of manifestos produced by all candidates in HOR elections, 1955-2009, from Diet Library.

- Poor quality of paper meant **difficulties extracting digitized text**. OCR didn't work → transcribed by hand.

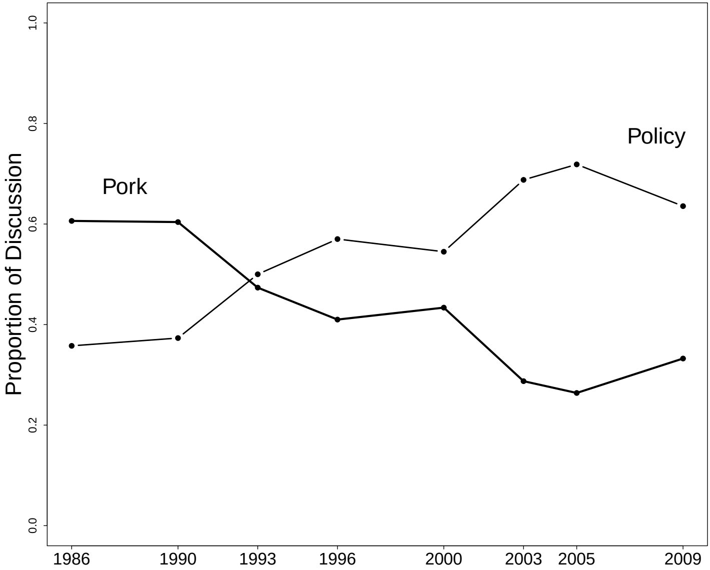
- Used `MeCab` implemented in `R` to create TDM → ran **LDA**.

Topic Distribution over Words

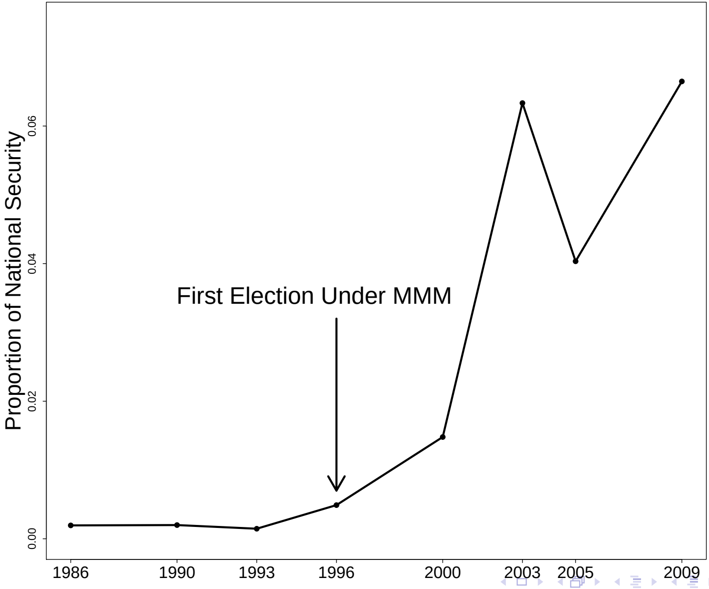


	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	改革	年金	推進	区	政治	日本
2	郵政	円	整備	政策	改革	国
3	民営	廃止	図る	地域	国民	外交
4	小泉	改革	つとめる	まち	企業	国家
5	構造	兆	社会	鹿児島	自民党	社会
6	政府	実現	対策	全力	日本	国民
7	官	無駄	振興	選挙	共産党	保障
8	推進	日本	充実	国政	献金	安全
9	民	増税	促進	作り	金権	地域
10	自民党	削減	安定	横浜	党	拉致
11	日本	一元化	確立	対策	選挙	経済
12	制度	政権	企業	中小	禁止	守る
13	民間	子供	実現	発電	憲法	問題
14	年金	地域	中小	推進	腐敗	北朝鮮
15	実現	ひと	育成	エネルギー	団体	教育
16	進める	サラリーマン	制度	企業	区	責任
17	断行	制度	政治	声	ソ連	力
18	地方	議員	地域	実現	守る	創る
19	止める	金	福祉	活性	平和	安心
20	保障	民主党	事業	自民党	円	目指す
21	財政	年間	改革	地方	反対	誇り
22	作る	一掃	確保	尽くす	真	憲法
23	賛成	郵政	強化	商店	是正	可能
24	社会	道路	教育	いかす	一掃	道
25	国民	交代	施設	全国	悪政	未来
26	公務員	社会保険庁	生活	政党	抜本	ひと
27	力	月額	支援	ひと	定数	再生
28	経済	手当	環境	支援	政党	将来
29	国	談合	発展	経済	金丸	解決
30	安心	支援	施策	福祉	改革	基本

Change in proportion of ‘Pork’ and ‘Policy’



Change in proportion of 'National Security' Topic



IV. Next Steps in Text as Data

Next Steps

- Moving beyond unigram representations: embeddings look promising.
- Thinking carefully about how to handle non-Latin scripts and their encodings.
- Working on social science specific training of models (e.g. for sentiment dictionaries and 'sophistication' measures).
- Bringing more supervision to unsupervised problems, like (structural) topic models.

Thanks!

Amy Catalinac

Ph.D., Harvard, 2011.

Assistant Professor, Department of Politics, New York University

Contact

https://scholar.harvard.edu/amycatalinac/home

amy.catalinac@nyu.edu