Computer-assisted content analysis 2

Will Lowe Princeton University
James Lo University of Southern California

Menu

Session 1: Classical Content Analysis

Session 2:

- Document classification
  - The direct approach to classification
  - The indirect approach to classification
  - Evaluation and interpretation
  - Topic models

Session 3: Scaling Models

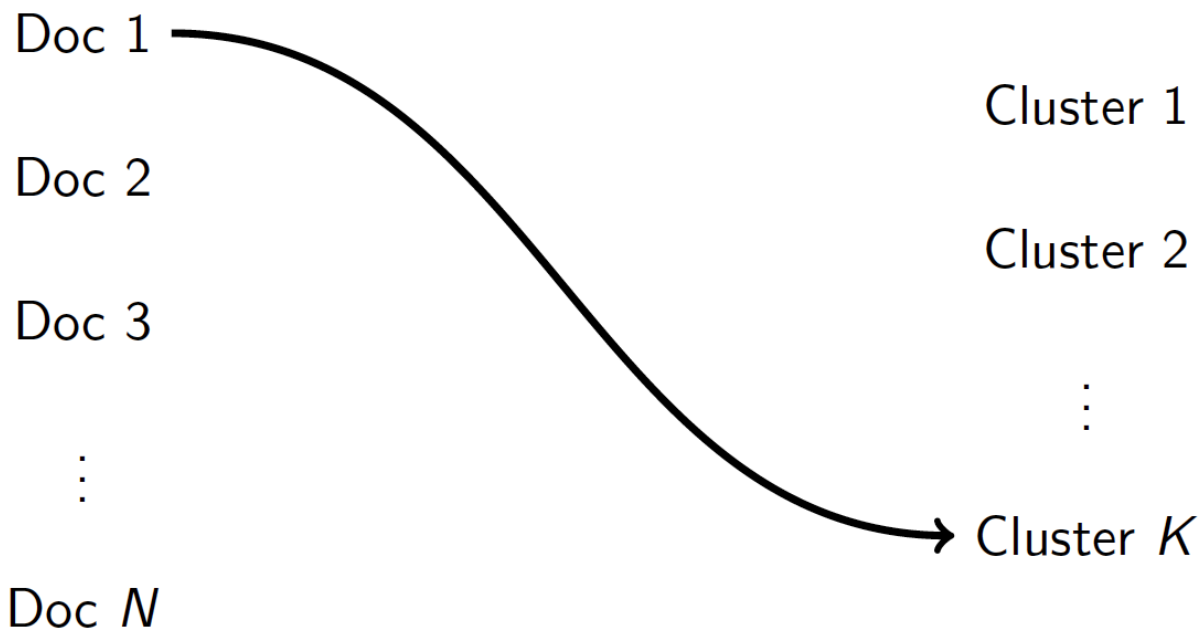
Classification Approaches when Categories are Known

We often want to read text to put them into categories

Examples:

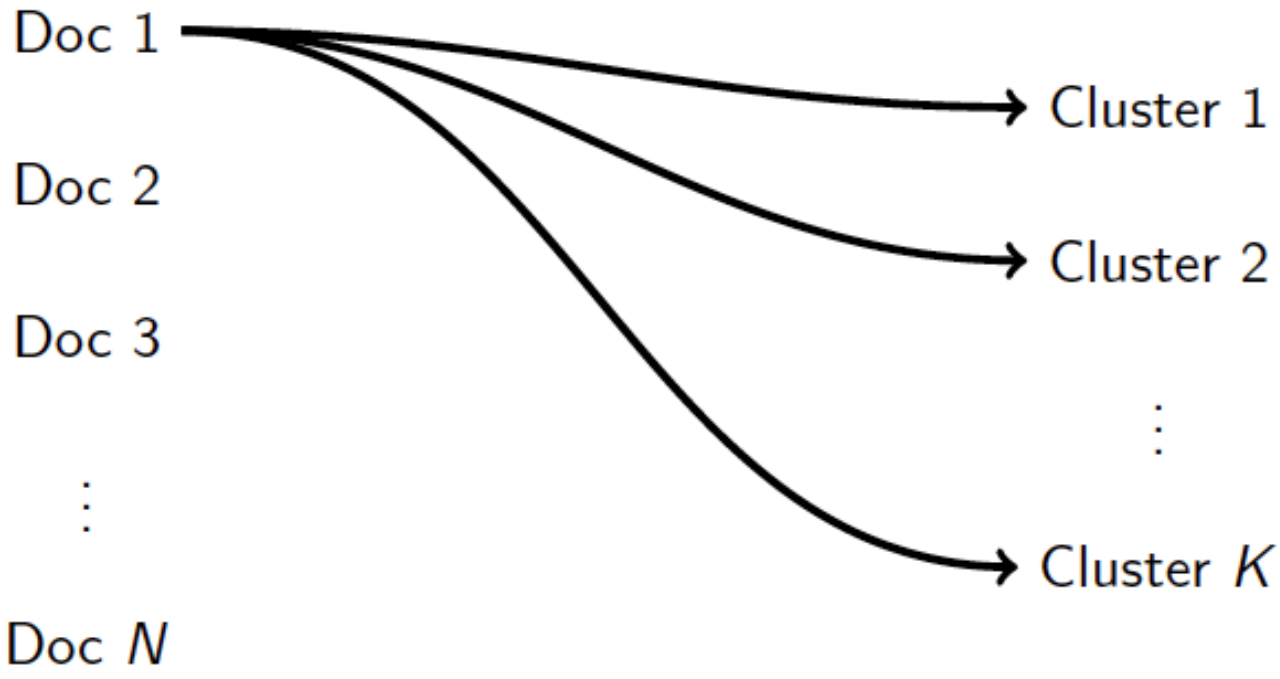
1. Are campaign advertisements positive or negative?
2. What policy areas do newspaper editorials cover?
3. Are international statements belligerent or peaceful?
4. Do court letters liberal or conservative?
5. Is this email spam?

Classification



For classification, each document belongs to a single cluster

Topic Models



In LDA, each document can contain multiple topics

Classification Approaches when Categories are Known

Yesterday: Classification using dictionary approach.

An alternative is **supervised machine learning** methods:

1. coders categorize a set of documents by hand
2. the algorithm "learns" how to sort the documents in categories
3. characteristics of training set are used to assign new documents to categories.

   Naive Bayes, Maximum Entropy, Support Vector Machines, Neural Networks, Bagging, Boosting, . . .

Classification Approaches when Categories are Known

Let θ_k be the *probability* that Z=k for each document

- Words are denoted $\{W\}$
- Each document has a *single* topic Z
- Some topic labels are observed
- We essentially want $P(Z \mid \{W\}) = \theta_{z|w}$

Supervised machine learning

- Tries to give good predictions of observed Y given X
- Naive Bayes, Maximum Entropy, Support Vector Machines, Neural Networks, Bagging, Boosting, . . .

Classification approaches

As before, we have two approaches

Discriminative: Model $P(Z \mid \{W\}) = \theta_{z|w}$ directly

Generative: Model $P(\{W\} \mid Z)$ and $P(Z) = \theta_z$, then get $P(Z \mid \{W\}) = \theta_{z|w}$ via Bayes theorem

Bayes theorem is about inverse probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

If $P(\{W\} \mid Z)$ known, estimate $P(Z \mid \{W\})$

Either way…

Desirable classification outcome:

	$P(Z = \text{`Domestic'} \mid \{W\}_d)$	$P(Z = \text{`Foreign'} \mid \{W\}_d)$
D_1	0.75	0.25
D_2	0.82	0.18
…	…	…
D_{N-1}	0.02	0.98
D_N	0.45	0.55

where $\theta_{z|w} = P(Z = z \mid \{W\})$

The Basic Steps

1. Construct a training set (a) create a coding scheme (b) select documents (ideally randomly sampled)
2. Apply the supervised learning method to learn features of a training set and infer labels
3. Validate and classify remaining documents

We will do this in the lab together with NY Times titles

An Applied Example: Affirmative Action

Evans et al. (2007) apply naive Bayes to legal text

- Affirmative action programs at university level

Examine Amicus Curiae briefs sent to appellate court

- Can words tell us direction of AC brief?
- Are the briefs liberal or conservative?

Naive Bayes

Naive Bayes is a relatively old ($\sim$1975) classification method

Want to guess if document j is liberal/conservative based on its word profile $\{W\}_j$.

Probability can be derived by applying Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(Z = \text{'Lib'} \mid \{W\}_j) = \frac{P(\{W\}_j \mid Z = \text{'Lib'})\, P(Z = \text{'Lib'})}{P(\{W\}_j)}$$

Naive Bayes

$$P(Z = \text{'Lib'} \mid \{W\}_j) \propto P(\{W\}_j \mid Z = \text{'Lib'})\, P(Z = \text{'Lib'})$$

We can drop $P(\{W\}_j)$ since it is constant across categories.

Given a representative training set, estimating $P(Z = L)$ is easy:

$$\hat{P}(Z = \text{'Lib'}) = \frac{\text{\# training docs that are liberal}}{\text{\# of training docs}}$$

Naive Bayes

Estimating the probability that a word profile $\{W\}_j$ occurs given that the document is liberal $P(\{W\}_j \mid Z = \text{'Lib'})$ is more challenging, because any one word profile is likely to occur only once.

Assumption: words are assumed to be generated *independently* given the category Z

$$P(\{W\}_j \mid Z = \text{'Lib'}) = \prod_i P(W_i \mid Z = \text{'Lib'})$$

$$P(\text{'Affirmative Action'} \mid Z = \text{'Lib'}) = P(\text{'Affirmative'} \mid Z = \text{'Lib'}) \cdot P(\text{'Action'} \mid Z = \text{'Lib'})$$

Naive Bayes

With this assumption, we can estimate the probability of observing a word *i* given that the document is liberal: proportion of word *i* in liberal training set.

The classifier then chooses the class Z (Liberal or Conservative) with the highest aggregate probability.

Note that every new word adds a bit of information that re-adjusts the conditional probabilities.

Naive Bayes

Note that with two classes (here: liberal and conservative) this has a rather neat interpretation:

$$\frac{P(Z = \text{'Lib'} \mid \{W\}_j)}{P(Z = \text{'Con'} \mid \{W\}_j)} =$$

$$\prod_i \frac{P(W_i \mid Z = \text{'Lib'})}{P(W_i \mid Z = \text{'Con'})} \times \frac{P(Z = \text{'Lib'})}{P(Z = \text{'Con'})}$$

Logging this probability ratio, every new word *adds* a bit of information that pushes the ratio above or below 0

Naive Bayes

Example: Naive Bayes with only word class 'discriminat*'.

$$P(W = \text{'discriminat*'} \mid Z = \text{'Lib'}) = (26 + 13)/(20002 + 18722) \approx 0.001$$
$$P(W = \text{'discriminat*'} \mid Z = \text{'Con'}) = (70 + 48)/(17368 + 17698) \approx 0.003$$

Assume that liberal and conservative supporting briefs are equally likely (true in the training set)

$$\frac{P(Z = \text{'Lib'})}{P(Z = \text{'Con'})} = 1$$

Last step: calculate posterior classification probabilities for a new document (based on occurrence of this word).

Naive Bayes

Amicus brief from ‘King County Bar Association’ containing 3667 words and 4 matches to disciminat*.

```
that "the state shall not [discriminate] against, or grant preferential treatment
the lingering effects of racial [discrimination] against minority groups in this
remedy the effects of societal [discrimination]. Another four Justices (Stevens
that "the state shall not [discriminate] against, or grant preferential treatment
```

Naive Bayes

A priori, the probabilities are…

Probability that we observe the word discriminat* 4 out of 3667 times if the document is liberal:

```
> dbinom(4, size=3667, prob=0.001007127)
[1] 0.1930602
```

Probability that we observe the word discriminat* 4 out of 3667 times if the document is conservative:

```
> dbinom(4, size=3667, prob=0.003365083)
[1] 0.004188261
```

Logged probability ratio = 3.83

Naive Bayes

Conclusion: Seeing 4 instances of discriminat* gives the posterior classification probabilities

$$\theta_{\text{liberal}} = \frac{0.193}{0.193+0.004} = 0.979$$

$$\theta_{\text{conservative}} = 1\text{-}0.979=0.021$$

This is *quite* confident

…but other words will be less loaded or push the other way

Evaluating Classifiers: Accuracy, Precision and Recall

		Machine	
		Liberal	Conservative
Training Data	Liberal	40	10
	Conservative	40	60

Accuracy = (40+60)/150 = .66

Precision ($Z_{machine}$=Liberal) = 40/(40+40) = 0.5

Precision ($Z_{machine}$=Cons) = 60/(10+60) = 0.86

Recall ($Z_{training}$=Liberal) = 40/(40+10) = 0.80

Recall ($Z_{training}$=Cons) = 60/(40+60) = 0.60

Latent Dirichlet Allocation

What if documents don't belong to only one category?

- Each document is a mixture over topics
- Each topic is a mixture over words

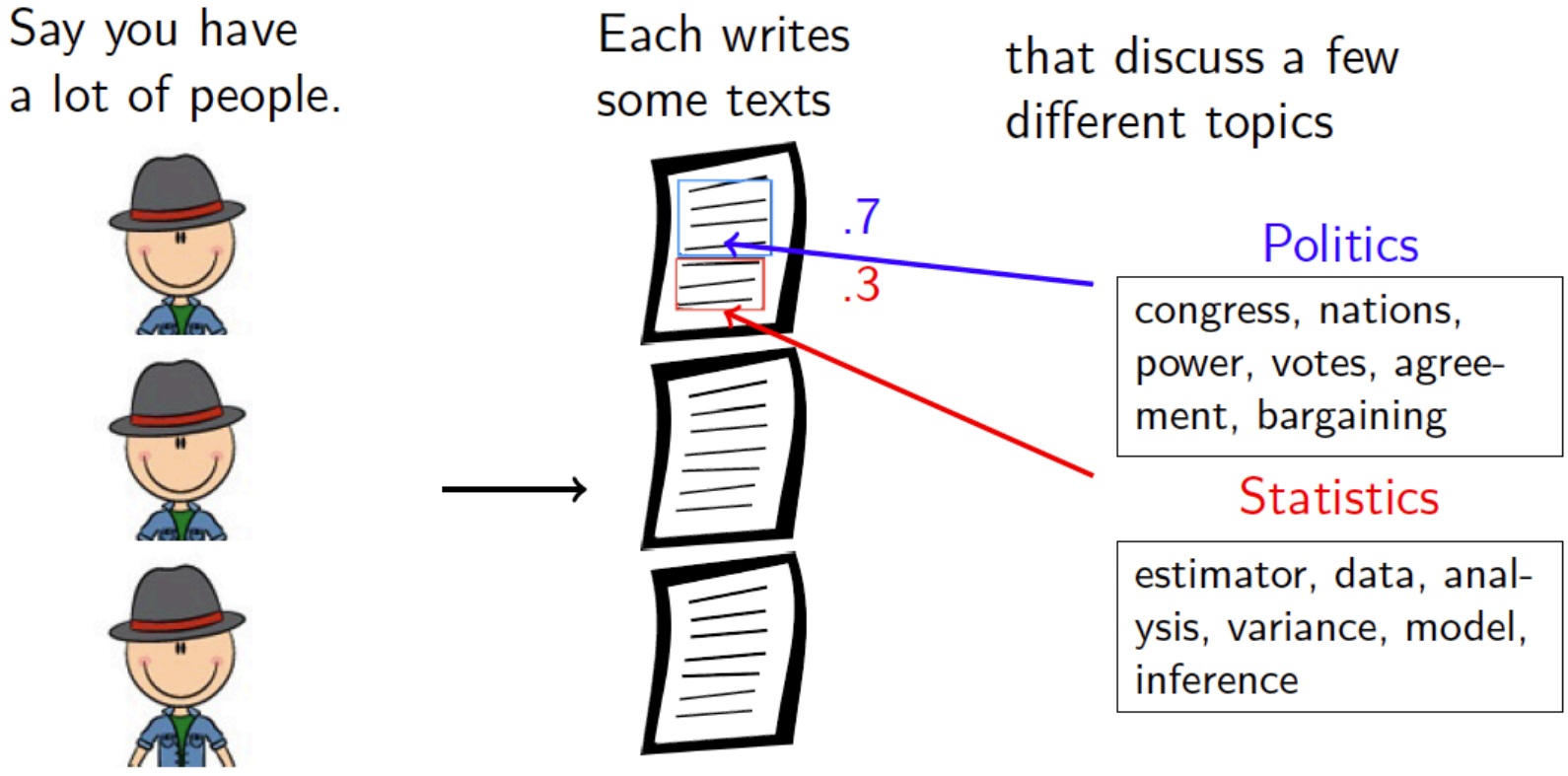
Using LDA gives us:

- Distribution of words for each topic (β)
- Proportion of a document in each topic (θ)

Setting up LDA

- Still uses bag of words
- Number of topics fixed ex ante
- Topics examined ex post – unsupervised learning

Visualizing Topic Models



The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words
2. The proportion of each document in each topic

Courtsey of Brandon Stewart

Latent Dirichlet Allocation Components

Given dimensions:

- **N** documents, **J** different topics, **K** unique words

We want the following matrices:

- $X = N \times K$ document-term matrix (observed)
- $\theta = N \times K$ matrix with row $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$
  - θ_{ik}: Proportion of document i allocated to topic k
- $\beta = K \times J$ matrix with row $\beta_k = (\beta_{k1}, \ldots, \beta_{kJ})$
  - β_{kj}: Probability of using word J, if topic k is chosen
- α = K length population prior for θ

Objective function: f(X, β, θ, α)

LDA Generative Model

When writing a word (m) for document i:

$$
\begin{aligned}
\mathrm{p}(\theta_i|a) &= \text{Dirichlet}(a) && \text{(Pick potential topics } \theta_i\text{)} \\
\mathrm{p}(z_{im}|\theta_i) &= \text{Multinomial}(1, \theta_i) && \text{(Pick the topic z to discuss)} \\
x_{im}|\beta_k, z_{im} &= \text{Multinomial}(1, \beta_k) && \text{(Pick a word from topic } z_{im}\text{)} \\
\mathrm{p}(\beta_k) &= \text{Dirichlet}(\eta) && \text{(Prior on topics)}
\end{aligned}
$$

with

$$\mathrm{p}(\beta, \theta, \mathrm{Z}, a \mid \mathrm{X}) \propto \mathrm{p}(\beta \mid \eta) \cdot \mathrm{p}(\theta|a) \cdot \mathrm{p}(\mathrm{Z} \mid \theta) \cdot \mathrm{p}(\mathrm{X} \mid \beta, \mathrm{Z})$$

Intuition on LDA

$p(\beta, \theta, Z, \alpha \mid X) \propto p(\beta \mid \eta) \cdot p(\theta|\alpha) \cdot p(Z \mid \theta) \cdot p(X \mid \beta, Z)$

Remember rows of β and θ must sum to 1!

$p(X \mid \beta, Z)$: Favors co-occurring words, segregated topics

- $\beta = (0.5,0.5,0,0)$ vs $\beta = (0.25,0.25,0.25,0.25)$

$p(Z \mid \theta)$: Higher if θ is concentrated (i.e. fewer potential topics)

- $\theta = (0.5,0.5, 0, 0)$ vs. $\theta = (0.25, 0.25, 0.25, 0.25)$

Joint distribution favors sparse topics, small topic clusters

But data often need a larger number of topics to assign small topic clusters to data

Japanese Campaign Manifestos (Catalinac 2016)

Examines the effect of electoral reform

- Before 1994: SNTV-MMD (more intraparty competition)
- After 1994: Mixed member majoritarian (PR + SMD)

LDA on N=7497 candidate manifestos for Diet 1950-2009

- Hand transcribed, OCR failed
- More complicated in Japanese

Characterizing campaigns across 50+ years

- What do candidates talk about?
- How did electoral reform change incentives?
- Why increasing interest in militaristic foreign action?

Japanese Manifestos

新宿・千代田・港区の皆さま！ より たしかな 未来を！

あなたの1票を生かします！

大つかゆうじ

前衆院土地問題特別委員長
自民党都連政調会長

身近な問題から着実に解決し、
子どもたちの笑顔に満ちた
たしかな日本の未来づくりを目指します。

民主化の進む国際社会において日本の役割はこれまでにない重要な段階を迎えています。いまこそ、たしかな日本の未来の実現のために、グローバルな視点から日本の政治・経済を見直し、国際社会での確たるスタンスを保っていかねばなりません。

そのためには、国内政局の安定を基本に、自らの政治姿勢を正し、'90年代にふさわしい新しい政治を構築せねばなりません。

大つかゆうじは、持てる力の限りを尽くして、たしかな日本の未来の実現のために全力投球の覚悟です。特に、都市議員として相続税の見直しをはじめ、抜本的な大都市土地政策を確立し、若者に夢と希望を与える住宅をつくることに全力を挙げる所存です。若者の活力こそが、21世紀の日本のようたしかな未来の実現につながると確信しています。ぜひ、大つかゆうじをご支援ください。

国民の納得する本当の税制改革を！
相続税見直しで定住人口の確保を！
土地有効利用で都心に若者の住宅を！

子どもたちに夢と希望を与える教育の実現！
21世紀を担う子どもたちに夢と希望を！個性と自主性を尊重したきめ細かい教育施策を実施。家庭教育にもスポット。

内需拡大を柱に商店、中小企業の活性化
交通網の整備、都市再開発を推進、多極多心型の都市づくり。先端技術の導入や市場開拓など緊急融資の実施！

女性新時代の実現
家族と共にある新しい母親像を提言！育児休業、再教育、保育施設の完備など女性が安心して社会参加できる環境を！

「生きがいある第二の人生」づくり
人生80年時代。予防医学、健康増進、成人学校など積極的な人生設計／年金、医療費の安定。寝たきり老人への施策。

全力で取り組んでいます！

私のこれまでの実績など

○地下鉄6・7・12号線建設に大きな貢献／13号線実現に奮闘中／
○都市再開発プロジェクトアークヒルズ建設をはじめ各地に計画を推進／
○市場整備計画の推進と予算化でバックアップ／
○マンション議連会長として永住できる快適な住環境づくりを推進中／
○都心の駐車場対策に尽力、事故撲滅運動を展開交通渋滞解消に全力／

私たちのくらしに、なくてはならない政治家です。

都政、区政と国政のパイプ役としての大つかさんは、抜群の行動派です。鋭い洞察力と明快な判断力を持つ大つかさんは、一区ではもっとも信頼と実績に裏付けられた政治家です。これまで、衆院土地問題に関する特別委員長として、国政レベルの土地施策を積極的に推進する傍ら、若者たちを都心に呼び戻すための、あるいは、高齢社会への各種福祉施策など鋭意邁進中です。

どうぞ、皆さま方のあたたかいご支援を大つかゆうじにお願いいたします。

大つかゆうじさんを推薦します。

鈴木俊一（東京都知事）
山本克忠（新宿区長）
木村 茂（千代田区長）
山田敬治（港区長）

大つかゆうじ 略歴

議員歴 東京都議会議員2期、衆議院議員5期
役職歴 国土政務次官、文部政務次官、衆議院法務委員長、検察官適格審査会委員、自民党調査局長、文教部会長、教育改革に関する調査会会長代理
現役職 衆議院土地問題特別委員会委員長、党文教制度調査会副会長、党都連政調会長、慶応義塾大学評議員、(財)国際連合大学協会理事他

Japanese Manifesto Topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
改革	年金	推進	区	政治	日本
郵政	円	整備	政策	改革	国
民営	廃止	図る	地域	国民	外交
小泉	改革	つとめる	まち	企業	国家
構造	兆	社会	鹿児島	自民党	社会
政府	実現	対策	全力	日本	国民
官	無駄	振興	選挙	共産党	保障
推進	日本	充実	国政	献金	安全
民	増税	促進	作り	金権	地域
自民党	削減	安定	横浜	党	拉致
日本	一元化	確立	対策	選挙	経済
制度	政権	企業	中小	禁止	守る
民間	子供	実現	発電	憲法	問題
年金	地域	中小	推進	腐敗	北朝鮮
実現	ひと	育成	エネルギー	団体	教育
進める	サラリーマン	制度	企業	区	責任
断行	制度	政治	声	ソ連	力
地方	議員	地域	実現	守る	創る
止める	金	福祉	活性	平和	安心
保障	民主党	事業	自民党	円	目指す
財政	年間	改革	地方	反対	誇り
作る	一掃	確保	尽くす	真	憲法
賛成	郵政	強化	商店	是正	可能
社会	道路	教育	いかす	一掃	道
国民	交代	施設	全国	悪政	未来
公務員	社会保険庁	生活	政党	抜本	ひと
力	月額	支援	ひと	定数	再生
経済	手当	環境	支援	政党	将来
国	談合	発展	経済	金丸	解決
安心	支援	施策	福祉	改悪	基本
Postal privatization	Reducing Wasteful Public Spending	Pork for the District	Policies for the district	Political Reform	National Security Policy

Manifesto Content over Time

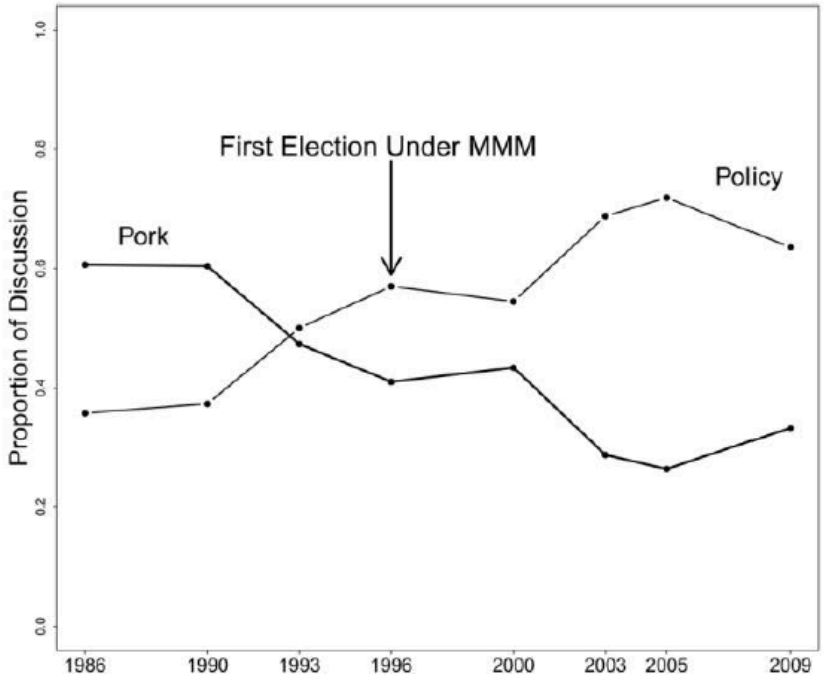


Figure 1. LDP candidates switched to more policy and less pork in the 1993 election and continued with this strategy under MMM. This figure plots the mean proportions of discussion devoted to pork and policy, respectively, in the 2,355 manifestos produced by LDP candidates in these eight elections.

Foreign Policy Content

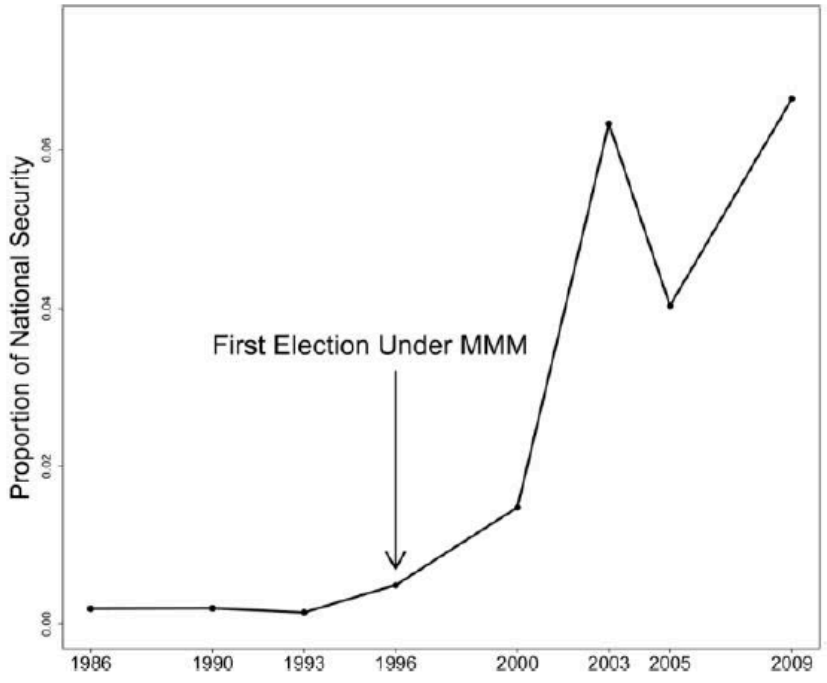


Figure 2. LDP candidates increased their discussion of national security in the first election under MMM. This figure plots the mean proportion of discussion comprised of national security in the 2,355 manifestos produced by LDP candidates in these eight elections.

Wrapping up

For single category classification

- Discriminative: Model $P(Z \mid \{W\}, \beta) = \theta_{z|w}$ directly
- Generative: Model $P(\{W\} \mid Z, \beta)$ and $P(Z) = \theta_z$, then get $P(Z \mid \{W\}, \beta) = \theta_{z|w}$ via Bayes theorem

For multiple categories

- Latent Dirichlet Allocation
- More generally used for discovery of topics

Tomorrow's Lab

- Classifying NY Times articles by topic