Text as Data

Justin Grimmer

Professor
Department of Political Science
Stanford University

April 30th, 2019

Stylometry$\rightsquigarrow$ Who Wrote Disputed Federalist Papers?

Federalist papers $\rightsquigarrow$ Mosteller and Wallace (1963)

- Persuade citizens of New York State to adopt constitution
- Canonical texts in study of American politics
- 77 essays
  - Published from 1787-1788 in Newspapers
  - And under the name Publius, anonymously

Who Wrote the Federalist papers?

- Jay wrote essays 2, 3, 4,5, and 64
- Hamilton: wrote 43 papers
- Madison: wrote 12 papers

Disputed: Hamilton or Madison?

- Essays: 49-58, 62, and 63
- Joint Essays: 18-20

Task: identify authors of the disputed papers.

Task: Classify papers as Hamilton or Madison using dictionary methods

Setting up the Analysis

Training $\rightsquigarrow$ papers Hamilton, Madison are known to have authored
Test $\rightsquigarrow$ unlabeled papers
Preprocessing:

- Hamilton/Madison both discuss similar issues
- Differ in extent they use stop words
- Focus analysis on the stop words

Setting up the Analysis

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N) = (\text{Hamilton}, \text{Hamilton}, \text{Madison}, ..., \text{Hamilton})$
  $N \times 1$ matrix with author labels
- Define the number of words in federalist paper i as num_i

$$\boldsymbol{X} = \begin{pmatrix} \frac{1}{\text{num}_1} & \frac{2}{\text{num}_1} & \frac{0}{\text{num}_1} & \cdots & \frac{3}{\text{num}_1} \\ \frac{0}{\text{num}_2} & \frac{1}{\text{num}_2} & \frac{0}{\text{num}_2} & \cdots & \frac{0}{\text{num}_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{0}{\text{num}_N} & \frac{0}{\text{num}_N} & \frac{1}{\text{num}_N} & \cdots & \frac{0}{\text{num}_N} \end{pmatrix}$$

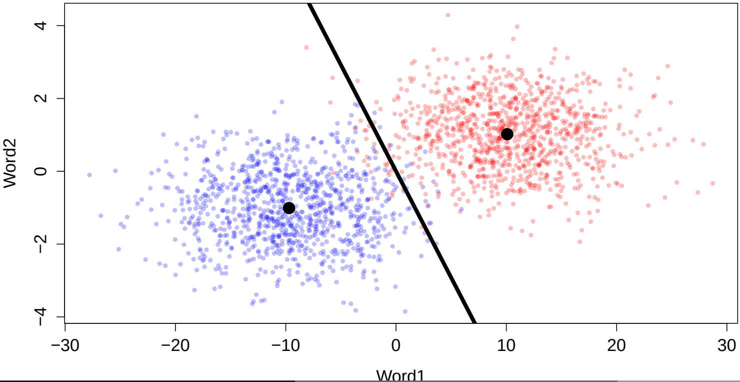
  $N \times J$ counting stop word usage rate
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$
  Word weights.

Objective Function

Heuristically: find $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \ldots, \theta_J^*)$ used to create score

$$p_i = \sum_{j=1}^{J} \theta_j^* X_{ij}$$

that maximally discriminates between categories



Objective Function

Define:

$$\boldsymbol{\mu}_{\text{Madison}} = \frac{1}{N_{\text{Madison}}} \sum_{i=1}^{N} I(Y_i = \text{Madison}) \boldsymbol{X}_i$$

$$\boldsymbol{\mu}_{\text{Hamilton}} = \frac{1}{N_{\text{Hamilton}}} \sum_{i=1}^{N} I(Y_i = \text{Hamilton}) \boldsymbol{X}_i$$

Objective Function

We can then define functions that describe the "projected" mean and variance for each author

$$
\begin{aligned}
g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison}) &= \frac{1}{N_{\text{Madison}}} \sum_{i=1}^{N} I(Y_i = \text{Madison}) \boldsymbol{\theta}^{'} \boldsymbol{X}_i = \boldsymbol{\theta}^{'} \boldsymbol{\mu}_{\text{Madison}} \\
g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) &= \frac{1}{N_{\text{Hamilton}}} \sum_{i=1}^{N} I(Y_i = \text{Hamilton}) \boldsymbol{\theta}^{'} \boldsymbol{X}_i = \boldsymbol{\theta}^{'} \boldsymbol{\mu}_{\text{Hamilton}} \\
s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison}) &= \sum_{i=1}^{N} I(Y_i = \text{Madison}) (\boldsymbol{\theta}^{'} \boldsymbol{X}_i - \boldsymbol{\theta}^{'} \boldsymbol{\mu}_{\text{Madison}})^2 \\
s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) &= \sum_{i=1}^{N} I(Y_i = \text{Hamilton}) (\boldsymbol{\theta}^{'} \boldsymbol{X}_i - \boldsymbol{\theta}^{'} \boldsymbol{\mu}_{\text{Hamilton}})^2
\end{aligned}
$$

Objective Function $\rightsquigarrow$ Optimization

$$
\begin{aligned}
f(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}) &= \frac{(g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) - g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison}))^2}{s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) + s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison})} \\
&= \frac{\left(\boldsymbol{\theta}^{'}(\boldsymbol{\mu}_{\text{Hamilton}} - \boldsymbol{\mu}_{\text{Madison}})\right)^2}{\text{Scatter}_{\text{Hamilton}} + \text{Scatter}_{\text{Madison}}}
\end{aligned}
$$

Optimization $\rightsquigarrow$ find $\boldsymbol{\theta}^{*}$ to maximize $f(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y})$, assuming independence across dimensions.
(Fisher's) Linear Discriminant Analysis

Optimization$\rightsquigarrow$ Word Weights

For each word j, construct weight θ_j^*,

$$\begin{aligned}
\mu_{j,\text{Hamilton}} &= \frac{\sum_{i=1}^{N} I(Y_i = \text{Hamilton})X_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N} I(Y_i = \text{Hamilton})X_{ij}} \\
\mu_{j,\text{Madison}} &= \frac{\sum_{i=1}^{N} I(Y_i = \text{Madison})X_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N} I(Y_i = \text{Madison})X_{ij}} \\
\sigma^2_{j,\text{Hamilton}} &= \text{Var}(X_{i,j}|\text{Hamilton}) \\
\sigma^2_{j,\text{Madison}} &= \text{Var}(X_{i,j}|\text{Madison})
\end{aligned}$$

We can then generate weight θ_j^* as

$$\theta_j^* = \frac{\mu_{j,\text{Hamilton}} - \mu_{j,\text{Madison}}}{\sigma^2_{j,\text{Hamilton}} + \sigma^2_{j,\text{Madison}}}$$

Optimization $\rightsquigarrow$ Trimming the Dictionary

- Trimming weights: Focus on discriminating words (very simple regularization)
- Cut off: For all $|\theta_j^*| < 0.025$ set $\theta_j^* = 0$.

Classification $\rightsquigarrow$ Determining Authorship

For each disputed document i, compute discrimination statistic

$$p_i = \sum_{j=1}^{J} \theta_j^* X_{ij}$$

$p_i \rightsquigarrow$ classification (linear discriminator)

- Above midpoint in training set $\rightarrow$ Hamilton text
- Below midpoint in training set $\rightarrow$ Madison text

Findings: Madison is the author of the disputed federalist papers.

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)

Inferring Separating Words

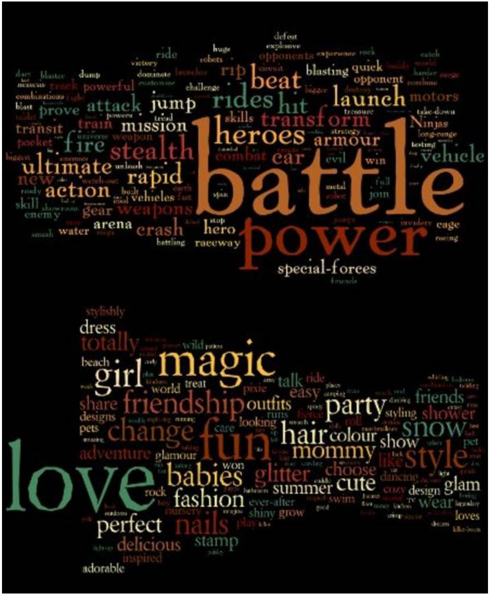
Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising

Inferring Separating Words



Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups$\rightsquigarrow$ Labeling output from Clustering/Topic Models

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups$\rightsquigarrow$ Labeling output from Clustering/Topic Models

Vague and Difficult to derive before hand

Congressional Language Across Sources

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
    - Yes: press releases have different purposes, targets, and need not relate to official business

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
    - Yes: press releases have different purposes, targets, and need not relate to official business
    - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
    - Yes: press releases have different purposes, targets, and need not relate to official business
    - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
- One Answer: texts used for different purposes

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
- One Answer: texts used for different purposes
- Partial answer: identify words that distinguish press releases and floor speeches

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

Mutual Information

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty
        - Unrelated: Conditional uncertainty = uncertainty
        - Perfect predictor: Conditional uncertainty = 0
- Mutual information(X_j): uncertainty - conditional uncertainty (X_j)

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty
        - Unrelated: Conditional uncertainty = uncertainty
        - Perfect predictor: Conditional uncertainty = 0
- Mutual information(X_j): uncertainty - conditional uncertainty (X_j)
    - Maximum: Uncertainty $\rightarrow$ X_j is perfect predictor

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0
- Mutual information(X_j): uncertainty - conditional uncertainty (X_j)
  - Maximum: Uncertainty $\rightarrow X_j$ is perfect predictor
  - Minimum: $0 \rightarrow X_j$ fails to separate speeches and floor statements

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre}, \text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre}, \text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits

A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits
- Maximum: $\Pr(\text{Press}) = \Pr(\text{Speech}) = 0.5$

A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits
- Maximum: $\Pr(\text{Press}) = \Pr(\text{Speech}) = 0.5$
- Minimum: $\Pr(\text{Press}) \rightarrow 0$ (or $\Pr(\text{Press}) \rightarrow 1$)

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

$$H(\text{Doc}|X_j) = -\sum_{s=0}^{1} \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

$$H(\text{Doc}|X_j) \;=\; -\sum_{s=0}^{1} \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

- Maximum: X_j unrelated to Press Releases/Floor Speeches

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

$$H(\text{Doc}|X_j) \;=\; -\sum_{s=0}^{1} \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

- Maximum: X_j unrelated to Press Releases/Floor Speeches
- Minimum: X_j is a perfect predictor of press release/floor speech

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

Bigger mutual information $\Rightarrow$ better discrimination

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) \quad = \quad H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

Bigger mutual information $\Rightarrow$ better discrimination

Objective function and optimization$\rightsquigarrow$ estimate probabilities that we then place in mutual information

A Method for Identifying Distinguishing Words

Formula for mutual information
(based on ML estimates of probabilities)

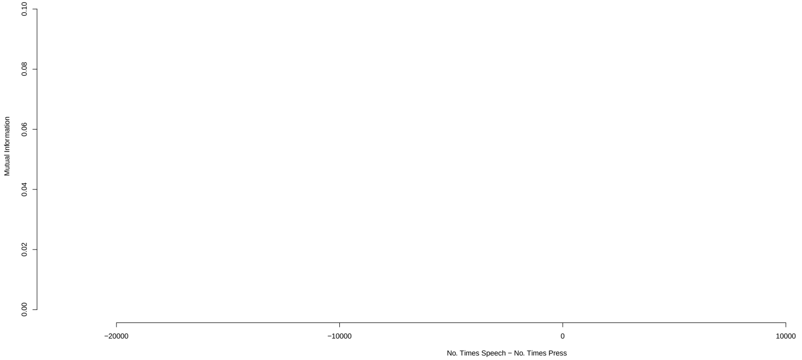
$$
\begin{aligned}
n_p &= \text{Number Press Releases} \\
n_s &= \text{Number of Speeches} \\
D &= n_p + n_s \\
n_j &= \sum_{i=1}^{D} X_{i,j} \qquad \text{(No. docs } X_j \text{ appears)} \\
n_{-j} &= \text{ No. docs } X_j \text{ does not appear} \\
n_{j,p} &= \text{No. press and } X_j \\
n_{j,s} &= \text{No. speech and } X_j \\
n_{-j,p} &= \text{No. press and not } X_j \\
n_{-j,s} &= \text{No. speech and not } X_j
\end{aligned}
$$

A Method for Identifying Distinguishing Words

Formula for Mutual Information

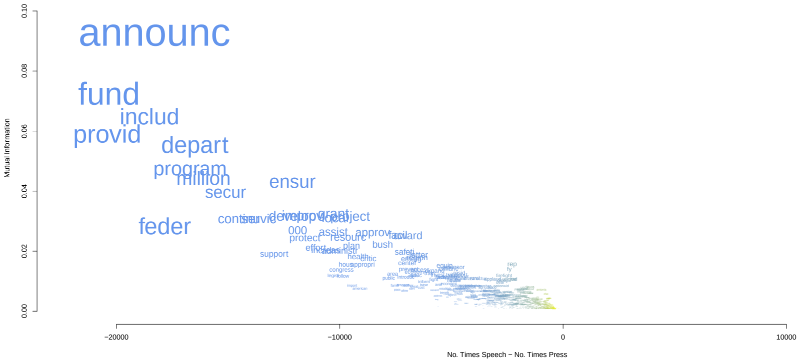
$$\begin{aligned} \mathrm{MI}(X_j) &= \frac{n_{j,p}}{D}\log_2 \frac{n_{j,p}D}{n_j n_p} + \frac{n_{j,s}}{D}\log_2 \frac{n_{j,s}D}{n_j n_s} \\ &\quad + \frac{n_{-j,p}}{D}\log_2 \frac{n_{-j,p}D}{n_{-j} n_p} + \frac{n_{-j,s}}{D}\log_2 \frac{n_{-j,s}D}{n_{-j} n_s}. \end{aligned}$$

What's Different About Press Releases



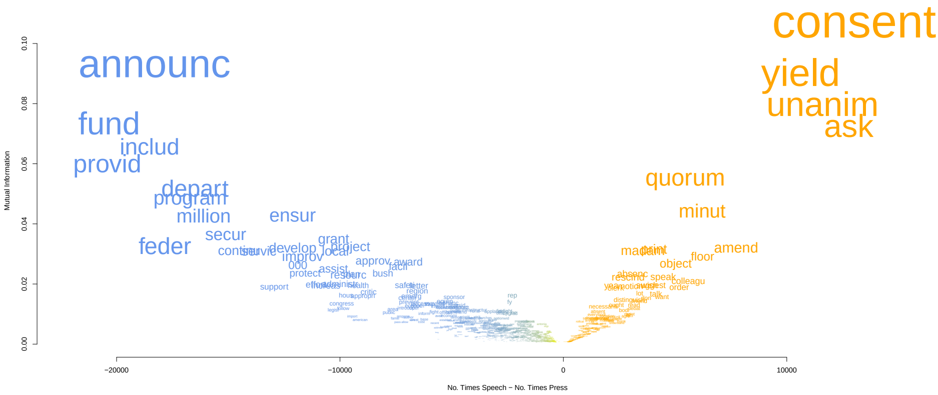
What's Different?

What's Different About Press Releases



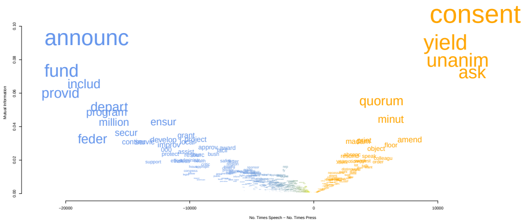
What's Different?

What's Different About Press Releases



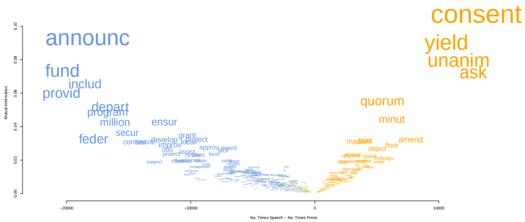
What's Different?

What's Different About Press Releases



What's Different?

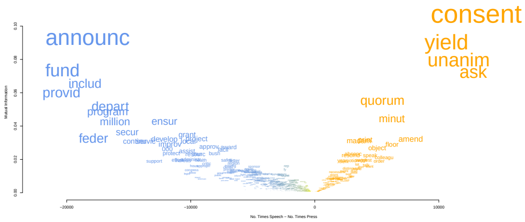
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming

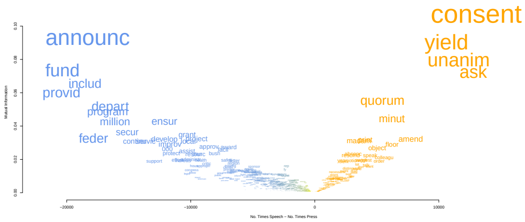
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words

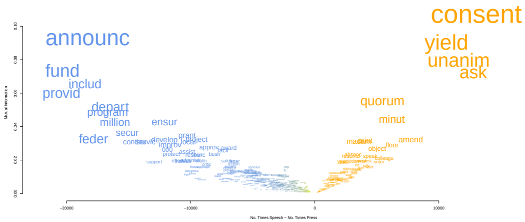
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification

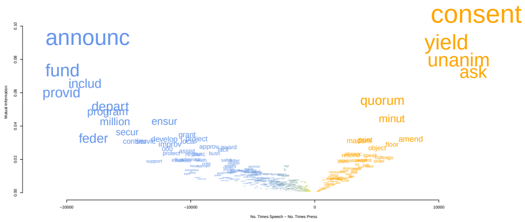
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches

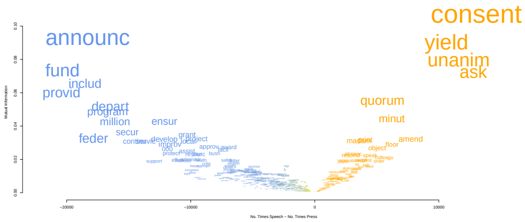
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches

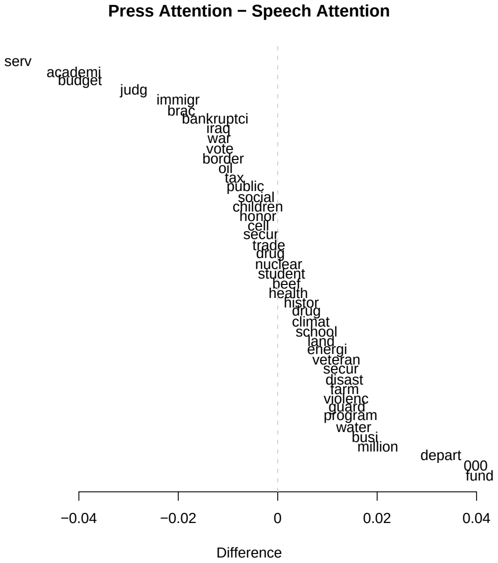
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches
- Procedural: 0% Press Releases, 44% Floor Speeches

What's Different About Press Releases



Fighting' Words$\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)$\rightsquigarrow$ what makes a word partisan?

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance

Fightin' Words$\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)$\rightsquigarrow$ what makes a word partisan?

Argue for using Log Odds Ratio, weighted by variance

Recall: For some event E and F

Fightin' Words $\leadsto$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\leadsto$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$P(E) = 1 - P(E^c)$$

Fightin' Words $\leadsto$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\leadsto$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$\begin{aligned} P(E) &= 1 - P(E^c) \\ \text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \end{aligned}$$

Fightin' Words$\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)$\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1-P(E))}}{\frac{P(F)}{1-P(F)}}
\end{aligned}$$

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$
\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1 - P(E))}}{\frac{P(F)}{1 - P(F)}} \\
\text{Log Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{aligned}
$$

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$
\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1 - P(E))}}{\frac{P(F)}{1 - P(F)}} \\
\text{Log Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{aligned}
$$

Strategy $\rightsquigarrow$ Construct objective function on *proportions* (and then calculate log-odds)

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$
\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1-P(E))}}{\frac{P(F)}{1-P(F)}} \\
\text{Log Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{aligned}
$$

Strategy $\rightsquigarrow$ Construct objective function on *proportions* (and then calculate log-odds)

Objective Function

Suppose we're interested in how a word separates partisan speech.
$\boldsymbol{Y} = (\text{Republican}, \text{Republican}, \text{Democrat}, \ldots, \text{Republican})$
$\boldsymbol{X} =$ Unnormalized matrix of word counts $N \times J$
Define

$$\begin{aligned}
\boldsymbol{x}_{\text{Republican}} &= \Big(\sum_{i=1}^{N} I(Y_i = \text{Republican})X_{i1}, \sum_{i=1}^{N} I(Y_i = \text{Republican})X_{i2}, \\
&\quad \ldots, \sum_{i=1}^{N} I(Y_i = \text{Republican})X_{iJ}\Big)
\end{aligned}$$

with $N_{\text{Republican}} =$ Total number of Republican words

Objective Function

Objective Function

$$\boldsymbol{\pi}_{\text{Republican}} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Objective Function

$$
\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}} | \boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}
$$

Objective Function

$$
\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}} | \boldsymbol{\pi}_{\text{Republican}} & \sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}
$$

This implies an objective function on $\boldsymbol{\pi}$,

Objective Function

$$\begin{aligned} \boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}}) \end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y})$$

Objective Function

$$\begin{aligned} \boldsymbol{\pi}_{\text{Republican}} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_{\text{Republican}} | \boldsymbol{\pi}_{\text{Republican}} & \sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}}) \end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) & \propto p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\ & \propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}} \end{aligned}$$

Objective Function

$$\begin{aligned}\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})\end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned}p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\ &\propto \frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}}\end{aligned}$$

$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y})$ is a Dirichlet distribution:

Objective Function

$$\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\
&\propto \frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}}
\end{aligned}$$

$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y})$ is a Dirichlet distribution:

$$\pi^{*}_{\text{Republican},j} = \frac{x_{\text{Republican},j} + \alpha_j}{N_{\text{Republican}} + \sum_{j=1}^{J}\alpha_j}$$

Objective Function

$$\begin{aligned} \boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}}) \end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\ &\propto \frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}} \end{aligned}$$

$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y})$ is a Dirichlet distribution:

$$\pi^{*}_{\text{Republican},j} = \frac{x_{\text{Republican},j} + \alpha_j}{N_{\text{Republican}} + \sum_{j=1}^{J}\alpha_j}$$

Calculating Log Odds Ratio

Define $\log \text{Odds Ratio}_j$ as

$$\log \text{Odds Ratio}_j = \log\left(\frac{\pi_{\text{Republican},j}}{1-\pi_{\text{Republican},j}}\right) - \log\left(\frac{\pi_{\text{Democratic},j}}{1-\pi_{\text{Democratic},j}}\right)$$

$$\begin{aligned}
\text{Var}(\log \text{Odds Ratio}_j) &\approx \frac{1}{x_{jD}+\alpha_j} + \frac{1}{x_{jR}+\alpha_j} \\
\text{Std. Log Odds}_j &= \frac{\log \text{Odds Ratio}_j}{\sqrt{\text{Var}(\log \text{Odds Ratio}_j)}}
\end{aligned}$$

Applying the Model

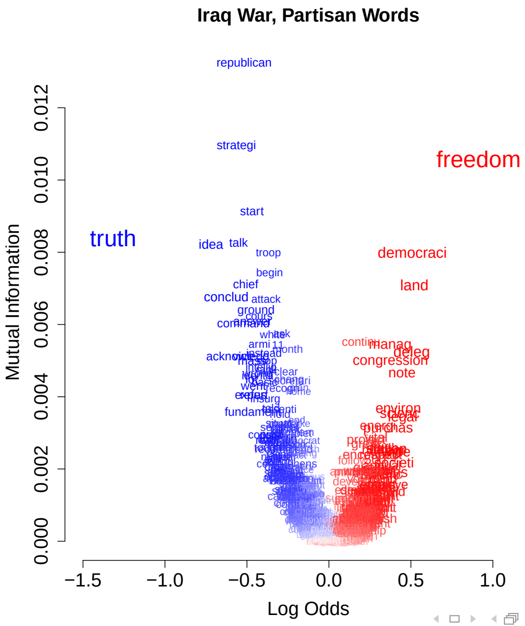
`https://gist.github.com/thiagomarzagao/5851207`

How do Republicans and Democrats differ in debate?

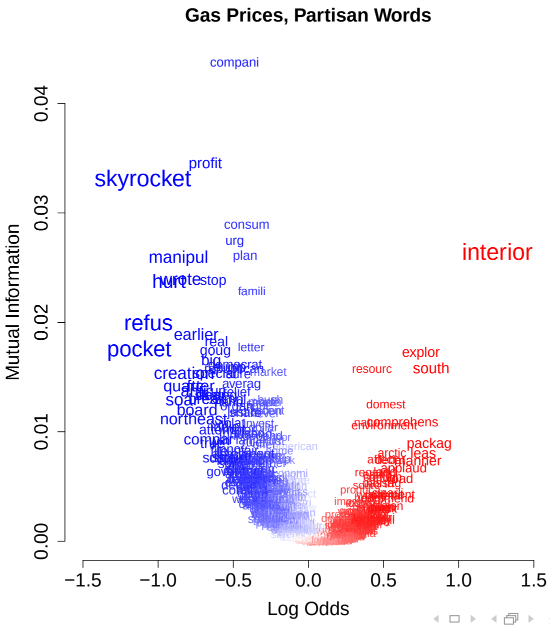
Condition on topic and examine word usage

- Press Releases (64,033)
- Topic Coded
- Given press release is about topic, what are the features that distinguish Republican and Democratic language?

Mutual Information, Standardized Log Odds



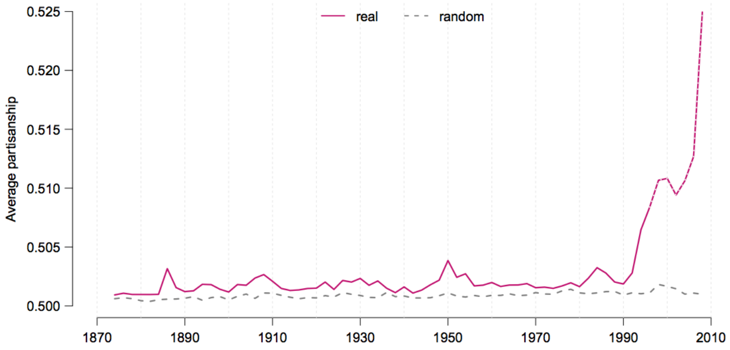
Mutual Information, Standardized Log Odds



Gentzkow, Shapiro, and Taddy (2017): Rhetorical Polarization

Figure 3: Average partisanship of speech, penalized estimates

Panel A: Preferred specification



Discovery

Where do concepts/ideas/questions come from?

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from
  does

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from
  does.not

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from
  does.not.matter