Political Science 452: Text as Data

Justin Grimmer

Assistant Professor
Department of Political Science
Stanford University

April 27th, 2011

Where We've Been, Where We're Going

- Class 1: Finding Text Data
- Class 2: Representing Texts Quantitatively
- Class 3: Dictionary Methods for Classification
- Class 4: Comparing Language Across Groups
- Class 5: Texts in Space
- Class 6: Clustering
- Class 7: Topic models
- Class 8: Supervised methods for classification
- Class 9: Ensemble methods for classification
- Class 10: Scaling Speech

Question (from email received 1 hour ago):

```
I'm curious if you have ever used mechanical turk for
coding of data (e.g., from text).  Any experience with
that?  Thoughts?
```

Question (from email received 1 hour ago):

```
I'm curious if you have ever used mechanical turk for
coding of data (e.g., from text).  Any experience with
that?  Thoughts?
```

How is Homework Going? Class? What Can I do to help you?

More About R Code

How to write to a file in R

Many method, easiest: sink

```
> sink('Test.txt')
> print('This is a great tool')
> sink()
```

Congressional Language Across Sources

Congregational Language Across Sources

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

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

Today's Lecture: How to identify those words?

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

Method 1: Mutual Information

A Method for Identifying Distinguishing Words

Method 1: Mutual Information

- Unconditional uncertainty (entropy):

A Method for Identifying Distinguishing Words

Method 1: Mutual Information

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

A Method for Identifying Distinguishing Words

Method 1: Mutual Information

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

A Method for Identifying Distinguishing Words

Method 1: Mutual Information

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

A Method for Identifying Distinguishing Words

Method 1: 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

Method 1: 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

Method 1: 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 (w) (conditional entropy)

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w
    - Randomly sample a press release

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w
    - Randomly sample a press release
    - Guess press release/floor statement

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
  - Condition on presence of word w
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
  - Condition on presence of word w
  - 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

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w
    - 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(w): uncertainty - conditional uncertainty (w)

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w
    - 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(w): uncertainty - conditional uncertainty (w)
    - Maximum: Uncertainty $\rightarrow$ w is perfect predictor

A Method for Identifying Distinguishing Words

Method 1: 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 (w) (conditional entropy)
    - Condition on presence of word w
    - 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(w): uncertainty - conditional uncertainty (w)
    - Maximum: Uncertainty $\rightarrow$ w is perfect predictor
    - Minimum: 0 $\rightarrow$ w 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 $H(k)$

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech
- Define entropy $H(k)$

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

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 $H(k)$

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

- $\log_2$? 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 $H(k)$

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

- $\log_2$? 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 $H(k)$

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

- $\log_2$? 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 w_j

A Method for Identifying Distinguishing Words

- Consider presence/absence of word w_j
- Define conditional entropy $H(k|w_j)$

A Method for Identifying Distinguishing Words

- Consider presence/absence of word w_j
- Define conditional entropy $H(k|w_j)$

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

A Method for Identifying Distinguishing Words

- Consider presence/absence of word w_j
- Define conditional entropy $H(k|w_j)$

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

- Maximum: w_j unrelated to Press Releases/Floor Speeches

A Method for Identifying Distinguishing Words

- Consider presence/absence of word w_j
- Define conditional entropy $H(k|w_j)$

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

- Maximum: w_j unrelated to Press Releases/Floor Speeches
- Minimum: w_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(w_j) as

A Method for Identifying Distinguishing Words

- Define Mutual Information(w_j) as

$$\text{Mutual Information}(w_j) = H(k) - H(k|w_j)$$

A Method for Identifying Distinguishing Words

- Define Mutual Information(w_j) as

$$\text{Mutual Information}(w_j) = H(k) - H(k|w_j)$$

- Maximum: entropy $\Rightarrow H(k|w_j) = 0$

A Method for Identifying Distinguishing Words

- Define Mutual Information(w_j) as

$$\text{Mutual Information}(w_j) \quad = \quad H(k) - H(k|w_j)$$

- Maximum: entropy $\Rightarrow H(k|w_j) = 0$
- Minimum: $0 \Rightarrow H(k|w_j) = H(k)$.

A Method for Identifying Distinguishing Words

- Define Mutual Information(w_j) as

$$\text{Mutual Information}(w_j) \quad = \quad H(k) - H(k|w_j)$$

- Maximum: entropy $\Rightarrow H(k|w_j) = 0$
- Minimum: $0 \Rightarrow H(k|w_j) = H(k)$.

Bigger mutual information $\Rightarrow$ better discrimination

A Method for Identifying Distinguishing Words

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

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

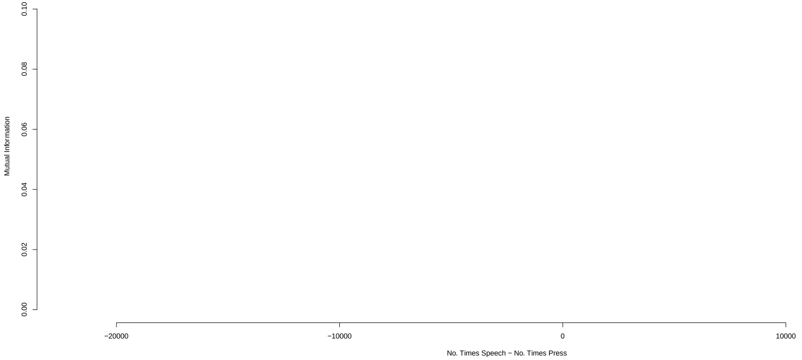
A Method for Identifying Distinguishing Words

Formula for Mutual Information

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

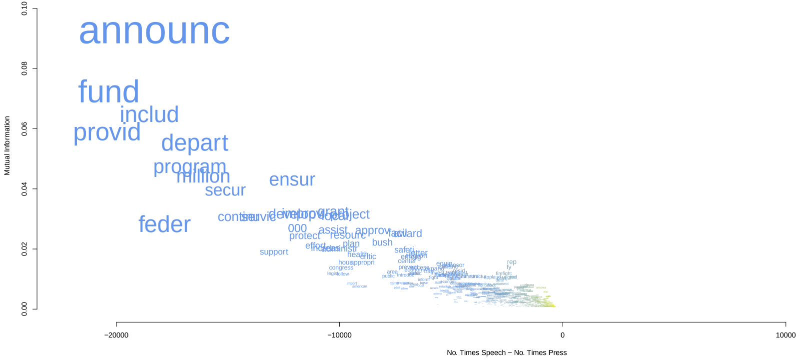
(Page 258, 259 of this document
`http://stanford.edu/~jgrimmer/RepStyle.pdf` for more information
)

What's Different About Press Releases



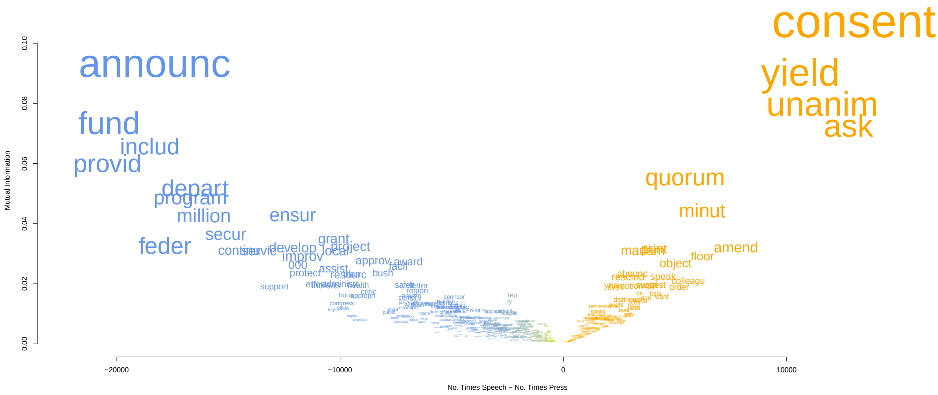
What's Different?

What's Different About Press Releases



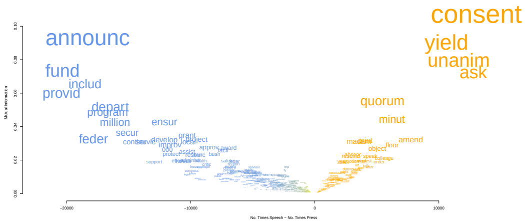
What's Different?

What's Different About Press Releases



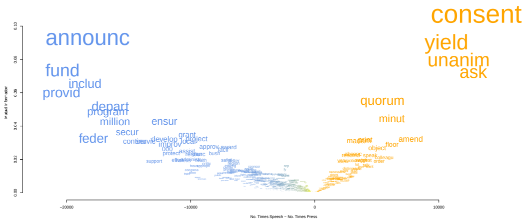
What's Different?

What's Different About Press Releases



What's Different?

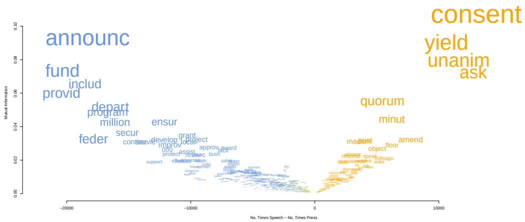
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming

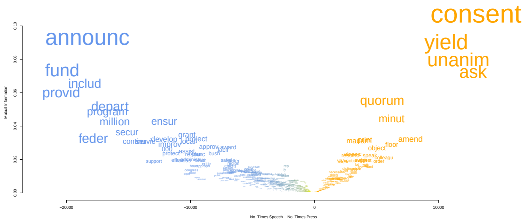
What's Different About Press Releases



What's Different?

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

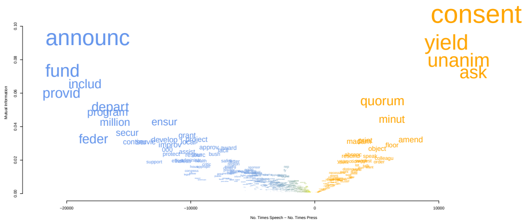
What's Different About Press Releases



What's Different?

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

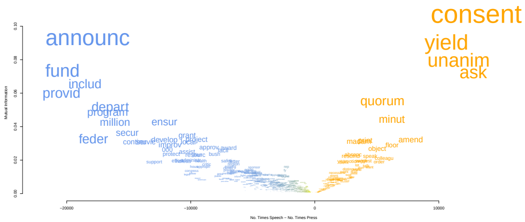
What's Different About Press Releases



What's Different?

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

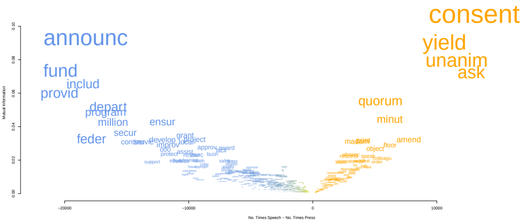
What's Different About Press Releases



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

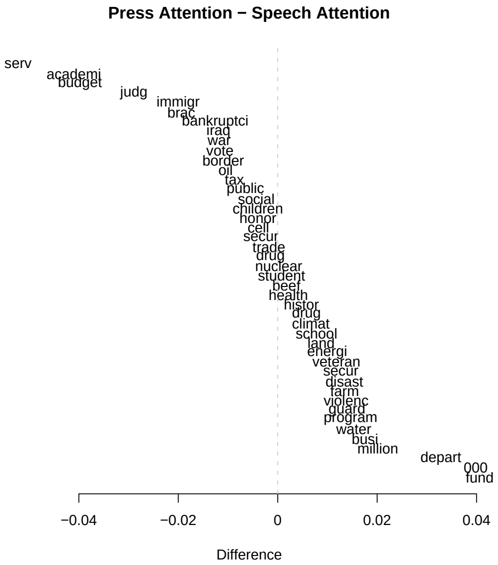
What's Different About Press Releases



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



General Idea

General Idea

- What we know: document labels

General Idea

- What we know: document labels
  - Certain
  - Complete

General Idea

- What we know: document labels
  - Certain
  - Complete
- Inference: discriminating words

General Idea

- What we know: document labels
  - Certain
  - Complete
- Inference: discriminating words
  - Words that separate classes

General Idea

- What we know: document labels
  - Certain
  - Complete
- Inference: discriminating words
  - Words that separate classes
- All methods presented today:

General Idea

- What we know: document labels
    - Certain
    - Complete
- Inference: discriminating words
    - Words that separate classes
- All methods presented today:
    - Know labels

General Idea

- What we know: document labels
  - Certain
  - Complete
- Inference: discriminating words
  - Words that separate classes
- All methods presented today:
  - Know labels
  - Infer words

General Idea

- What we know: document labels
    - Certain
    - Complete
- Inference: discriminating words
    - Words that separate classes
- All methods presented today:
    - Know labels
    - Infer words
- All methods last week:

General Idea

- What we know: document labels
  - Certain
  - Complete
- Inference: discriminating words
  - Words that separate classes
- All methods presented today:
  - Know labels
  - Infer words
- All methods last week:
  - Know words

General Idea

- What we know: document labels
  - Certain
  - Complete
- Inference: discriminating words
  - Words that separate classes
- All methods presented today:
  - Know labels
  - Infer words
- All methods last week:
  - Know words
  - Infer labels

Why Infer Separating Words?

Why do we care?

Why Infer Separating Words?

Why do we care?
Social Science Inference:

Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language

Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language
- Differences in Liberal, Conservative Language

Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language
- Differences in Liberal, Conservative Language
- Differences in Campaign Agendas

Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language
- Differences in Liberal, Conservative Language
- Differences in Campaign Agendas
- Different Advice to Muslim and Christian Kings

Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language
- Differences in Liberal, Conservative Language
- Differences in Campaign Agendas
- Different Advice to Muslim and Christian Kings
- Recommendation Letters for Men and Women?

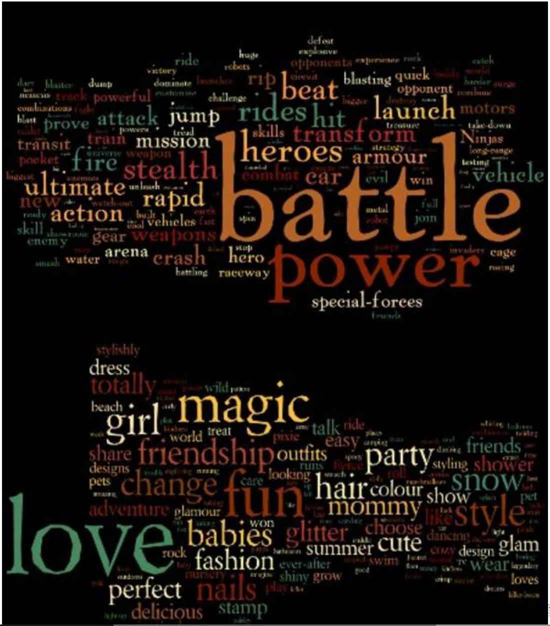
Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language
- Differences in Liberal, Conservative Language
- Differences in Campaign Agendas
- Different Advice to Muslim and Christian Kings
- Recommendation Letters for Men and Women?
- Toy Advertising for Boys and Girls?

Why Infer Separating Words?



Why Infer Separating Words?

Why do we care?

Social Science Inference:

- Differences in Republican, Democrat Language
- Differences in Liberal, Conservative Language
- Differences in Campaign Agendas
- Different Advice to Muslim and Christian Kings
- Recommendation Letters for Men and Women?
- Toy Advertising for Boys and Girls?
- Beginning of Inference

Why Infer Separating Words?

Why Infer Separating Words?

Labeling

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

- Training Set: use documents to identify separating words

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

- Training Set: use documents to identify separating words
- Test Set: Validate separating words

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

- Training Set: use documents to identify separating words
- Test Set: Validate separating words

Improve Supervised Learning Classification (Weeks 8, 9)

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

- Training Set: use documents to identify separating words
- Test Set: Validate separating words

Improve Supervised Learning Classification (Weeks 8, 9)

- Usually: more information, better

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

- Training Set: use documents to identify separating words
- Test Set: Validate separating words

Improve Supervised Learning Classification (Weeks 8, 9)

- Usually: more information, better
- Reality: rare words can cause over fitting

Why Infer Separating Words?

Labeling

- Methods to estimate classes (Week 6 and 7)
- Label classes: why grouped together?

Dictionary Creation

- Training Set: use documents to identify separating words
- Test Set: Validate separating words

Improve Supervised Learning Classification (Weeks 8, 9)

- Usually: more information, better
- Reality: rare words can cause over fitting
- Feature selection: one method to mitigate over fitting

Methods for Inference/Labeling

Methods for Inference/Labeling

Task 1: Well defined

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Take 2: Vague

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Take 2: Vague

- Generate intuition about differences

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Take 2: Vague

- Generate intuition about differences
- Use this intuition then to investigate claims

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Take 2: Vague

- Generate intuition about differences
- Use this intuition then to investigate claims
- Intuition very hard to formalize

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Take 2: Vague

- Generate intuition about differences
- Use this intuition then to investigate claims
- Intuition very hard to formalize

$\rightsquigarrow$ Very difficult (impossible) to derive optimal method a priori

Methods for Inference/Labeling

Task 1: Well defined

- Identifying maximally discriminating words
- Objective function for discrimination
- Identify each word's discrimination

$\rightsquigarrow$ We can derive optimal method, given objective

Take 2: Vague

- Generate intuition about differences
- Use this intuition then to investigate claims
- Intuition very hard to formalize

$\rightsquigarrow$ Very difficult (impossible) to derive optimal method a priori

Be skeptical!

Running Example

How Do Democrat and Republican Arguments About the Iraq War Differ?

- Assume: Identified set of documents (press releases) about Iraq War
- Speaker labels: know who (Democrat, Republican) issued press release
- Inferential Goal: framing–considerations Democrats and Republicans use when discussing war

Present simple methods, show similarity.
The example already has stop words and some names removed.

Methods for Identifying Words

(Following steps are from Fightin' Words)
Difference in word frequency:
For each word j compute

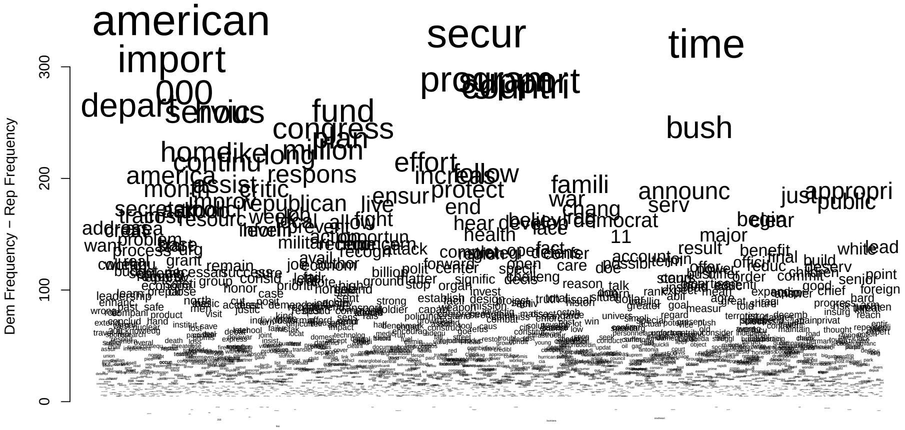
$$
\begin{aligned}
n_{jD} &= \text{No. times used in Dem Documents} \\
n_{jR} &= \text{No. times used in Rep Documents}
\end{aligned}
$$

Difference $= n_{jD} - n_{jR}$

Methods for Identifying Words

(Following steps are from Fightin' Words)

Difference $= n_{jD} - n_{jR}$



Methods for Identifying Words

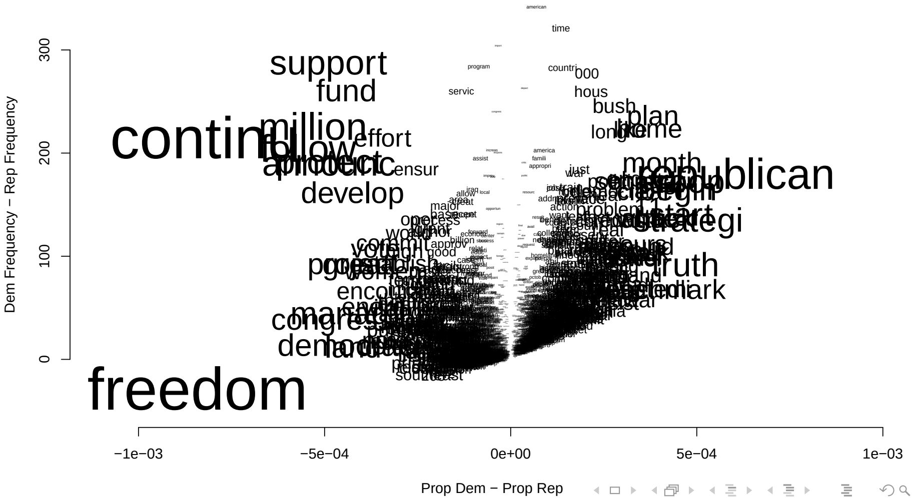
Differences in Word Proportions:
For each word j compute

$$\begin{aligned}
p_{jD} &= \frac{n_{jD}}{n_D} \\
&= \text{Proportion of Dem words that are } j \\
p_{jR} &= \frac{n_{jR}}{n_R} \\
&= \text{Proportion of Rep words that are } j
\end{aligned}$$

Difference $= p_{jD} - p_{jR}$

Methods for Identifying Words

Difference $= p_{jD} - p_{jR}$



Methods for Identifying Words

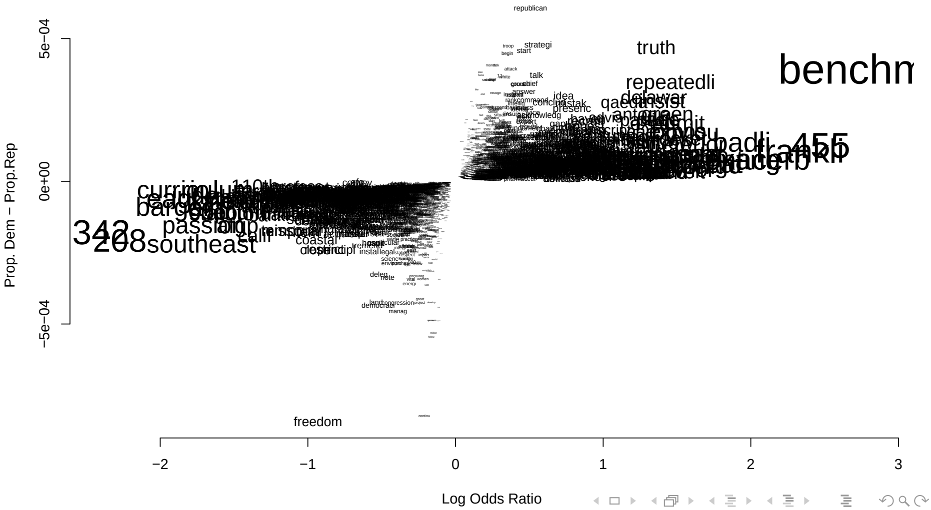
Log Odds Ratio:
For each word j compute:

$$\begin{aligned}
\text{Odds}_{jD} &= \frac{p_{jD}}{1 - p_{jD}} \\
\text{Odds}_{jR} &= \frac{p_{jR}}{1 - p_{jR}} \\
\text{Odds Ratio}_{j} &= \frac{\text{Odds}_{jD}}{\text{Odds}_{jR}}
\end{aligned}$$

$\log \text{Odds Ratio}_j = \log \text{Odds}_{jD} - \log \text{Odds}_{jR}$

Methods for Identifying Words

$$\log \text{Odds Ratio}_j = \log \text{Odds}_{jD} - \log \text{Odds}_{jR}$$



Methods for Identifying Words

Problem: What to Do With Dem (GOP) Only Words?
If Only Dems Use Words:

$$
\begin{aligned}
p_{jR} &= \frac{0}{n_R} \\
\text{Odds}_{jR} &= \frac{0}{1} \\
\log \text{Odds}_{jR} &= \log 0 - \log 1
\end{aligned}
$$

What should we do?

Methods for Identifying Words

Solution: "add" a little, but in a principled way

Methods for Identifying Words

Solution: "add" a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Methods for Identifying Words

Solution: "add" a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Notation:

$$\begin{aligned}
\mathbf{p}_R &= (p_{1R}, p_{2R}, \ldots, p_{NR}) \\
\mathbf{p}_D &= (p_{1D}, p_{2D}, \ldots, p_{ND})
\end{aligned}$$

Methods for Identifying Words

Solution: "add" a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Notation:

$$
\begin{aligned}
\mathbf{p}_R &= (p_{1R}, p_{2R}, \ldots, p_{NR}) \\
\mathbf{p}_D &= (p_{1D}, p_{2D}, \ldots, p_{ND}) \\
\mathbf{y}_D &\sim \text{Multinomial}(n_D, \mathbf{p}_D)
\end{aligned}
$$

Methods for Identifying Words

Solution: "add" a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Notation:

$$\begin{aligned}
\mathbf{p}_R &= (p_{1R}, p_{2R}, \ldots, p_{NR}) \\
\mathbf{p}_D &= (p_{1D}, p_{2D}, \ldots, p_{ND}) \\
\mathbf{y}_D &\sim \text{Multinomial}(n_D, \mathbf{p}_D)
\end{aligned}$$

Prior

$$\begin{aligned}
\mathbf{p}_R &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\mathbf{p}_D &\sim \text{Dirichlet}(\boldsymbol{\alpha})
\end{aligned}$$

Methods for Identifying Words

Solution: "add" a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Notation:

$$\begin{aligned}
\mathbf{p}_R &= (p_{1R}, p_{2R}, \ldots, p_{NR}) \\
\mathbf{p}_D &= (p_{1D}, p_{2D}, \ldots, p_{ND}) \\
\mathbf{y}_D &\sim \text{Multinomial}(n_D, \mathbf{p}_D)
\end{aligned}$$

Prior

$$\begin{aligned}
\mathbf{p}_R &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\mathbf{p}_D &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_N)
\end{aligned}$$

Methods for Identifying Words

Solution: “add” a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Notation:

$$
\begin{aligned}
\mathbf{p}_R &= (p_{1R}, p_{2R}, \ldots, p_{NR}) \\
\mathbf{p}_D &= (p_{1D}, p_{2D}, \ldots, p_{ND}) \\
\mathbf{y}_D &\sim \text{Multinomial}(n_D, \mathbf{p}_D)
\end{aligned}
$$

Prior

$$
\begin{aligned}
\mathbf{p}_R &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\mathbf{p}_D &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_N)
\end{aligned}
$$

Before showing why this “adds” a little.

Methods for Identifying Words

Solution: "add" a little, but in a principled way

We need a model!: Intro to Bayes in 10 minutes

Notation:

$$
\begin{aligned}
\mathbf{p}_R &= (p_{1R}, p_{2R}, \ldots, p_{NR}) \\
\mathbf{p}_D &= (p_{1D}, p_{2D}, \ldots, p_{ND}) \\
\mathbf{y}_D &\sim \text{Multinomial}(n_D, \mathbf{p}_D)
\end{aligned}
$$

Prior

$$
\begin{aligned}
\mathbf{p}_R &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\mathbf{p}_D &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_N)
\end{aligned}
$$

Before showing why this "adds" a little.

Let me teach you how to Dirichlet

Dirichlet Distribution

Distribution over proportions.

$$
\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
p(\boldsymbol{\pi}|\boldsymbol{\alpha}) &= \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma\alpha_j} \prod_{j=1}^{N} \pi_j^{\alpha_j - 1}
\end{aligned}
$$

Facts:

$$
\begin{aligned}
E[\pi_j] &= \frac{\alpha_j}{\sum_{k=1}^{N} \alpha_k} \\
\text{Variance}[\pi_j] &= \frac{E[\pi_j](1 - E[\pi_j])}{\sum_{k=1}^{N}(\alpha_k) + 1}
\end{aligned}
$$

Conjugate to Multinomial : easily apply to the model

Methods for Identifying Separating Words

$$\begin{aligned}
\mathbf{p}_D|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\mathbf{y}_D|\mathbf{p}_D &\sim \text{Multinomial}(n_D, \mathbf{p}_D)
\end{aligned}$$

Conjugacy implies

$$\begin{aligned}
\mathbf{p}_D|\mathbf{y}_D, \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}^{'}) \\
\alpha_j^{'} &= y_{jD} + \alpha_j \\
E[p_{j,D}] &= \frac{y_{jD} + \alpha_j^{'}}{n_D + \sum_{k=1}^{N} \alpha_k^{'}}
\end{aligned}$$

Smoothing (borrowing information): easy to understand in Bayesian framework, take Simon's class

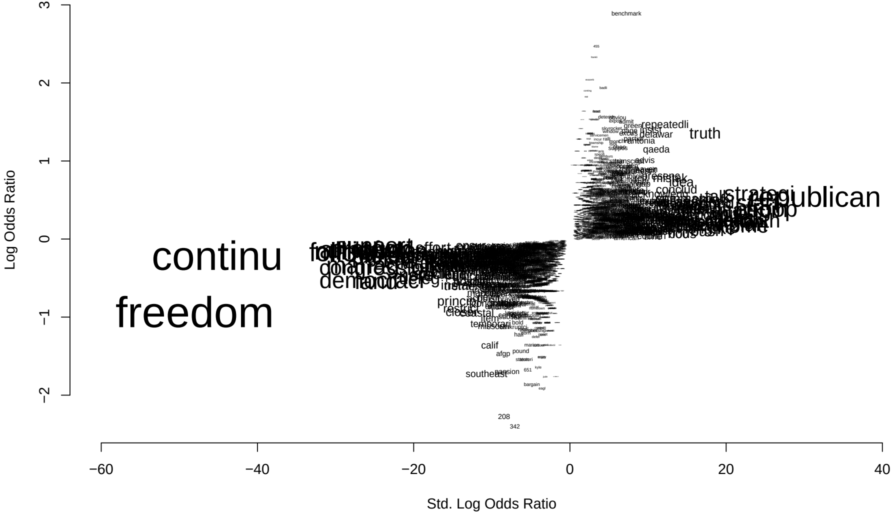
Methods for Identifying Separating Words

Now, we can compute all log-odds.
But same problem: rare words dominate
Solution: include estimate of variance

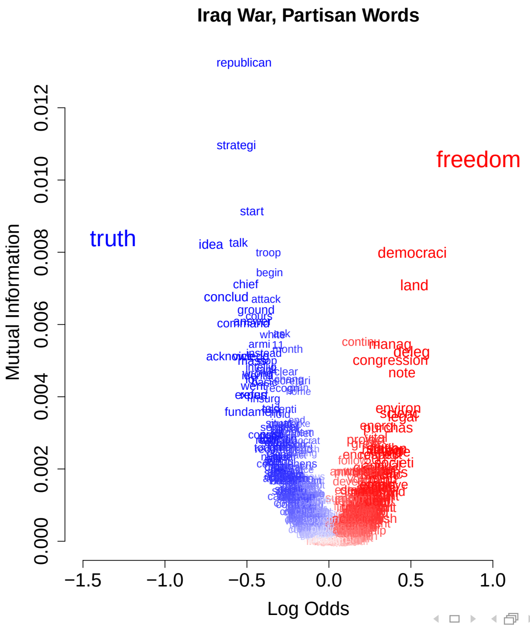
$$\begin{aligned} \text{Var}(\log \text{Odds Ratio}_j) &\approx \frac{1}{y_{jD} + \alpha_j} + \frac{1}{y_{jR} + \alpha_j} \\ \text{Std. Log Odds}_j &= \frac{\log \text{Odds Ratio}_j}{\sqrt{\text{Var}(\log \text{Odds Ratio}_j)}} \end{aligned}$$

Analogues from Contingency Tables
Key Idea:
Systematic or Random Difference

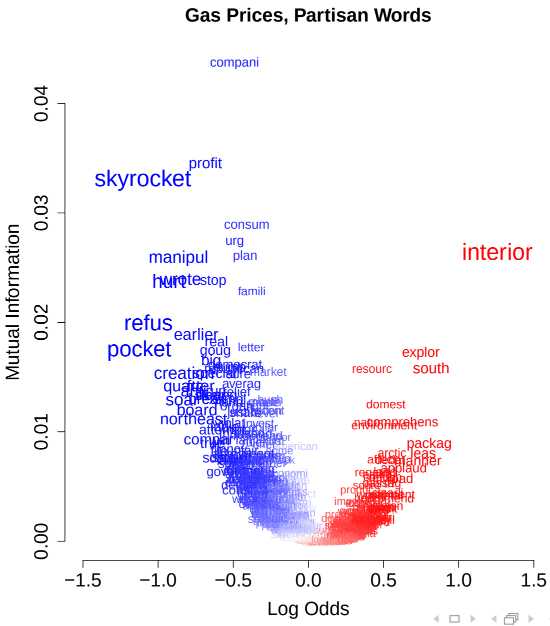
Methods For Identifying Words



Mutual Information, Standardized Log Odds



Mutual Information, Standardized Log Odds



Methods for Identifying Words

There are many other similar methods

- Difference in standardized proportions
- χ^2 statistics
- Pointwise Mutual Information
- ...

Characteristics:

- Definition of separation
- Word by word test of separation
- Providing rank ordering of words
- Best Method: depends on context, intuition provided

Moving Forward

- Considered word by word methods solely
- During supervised classification, we will consider joint separability
  - Conditional on other words, how much more information does this word provide

Next Week:

- Geometry of texts
- Foundation for clustering
- topic modeling
- supervised classification