

# Introduction to Machine Learning for Social Science

## Class 11: Distinctive / Discriminating Words

Rochelle Terman

Postdoctoral Fellow  
Center for International Security Cooperation  
Stanford University

February 20, 2018

## Last class: Dictionary methods

- Use lists of words to score documents, e.g. positive or negative.

Last class: Dictionary methods

- Use lists of words to score documents, e.g. positive or negative.

Problem: How do generate dictionaries?

## Last class: Dictionary methods

- Use lists of words to score documents, e.g. positive or negative.

## Problem: How do generate dictionaries?

- Manually

## Last class: Dictionary methods

- Use lists of words to score documents, e.g. positive or negative.

## Problem: How do generate dictionaries?

- Manually
- Crowd sourcing

## Last class: Dictionary methods

- Use lists of words to score documents, e.g. positive or negative.

## Problem: How do generate dictionaries?

- Manually
- Crowd sourcing
- **Statistical methods**

## Last class: Dictionary methods

- Use lists of words to score documents, e.g. positive or negative.

## Problem: How do generate dictionaries?

- Manually
- Crowd sourcing
- **Statistical methods**  $\rightsquigarrow$  **discriminating words**

# Discriminating Words

**Goal:** Find words that distinguish one group of texts from another group of texts.



# Discriminating Words

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language

# Discriminating Words

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing Liberal, Conservative books  $\rightsquigarrow$  **Ideological** language

# Discriminating Words

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing Liberal, Conservative books  $\rightsquigarrow$  **Ideological** language
- Comparing in Toy advertising  $\rightsquigarrow$  **Gendered** language



**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing liberal, conservative media  $\rightsquigarrow$  **Ideological** language
- Comparing in toy advertising  $\rightsquigarrow$  **Gendered** language

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing liberal, conservative media  $\rightsquigarrow$  **Ideological** language
- Comparing in toy advertising  $\rightsquigarrow$  **Gendered** language

**Reasons:**

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing liberal, conservative media  $\rightsquigarrow$  **Ideological** language
- Comparing in toy advertising  $\rightsquigarrow$  **Gendered** language

**Reasons:**

- 1) Interesting in their own right

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing liberal, conservative media  $\rightsquigarrow$  **Ideological** language
- Comparing in toy advertising  $\rightsquigarrow$  **Gendered** language

**Reasons:**

- 1) Interesting in their own right
- 2) Create custom dictionaries for classification task.



**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing liberal, conservative media  $\rightsquigarrow$  **Ideological** language
- Comparing in toy advertising  $\rightsquigarrow$  **Gendered** language

**Reasons:**

- 1) Interesting in their own right
- 2) Create custom dictionaries for classification task.
- 3) Feature selection: inclusion of features in some subsequent analysis

**Goal:** Find words that distinguish one group of texts from another group of texts.

- Comparing Republican, Democratic speeches  $\rightsquigarrow$  **Partisan** language
- Comparing liberal, conservative media  $\rightsquigarrow$  **Ideological** language
- Comparing in toy advertising  $\rightsquigarrow$  **Gendered** language

**Reasons:**

- 1) Interesting in their own right
- 2) Create custom dictionaries for classification task.
- 3) Feature selection: inclusion of features in some subsequent analysis

**Method:** Distinctive / Discriminating / Separating word scores.

# Preparing the Corpus

Our corpus: 6 novels by two authors, Jane Austen and Charlotte Bronte.

First create a DTM of the corpus.

# What does “distinctive” mean?

- **Goal:** find words (or features) distinctive to each corpus.

To the R code!

# What does “distinctive” mean?

- **Goal:** find words (or features) distinctive to each corpus.
- Requires a decision about what “distinctive” means.

To the R code!

# What does “distinctive” mean?

- **Goal:** find words (or features) distinctive to each corpus.
- Requires a decision about what “distinctive” means.
- There are a variety of definitions that we might use.

To the R code!

# Option 1: Unique usage

- Distinctive = exclusive

# Option 1: Unique usage

- **Distinctive = exclusive**
- If Bront uses the word “access” and Austen never does, we should count “access” as distinctive.



# Option 1: Unique usage

- **Distinctive = exclusive**
- If Bront uses the word “access” and Austen never does, we should count “access” as distinctive.
- These words tend not to be terribly interesting or informative

## Option 2: Difference in Frequencies

- Distinctive = difference in frequency

## Option 2: Difference in Frequencies

- Distinctive = difference in frequency
- Compare the number of times each author uses a word

## Option 2: Difference in Frequencies

- **Distinctive = difference in frequency**
- Compare the number of times each author uses a word
- Find the largest absolute difference.

## Option 2: Difference in Frequencies

- **Distinctive = difference in frequency**
- Compare the number of times each author uses a word
- Find the largest absolute difference.
- Doesn't take into account difference in total words.

## Option 3: Difference in Averages

- Distinctive = difference in rates

## Option 3: Difference in Averages

- Distinctive = difference in rates
- Compare the average rate each author uses a word

## Option 3: Difference in Averages

### 1 Normalize DTM from counts to proportions:

For each word  $p$  in an arbitrary corpus  $c$ :

$$\mu_p = \frac{\sum_{i=1}^N p_i}{T}$$

where  $p_i$  is the number of times a  $p$  appears in document  $i$ ,  $N$  is the total number of documents in  $c$  and  $T$  is the total number of words in  $c$ .



## Option 3: Difference in Averages

- 1 Normalize DTM from counts to proportions:

For each word  $p$  in an arbitrary corpus  $c$ :

$$\mu_p = \frac{\sum_{i=1}^N p_i}{T}$$

where  $p_i$  is the number of times a  $p$  appears in document  $i$ ,  $N$  is the total number of documents in  $c$  and  $T$  is the total number of words in  $c$ .

- 2 Take the difference between one author's proportion of a word and another's proportion of the same word.

$$\theta_p = \mu_{p,Bronte} - \mu_{p,Austen}$$

## Option 3: Difference in Averages

- 1 Normalize DTM from counts to proportions:

For each word  $p$  in an arbitrary corpus  $c$ :

$$\mu_p = \frac{\sum_{i=1}^N p_i}{T}$$

where  $p_i$  is the number of times a  $p$  appears in document  $i$ ,  $N$  is the total number of documents in  $c$  and  $T$  is the total number of words in  $c$ .

- 2 Take the difference between one author's proportion of a word and another's proportion of the same word.

$$\theta_p = \mu_{p,Bronte} - \mu_{p,Austen}$$

- 3 Find words with highest absolute difference.

# Difference in Averages: Problems

- Favors more frequent words.

# Difference in Averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Bronte); 25/1000 (Austen)  $\rightsquigarrow$  Score: 5/1000.

# Difference in Averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Bronte); 25/1000 (Austen)  $\rightsquigarrow$  Score: 5/1000.
- Word 2: 5/1000 (Bronte); .1/1000 (Austen)  $\rightsquigarrow$  Score: 4.9/1000.

# Difference in Averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Bronte); 25/1000 (Austen)  $\rightsquigarrow$  Score: 5/1000.
- Word 2: 5/1000 (Bronte); .1/1000 (Austen)  $\rightsquigarrow$  Score: 4.9/1000.
- Ignores cases when one author uses a word frequently and a another author barely uses it.

# Difference in Averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Bronte); 25/1000 (Austen)  $\rightsquigarrow$  Score: 5/1000.
- Word 2: 5/1000 (Bronte); .1/1000 (Austen)  $\rightsquigarrow$  Score: 4.9/1000.
- Ignores cases when one author uses a word frequently and a another author barely uses it.
- More generally: Differences in rates of frequent words  $>$  Differences in rates of rare words.

# Difference in Averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Bronte); 25/1000 (Austen)  $\rightsquigarrow$  Score: 5/1000.
- Word 2: 5/1000 (Bronte); .1/1000 (Austen)  $\rightsquigarrow$  Score: 4.9/1000.
- Ignores cases when one author uses a word frequently and a another author barely uses it.
- More generally: Differences in rates of frequent words  $>$  Differences in rates of rare words.



# Difference in Averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Bronte); 25/1000 (Austen)  $\rightsquigarrow$  Score: 5/1000.
- Word 2: 5/1000 (Bronte); .1/1000 (Austen)  $\rightsquigarrow$  Score: 4.9/1000.
- Ignores cases when one author uses a word frequently and a another author barely uses it.
- More generally: Differences in rates of frequent words  $>$  Differences in rates of rare words.

Adjustment: Divide the difference in authors' average rates by the average rate across all authors.

# Other options

Other metrics for “distinctiveness”:

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

How do we choose?

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

How do we choose?

- Depends on context, goal



# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

How do we choose?

- Depends on context, goal
- Classification  $\rightsquigarrow$  accuracy, precision, recall

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

How do we choose?

- Depends on context, goal
- Classification  $\rightsquigarrow$  accuracy, precision, recall
- Qualitative inference  $\rightsquigarrow$  face validity, convergence, etc.

# Other options

Other metrics for “distinctiveness”:

- Standardized mean difference (take into account variability)
- Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009)
- Many more!

How do we choose?

- Depends on context, goal
- Classification  $\rightsquigarrow$  accuracy, precision, recall
- Qualitative inference  $\rightsquigarrow$  face validity, convergence, etc.
- More on this later (at the end of slides)

# Why do we care?

## 1 Qualitative inference comparing 2 groups

# Why do we care?

- 1 Qualitative inference comparing 2 groups
- 2 Create custom dictionaries for classification task

# Stylometry: Who Wrote Disputed Federalist Papers?

# Stylometry: Who Wrote Disputed Federalist Papers?

## Federalist Papers:

- Canonical texts in study of American politics
- Designed to persuade citizens of New York to adopt constitution
- 77 essays, published from 1787-1799 in newspapers, published **anonymously** under the name Publius.

# Stylometry: Who Wrote Disputed Federalist Papers?

## Federalist Papers:

- Canonical texts in study of American politics
- Designed to persuade citizens of New York to adopt constitution
- 77 essays, published from 1787-1799 in newspapers, published **anonymously** under the name Publius.

Who wrote the Federalist papers? (Hostler and Wallace (1963))



# Stylometry: Who Wrote Disputed Federalist Papers?

## Federalist Papers:

- Canonical texts in study of American politics
- Designed to persuade citizens of New York to adopt constitution
- 77 essays, published from 1787-1799 in newspapers, published **anonymously** under the name Publius.

## Who wrote the Federalist papers? (Hostler and Wallace (1963))

- Jay: wrote 5 essays
- Hamilton: wrote 43 papers
- Madison: wrote 12 papers
- **Disputed (Hamilton or Madison?):** Essays 49-58, 62, and 63

# Stylometry: Who Wrote Disputed Federalist Papers?

## Federalist Papers:

- Canonical texts in study of American politics
- Designed to persuade citizens of New York to adopt constitution
- 77 essays, published from 1787-1799 in newspapers, published **anonymously** under the name Publius.

## Who wrote the Federalist papers? (Hostler and Wallace (1963))

- Jay: wrote 5 essays
- Hamilton: wrote 43 papers
- Madison: wrote 12 papers
- **Disputed (Hamilton or Madison?):** Essays 49-58, 62, and 63

**Task:** Identify authors of disputed papers

**Method:** Classify papers as Hamilton or Madison using dictionary methods

# Stylometry: Who Wrote Disputed Federalist Papers?

**Training**  $\rightsquigarrow$  papers Hamilton, Madison are known to have authored

**Test**  $\rightsquigarrow$  disputed (i.e. unlabeled) papers

**Preprocessing:**

- Hamilton/Madison discuss similar themes
- Differ on the extent they use **stop words**
- Focus analysis on the stop words

# Word Weights: Standardized Mean Difference

For each word  $p$ , construct weight  $\theta_p^*$ ,

$\mu_{p,\text{Hamilton}}$  = Rate( $p$ ) in subcorpus of Hamilton docs

$\mu_{p,\text{Madison}}$  = Rate( $p$ ) in subcorpus of Madison docs

$\sigma_{p,\text{Hamilton}}^2$  = Var( $p$ ) in subcorpus of Hamilton docs

$\sigma_{p,\text{Madison}}^2$  = Var( $p$ ) in subcorpus of Madison docs

We can then generate weight  $\theta_p^*$  as

$$\theta_p^* = \frac{\mu_{p,\text{Hamilton}} - \mu_{p,\text{Madison}}}{\sqrt{\sigma_{p,\text{Hamilton}}^2 + \sigma_{p,\text{Madison}}^2}}$$

# Trimming the Dictionary

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

# Classification $\rightsquigarrow$ Determining Authorship

For each disputed document  $i$ , compute discrimination statistic

$$Y_i = \sum_{p=1}^P \theta_p^* X_{ip}$$

$Y_i \rightsquigarrow$  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.

# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?

# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?



# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we evaluate our findings?

# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we evaluate our findings?

Three evaluation strategies

# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we evaluate our findings?

## Three evaluation strategies

- Face validity (do these results make sense?)

# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we evaluate our findings?

## Three evaluation strategies

- Face validity (do these results make sense?)
- Convergence (do different metrics lead to the same result?)

# Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we evaluate our findings?

## Three evaluation strategies

- Face validity (do these results make sense?)
- Convergence (do different metrics lead to the same result?)
- "Gold Standard" (do our results align with human coding?)