Text Analysis for Legal Practice

(Class 2)

Charles Crabtree & Kevin L. Cope

January 14, 2020

Class outline

1. Data management with texts.
2. Understanding your text data.

Class outline

1. Data management with texts.
2. Understanding your text data.
   - Word clouds.

Class outline

1. Data management with texts.
2. Understanding your text data.
    - Word clouds.
    - Discriminating terms.

Class outline

1. Data management with texts.
2. Understanding your text data.
   - Word clouds.
   - Discriminating terms.
   - Word frequencies and associations.

Class outline

1. Data management with texts.
2. Understanding your text data.
    - Word clouds.
    - Discriminating terms.
    - Word frequencies and associations.
    - DTMs.

Class outline

1. Data management with texts.
2. Understanding your text data.
    - Word clouds.
    - Discriminating terms.
    - Word frequencies and associations.
    - DTMs.

Class outline

1. Data management with texts.
2. Understanding your text data.
   3. Word clouds.
   4. Discriminating terms.
   5. Word frequencies and associations.
   6. DTMs.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches ⇝ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts ⇝ Different Style/Substantive Focus.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
  - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
    - Comparing Rep., Dem. speeches ⇝ Partisan language.
    - Comparing Judicial Opinions Between Judges/Courts ⇝ Different Style/Substantive Focus.
    - Comparing Contracts From Different Jurisdictions/Over Time ⇝ Different Terms (e.g., liquidated damages).

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
  - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
    - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
    - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
    - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).
- Reasons:

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
  - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).
- Reasons:
  1. Interesting in their own right.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
  - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).
- Reasons:
  1. Interesting in their own right.
  2. Create custom dictionaries for a classification task.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
    - Comparing Rep., Dem. speeches ⇝ Partisan language.
    - Comparing Judicial Opinions Between Judges/Courts ⇝ Different Style/Substantive Focus.
    - Comparing Contracts From Different Jurisdictions/Over Time ⇝ Different Terms (e.g., liquidated damages).
- Reasons:
    1. Interesting in their own right.
    2. Create custom dictionaries for a classification task.
    3. Feature selection: inclusion of features in some subsequent analysis.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
  - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).
- Reasons:
  1. Interesting in their own right.
  2. Create custom dictionaries for a classification task.
  3. Feature selection: inclusion of features in some subsequent analysis.
- Method: Distinctive / Discriminating / Separating word scores.

Discriminating words

- Goal: Find words that distinguish one group of texts from another group of texts.
  - Comparing Rep., Dem. speeches $\rightsquigarrow$ Partisan language.
  - Comparing Judicial Opinions Between Judges/Courts $\rightsquigarrow$ Different Style/Substantive Focus.
  - Comparing Contracts From Different Jurisdictions/Over Time $\rightsquigarrow$ Different Terms (e.g., liquidated damages).
- Reasons:
  1. Interesting in their own right.
  2. Create custom dictionaries for a classification task.
  3. Feature selection: inclusion of features in some subsequent analysis.
- Method: Distinctive / Discriminating / Separating word scores.

What does ’distinctive’ means?

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

What does ’distinctive’ means?

- 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.

What does ’distinctive’ means?

- 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.

Option 1: Unique usage

- Distinctive = exclusive.
- If Trump uses the word ‘bigly’ and Obama never does, we should count ‘bigly’ as distinctive.

Option 1: Unique usage

- Distinctive = exclusive.
- If Trump uses the word ‘bigly’ and Obama never does, we should count ‘bigly’ as distinctive.
- These words tend not to be terribly interesting or informative.

Option 1: Unique usage

- Distinctive = exclusive.
- If Trump uses the word ‘bigly’ and Obama never does, we should count ‘bigly’ as distinctive.
- These words tend not to be terribly interesting or informative.
- More likely to capture differences in linguistic style than content.

Option 1: Unique usage

- Distinctive = exclusive.
- If Trump uses the word ‘bigly’ and Obama never does, we should count ‘bigly’ as distinctive.
- These words tend not to be terribly interesting or informative.
- More likely to capture differences in linguistic style than content.

Option 2: Difference in frequencies

- Distinctive = difference in frequency.
- Compare the number of times each author, Trump or Obama, uses a word.

Option 2: Difference in frequencies

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

Option 2: Difference in frequencies

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

Option 2: Difference in frequencies

- Distinctive = difference in frequency.
- Compare the number of times each author, Trump or Obama, 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.
- Compare the average rate each author uses a word.

Option 3: Difference in averages

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

Option 3: Difference in averages

- 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.

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

$$\theta_p = \mu_{p,Trump} - \mu_{p,Obama}$$

- Find words with highest absolute difference.

Differences in averages: Problems

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

Differences in averages: Problems

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

Differences in averages: Problems

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

Differences in averages: Problems

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

Differences in averages: Problems

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

Differences in averages: Problems

- Favors more frequent words.
- Word 1: 30/1000 (Trump); 25/1000 (Obama) $\rightsquigarrow$ Score: 5/1000.
- Word 2: 5/1000 (Trump); 1/1000 (Obama) $\rightsquigarrow$ 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’:
    - Standardized mean difference (account for variability).

Other options

- Other metrics for ‘distinctiveness’:
    - Standardized mean difference (account for variability).
    - Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009).

Other options

- Other metrics for ‘distinctiveness’:
  - Standardized mean difference (account for variability).
  - Standard Log Odds (used in Monroe, Colaresi, and Quinn, 2009).
  - Many more!

Other options

- Other metrics for ‘distinctiveness’:
    - Standardized mean difference (account for 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 (account for 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 (account for 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 (account for 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 (account for 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).

Other options

- Other metrics for 'distinctiveness':
  - Standardized mean difference (account for 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?

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

Why do we care?

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

Stylometry: Who Wrote Disputed Federalist Papers?

- Federalist Papers:
    1. Canonical texts in study of American politics.

Stylometry: Who Wrote Disputed Federalist Papers?

- Federalist Papers:
    1. Canonical texts in study of American politics.
    2. Designed to persuade citizens of New York to adopt constitution.

Stylometry: Who Wrote Disputed Federalist Papers?

- Federalist Papers:
    1. Canonical texts in study of American politics.
    2. Designed to persuade citizens of New York to adopt constitution.
    3. 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:
  1. Canonical texts in study of American politics.
  2. Designed to persuade citizens of New York to adopt constitution.
  3. 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:
  1. Canonical texts in study of American politics.
  2. Designed to persuade citizens of New York to adopt constitution.
  3. 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?

- Tasks: Identify authors of disputed papers.
- Method: Classify papers as Hamilton or Madison using dictionary methods.
- Training data $\rightsquigarrow$ Hamilton, Madison are known to have authored.
- Test data $\rightsquigarrow$ 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 θ_p,

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

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

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

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

- We can then generate weight θ_p as

$$\theta_p = \frac{\mu_{p,Hamilton} - \mu_{p,Madison}}{\sigma^2_{p,Hamilton} + \sigma^2_{p,Madison}}$$

Trimming the dictionary

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

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 for 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 $\rightsquigarrow$ Hamilton text.
  - Below midpoint in training set $\rightsquigarrow$ 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?
- How do we chose between dictionaries?

Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different ‘distinctive word’ metrics?
- How do we chose 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 chose 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 chose 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?)

Evaluation for Dictionary Methods & Distinctive Words

- How do we choose between different ‘distinctive word’ metrics?
- How do we chose 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?)

Let's Program

Launch RStudio!

kcope@law.virginia.edu | crabtree@dartmouth.edu