Text Analysis for Legal Practice

(Class 1)

Charles Crabtree & Kevin L. Cope

January 17, 2022

charlescrabtree.com/materials.zip

Instructors

- Instructor: Kevin Cope
  - Associate Professor, University of Virginia School of Law

Instructors

- Instructor: Kevin Cope
    - Associate Professor, University of Virginia School of Law
- Instructor: Charles Crabtree

Instructors

- Instructor: Kevin Cope
  - Associate Professor, University of Virginia School of Law
- Instructor: Charles Crabtree
  - Visiting Assistant Professor, Shorenstein Asia-Pacific Research Center, Stanford University
  - Assistant Professor, Department of Government, Dartmouth College

Instructors

- Instructor: Kevin Cope
  - Associate Professor, University of Virginia School of Law
- Instructor: Charles Crabtree
  - Visiting Assistant Professor, Shorenstein Asia-Pacific Research Center, Stanford University
  - Assistant Professor, Department of Government, Dartmouth College

Learning Goals

- Understand how computational text analysis is changing legal research and practice.
- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.

Learning Goals

- Understand how computational text analysis is changing legal research and practice.
- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.
- Immediate Goals:
    1. Understand the primary tools used in automatic text analysis.

Learning Goals

- Understand how computational text analysis is changing legal research and practice.
- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.
- Immediate Goals:
  1. Understand the primary tools used in automatic text analysis.
  2. Locate and turn texts into data.

Learning Goals

- Understand how computational text analysis is changing legal research and practice.
- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.
- Immediate Goals:
  1. Understand the primary tools used in automatic text analysis.
  2. Locate and turn texts into data.
  3. Conduct simple descriptive analyses and measurements.

Learning Goals

- Understand how computational text analysis is changing legal research and practice.

- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.

- Immediate Goals:

  1. Understand the primary tools used in automatic text analysis.
  2. Locate and turn texts into data.
  3. Conduct simple descriptive analyses and measurements.
  4. Know where to turn for additional resources.

Learning Goals

- Understand how computational text analysis is changing legal research and practice.

- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.

- Immediate Goals:

  1. Understand the primary tools used in automatic text analysis.
  2. Locate and turn texts into data.
  3. Conduct simple descriptive analyses and measurements.
  4. Know where to turn for additional resources.
  5. Learn how to develop your own text-analysis-based research projects.

Learning Goals

- Understand how computational text analysis is changing legal research and practice.
- Larger Ambition: Introduce students to the wonders of automatic text analysis, provide a framework for future learning, and encourage the development of interesting text-as-data research projects.
- Immediate Goals:
    1. Understand the primary tools used in automatic text analysis.
    2. Locate and turn texts into data.
    3. Conduct simple descriptive analyses and measurements.
    4. Know where to turn for additional resources.
    5. Learn how to develop your own text-analysis-based research projects.

Legal Applications

- Search algorithms.
- Developing empirical arguments about the correct meaning of statutory/regulatory/contract terms based on usage in thousands of texts (See, e.g., Wilson v. Safelight Group, Inc., No. 18-3408 (6th Cir. July 10, 2019).

Legal Applications

- Search algorithms.
- Developing empirical arguments about the correct meaning of statutory/regulatory/contract terms based on usage in thousands of texts (See, e.g., Wilson v. Safelight Group, Inc., No. 18-3408 (6th Cir. July 10, 2019).
- Machine learning for reviewing documents pursuant to e-discovery requests.

Legal Applications

- Search algorithms.
- Developing empirical arguments about the correct meaning of statutory/regulatory/contract terms based on usage in thousands of texts (See, e.g., Wilson v. Safelight Group, Inc., No. 18-3408 (6th Cir. July 10, 2019).
- Machine learning for reviewing documents pursuant to e-discovery requests.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.
- Newspapers can be used to learn about what events attract media attention and even about the events themselves.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.
- Newspapers can be used to learn about what events attract media attention and even about the events themselves.
- Social media can provide insight into public/elite opinion, communication, attitude formation, and political/legal events.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.
- Newspapers can be used to learn about what events attract media attention and even about the events themselves.
- Social media can provide insight into public/elite opinion, communication, attitude formation, and political/legal events.
- For the purpose of constitutional interpretation, we can learn the 'original meaning' of old terms and phrases based on their usage in historical texts.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.
- Newspapers can be used to learn about what events attract media attention and even about the events themselves.
- Social media can provide insight into public/elite opinion, communication, attitude formation, and political/legal events.
- For the purpose of constitutional interpretation, we can learn the 'original meaning' of old terms and phrases based on their usage in historical texts.
- Changes in judicial-opinion-writing style can show how much influence law clerks have on opinion content.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.
- Newspapers can be used to learn about what events attract media attention and even about the events themselves.
- Social media can provide insight into public/elite opinion, communication, attitude formation, and political/legal events.
- For the purpose of constitutional interpretation, we can learn the ‘original meaning’ of old terms and phrases based on their usage in historical texts.
- Changes in judicial-opinion-writing style can show how much influence law clerks have on opinion content.

Research Applications in Law and Political Science

- Political speeches can shed light on the strategic communications of government leaders.
- Electoral manifestos can illuminate what parties value and how they view the world.
- Newspapers can be used to learn about what events attract media attention and even about the events themselves.
- Social media can provide insight into public/elite opinion, communication, attitude formation, and political/legal events.
- For the purpose of constitutional interpretation, we can learn the 'original meaning' of old terms and phrases based on their usage in historical texts.
- Changes in judicial-opinion-writing style can show how much influence law clerks have on opinion content.

Course Focus

- Courses provides a practical introduction to automatic text analysis. Does **not cover** technical details. Many great technical guides are available online.
- Courses introduces some of the most commonly used tools. Many tools **not described**. This is a very active research area and more programs are available all the time.

Course Focus

- Courses provides a practical introduction to automatic text analysis. Does **not cover** technical details. Many great technical guides are available online.
- Courses introduces some of the most commonly used tools. Many tools **not described**. This is a very active research area and more programs are available all the time.

Course Format

- Partially “flipped classroom.”
- 1/2 lecture, 1/2 lab
- If your laptop/program/code isn’t working, please let us know and team up with another student until it’s resolved.

Course Format

- Partially “flipped classroom.”
- 1/2 lecture, 1/2 lab
- If your laptop/program/code isn’t working, please let us know and team up with another student until it’s resolved.

Course Outline

- Day 1: Course intro; Overview of text analysis methods and relevance to legal research and practice; how to acquire and preprocess texts.
- Day 2: Data management with texts; understanding your text data.

Course Outline

- Day 1: Course intro; Overview of text analysis methods and relevance to legal research and practice; how to acquire and preprocess texts.
- Day 2: Data management with texts; understanding your text data.
- Day 3: Dictionary methods.

Course Outline

- Day 1: Course intro; Overview of text analysis methods and relevance to legal research and practice; how to acquire and preprocess texts.
- Day 2: Data management with texts; understanding your text data.
- Day 3: Dictionary methods.
- Day 4: Topic modeling; Practical applications in the study of law

Course Outline

- Day 1: Course intro; Overview of text analysis methods and relevance to legal research and practice; how to acquire and preprocess texts.
- Day 2: Data management with texts; understanding your text data.
- Day 3: Dictionary methods.
- Day 4: Topic modeling; Practical applications in the study of law

Logistics

- Grading
  - 80% – Short paper using the methods we develop in the course

Logistics

- Grading
  - 80% – Short paper using the methods we develop in the course
  - 20% – Class participation

Logistics

- Grading
  - 80% – Short paper using the methods we develop in the course
  - 20% – Class participation
- Office Hours - 12:00–1:00 P.M., 4:00–4:30 P.M.

Logistics

- Grading
  - 80% – Short paper using the methods we develop in the course
  - 20% – Class participation
- Office Hours - 12:00–1:00 P.M., 4:00–4:30 P.M.

Class Outline

1 Introduction to automatic text analysis.

2 Turn texts into data.

Class Outline

1. Introduction to automatic text analysis.
2. Turn texts into data.
3. Preprocess texts.

Class Outline

1. Introduction to automatic text analysis.
2. Turn texts into data.
3. Preprocess texts.
4. Conduct basic descriptive analyses.

Class Outline

1. Introduction to automatic text analysis.
2. Turn texts into data.
3. Preprocess texts.
4. Conduct basic descriptive analyses.

Typology of Legal Research

	Quantitative	Qualitative
External	Empirical legal scholarship (Holmes's person of the future)	Legal history
Internal	Computational text analysis	Traditional legal scholarship

Why?

- We care about the written word.
- A universe of human activity is encoded in texts.

Why?

- We care about the written word.
- A universe of human activity is encoded in texts.
- Nearly all work relies on the understanding of texts.

Why?

- We care about the written word.
- A universe of human activity is encoded in texts.
- Nearly all work relies on the understanding of texts.
- Researchers and other professionals have informally been using texts as data.

Why?

- We care about the written word.
- A universe of human activity is encoded in texts.
- Nearly all work relies on the understanding of texts.
- Researchers and other professionals have informally been using texts as data.
- Analyzing large corpora of texts costly.

Why?

- We care about the written word.
- A universe of human activity is encoded in texts.
- Nearly all work relies on the understanding of texts.
- Researchers and other professionals have informally been using texts as data.
- Analyzing large corpora of texts costly.
- Computers reduce these costs.

Why?

- We care about the written word.
- A universe of human activity is encoded in texts.
- Nearly all work relies on the understanding of texts.
- Researchers and other professionals have informally been using texts as data.
- Analyzing large corpora of texts costly.
- Computers reduce these costs.

Where to Find Texts

- Online databases (LexisNexis, U.N., Comparative Manifesto Project, replication archives).
- Websites (scraping, APIs).

Where to Find Texts

- Online databases (LexisNexis, U.N., Comparative Manifesto Project, replication archives).
- Websites (scraping, APIs).
- R data packages (CRAN or GitHub).

Where to Find Texts

- Online databases (LexisNexis, U.N., Comparative Manifesto Project, replication archives).
- Websites (scraping, APIs).
- R data packages (CRAN or GitHub).
- Archives (scan and OCR documents).

Where to Find Texts

- Online databases (LexisNexis, U.N., Comparative Manifesto Project, replication archives).
- Websites (scraping, APIs).
- R data packages (CRAN or GitHub).
- Archives (scan and OCR documents).

Where to Find Texts

Our Sources

- State constitutions.
- National constitutions.

Our Sources

- State constitutions.
- National constitutions.
- U.S. Supreme Court Oral Arguments.

Our Sources

- State constitutions.
- National constitutions.
- U.S. Supreme Court Oral Arguments.
- U.S. Case Law. (lawcorpus.byu.edu/cusc/concordances)

Our Sources

- State constitutions.
- National constitutions.
- U.S. Supreme Court Oral Arguments.
- U.S. Case Law. (lawcorpus.byu.edu/cusc/concordances)
- Current U.S. Code (bit.ly/3KiLVsK).

Our Sources

- State constitutions.
- National constitutions.
- U.S. Supreme Court Oral Arguments.
- U.S. Case Law. (lawcorpus.byu.edu/cusc/concordances)
- Current U.S. Code (bit.ly/3KiLVsK).

Data

- We want machine readable text.
  - Ideal: Plain text files (.txt or .csv), or data directly from APIs.

Data

- We want machine readable text.
  - Ideal: Plain text files (.txt or .csv), or data directly from APIs.
  - Common: PDFs, HTML files, Word docs.

Data

- We want machine readable text.
  - Ideal: Plain text files (.txt or .csv), or data directly from APIs.
  - Common: PDFs, HTML files, Word docs.
- We also normally want meta-data (author/speaker, date, section, headline, tags).

Data

- We want machine readable text.
  - Ideal: Plain text files (.txt or .csv), or data directly from APIs.
  - Common: PDFs, HTML files, Word docs.
- We also normally want meta-data (author/speaker, date, section, headline, tags).
- We'll preprocess texts to improve ease and accuracy of analysis.

Data

- We want machine readable text.
  - Ideal: Plain text files (.txt or .csv), or data directly from APIs.
  - Common: PDFs, HTML files, Word docs.
- We also normally want meta-data (author/speaker, date, section, headline, tags).
- We'll preprocess texts to improve ease and accuracy of analysis.

Ethos

- All approaches to automatic text analysis are wrong, but some are useful (Grimmer and Stewart 2013).
- Methods supplement and augment human abilities.

Ethos

- All approaches to automatic text analysis are wrong, but some are useful (Grimmer and Stewart 2013).
- Methods supplement and augment human abilities.
- Cannot replace humans.

Ethos

- All approaches to automatic text analysis are wrong, but some are useful (Grimmer and Stewart 2013).
- Methods supplement and augment human abilities.
- Cannot replace humans.
- No computational silver bullet.

Ethos

- All approaches to automatic text analysis are wrong, but some are useful (Grimmer and Stewart 2013).
- Methods supplement and augment human abilities.
- Cannot replace humans.
- No computational silver bullet.
- Iterative process with repeated validation.

Ethos

- All approaches to automatic text analysis are wrong, but some are useful (Grimmer and Stewart 2013).
- Methods supplement and augment human abilities.
- Cannot replace humans.
- No computational silver bullet.
- Iterative process with repeated validation.

Methods Overview

- Two general approaches to automatic text analysis.

  1. Supervised methods: We know what should be in the texts, and use computers to extend our understanding to a larger population of unseen documents.

  2. Unsupervised methods: We do not know the structure or contents of the texts beforehand. Instead, we use the model to discover a structure that best explains (or summarizes) the documents.

Methods Overview

- Two general approaches to automatic text analysis.

    1. Supervised methods: We know what should be in the texts, and use computers to extend our understanding to a larger population of unseen documents.
    2. Unsupervised methods: We do not know the structure or contents of the texts beforehand. Instead, we use the model to discover a structure that best explains (or summarizes) the documents.

Supervised Methods

- Hand coding is used to train, or supervise, statistical models to classify texts in pre-determined categories.

  1. Set of known categories.
     - Positive tone, negative tone.
     - About South Korea, Japan, United States.

Supervised Methods

- Hand coding is used to train, or supervise, statistical models to classify texts in pre-determined categories.

  1. Set of known categories.
     - Positive tone, negative tone.
     - About South Korea, Japan, United States.
  2. Set of hand-coded documents.
     - Coding done by human coders.
     - Training Set: documents we'll use to learn how to code.
     - Validation Set: documents we'll use to learn how well we code.

Supervised Methods

- Hand coding is used to train, or supervise, statistical models to classify texts in pre-determined categories.

  1. Set of known categories.
     - Positive tone, negative tone.
     - About South Korea, Japan, United States.
  2. Set of hand-coded documents.
     - Coding done by human coders.
     - Training Set: documents we'll use to learn how to code.
     - Validation Set: documents we'll use to learn how well we code.
  3. Set of unlabeled documents that we want to classify.

Supervised Methods

- Hand coding is used to train, or supervise, statistical models to classify texts in pre-determined categories.

  1. Set of known categories.
     - Positive tone, negative tone.
     - About South Korea, Japan, United States.
  2. Set of hand-coded documents.
     - Coding done by human coders.
     - Training Set: documents we'll use to learn how to code.
     - Validation Set: documents we'll use to learn how well we code.
  3. Set of unlabeled documents that we want to classify.
  4. Method to generalize from hand coding to unlabeled documents (dictionary methods, logistic regression, naive bayes).

Supervised Methods

- Hand coding is used to train, or supervise, statistical models to classify texts in pre-determined categories.

  1. Set of known categories.
     - Positive tone, negative tone.
     - About South Korea, Japan, United States.
  2. Set of hand-coded documents.
     - Coding done by human coders.
     - Training Set: documents we'll use to learn how to code.
     - Validation Set: documents we'll use to learn how well we code.
  3. Set of unlabeled documents that we want to classify.
  4. Method to generalize from hand coding to unlabeled documents (dictionary methods, logistic regression, naive bayes).
  5. Validate by comparing predicted label to hand-coded label.

Supervised Methods

- Hand coding is used to train, or supervise, statistical models to classify texts in pre-determined categories.

  1. Set of known categories.
     - Positive tone, negative tone.
     - About South Korea, Japan, United States.
  2. Set of hand-coded documents.
     - Coding done by human coders.
     - Training Set: documents we'll use to learn how to code.
     - Validation Set: documents we'll use to learn how well we code.
  3. Set of unlabeled documents that we want to classify.
  4. Method to generalize from hand coding to unlabeled documents (dictionary methods, logistic regression, naive bayes).
  5. Validate by comparing predicted label to hand-coded label.

Unsupervised Methods

- Discover new ways of organizing texts that are useful, but maybe understudied or unknown.

    1. Set of unlabeled documents that we want to classify.

Unsupervised Methods

- Discover new ways of organizing texts that are useful, but maybe understudied or unknown.

    1. Set of unlabeled documents that we want to classify.
    2. Method to discover categories and then classify (label) documents into those categories (k-means clustering, topic models).

Unsupervised Methods

- Discover new ways of organizing texts that are useful, but maybe understudied or unknown.

  1. Set of unlabeled documents that we want to classify.
  2. Method to discover categories and then classify (label) documents into those categories (k-means clustering, topic models).
  3. Interpret labels assigned to categories and understand what they mean.

Unsupervised Methods

- Discover new ways of organizing texts that are useful, but maybe understudied or unknown.
  1. Set of unlabeled documents that we want to classify.
  2. Method to discover categories and then classify (label) documents into those categories (k-means clustering, topic models).
  3. Interpret labels assigned to categories and understand what they mean.

Methods Covered

1 Preprocessing

2 Dictionary methods / sentiment analysis (Supervised)

Methods Covered

1. Preprocessing
2. Dictionary methods / sentiment analysis (Supervised)

First Steps

1. Launch RStudio
2. Download scripts from github.com/cdcrabtree/uva-2022
3. Load 00begin.R into RStudio.
4. Install necessary packages.

Preprocessing

- Objective: Prepare texts for automatic text analysis.
- Primary Problem: Texts were designed for manual reading.

Preprocessing

- Objective: Prepare texts for automatic text analysis.
- Primary Problem: Texts were designed for manual reading.
- Secondary Problem: Many texts are pretty ‘dirty’.

Preprocessing

- Objective: Prepare texts for automatic text analysis.
- Primary Problem: Texts were designed for manual reading.
- Secondary Problem: Many texts are pretty ‘dirty’.
- Solution: Text manipulation and pre-processing.

Preprocessing

- Objective: Prepare texts for automatic text analysis.
- Primary Problem: Texts were designed for manual reading.
- Secondary Problem: Many texts are pretty 'dirty'.
- Solution: Text manipulation and pre-processing.
- Important decisions: What words matter in a text, what is the right unit of analysis, how to analyze non-English texts.

Preprocessing

- Objective: Prepare texts for automatic text analysis.
- Primary Problem: Texts were designed for manual reading.
- Secondary Problem: Many texts are pretty ‘dirty’.
- Solution: Text manipulation and pre-processing.
- Important decisions: What words matter in a text, what is the right unit of analysis, how to analyze non-English texts.

Key Terms

- Corpus / document
- Preprocessing

Key Terms

- Corpus / document
- Preprocessing
- Tokens, grams

Key Terms

- Corpus / document
- Preprocessing
- Tokens, grams
- Stemming / Lemmatize

Key Terms

- Corpus / document
- Preprocessing
- Tokens, grams
- Stemming / Lemmatize
- Bag of Words

Key Terms

- Corpus / document
- Preprocessing
- Tokens, grams
- Stemming / Lemmatize
- Bag of Words

Preparing a Corpus

- A corpus (pl: corpora) is a collection of texts, usually stored electronically, and from which we perform our analysis. A corpus might be a collection of news articles from Reuters or the published works of Shakespeare.
- Within each corpus we will have separate articles, stories, volumes, each treated as a separate entity or record. Each unit is called a document.

Preparing a Corpus

- A corpus (pl: corpora) is a collection of texts, usually stored electronically, and from which we perform our analysis. A corpus might be a collection of news articles from Reuters or the published works of Shakespeare.

- Within each corpus we will have separate articles, stories, volumes, each treated as a separate entity or record. Each unit is called a document.

- Documents come in a variety of formats, but plain text is best (e.g. .txt, .csv).

Preparing a Corpus

- A corpus (pl: corpora) is a collection of texts, usually stored electronically, and from which we perform our analysis. A corpus might be a collection of news articles from Reuters or the published works of Shakespeare.

- Within each corpus we will have separate articles, stories, volumes, each treated as a separate entity or record. Each unit is called a document.

- Documents come in a variety of formats, but plain text is best (e.g. .txt, .csv).

- Corpora often come with metadata (author, date, label).

Preparing a Corpus

- A corpus (pl: corpora) is a collection of texts, usually stored electronically, and from which we perform our analysis. A corpus might be a collection of news articles from Reuters or the published works of Shakespeare.
- Within each corpus we will have separate articles, stories, volumes, each treated as a separate entity or record. Each unit is called a document.
- Documents come in a variety of formats, but plain text is best (e.g. .txt, .csv).
- Corpora often come with metadata (author, date, label).
- Bag of Words

Preparing a Corpus

- A corpus (pl: corpora) is a collection of texts, usually stored electronically, and from which we perform our analysis. A corpus might be a collection of news articles from Reuters or the published works of Shakespeare.
- Within each corpus we will have separate articles, stories, volumes, each treated as a separate entity or record. Each unit is called a document.
- Documents come in a variety of formats, but plain text is best (e.g. .txt, .csv).
- Corpora often come with metadata (author, date, label).
- Bag of Words

Data Structures

- Traditional: Each document a row, one column for text, and other columns for metadata.
- Tidy: Each document-word a row, one column for each word in a text, and other columns for metadata.

Data Structures

- Traditional: Each document a row, one column for text, and other columns for metadata.
- Tidy: Each document-word a row, one column for each word in a text, and other columns for metadata.
- We'll use **both**. Most researchers probably should, converting between formats when necessary.

Data Structures

- Traditional: Each document a row, one column for text, and other columns for metadata.
- Tidy: Each document-word a row, one column for each word in a text, and other columns for metadata.
- We'll use **both**. Most researchers probably should, converting between formats when necessary.

Preprocessing Texts

- Goal is to reduce noise and amplify signal.
- **MANY** preprocessing recipes. This can be a problem (Denny and Spirling 2018).

Preprocessing Texts

- Goal is to reduce noise and amplify signal.
- **MANY** preprocessing recipes. This can be a problem (Denny and Spirling 2018).

1 Clean data. Takes a long time. Sometimes iterative.

2 Remove capitalization, punctuation, numbers, extra white space.

3 Discard word order (bag of words assumption).

4 Discard stop words.

5 Combine similar terms (stem, lemmatize).

5 Output: Document-Term Matrix, each element counts occurrence of a particular term in a particular document.

Preprocessing Texts

- Goal is to reduce noise and amplify signal.
- **MANY** preprocessing recipes. This can be a problem (Denny and Spirling 2018).

    1 Clean data. Takes a long time. Sometimes iterative.

    2 Remove capitalization, punctuation, numbers, extra white space.

    3 Discard word order (bag of words assumption).

    4 Discard stop words.

    5 Combine similar terms (stem, lemmatize).

    5 Output: Document-Term Matrix, each element counts occurrence of a particular term in a particular document.

Remove Capitalization and Punctuation

- **Assumption: Capitalization/punctuation does not provide useful information.**

```
Freedom means the supremacy of human rights
everywhere. Our support goes to those who struggle to
gain those rights and keep them.
```

Remove Capitalization and Punctuation

- **Assumption: Capitalization/punctuation does not provide useful information.**

```
Freedom means the supremacy of human rights
everywhere. Our support goes to those who struggle to
gain those rights and keep them.
```

```
freedom means the supremacy of human rights
everywhere our support goes to those who struggle to
gain those rights and keep them
```

Remove Capitalization and Punctuation

- **Assumption: Capitalization/punctuation does not provide useful information.**

  ```
  Freedom means the supremacy of human rights
  everywhere. Our support goes to those who struggle to
  gain those rights and keep them.
  ```

  ```
  freedom means the supremacy of human rights
  everywhere our support goes to those who struggle to
  gain those rights and keep them
  ```

- Also might want to remove extra space, symbols, and numbers.

Remove Capitalization and Punctuation

- **Assumption: Capitalization/punctuation does not provide useful information.**

  ```
  Freedom means the supremacy of human rights
  everywhere. Our support goes to those who struggle to
  gain those rights and keep them.
  ```

  ```
  freedom means the supremacy of human rights
  everywhere our support goes to those who struggle to
  gain those rights and keep them
  ```

- Also might want to remove extra space, symbols, and numbers.

Discard Word Order

- **Assumption: Word order doesn't Matter. information.**

```
freedom means the supremacy of human rights
everywhere our support goes to those who struggle to
gain those rights and keep them
```

Discard Word Order

- **Assumption: Word order doesn't Matter. information.**

  freedom means the supremacy of human rights everywhere our support goes to those who struggle to gain those rights and keep them

  [and, everywhere, freedom, gain, goes, human, keep, means, of, our, rights, rights, struggle, support, supremacy, the, them, those, those, to, to, who]

Discard Word Order

- **Assumption: Word order doesn't Matter. information.**

  freedom means the supremacy of human rights everywhere our support goes to those who struggle to gain those rights and keep them

  [and, everywhere, freedom, gain, goes, human, keep, means, of, our, rights, rights, struggle, support, supremacy, the, them, those, those, to, to, who]

  [rights, goes, struggle, keep, who, rights, freedom, them, our, supremacy, the, to, means, human, everywhere, gain, and, to, of, those, support, those]

Discard Word Order

- **Assumption: Word order doesn't Matter. information.**

  freedom means the supremacy of human rights everywhere our support goes to those who struggle to gain those rights and keep them

  [and, everywhere, freedom, gain, goes, human, keep, means, of, our, rights, rights, struggle, support, supremacy, the, them, those, those, to, to, who]

  [rights, goes, struggle, keep, who, rights, freedom, them, our, supremacy, the, to, means, human, everywhere, gain, and, to, of, those, support, those]

Tokenization

- Unigrams: [rights, goes, struggle, keep, who, rights, freedom, them, our, supremacy, the, to, means, human, everywhere, gain, and, to, of, those, support, those]

Tokenization

- Unigrams: [rights, goes, struggle, keep, who, rights, freedom, them, our, supremacy, the, to, means, human, everywhere, gain, and, to, of, those, support, those]

- Bigrams: [freedom means, means the,the supremacy, supremacy of,of human, human rights, rights everywhere, everywhere our, our support, support goes, …]

Tokenization

- Unigrams: [rights, goes, struggle, keep, who, rights, freedom, them, our, supremacy, the, to, means, human, everywhere, gain, and, to, of, those, support, those]

  Bigrams: [freedom means, means the,the supremacy, supremacy of,of human, human rights, rights everywhere, everywhere our, our support, support goes, …]

  Trigrams: [freedom means the, means the supremacy, the supremacy of, supremacy of human, of human rights, human rights everywhere, rights everywhere our, everywhere our support, our support goes, support goes to, …]

Tokenization

- Unigrams: [rights, goes, struggle, keep, who, rights, freedom, them, our, supremacy, the, to, means, human, everywhere, gain, and, to, of, those, support, those]

  Bigrams: [freedom means, means the,the supremacy, supremacy of,of human, human rights, rights everywhere, everywhere our, our support, support goes, …]

  Trigrams: [freedom means the, means the supremacy, the supremacy of, supremacy of human, of human rights, human rights everywhere, rights everywhere our, everywhere our support, our support goes, support goes to, …]

???

- **How does this work?**

???

- **How does this work?**

  Speech involves …

???

- **How does this work?**

  Speech involves …

  Irony. Thanks Obama!

  Subtle negation. I don’t not like you.

  Order. Peace means no more war != War means no more peace.

???

- **How does this work?**

  Speech involves ...

  Irony. Thanks Obama!

  Subtle negation. I don't not like you.

  Order. Peace means no more war != War means no more peace.

???

- At least three replies.

1 Maybe it doesn't work. Text and task specific.

???

- At least three replies.

1. Maybe it doesn't work. Text and task specific.
2. Words imply what a text is about. Themes can be understood, if not necessarily nuances.

???

- At least three replies.

1. Maybe it doesn't work. Text and task specific.
2. Words imply what a text is about. Themes can be understood, if not necessarily nuances.
3. It just works. Bag-of-words assumption validated in a large number of papers across the sciences.

???

- At least three replies.

1. Maybe it doesn't work. Text and task specific.
2. Words imply what a text is about. Themes can be understood, if not necessarily nuances.
3. It just works. Bag-of-words assumption validated in a large number of papers across the sciences.

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

- Add “noise” to documents (without conveying much information).

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

- Add “noise” to documents (without conveying much information).
- Discarding stop words $\rightsquigarrow$ focus on substantive words (maximizes signal).

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

- Add “noise” to documents (without conveying much information).
- Discarding stop words ⇝ focus on substantive words (maximizes signal).
  - **Caution:** Might need to use custom stopword lists.

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

- Add “noise” to documents (without conveying much information).
- Discarding stop words ⇝ focus on substantive words (maximizes signal).
  - **Caution:** Might need to use custom stopword lists.
  - Many English language stop lists include gender pronouns (Monroe, Colaresi, and Quinn 2008).

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

- Add “noise” to documents (without conveying much information).
- Discarding stop words ⇝ focus on substantive words (maximizes signal).
  - **Caution:** Might need to use custom stopword lists.
  - Many English language stop lists include gender pronouns (Monroe, Colaresi, and Quinn 2008).
  - Many also include negations.

Discard Stop Words

- Stop Words: English Language place holding words.

  `the, it, if, a, able, at, be, because, …`

- Add “noise” to documents (without conveying much information).
- Discarding stop words $\rightsquigarrow$ focus on substantive words (maximizes signal).
  - **Caution:** Might need to use custom stopword lists.
  - Many English language stop lists include gender pronouns (Monroe, Colaresi, and Quinn 2008).
  - Many also include negations.

Combine Similar Terms

- Reduce dimensionality further. Boost signal.
- Combine similar terms (tense and number).

Combine Similar Terms

- Reduce dimensionality further. Boost signal.
- Combine similar terms (tense and number).
    - Words used to refer to same basic concept.

Combine Similar Terms

- Reduce dimensionality further. Boost signal.
- Combine similar terms (tense and number).
  - Words used to refer to same basic concept.

    `family, families, familial` ⇝ `famili`

Combine Similar Terms

- Reduce dimensionality further. Boost signal.
- Combine similar terms (tense and number).
  - Words used to refer to same basic concept.

    `family, families, familial` ⇝ `famili`
  - Many-to-one mapping from words to stem/lemma

Combine Similar Terms

- Reduce dimensionality further. Boost signal.
- Combine similar terms (tense and number).
  - Words used to refer to same basic concept.

    `family, families, familial` ⇝ `famili`
  - Many-to-one mapping from words to stem/lemma

Stemming v. Lemmatizing

- Stemming.

Stemming v. Lemmatizing

- Stemming.
  - Simplistic algorithms.

Stemming v. Lemmatizing

- Stemming.
  - Simplistic algorithms.

    Chop off end of word.

Stemming v. Lemmatizing

- Stemming.
    - Simplistic algorithms.

      Chop off end of word.
    - Porter stemmer, Lancaster stemmer, Snowball stemmer.

Stemming v. Lemmatizing

- Stemming.
  - Simplistic algorithms.

    Chop off end of word.
  - Porter stemmer, Lancaster stemmer, Snowball stemmer.
- Lematizing.

Stemming v. Lemmatizing

- Stemming.
  - Simplistic algorithms.

    Chop off end of word.
  - Porter stemmer, Lancaster stemmer, Snowball stemmer.
- Lematizing.
  - Condition on part of speech (noun, verb, etc).

Stemming v. Lemmatizing

- Stemming.
  - Simplistic algorithms.

    Chop off end of word.
  - Porter stemmer, Lancaster stemmer, Snowball stemmer.
- Lematizing.
  - Condition on part of speech (noun, verb, etc).
  - Verify result is a word.

Stemming v. Lemmatizing

- Stemming.
  - Simplistic algorithms.

    Chop off end of word.
  - Porter stemmer, Lancaster stemmer, Snowball stemmer.
- Lematizing.
  - Condition on part of speech (noun, verb, etc).
  - Verify result is a word.

Other Common Steps

- Remove sparse terms.
- Remove common terms.

Other Common Steps

- Remove sparse terms.
- Remove common terms.
- Remove other terms (proper nouns, technical language, website addresses, emojis).

Other Common Steps

- Remove sparse terms.
- Remove common terms.
- Remove other terms (proper nouns, technical language, website addresses, emojis).
- Weight some terms more than others (tf-idf).

Other Common Steps

- Remove sparse terms.
- Remove common terms.
- Remove other terms (proper nouns, technical language, website addresses, emojis).
- Weight some terms more than others (tf-idf).

Applied Example

Freedom means the supremacy of human rights everywhere. Our support goes to those who struggle to gain those rights and keep them.

Step 1: Remove capitalization and punctuation

Applied Example

```
Freedom means the supremacy of human rights
everywhere. Our support goes to those who struggle to
gain those rights and keep them.
```

Step 1: Remove capitalization and punctuation

```
freedom means the supremacy of human rights
everywhere our support goes to those who struggle to
gain those rights and keep them
```

Applied Example

```
Freedom means the supremacy of human rights
everywhere. Our support goes to those who struggle to
gain those rights and keep them.
```

Step 1: Remove capitalization and punctuation

```
freedom means the supremacy of human rights
everywhere our support goes to those who struggle to
gain those rights and keep them
```

Applied Example

Step 2: Tokenize

```
freedom, means, the, supremacy, of, human, rights,
everywhere, our, support, goes, to, those, who,
struggle, to, gain, those, rights, and, keep, them
```

Applied Example

Step 2: Tokenize

```
freedom, means, the, supremacy, of, human, rights,
everywhere, our, support, goes, to, those, who,
struggle, to, gain, those, rights, and, keep, them
```

Step 3: Remove stop words

Applied Example

Step 2: Tokenize

freedom, means, the, supremacy, of, human, rights, everywhere, our, support, goes, to, those, who, struggle, to, gain, those, rights, and, keep, them

Step 3: Remove stop words

freedom, means, supremacy, human, rights, support, struggle, gain, rights

Applied Example

Step 2: Tokenize

```
freedom, means, the, supremacy, of, human, rights,
everywhere, our, support, goes, to, those, who,
struggle, to, gain, those, rights, and, keep, them
```

Step 3: Remove stop words

```
freedom, means, supremacy, human, rights, support,
struggle, gain, rights
```

Step 4: Stem words

Applied Example

Step 2: Tokenize

freedom, means, the, supremacy, of, human, rights, everywhere, our, support, goes, to, those, who, struggle, to, gain, those, rights, and, keep, them

Step 3: Remove stop words

freedom, means, supremacy, human, rights, support, struggle, gain, rights

Step 4: Stem words

freedom, mean, supremaci, human, right, support, struggl, gain, right

Applied Example

Step 2: Tokenize

freedom, means, the, supremacy, of, human, rights, everywhere, our, support, goes, to, those, who, struggle, to, gain, those, rights, and, keep, them

Step 3: Remove stop words

freedom, means, supremacy, human, rights, support, struggle, gain, rights

Step 4: Stem words

freedom, mean, supremaci, human, right, support, struggl, gain, right

Applied Example

Step 5: Create count vector

Stem	Count
freedom	1
mean	1
supremaci	1
human	1
right	1
support	1
struggle	1
gain	1
right	1

Applied Example

Step 6: Create Document-Term matrix

- $\boldsymbol{x} = NxP$ matrix.

Applied Example

Step 6: Create Document-Term matrix

- $\boldsymbol{X} = NxP$ matrix.
    - N = Number of documents.

Applied Example

Step 6: Create Document-Term matrix

- $\boldsymbol{X} = NxP$ matrix.
  - N = Number of documents.
  - P = Number of features.

Applied Example

Step 6: Create Document-Term matrix

- $\boldsymbol{X} = NxP$ matrix.
  - $N =$ Number of documents.
  - $P =$ Number of features.
- $\boldsymbol{X}$ is main input for many automatic text analyses.

Applied Example

Step 6: Create Document-Term matrix

- $\boldsymbol{X} = NxP$ matrix.
  - N = Number of documents.
  - P = Number of features.
- $\boldsymbol{X}$ is main input for many automatic text analyses.

Let's Program

Launch RStudio and we can get started programming.

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