Text Cleaning: An Introduction with R

Social Science Data Analytics Workshop

Caleb Lucas (@calebjlucas)

July 22, 2020

Michigan State University

Introduction

- A sentence (or two) about me

Introduction

- A sentence (or two) about me
- Goals for today:

Introduction

- A sentence (or two) about me
- Goals for today:
  - Understand why text cleaning matters

Introduction

- A sentence (or two) about me
- Goals for today:
  - Understand why text cleaning matters
  - Clean messy text using R

Introduction

- A sentence (or two) about me
- Goals for today:
    - Understand why text cleaning matters
    - Clean messy text using R
    - Process/prepare text for analysis using R

Introduction

- A sentence (or two) about me
- Goals for today:
    - Understand why text cleaning matters
    - Clean messy text using R
    - Process/prepare text for analysis using R
    - Plan to attend the workshop tomorrow (same time, same place) that covers the next step- text analysis!

Text Analysis- what is it?

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.
    - Make inferences about things we can't directly observe like ideology

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.
    - Make inferences about things we can't directly observe like ideology
- Or to categorize texts into different classes

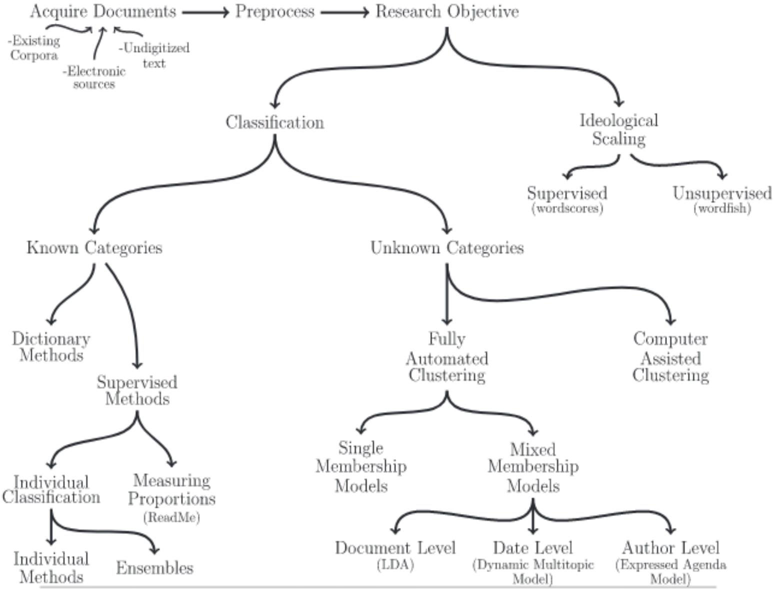


Fig. 1 An overview of text as data methods.

Text Analysis

- Countless applications:

Text Analysis

- Countless applications:
    - Scale the ideology of politicians

Text Analysis

- Countless applications:
  - Scale the ideology of politicians
  - Measure relationships between rebel groups using speech about each other

Text Analysis

- Countless applications:
  - Scale the ideology of politicians
  - Measure relationships between rebel groups using speech about each other
  - Assist psychological assessments of patients using open-ended questions

Text Analysis

- Countless applications:
    - Scale the ideology of politicians
    - Measure relationships between rebel groups using speech about each other
    - Assist psychological assessments of patients using open-ended questions
- Focus of published text models is on what readers want - the question, the math, the results, etc

Text Analysis

- Countless applications:
    - Scale the ideology of politicians
    - Measure relationships between rebel groups using speech about each other
    - Assist psychological assessments of patients using open-ended questions
- Focus of published text models is on what readers want - the question, the math, the results, etc
    - Text cleaning is rarely discussed in detail in published papers

Text Analysis

- Countless applications:
  - Scale the ideology of politicians
  - Measure relationships between rebel groups using speech about each other
  - Assist psychological assessments of patients using open-ended questions
- Focus of published text models is on what readers want - the question, the math, the results, etc
  - Text cleaning is rarely discussed in detail in published papers
  - ... but it can affect results/findings and is hard!

Textual Data

Textual Data

- Tons of textual data sources

Textual Data

- Tons of textual data sources
  - Open-ended survey responses

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles
  - Historical legal documents

Textual Data

- Tons of textual data sources
    - Open-ended survey responses
    - News articles
    - Historical legal documents
    - Social media posts

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles
  - Historical legal documents
  - Social media posts
  - ... many more

Textual Data

- Tons of textual data sources
    - Open-ended survey responses
    - News articles
    - Historical legal documents
    - Social media posts
    - … many more
- Text data is not typically formatted nicely for us in nature

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles
  - Historical legal documents
  - Social media posts
  - … many more
- Text data is not typically formatted nicely for us in nature
  - Dirty documents OCR'ed

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles
  - Historical legal documents
  - Social media posts
  - … many more
- Text data is not typically formatted nicely for us in nature
  - Dirty documents OCR'ed
  - Messy scraped web pages

Textual Data

- Tons of textual data sources
    - Open-ended survey responses
    - News articles
    - Historical legal documents
    - Social media posts
    - … many more
- Text data is not typically formatted nicely for us in nature
    - Dirty documents OCR’ed
    - Messy scraped web pages
    - Poorly formatted web input forms

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles
  - Historical legal documents
  - Social media posts
  - … many more
- Text data is not typically formatted nicely for us in nature
  - Dirty documents OCR'ed
  - Messy scraped web pages
  - Poorly formatted web input forms
  - Tweets with emojis, urls, etc.

Textual Data

- Tons of textual data sources
  - Open-ended survey responses
  - News articles
  - Historical legal documents
  - Social media posts
  - … many more
- Text data is not typically formatted nicely for us in nature
  - Dirty documents OCR'ed
  - Messy scraped web pages
  - Poorly formatted web input forms
  - Tweets with emojis, urls, etc.

  ↪ Need to clean and prepare for statistical modeling

Text Cleaning

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
    - Help computer know “msu” and “MSU!” are the same

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
  - Help computer know “msu” and “MSU!” are the same
    - msu = \u006d\u0073\u0075

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
  - Help computer know “msu” and “MSU!” are the same
    - msu = \u006d\u0073\u0075
    - MSU! = \u004d\u0053\u0055\u0021

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
  - Help computer know “msu” and “MSU!” are the same
    - msu = \u006d\u0073\u0075
    - MSU! = \u004d\u0053\u0055\u0021
  - Reduce the corpus to meaningful words

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
  - Help computer know “msu” and “MSU!” are the same
    - msu = \u006d\u0073\u0075
    - MSU! = \u004d\u0053\u0055\u0021
  - Reduce the corpus to meaningful words
- Target: punctuation, numbers, lowercasing, reducing words, stopwords, n-grams, infrequent terms

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
  - Help computer know “msu” and “MSU!” are the same
    - msu = \u006d\u0073\u0075
    - MSU! = \u004d\u0053\u0055\u0021
  - Reduce the corpus to meaningful words
- Target: punctuation, numbers, lowercasing, reducing words, stopwords, n-grams, infrequent terms
- How we go about this can have down-stream effects

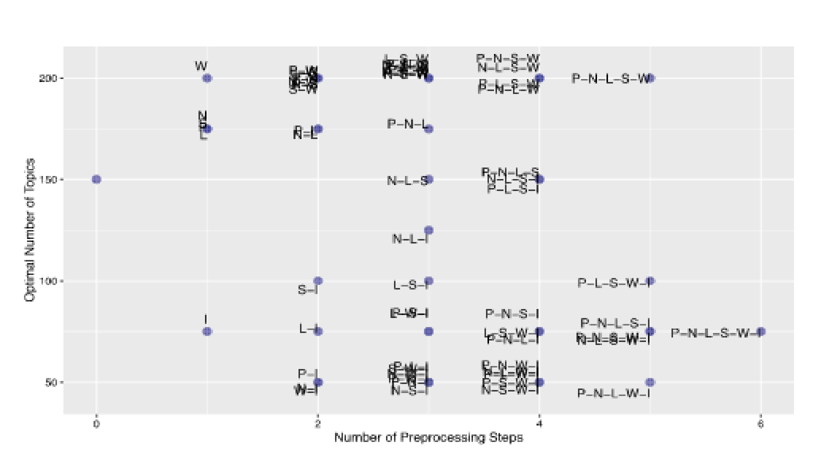
Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
    - Help computer know "msu" and "MSU!" are the same
        - msu = \u006d\u0073\u0075
        - MSU! = \u004d\u0053\u0055\u0021
    - Reduce the corpus to meaningful words
- Target: punctuation, numbers, lowercasing, reducing words, stopwords, n-grams, infrequent terms
- How we go about this can have down-stream effects
    - Different cleaning procedures = different results

Text Cleaning

- Goal: use substantive knowledge to strip text of unhelpful features
  - Help computer know “msu” and “MSU!” are the same
    - msu = \u006d\u0073\u0075
    - MSU! = \u004d\u0053\u0055\u0021
  - Reduce the corpus to meaningful words
- Target: punctuation, numbers, lowercasing, reducing words, stopwords, n-grams, infrequent terms
- How we go about this can have down-stream effects
  - Different cleaning procedures = different results
  - 7 binary preprocessing steps = 128 possible models

Denny and Spirling (2018)



General Text Cleaning Steps

1. Fix representational issues

General Text Cleaning Steps

1. Fix representational issues
    - Expand contractions, expand abbreviations, make lowercase, etc.

General Text Cleaning Steps

1. Fix representational issues
    - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words

General Text Cleaning Steps

1. Fix representational issues
   - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words
   - Remove common words (‘stopwords’) like *the*

General Text Cleaning Steps

1. Fix representational issues
   - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words
   - Remove common words (‘stopwords’) like *the*
3. Remove ‘dirty’ characters/text

General Text Cleaning Steps

1. Fix representational issues
   - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words
   - Remove common words (‘stopwords’) like *the*
3. Remove ‘dirty’ characters/text
   - Correct spelling, remove numbers, etc.

General Text Cleaning Steps

1. Fix representational issues
   - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words
   - Remove common words (‘stopwords’) like *the*
3. Remove ‘dirty’ characters/text
   - Correct spelling, remove numbers, etc.
4. Analysis-specific steps

General Text Cleaning Steps

1. Fix representational issues
    - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words
    - Remove common words (‘stopwords’) like *the*
3. Remove ‘dirty’ characters/text
    - Correct spelling, remove numbers, etc.
4. Analysis-specific steps
    - Normalize synonyms, remove parentheticals, etc.

General Text Cleaning Steps

1. Fix representational issues
   - Expand contractions, expand abbreviations, make lowercase, etc.
2. Keep meaningful words
   - Remove common words ('stopwords') like *the*
3. Remove 'dirty' characters/text
   - Correct spelling, remove numbers, etc.
4. Analysis-specific steps
   - Normalize synonyms, remove parentheticals, etc.
   - Researchers typically 'fall into' these steps as they analyze - leads to important point...

Text Cleaning Tips

- Cleaning is not a series of steps

Text Cleaning Tips

- Cleaning is not a series of steps
    - Clean *your* corpus

Text Cleaning Tips

- Cleaning is not a series of steps
  - Clean *your* corpus
  - Should be a continual *process*

Text Cleaning Tips

- Cleaning is not a series of steps
    - Clean *your* corpus
    - Should be a continual *process*
    - Clean, inspect, clean, inspect, analyze, clean, ...

Text Cleaning Tips

- Cleaning is not a series of steps
    - Clean *your* corpus
    - Should be a continual *process*
    - Clean, inspect, clean, inspect, analyze, clean, …
    - Text is messier than most other forms of data, takes more time/effort to prepare

Text Cleaning Tips

- Cleaning is not a series of steps
  - Clean *your* corpus
  - Should be a continual *process*
  - Clean, inspect, clean, inspect, analyze, clean, …
  - Text is messier than most other forms of data, takes more time/effort to prepare
  - Take time to read the text (yes!) before/during/after

Text Cleaning Tips

- Cleaning is not a series of steps
    - Clean *your* corpus
    - Should be a continual *process*
    - Clean, inspect, clean, inspect, analyze, clean, …
    - Text is messier than most other forms of data, takes more time/effort to prepare
    - Take time to read the text (yes!) before/during/after
- Most researchers ‘do the steps’ and then proceed to their analysis

Text Cleaning Tips

- Cleaning is not a series of steps
    - Clean *your* corpus
    - Should be a continual *process*
    - Clean, inspect, clean, inspect, analyze, clean, ...
    - Text is messier than most other forms of data, takes more time/effort to prepare
    - Take time to read the text (yes!) before/during/after
- Most researchers ‘do the steps’ and then proceed to their analysis
- Crucial to always be in cleaning mode

Text Cleaning Tips

- Cleaning is not a series of steps
    - Clean *your* corpus
    - Should be a continual *process*
    - Clean, inspect, clean, inspect, analyze, clean, ...
    - Text is messier than most other forms of data, takes more time/effort to prepare
    - Take time to read the text (yes!) before/during/after
- Most researchers 'do the steps' and then proceed to their analysis
- Crucial to always be in cleaning mode
    - More confidence you have the 'right' data

Text Cleaning Tips

- Cleaning is not a series of steps
  - Clean *your* corpus
  - Should be a continual *process*
  - Clean, inspect, clean, inspect, analyze, clean, ...
  - Text is messier than most other forms of data, takes more time/effort to prepare
  - Take time to read the text (yes!) before/during/after
- Most researchers 'do the steps' and then proceed to their analysis
- Crucial to always be in cleaning mode
  - More confidence you have the 'right' data
  - Limits chance of weird data/results, which is easy to spot with text data

Text Cleaning Examples

1. So… I JUST GOT ACCEPTED TO MICHIGAN STATE ☻ ☺

Text Cleaning Examples

1. So... I JUST GOT ACCEPTED TO MICHIGAN STATE ☻ ☺
    - so i just got accepted to msu

Text Cleaning Examples

1. So… I JUST GOT ACCEPTED TO MICHIGAN STATE ☻ ☺
    - so i just got accepted to msu
2. Check out this study by MSU profs- bitly.com/123

Text Cleaning Examples

1. So... I JUST GOT ACCEPTED TO MICHIGAN STATE ☻ ☺
    - so i just got accepted to msu
2. Check out this study by MSU profs- bitly.com/123
    - check out this study by msu profs [professors]

Text Cleaning Examples

1. So... I JUST GOT ACCEPTED TO MICHIGAN STATE ☻ ☺
   - so i just got accepted to msu
2. Check out this study by MSU profs- bitly.com/123
   - check out this study by msu profs [professors]
3. The plans by Mich. State U. profs for a cheap ventilator are GREAT y'all

Text Cleaning Examples

1. So... I JUST GOT ACCEPTED TO MICHIGAN STATE ☻ ☺
   - so i just got accepted to msu
2. Check out this study by MSU profs- bitly.com/123
   - check out this study by msu profs [professors]
3. The plans by Mich. State U. profs for a cheap ventilator are GREAT y’all
   - the plans by msu profs [professors] for a cheap ventilator are great yall [you all]

Text Processing

Text Processing

Ok, you cleaned the text up... now what?

Stemming

- Stemming is a form of *word reduction*

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - 'ing,' 'ed,' 'es,' etc.

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - ‘ing,’ ‘ed,’ ‘es,’ etc.
  - learns, learning, learned → learn

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - ‘ing,’ ‘ed,’ ‘es,’ etc.
  - learns, learning, learned → learn
  - boy’s, boys → boy

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - ‘ing,’ ‘ed,’ ‘es,’ etc.
  - learns, learning, learned → learn
  - boy’s, boys → boy
  - ties → ti

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - ‘ing,’ ‘ed,’ ‘es,’ etc.
    - learns, learning, learned → learn
    - boy’s, boys → boy
    - ties → ti
    - easily → easili

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - ‘ing,’ ‘ed,’ ‘es,’ etc.
    - learns, learning, learned → learn
    - boy’s, boys → boy
    - ties → ti
    - easily → easili
- This reduces the corpus’ dimensions

Stemming

- Stemming is a form of *word reduction*
- Generally chops off inflections - 'ing,' 'ed,' 'es,' etc.
  - learns, learning, learned → learn
  - boy's, boys → boy
  - ties → ti
  - easily → easili
- This reduces the corpus' dimensions
- Acknowledges "run" and "runs" are different versions of the same word

Lemmatization

- Lemmatization returns a word’s ‘dictionary’ form

Lemmatization

- Lemmatization returns a word's 'dictionary' form
    - This is called a 'lemma'

Lemmatization

- Lemmatization returns a word's 'dictionary' form
  - This is called a 'lemma'
- Not just word reduction

Lemmatization

- Lemmatization returns a word's 'dictionary' form
  - This is called a 'lemma'
- Not just word reduction
  - saw → see

Lemmatization

- Lemmatization returns a word's 'dictionary' form
    - This is called a 'lemma'
- Not just word reduction
    - saw → see
    - geese → goose

Lemmatization

- Lemmatization returns a word's 'dictionary' form
    - This is called a 'lemma'
- Not just word reduction
    - saw → see
    - geese → goose
    - easily → easy

Lemmatization

- Lemmatization returns a word's 'dictionary' form
    - This is called a 'lemma'
- Not just word reduction
    - saw → see
    - geese → goose
    - easily → easy
- This also reduces the corpus' dimensions

Lemmatization

- Lemmatization returns a word's 'dictionary' form
    - This is called a 'lemma'
- Not just word reduction
    - saw → see
    - geese → goose
    - easily → easy
- This also reduces the corpus' dimensions
- More computationally expensive

Lemmatization

- Lemmatization returns a word's 'dictionary' form
    - This is called a 'lemma'
- Not just word reduction
    - saw → see
    - geese → goose
    - easily → easy
- This also reduces the corpus' dimensions
- More computationally expensive
- Not available in every language

Data Format

- We typically format text data as a document term matrix
  - Rows = documents, columns = terms
- *Terms* are typically single words, but can be other things that we *tokenize* the text into
  - *Tokens* are just units of the text used in the analysis
  - We can tokenize text into sentences, paragraphs, n-grams (collections of words or sentences), etc.

	Token_1	Token_2	…	Token_n
Doc_1	0	0	.	0
Doc_2	5	0	.	3
…	.	.	.	.
Doc_n	1	0	.	0

Zoom Back Out

- Cleaning text is a messy process with many steps

1. Use your knowledge to clean your corpus
2. Assess the effect of other choices on your model

Zoom Back Out

- Cleaning text is a messy process with many steps
- There is not a predetermined set of steps

1. Use your knowledge to clean your corpus
2. Assess the effect of other choices on your model

Zoom Back Out

- Cleaning text is a messy process with many steps
- There is not a predetermined set of steps
- Somewhat different than other types of data

1. Use your knowledge to clean your corpus
2. Assess the effect of other choices on your model

Zoom Back Out

- Cleaning text is a messy process with many steps
- There is not a predetermined set of steps
- Somewhat different than other types of data
    - A great deal of ad hoc decisions

1. Use your knowledge to clean your corpus
2. Assess the effect of other choices on your model

Zoom Back Out

- Cleaning text is a messy process with many steps
- There is not a predetermined set of steps
- Somewhat different than other types of data
    - A great deal of ad hoc decisions
    - Not obvious object type conversions, etc.

1. Use your knowledge to clean your corpus
2. Assess the effect of other choices on your model

R sesh!

R sesh!

https://github.com/caleblucas/text_cleaning