Introduction to R & Data for Humanities

1 July 2021

Why learn computational text analysis?

- Data literacy is at the heart of research
- Data skills are in demand
- Addressing algorithmic bias

If you cannot read, manipulate, and interpret data, you are overlooking an incredibly rich source for understanding the human condition.

Why learn text analysis with ‘R’ and/or ‘Python’?

How long will it take me to learn text analysis?

What kind of data do I need?

- Machine-readable text
- Lots of it

Current methods in text analysis

1. What are these texts about?
   Word Frequency, Collocation, Topic Analysis, Significant Terms

2. How are these texts connected?
   Concordance, Network Analysis

3. What emotions are expressed?
   Sentiment Analysis

4. What key names can I find?
   Named Entity Recognition

5. Which of these texts are similar?
   Clustering, Supervised Machine Learning, Authorship Attribution

1. What are these texts about?

Word Frequency (Beginner)
Counting the frequency of each word in each text. This includes the Bag of Words approach.

Example: "Which of these texts focus on women?"

Collocation (Beginner)
Examining where two significant words occur close to one another.

Example: "Where are women mentioned in relation to home ownership?"

Topic Analysis (or Topic Modeling) (Intermediate)
Discovering the topics within a group of texts.

Example: "What are the most frequent topics discussed in this newspaper?"

Significant Terms (or TF-IDF) (Intermediate)
Finding the significant words within a text.

Example: "What language is most significant within 1970s political speech?

2. How are these texts connected?

Concordance (Beginner)
Where is this word or phrase used in every document?

Example: “Which journal articles mention Maya Angelou’s phrase, ‘If you’re for the right thing, then you do it without thinking.’”

Network Analysis (Intermediate)
How are these terms connected?

Example: “What local communities formed around civil rights in 1963”

3. What emotions are expressed?

Sentiment Analysis (Intermediate)
Is the language used happy, angry, or confused?

Example: “How do these presidential speeches describe the second amendment?”

4. What key names can I find?

Named Entity Recognition (Intermediate)
List every example of a kind of entity.

Example: “What are the geographic locations mentioned by Marie de France?”

5. Which of these texts are similar?

Clustering (Advanced)
Which texts are the most similar?

Example: "Is this play closer to comedy or tragedy?”

Supervised Machine Learning (Advanced)
Can we identify texts that are similar to this?

Example: "Are there other Jim Crow laws like these we have already identified?"

Authorship Attribution (Advanced)
Which texts are the most similar?

Example: “Did J. K. Rowling write *The Cuckoo’s Calling*?”

Current methods in text analysis

1. What are these texts about?
   Word Frequency, Collocation, Topic Analysis, **Significant Terms**

2. How are these texts connected?
   Concordance, Network Analysis

3. What emotions are expressed?
   Sentiment Analysis

4. What key names can I find?
   Named Entity Recognition

5. Which of these texts are similar?
   Clustering, Supervised Machine Learning, Authorship Attribution

The basics of R and Text Mining with R within the Tidyverse

- “The Tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar and data structures.” tidyverse.org (2018)

- The tidy text format: Text mining based on tidy data principles

- Burrows, J.F. (1987) *Computation into Criticism: A Study of Jane Austen's Novels*. Oxford: Oxford University Press

- R packages:
  - janeaustenr package (Silge 2016)
  - gutenbergr package (Robinson 2016)