Workshop Introduction to Text Analysis in R

Denise J. Roth

6th Annual COMPTEXT Conference 2024, Vrije Universiteit Amsterdam

May 2nd, 2024

Introduction

- PhD Candidate in Political Communication
- Strategic Communication Group at Wageningen University & Research
- Work focusses on the politicization of science
- Mostly use R and Python
- denise.roth@wur.nl

Introduction

- Who are you?
- Where are you from?
- What does your research focus on?
- Do you have any previous experience in R or quantitative text analysis?
- What other program or software do you use?
- In what way would you like to use textual data in your research?

(Tentative) Agenda

- Part 1
  - 9:40 - 10:15: Getting to Know R and its Basics
  - 10:15 - 10:35: Text Pre-processing (Theory)
  - 10:35 - 10:45: Break
- Part 2
  - 10:45 - 11:05: Text Pre-processing (Practical Example)
  - 11:05 - 11:25: Exploring our Data (Topic Models)
  - 11:25 - 11:35: Break
- Part 3
  - 11:35 - 12:00: Lexical Sentiment Analysis
  - 12:00 - 12:00: Supervised Machine Learning (Automatic Classification)
  - 12:25 - 12:30: Wrap-Up & Lunch Break

Today's Goal

We only have 3 hours... ...and there are so many things we could cover

I would like to:

- Give you an overview of the essential basics
- Help you gain some first hands-on experience
- Provide you with well-documented and freely available resources to deepen your knowledge and practice your skills

Important to keep in mind that...

...struggling is part of it
...learning curve tends to be very steep
...nobody knows everthing from the top of their head

What is R? And why is it useful?

- Popular, free, open-source programming language
  - Primarily developed for statistical computing and data visualization
- Flexible and extensible
- Allows handling large amounts of data
- Online resources

R Studio

- Integrated development environment (IDE)
- Very user-friendly
- Facilitates importing, managing and wrangling data
- Allows creation of reproducible research products using R Markdown

R for Text Analysis

- offers numerous packages tailored for text analysis
- Very user-friendly
- vibrant community providing ample resources and assistance
- statistical capabilities empower advanced text analysis techniques
- powerful visualization tools for creating insightful plots and charts to visualize text analysis results

(Some) relevant packages/libraries

- **tidyverse**: provides a cohesive suite of packages for data manipulation, visualization, and analysis, streamlining workflows with a consistent and intuitive syntax
- **quanteda**: extensive toolkit for quantitative text analysis, offering robust tools for tasks like text preprocessing, corpus management, and advanced analysis techniques within a unified framework
- **tidytext**: offers specialized functions and methods tailored for text mining and analysis within the tidyverse framework, simplifying the process of handling and analyzing text data

How do we analyze text?

- We usually do computations of any kind on numbers
  - But we are interested in text!
- Thus, we need a manner in which we can represent our text in numbers

Document-Term Matrix (DTM)

- Commonly used representation of text
- Represents our corpus(or our set of documents) as a matrix
  - Each row represents a document
  - Each column represents a term (word)
  - Numbers in each of the cells show how often that word is present in the document

Document-Term Matrix (DTM)

Example
Document 1: "The door was opened by the man in the green shirt"
Document 2: "The man left through the green door"

Doc	the	door	was	opened	by	man	in	green	shirt	left	through
1	3	1	1	1	1	1	1	1	0	0	0
2	2	1	0	0	0	1	0	1	0	1	1

Figure: Document Term Matrix

DTM as a Bag of Words

- Discards quite some information from the text
- All words are "put in a big" without looking at full sentences or the specific contexts of the words
- Can be used for many different text analyses
- However, some analyses may require richer matrix-representations such as word-pairs or automatic syntactic analysis

Unsupervised Machine Learning: Topic Models

- Latent Dirichlet Allocation (LDA) Topic Model is a statistical model used to discover underlying topics in a collection of text documents
- assumes that each document is a mixture of topics, and each topic is a mixture of words
- goal of LDA is to uncover these latent topics and the distribution of words within each topic
- can help uncover hidden patterns, relationships, and themes within text data

Lexical Sentiment Analysis

- We are often interest in the sentiment of our textual data
  - Is a given generally positive or negative?
- Historically, sentiment analysis was done using dictionaries
  - Dictionaries were compiled for specific tasks

Dictionary Analysis in R

Main caveat:

- Very extensive dictionaries will have a high **recall**, but often suffer from low **precision**
- very short dictionary will often be very **precise**, but will have a low **recall**
- Require human **validation**

Supervised Machine Learning for Automatic Text Classification

- model learns patterns and relationships in labeled data
- goal is to assign predefined categories or labels to text documents based on their content
- involves training a model on a labeled dataset of text documents, where each document is associated with one or more categories or classes

Examples of Classifiers

- **Naïve Bayes Classifier**
  - simple probabilistic classifier based on Bayes' theorem
  - assumes that the features (attributes) are conditionally independent given the class label
- **Lasso (Least Absolute Shrinkage and Selection Operator) Classifier**
  - linear model that performs both variable selection and regularization
  - adds a penalty term to the cost function, forcing some feature coefficients to be exactly zero, effectively selecting a subset of the most relevant features
- **Support Vector Machine (SVM) Classifier**
  - finds the optimal hyperplane that separates different classes in the feature space with the maximum margin
  - can handle both linear and non-linear classification tasks through the use of different kernel functions

How to Choose the Right Classifier

- Accuracy
- Interpretability
- Data Complexity
- Computational Efficiency
- Robustness
- Scalability
- Overfitting

Where Do I Go from Here?

- Check out the GitHub repositories provided by some of the colleagues at the Communication Science Department of the VU Amsterdam
  - Many of the materials were used today
  - Much of the credit goes to Wouter van Atteveldt, Johannes Gruber, Philipp Masur, Kasper Welbers and many others who have and are continuously working on great materials to learn about and practice with computational tools in R
  - R Course Materials
- Make sure to have a look at the freely available book by Atteveld, Trilling and Arcila
  - Computational Analysis of Communication (2022)
- AI models such as ChatGPT can be helpful if you know how to ask the right questions
  - Check out AskGPT