

Text Mining and Sentiment Analysis

Practical Exercise with R-Studio

Today's Goal

- Learn tidy text mining in R.
- Explore term frequencies and visualizations.
- Discover word associations and networks.
- Practice with data about Coffee.

Jurgena Myftiu

jurgena.myftiu@unibg.it



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Scienze Aziendali

Key Concepts

- ❖ **Token:** smallest unit (word, phrase).
- ❖ **Tokenization:** splitting text into tokens.
- ❖ **Stop words:** very common words ("the", "and") we usually remove.
- ❖ **Tidy text format:** one token per row.



Libraries We Will Use

- ❖ **tidyverse**: Data manipulation.
- ❖ **tidytext**: Text tokenization and cleaning.
- ❖ **wordcloud**: Word clouds.
- ❖ **widyr**: Word associations.
- ❖ **igraph, ggraph**: Word networks.
- ❖ **RColorBrewer**: Color palettes.



Workflow

1. Import dataset.
2. Tokenize and clean text.
3. Visualize most frequent words (Bar Plot, Word Cloud).
4. Analyze word associations.
5. Build a word network.



Visualizations

Bar Plot: shows top frequent words.

Word Cloud: visual impact of frequent words.

- ❖ Word size = Frequency
- ❖ Color variation possible



Word Associations

- ❖ `pairwise_count()`: count co-occurrences of words.
- ❖ `pairwise_cor()`: compute correlation (phi coefficient).

Useful to understand strong thematic links between words!



Word Networks

- ❖ Nodes = Words.
- ❖ Edges = Strength of association.

Tools:

- **igraph**: Create graph.
- **ggraph**: Visualize network.



Exercise for you

The data set `cofee.csv` contains tweets related to coffee. Write R code to perform the following

1. Import the dataset and create a tibble named `coffee.tweets`.
2. Inspect the imported dataset.
3. Select the variable `n.doc` and `text`.
4. Convert the tibble to the tidy format and remove stopwords, creating a new tibble named `tidy.coffee`.
5. Produce the frequency table of words in `tidy.coffee`, named `coffee.freq`.
6. Create a wordcloud for the values in `coffee.freq`. What do you notice?
7. Create a custom stopwords tibble by adding to the tidy dataset `stop_words` the words "http", "https", "rt", "t.co", "ed", "amp", "coffee", "morning", "barista", "caffeine", "espresso", "coldbrew".
8. Create a new tibble named `tidy.coffee.2` by removing the custom stopwords. Further remove all words starting with "00" and the elements made by one digit.



Exercise for you

9. Produce the frequency table of words in `tidy.coffee.2`, named `coffee.freq.2`.
10. Create a wordcloud for the values in `coffee.freq.2`.
11. Explore the use of different colors for the plot. You can take a look at some available colors with `head(colors(), 50)`.
12. Explore the use of prebuilt color palettes, using the function `brewer.pal()`.
13. Build a frequency plot for the most frequent words in the dataset.
14. Explore word co-appearance using the function `pairwise_count()`. Which are the words that co-appear most often? Does that make sense?
15. Explore which are the words most often co-appearing with “latte”.
16. Compute the phi coefficient for words co-appearing with “latte”.



Challenge for You!

- ❖ Repeat correlation and network analysis starting from another coffee-related word.
- ❖ Examples: “shop”, “brand”, “pick”.

Explore and be creative!



Interactive Questions for You

- ❖ Why is it important to remove stop words?
 - Because they are extremely common words that don't add meaningful information for the analysis.
- ❖ What could it mean if two words are very highly correlated?
 - They often appear together, possibly indicating a strong thematic or contextual link.



- ❖ When would you prefer a bar plot over a word cloud?
 - When you want precise comparisons of word frequencies; bar plots allow exact reading of values.
- ❖ What risks do you see when building a word network with too many nodes?
 - The network becomes cluttered, unreadable, and difficult to interpret.
- ❖ Can you think of a scenario where you **should not** remove a word like "the" or "and"?
 - In syntactic or grammatical analysis where the presence of such words is important (e.g., studying sentence structure)

