Rule-Based Topic Classification of Newspaper Articles

Exercise

Prof. Dr. Nicolas Meseth

OVERVIEW

You have learned the process of **text tokenization** and the advantages it offers. By tokenizing the text data, we have organized it neatly (tidy) into a column `word`, where each entry represents an individual unit (atomic values). This enables us to perform various operations, including filtering, counting, grouping, and more.

In this exercise, your task is to apply tokenization to a collection of newspaper articles. Once tokenization is completed, you will proceed to conduct several analyses based on the obtained results.

Load the Articles

The articles are provided to you as R-data source (RDS) file:

```
articles <- readRDS("data/noz_articles/noz_articles_01-2023.rds")
```

Task 1: Tokenize Articles

Perform the steps required for tokenization of the articles. In the final result, you should have a tibble with one row corresponding to a token in an article.

1. Filter out irrelevant data. In this step, check if there are any articles with empty text or similar data quality issues.
2. Clean and normalize the text.
3. Split the text into tokens.

4. Remove any irrelevant and frequent stop words.

Save the result from the last step on a tibble named `articles_tokenized`.

Task 2: Deductive Topic Classification

Deductive topic classification involves formulating hypotheses regarding relevant topics and then searching for evidence within the data to either support or disprove these hypotheses.

1. Recall recent events and identify three topics that have been discussed in the media. These topics can encompass political issues, scientific advancements, ongoing crises, significant events, and more.
2. Create a dictionary and add keywords that you think would likely appear in articles about the three topics.
3. Apply the dictionary to the articles and calculate a score for each article and topic. Choose a suitable score from the set we introduced in class. Or you develop your own scoring system.
4. Introduce a threshold for your score and identify the articles or each of your topics. If an article has more than one topic assigned, keep only the topic with the highest score.
5. Review the classification results. See if you can refine your dictionary, either theory-driven (add more keywords from your knowledge or research) or data-driven (check frequent words in classified articles and add missing ones to your dictionary).
6. Visualize the distribution of scores for each topic as a boxplot. How would you interpret the result, and what actions could you take as a consequence?
7. Create a separate word cloud for each topic showing the most frequently used nouns in all articles from that topic. To identify nouns, create a second dictionary, or apply part-of-speech-tagging (POS) with spaCy.