Socio-Informatics 348
Practical 10

Submission Instructions

- Submit your completed practical as `studentnumber.qmd` on SocSciLearn.
- Submissions are checked for completeness, not correctness.
- At least 80% of exercises must be attempted to receive 1% towards AF assessment.
- Attendance of at least one practical session per week is required to earn the 1% for that week's practical.

Deadline

Friday 24 October, 17:00 (submit on SocSciLearn)

Chapters Covered:

- Silge: Chapters 1-3

Exercises

1. You are provided a dataset containing text, authors, dates, and sources. Convert this dataset into a tidy format using `unnest_tokens()`, ensuring that each token (word) is in a separate row while retaining the metadata (author, date, source).

   Here is the dataset:

```
text_data <- data.frame(
  line = 1:6,
  text = c("I love learning R programming",
          "The tidyverse package is amazing",
          "Sentiment analysis is fun!",
          "R makes data wrangling easy",
          "Text mining helps us understand language",
          "Tokenization breaks text into meaningful units"),
  author = c("Alice", "Bob", "Charlie", "Alice", "David", "Eve"),
  date = c("2024-01-15", "2024-01-16", "2024-01-17", "2024-01-18",
               "2024-01-19", "2024-01-20"),
    source = c("Blog", "Twitter", "Research Paper",
                "Blog", "News Article", "Twitter")
  )
```

2. Use the `janeaustenr` package to extract the text of Jane Austen's books and convert it to a tidy format. Remove common stop words from the dataset. Calculate and display the 10 most frequent words across all books.

3. Perform sentiment analysis on the provided dataframe of senator tweets (see file on SUNLearn) using the Bing sentiment lexicon and visualise the results. How does the count of positive and negative words vary across different senators? Which senators have the most negative sentiment according to the tweets?

4. Using the same senator tweets dataset, perform sentiment analysis on the dataframe of senator tweets using the NRC sentiment lexicon and visualise sentiment trends over time.

5. Analyse and visualise the most frequent words in the Jane Austen texts. Make sure that you remove stopwords.

6. Analyse the TF-IDF scores to identify the most distinctive words in each Jane Austen text.