MA22019 2025 - Problem Sheet 4

Text data analysis - Word frequency and sentiment

Overview

This week's problem sheet focuses on the text data analysis techniques covered in Sections 3.1-3.3.1 in the lecture notes. Exercises 1-3 are designed to help you with revising the content of the lecture from Week 5. You can check your solutions to these questions yourself by answering the Moodle quiz.

Tutorial Question 1 is similar to Exercises 1-3, while Tutorial Question 2 explores how well the type of sentiment analysis we introduced performs at assessing sentiment for product reviews. Should time permit, you may want to work on the Homework Question during the tutorial.

Your answer to the Homework Question can be submitted on Moodle to your tutor for feedback. The submission deadline is 17:00 on Thursday 13 March 2025. You should submit a single PDF or Word file that provides your R code, any created R output and all your comments.

You may want to load the following packages before starting the exercise:

```r
library( dplyr )
library( ggplot2 )
library( tidytext )
library( wordcloud )
library( stringr )
```

When working on a University PC, you have to first install the tidytext and wordcloud packages. The installation of the textdata package is quite cumbersome, and thus I provide the AFINN and Bing sentiment lexicons as .csv files for your convenience. You can load them using

```r
AFINN <- read.csv("AFINN Sentiment Lexicon.csv")
Bing <- read.csv("Bing Sentiment Lexicon.csv")
```

Exercise 1 - Short stories by Edgar Allan Poe

The file "Poe.csv" contains the book *The Works of Edgar Allan Poe - Volume 1* which includes eight short stories by the American author Edgar Allen Poe. Each row in the data set gives a line from the book and the name of the short story that line belongs to. To load the data, we use

```
Poe_raw <- read.csv("Poe.csv")
```

The following exercises are designed to help you with revising the steps we used for extracting the most common words for *Jane Eyre* in the lecture.

a) Extract the individual words from the text and remove any underscores.

b) Which are the five most common words (including stop words) across all stories?

c) For each short story, identify the word most commonly used within it (excluding stop words).

Exercise 2 - Baby names in the USA between 1880 and 2017

The file "Babynames.csv" provides a comprehensive record of the names given to newborns in the United States between 1880 and 2017 according to the Social Security Administration. For each name, year and sex, we are given the number of newborn babies given that name (as long as the name was used at least five times in that year). We are further provided with the proportion of newborn babies who were given that name amongst babies of the same sex.

a) What was the most common baby name over the period 1880-2017?

b) Create a word cloud which visualizes the 30 most common baby names (in terms of total number) for a girl between 1880 and 2017.

c) Explore how the popularity of the girl name "Mary" evolved over the period 1880-2017. When did its popularity peak?

Exercise 3 - Analysis of the book *Frankenstein*

The novel *Frankenstein* by Mary Shelley tells the story of a young scientist who creates a creature via an experiment and is subsequently horrified by what he has made. The book is provided in the file "Frankenstein.csv" and can be loaded using

```
Frankenstein_raw <- read.csv( "Frankenstein.csv", fileEncoding = "UTF-8" )
```

Perform an analysis that addresses the following questions for *Frankenstein*:

a) Which five words appear the most often (excluding stop words)?

b) Calculate the AFINN sentiment score (sum of the AFINN score of all words) for each chapter. Which chapter has the highest/lowest sentiment score?

Tutorial Question 1 - Word frequency analysis for *Les Misérables*

We want to analyse the book *Les Misérables* by Victor Hugo in terms of the words used therein. The file "LesMis.csv" contains the data as provided by Project Gutenberg, but with the metadata at the beginning and end already removed. Consider the following questions:

a) Extract the individual words from the text, and remove any stop words and underscores.

b) Find the 10 most common in *Les Misérables* (excluding stop words) and create a bar plot which visualizes their number of occurrences.

c) Create a word cloud for the 30 most frequently used words (excluding stop words).

Tutorial Question 2 - Amazon Reviews

One important application area for sentiment analysis concerns the analysis of customer reviews. To explore how good the AFINN lexicon performs at capturing the overall sentiment of a review, we consider 1,000 Amazon product reviews. The data is stored in the file "AmazonReviews.csv".

```
Reviews_raw <- read.csv( "AmazonReviews.csv" )
```

The data provides the rating ("1 Star" to "5 Stars") and the accompanying text for each review. We want to derive the sentiment for each review and then compare it to the given score to assess how well we perform at capturing the sentiment of the text.

a) Using the AFINN sentiment lexicon, for each review, calculate the average sentiment per word.

b) Plot the average sentiments calculated in part a) against the star ratings. What do you conclude?

Homework Question - The work of H.G. Wells

We want to analyze and compare the novels *The Time Machine* and *War of the Worlds* by the British author Herbert George Wells. The books are provided in the files "Time Machine.csv" and "War of Worlds.csv".

a) Compare the two books in terms of the words used within them.

b) How do "The Time Machine" and "The War of the Worlds" differ regarding their sentiment?