MA22019 2025 - Problem Sheet 5

Term frequency - inverse document frequency and topic modelling

Overview

This week's problem sheet focuses on the text data analysis techniques covered in Sections 3.3.2 and 3.4 of the lecture notes. Exercises 1-2 help you with revising the content of the lecture in Week 6. You can check your solutions to these questions yourself by answering the Moodle quiz.

Tutorial Question 1 is quite extensive and will require you to apply most of the techniques from Chapter 3. Tutorial Question 2 focuses on the function grep(), which was used before, but we never explored its full potential. 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 20 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:

```
library( dplyr )
library( ggplot2 )
library( tidytext )
library( stringr )
library( tidyr )
library( topicmodels )
```

When working on a University PC, you have to first install the tidytext package and any dependencies using

```
install.packages( "tidytext", dependencies = TRUE )
```

For the sentiment analysis you can load the sentiment lexicons using

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

Exercise 1 - Comparing Moby Dick and Robinson Crusoe

We want to consider the books *Moby Dick* and *The Life and Adventures of Robinson Crusoe*. The text for both books is provided in the file "AdventureBooks.csv" and we load it using

```
Books <- read.csv( "AdventureBooks.csv" )
```

Consider the following two questions:

a) Which five words (excluding stop words) are the most common in *Moby Dick*?

b) When considering a corpus which only includes *Moby Dick* and *The Life and Adventures of Robinson Crusoe*, which five words have the highest term frequency - inverse document frequency (tf-idf)?

Exercise 2 - Analysis of news articles

In Section 4.4 we applied topic modelling to articles published in the New York Times. We now study another example. The file "Articles.csv" on Moodle provides the text for 2692 news articles from 2015. The articles were published either in the "business" or the "sports" category. In this exercise you will first repeat the steps from Section 4.4, and then explore how well your fitted model performs at identifying whether an article belongs to the "business" or "sports" category.

a) Treading each article as a separate document, derive the document term matrix for the set of articles and store it as **Articles_dtm**.

With the document term matrix having been derived, we estimate the parameters of an LDA model with $K = 2$ topics using:

```
Articles_LDA <- LDA( Articles_dtm, k = 2, method="Gibbs", control = list(seed=2024) )
```

b) For each article, extract the proportions with which the different topics feature. Is there a difference in proportions between "business" and "sports" articles?

c) Which are the five most common words in each of the two topics?

Tutorial Question 1 - Comparing books

We are asked to explore and compare the books *Anne of Green Gables* by L.M. Montgomery and *Rebecca of Sunnybrook Farm* by Kate Douglas Wiggin. The two books are provided together in the file "Books Tutorial Question 1.csv" and the following information is provided:

- **text** - Text as printed in the book
- **title** - Book the text comes from
- **chapter** - Chapter the text belongs to

Perform the following tasks using the techniques described in Chapter 3 of the lecture notes.

a) Extract the two words with the highest term frequency-inverse document frequency for each book, with the corpus only containing *Anne of Green Gables* and *Rebecca of Sunnybrook Farm*.

b) Use sentiment analysis to explore how the emotional intent has evolved over the two books. How do the two books compare?

c) Suppose each book chapter is considered as a separate document (as we did in Section 3.4.3 in the lecture notes). Use Latent Dirichlet Allocation to derive $K = 2$ topics, and then study the estimated proportions provided by the model. What do you conclude?

d) Some scholars claim that *Anne of Green Gables* is patterned after *Rebecca of Sunnybrook Farm*. Discuss whether your results in parts a)-c) support this claim or not.

Tutorial Question 2 - The function grep()

We so far only used the function grep() when extracting the text data from the files provided by Project Gutenberg. Specifically, we used it to identify the lines containing the word "EBOOK", signalling the beginning and end of the book. The following exercise will require you to use grep() to identify all lines which contain a specific phrase:

The Police of Utopia sent us data on burglaries which were reported between 2015 and 2021, including a short description providing information on the number of criminals and their victims. Victims are classified into six groups: "young single", "young couple", "middle-aged single", "middle-aged couple", "elderly single" and "elderly couple". The data are available in the file "UtopiaCrimes.csv". Extract the following information using the functions grep():

a) For which group of people did the police record the most burglaries?

b) What is the proportion of burglaries that involved more than two criminals?

Homework Question - Sentiment Analysis vs Latent Dirichlet Allocation

So far we have used sentiment analysis to explore whether a statement has a positive or negative emotional intent. The purpose of this exercise is to apply sentiment analysis to another data set and to explore whether Latent Dirichlet Allocation (LDA) is able to identify differences in the language used for positive and negative reviews.

We will be working with customer reviews for British Airways. The reviews are stored in the file "British Airways Reviews.csv" and the following information is provided:

- **rating** - Score given by the customer (1="very poor", 10="very good")
- **country** - The country where the customer resides
- **review** - Written comment provided by the customer

Perform the following tasks using the techniques introduced in Chapter 3 of the lecture notes:

a) Derive a sentiment score for each review based on the written comments. Explore how the score you obtain compares with the numerical rating given by the customer.

b) Estimate a LDA model with $K = 2$ topics. Explore the relation between the proportions $\psi_{1,1}, \ldots, \psi_{N,1}$ and the numerical score given by the customer. In your analysis you should carefully consider which words to include in the analysis.

c) Based on your results in parts a) and b), discuss the performance of sentiment analysis and LDA in terms of identifying whether a review by a British Airways customer is more positive or more negative.