# Exercises: Strings in 'Alice in Wonderland' YSC2210 - DAVis with R

Michael T. Gastner

### 1 Introduction

A common task in computational text analysis is to study word frequency distributions. They can reveal content and style of the analysed text. Word frequency can also help us to classify text (e.g. by topic). In this exercise, we determine the most frequent words in Lewis Carroll's novel 'Alice's Adventures in Wonderland'. By the end of this exercise, we present the results as a word cloud similar to figure 1.

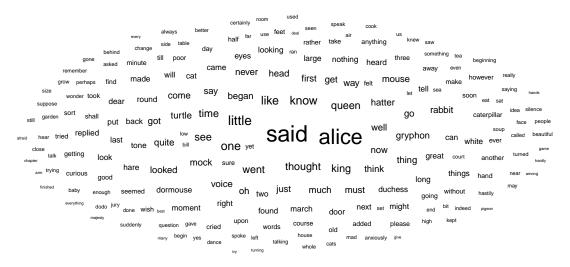


Figure 1: 200 most frequent words in 'Alice's Adventures in Wonderland.'

#### 2 Data

The text of 'Alice's Adventures in Wonderland' is freely available at https://www.gutenberg.org/ebooks/11. Download the text as 'Plain Text UTF-8' and save it in your project directory.

## 3 Objective

We apply functions from the **stringr** package to practise string parsing and regular expressions. We encounter some typical problems in text analysis (e.g. how to handle stop words). We also add word clouds to our plotting repertoire.

#### 4 Tasks

(1) Import the text of 'Alice's Adventures in Wonderland' as a single character string (i.e. a character vector of length 1) called alice.

- (2) Split the text in alice into lines. In this text, lines end with the regular expression "\r\n". Turn the result into a character vector with one element for each line of text.
- (3) Remove text from the beginning and end of the vector alice that is not part of the novel (e.g. front matter and information about Project Gutenberg).
- (4) Split alice into a vector in which each element is one word (i.e. separated from other words by whitespace).
- (5) Find the longest word(s) in alice. (There may be more than one word of the same maximum length.) Should these strings really be treated as single words? You may want to write a function find\_longest\_words() because we want to perform this task a few more times in later sub-tasks.
- (6) In the previous sub-task, you should have found that the largest 'word' contains em-dashes. An em-dash (—) is longer than a hyphen (-). Split elements in alice at em-dashes. What are now the longest words?
- (7) Is there a punctuation symbol at the end of the longest word you found in the previous sub-task? Use a regular expression to answer this question.
- (8) Remove punctuation symbols at the end of all words. Afterwards, confirm that now there are no words in alice ending with a punctuation symbol. What are now the longest words?
- (9) Find out whether alice contains words starting with a punctuation symbol.
- (10) Remove punctuation symbols from the start of all words. Afterwards, confirm that now there are no words in alice starting with a punctuation symbol.
- (11) Remove empty character strings from alice.
- (12) Change all curly quotes ' to straight quotes '.
- (13) Find all spellings of 'drink' with any combinations of upper and lower case letters. (You may need to search the World Wide Web for an elegant solution.) Do you think we should differentiate between words if they only differ in the letter case?
- (14) Turn all characters in alice into lower case characters.
- (15) What are the five most frequent words in alice? Are you surprised?
- (16) It is common practice in text mining to remove 'stop words', which are words that are common in almost every text written in English. Let us remove stop words from alice because, otherwise, our results would not reveal much information about 'Alice's Adventures in Wonderland.'
  - There is no single list of stop words that everybody agrees on, but the **tm** package contains a function **stopwords()** that returns a vector with common stop words. Remove the corresponding elements from alice.
- (17) What are the ten most frequent words now?
- (18) The **ggwordcloud** package contains a function **geom\_text\_wordcloud()** that adds a word-cloud geom to **ggplot2**. Read the documentation and run the examples shown there. Then make a word cloud that shows the 200 most frequent words in **alice**.
- (19) Briefly comment on the word cloud. What does it reveal about the content and style of 'Alice's Adventures in Wonderland'?