Text Mining: Assignment 1

Joris van Vugt, s4279859

September 7, 2016

1 Introduction

The goal of the assignment is to perform some *natural language processing* (NLP) on a text file generated by an *optical character recognition* (OCR) algorithm from an old Dutch text. The text should contain one senctence per line. However, spacing around page numbers and footnotes should be maintained. Sentences can be broken up over multiple lines, so these linebreaks should be removed. Even single words can be broken up across multiple lines and hyphenated. In this case, the hyphen should be removed.

2 Difficulties & Considerations

First of all, the text generated by OCR contains a lot of mistakes. Some common mistakes are:

- Individual letters are mistaken (e.g., an *'c'* instead of an *'e'*)
- Extra whitespace is added
- Non-standard characters are used (e.g. the *'soft-hyphen'* and various quotation marks)

Letters that are mistaken for another are usually not a problem for this task. However, this can be a problem in detecting page numbers or footnote marks. I have chosen to ignore this problem, as it does not cause many mistakes in the output.

Extra whitespace is also not usually a problem, but can be cause trouble when trying to distinguish abreviations from the end of a sentence. If a person's initials are seperated by a space (e.g., *"J. J. L. van Vugt"*) the same pattern occurs as at the start of a sentence: period and a space followed by a capital letter. To avoid this problem I check if the capital letter is not followed by another period.

Another issue has to do with OCR and character encoding. Apparently, the OCR had access to a wide range of *Unicode* characters. As a result, it uses some less common characters. The most important one is the soft-hyphen, which is used to break up a word over multiple lines. I have replaced all soft-hyphens with regular hyphens, so I don't need to write any special cases. I have left all other characters as-is.

The last decision I made is about titles. In the original text, titles are all capital letters with no punctuation. These titles should probably be kept as a single sentence. However, in the text generated by OCR the capital letters are sometimes mistaken for lower case letters, making it extremely difficult to distinguish them from normal text. Therefore, I have decided to not do anything special for titles. As a result they are prepended to the next line.

3 Implementation & Results

I have implemented the program in Python. The program loads a given text file, manipulates it using regular expressions and then saves it to another file. The algorithm works fairly well, although it fails to detect footnotes and still makes some mistakes with sentence splitting. Footnotes are very hard to distinguish from normal text, since they don't always start with a number after OCR and normal sentences might also start with a number. Wrongly split sentences are often the cause of an abreviation followed by another capital letter or lines that end in a non-alphanumeric or standard punctation character.