Computational Methods for Document Analysis 02

Mirco Dietrich
Maurice-Roman Isele

23. April 2016

Exercise 2

Exercise 3

One of the obvious problems was to identify the abbreviations used by the people reviewing. We solved this by looking at the data and furthermore considering all common abbreviations used in the English language. Additionally, any name abbreviations are possible, e.g. *"Mr. K."*, *"T.J."*, and so on.

A more complex problem was to decide whether an abbreviation is the end of the sentence or not. For example, in *"[...] mental insanity, suicide, etc. However, if you don't want [...]"* the abbreviation is the end of a sentence, but in *[...] acting, storyline, etc. of this were good [...]* it is not. You can't decide this by only considering punctuation and a list of all abbreviations.

You could say that an uppercase letter helps you identify the beginning of the next sentence in this scenario, but you have to consider 1) you always write names in uppercase and 2) you can't count on the user to use correct spelling and grammar.

The last problem was dealing with multiple dots/question marks/exclamation marks, and again deciding whether the sentence ended after multiple dots.

Exercise 4