Statistical Natural Language Processing
SS2016 Exercise 1

Junzhe Zhu, Xiaoyu Shen

May 3, 2016

1 Text Preprocessing

1.1 Text Preprocessing techniques

- Stop Word Removal : Stop words do not contribute to the context or content of textual documents.While their frequency of occurrence is high.
- Different character cases : The character has different cases gives the same meaning for the tokens we are counting.
- headers and footers : They may be in the every page and might not be the actual words.

1.2 Tokenizer for Zipf's Law

The book *The adventures of Tom Sawyer by Mark Twain*, as recommended in the assignment sheet. It has three different version, English, German and Finnish.And the UTF-8 encoded plain text files are used.

2 Zipf's law