Text Analysis Report

via Hatfield & Hogg's Text Analysis App

Date: 2023-04-06

Summary statistics

A corpus consisting of 18 documents was submitted for analysis. Text was tokenised by words and no stop-words were omitted from their contents.

This resulted in the total count of all tokens in the corpus being 4589, with the mean token count (words) per document being 254 with a standard deviation of 60.6434877.

The document with the highest token count (373 words) was 2673400.txt, whilst the document with the lowest token count (167 words) was 1067080.txt.

A summary of these statistics can be found in Table 1 below.

Table 1: Summary statistics for the current corpus.

Token type	Mean document token count	Std. dev.	Min.	Max.	Total corpus count
words	254	60.64349	167	373	4,589

Token frequency

Most frequent terms across the corpus

This plot presents the most common tokens in your text data by calculating their frequency (number of times occuring) (Fig. 1).

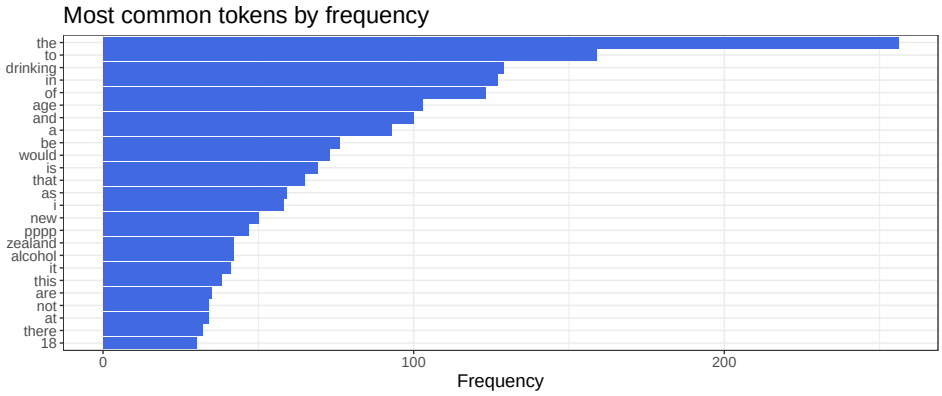

Figure 1: Frequency of most common tokens across the corpus.

Token frequency comparison

Fig. 2 plots the proportion of the total tokens that a particular token takes up in the single text compared to the corpus token proportions.

Tokens closer to the line have similar frequencies in both sets of texts.

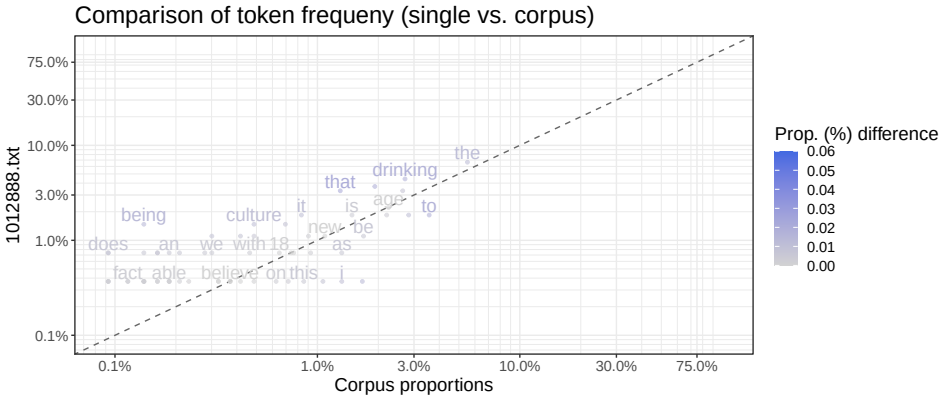

Figure 2: Comparison of token frequencies between a single selected text and the current corpus.

Zipf's Law

Zipf's law states that the frequency of a word appearing is inversely proportional to its rank.

$$\text{frequency} \propto \frac{1}{\text{rank}}$$

This inverse proportional relationship is visualised by plotting these values on log scales in Fig. 3.

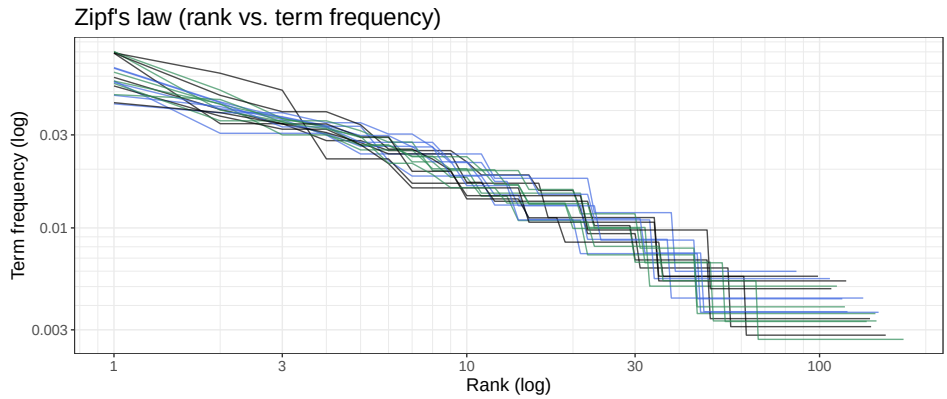

Figure 3: A demonstration of Zipf's Law: term frequency against rank on log−log scales.

Term frequency-inverse document frequency (tf-idf)

Term frequency-inverse document frequency (tf-idf) is a method used to quantify which words are important to a document.

Tf-idf is the product of a term's frequency and the inverse document frequency, producing a statistic which measures how important a term is to a document.

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

The method aims to decrease weighting of common terms and increase weighting of terms with low frequencies.

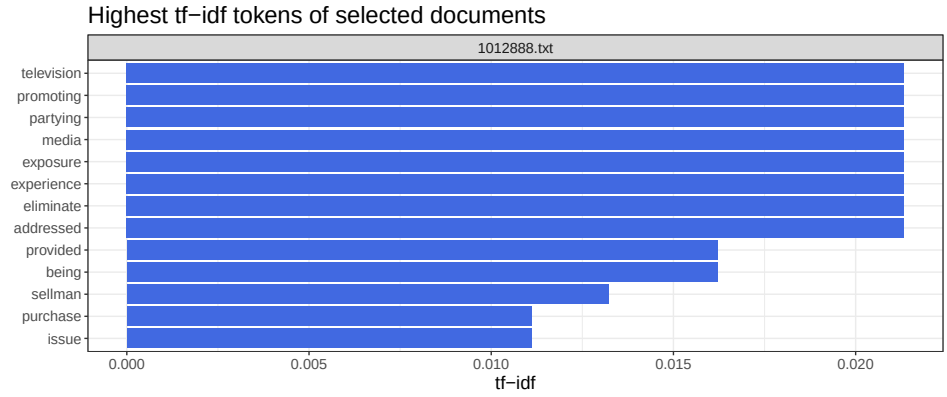

Figure 4: Most important terms according to term frequency−inverse document frequency analysis.

Tf-idf analysis shows some of the most important words to the document 1012888.txt are: addressed , eliminate , and experience .