Frequencies

Prof. Eric A. Suess

March 22, 2021

Introduction

A central question in text mining and natural language processing is how to quantify what a document is about.

Word Frequenies

Can we do this by looking at the words that make up the document? One measure of how important a word may be is its term frequency (tf), how frequently a word occurs in a document, as we examined in Chapter 1.

Document Frequencies

Another approach is to look at a term's inverse document frequency (idf), which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.

TF-IDF

The statistic tf-idf is intended to measure how important a word is to a document in a collection (or corpus) of documents, for example, to one novel in a collection of novels or to one website in a collection of websites.

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

The Wikipedia page about tf-idf is useful.

Zipf's law

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

The Wikipedia page about Zipf's law is very helpful.

Power law.

TF-IDF

The point of tf-idf; it identifies words that are important to one document within a collection of documents.