Text Mining and Analysis

Harper Owen

February 25, 2016

Project Overview

For my mini-project I decided to explore an extremely interesting lexical phenomenon. Called Zipf's Law, the rule states that for any language (even those that are completely untranslated), the rate of word frequency will follow a logarithmic pattern with the most second most frequent word being used 50% as much as the most frequent word, the third most frequent being used 33.3% as much, and so on. If this were plot on a logarithmic scale, it would be a straight downward sloping line. I wanted to explore how this phenomenon was changed based on type of language (literature, poetry, speeches, etc).

Implementation

To test the hypothesis that Zipfs law would apply to any type of text, I first had to download different types of texts. The ones I chose include:

Book
: Oscar Wilde's *A Picture of Dorian Gray*

Poetry
: TS Eliot's *Collected Poems*

Speeches
: Abraham Lincoln's *Complete Collection of Speeches & Letters ...*

After mining these from Project Gutenberg and pickling them, I processed each of the texts by stripping them of punctuation, converting to lowercase, and splitting them into words. To do this I used several modules including *collections* and *BeautifulSoup*. I then created a function called *histogram* to count the frequency of each word, and another function entitled *xaxispts* to take the length of the list of frequencies and output an index for each one so that I could plot the frequencies. After this, I implemented matplotlib to plot frequency vs. index on a log-log scale.

Results

The results I calculated were fairly consistent for all three data sets. Each showed a definitive downward trend, indicating that Zipf's law applied to all of the texts, regardless of form or style. This inidicates that not only do we write according to Zipf, but be also speak, and thus think in this pattern.

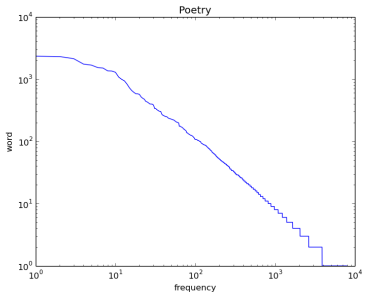

Figure 1: Graph of Book

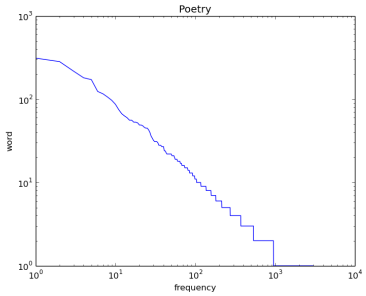

Figure 2: Graph of Poetry

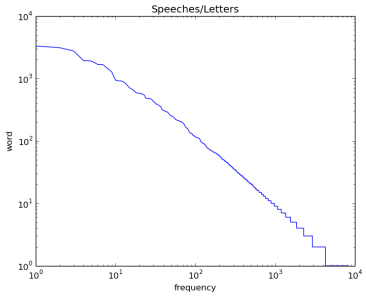

Figure 3: Graph of Speeches/Letters

Reflection

I think this project was an effective validation of my knowledge of dictionaries, lists, and functions. I feel that I wrote efficient code and used modules and other tools to the best of my ability. I was frustrated that I couldn't download different texts, such as blog posts, transcripts, and song lyrics, however due to a series of bad life choices I only tried to expand the project last night and the documentation for pattern refused to function. Overall I think this project went well in that I learned a lot about using python to interact with the internet and I created something that (at least I believe) is mildly interesting and scienficically useful.