Tokenization with Regex on a small subset of the Gutenberg project texts

Analyzing the effects of punctuation and case in tokenization on the frequency distribution of tokens

A report by Aime Cesaire Mugishawayo

Findings

1. **Most Frequent Word**

- I have found that `the` is often but not always the most common word in each individual text. When the text has been converted to lowercase and the punctuation removed, in 18 texts, `the` was most frequent 16 times. `and` was the other common word, appearing as most frequent in two texts and as the second-most common in nine.

2. **Impact of Lowercasing**

If the words are not lowercased before we count them, then, `the` is only the most common word in 14 texts. `to` is the most common in two texts and `and` and `I` in one text each.

3. **Frequency Graphs**

    1. The graphs of `frequency_counts` against `words by frequency` by the function `plot_frequency()` matches the graph of an exponential decay.
        - This finding suggests heavily skewed probability distribution of tokens in the overall Gutenberg project, and perhaps in the English language.
        - This is perhaps to be expected. The use of vocabulary in English tends to skew heavily towards the most common words, such that they appear orders of magnitude more times than what one would consider 'normal frequency' words.
    2. On the graphs of `plot_frequency_by_length()` , we see that when we sort the words by length, the plots look like exponential decay functions arranged in layers, with each subsequent layer shifted to the right. Each subsequent layer is shorter than or equal to the one to its left/before it.
        - There can be many theories for this. My theory is that at the smallest length range we have an exponential decay distribution that almost gets repeated at twice or thrice the size because those words are usually combinations or extensions (derived from) of the shorter ones. For example, the probability distribution of a verb will inform that of its -isms, -ments, -ations, and -allys (nouns and adverbs derived from it) as people likely use them as much (relative to size) as they use the root verb. So you can expect that if some words dominate the distribution between length one and three, their derivations will dominate at maybe length six to nine.
        - The effect tapers off as length increases.

4. **Impact of Removing Punctuation**

- Interestingly, the exponential decay shape holds for both 'punctuation-treated' and 'untreated' datasets, and it is smoother in the untreated ones i.e texts where a punctuation is not a token on its own. One reason may be that it breaks up the overwhelming probability mass of the highly probable words and give some of it to their punctuated siblings.
- The trend of "layers of distribution" is both visible in treated and untreated distributions, albeit having a smoother look with untreated texts.
- Removing the punctuation approximately doubles our maximum count as well as counts for most frequent words.