Information Retrieval and Data Mining (COMP0084)
Coursework 1

Abstract

In this coursework, Task1 would use the data in passage-collection.txt to generate plots of empirical distribution and Zipif's law distribution. Task 2 would use the data in candidate-passage-top1000.tsv to generate an inverted index for the collection. Task 3 would use test-queries.tsv and candidate-passages-top1000.tsv and two retrieval models would be implemented. Task 4 would use test-queries.tsv and candidate-passages-tsp1000.tsv and three query likelihood language models would be implemented.

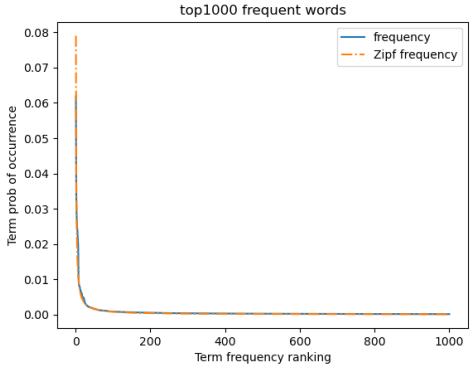

Figure 1: Empiricial and Zipf's law comparison

1 Task1 - Text statistics

1.1 Deliverable 1

In this section, the stop words in the dataset would not be removed, the string in documents would be converted to lowercase letters. After that, the punctuation in the documents would be removed. The preprocessed data would be stored in list format. The lemmatisation and stemming steps would not be selected in this task, which would be hidden in the source code.

To report the size of the identified index of terms, a function would be implemented to display the occurrences of the terms. The function would return a dictionary, the tokens would be the key and occurrences would be the value. As a result, there are totally 10061726 tokens and tokens and 174911 indentified terms.

In this task, the probability of occurrence (normalized frequency) against their frequency ranking as well as the one with Zipfian distribution would be plotted. the expression of Zipf's law could be displayed as following:

$$f(k; s, N) = \frac{k^{-s}}{\sum_{i=1}^{N}} \tag{1}$$

In this case, the s for text sets is 1. The comparison of the empirical and Zipf's law distribution could be shown in figure 1.

The log-log plot of the two distributions could be shown in figure 2.

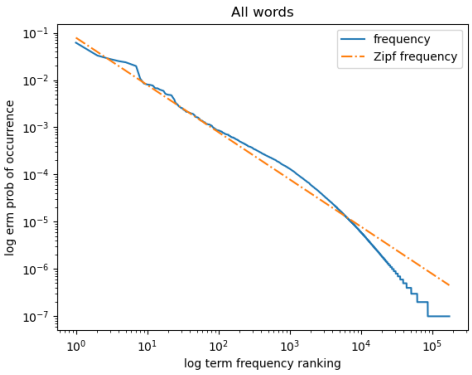

Figure 2: log-log of Empirical and Zipf's law comparison

According to the results, it is obvious that the Zipf law frequency predicts more high probability terms compared to the Empirical method. In the region $10^{0.5} - 10^{4}$, there are more terms according to the empirical method compared to the Zidf law. It is clear that the log-log plot of the Zipf law is proportional to k^{-s}. The trends of the two methods are the same.

When the step of removing stop words is applied, the results could be shown in figure 3.

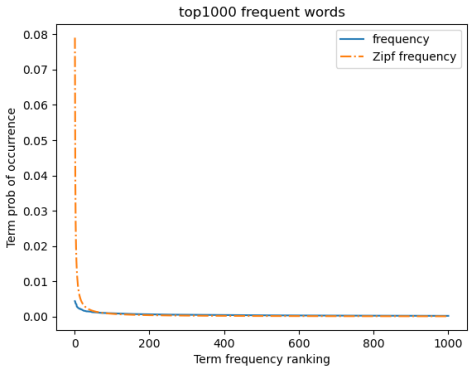

Figure 3: Empiricial and Zipf's law comparison with removing stop words

The log-log plot of the two distributions could be shown in figure 4.

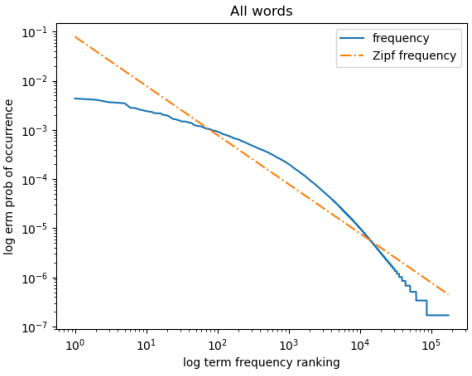

Figure 4: Empiricial and Zipf's law comparison with removing stop words

According to the results, it is obvious that the empirical distribution would be affected significantly and the Zipf's law distribution remains the same. For the empirical distribution, the probability of the terms with high ranking would be reduced due to the stop words removed.

2 Task 2 - Inverted index

2.1 Deliverable 3 & 4

To generate the inverted index, the **candidate-passages-top1000.tsv** file is used. There are serval steps including.

data preprocessing In this section, all the stop words in the data set would be removed, which is different from task 1. After removing the stop words, all the terms would be processed by lemmatization and stemming.

generating inverted index In this section, the inverted index would be generated. the **token**, **pid**, **count** information would be stored in a dictionary, which is the output of the function. The format of the dictionary could be displayed as **{token:{pid:counts}}** format.

3 Task 4 - Query likelihood language models

3.1 Deliverable 11

Which language model do you expect to work better? According to the concepts of the three smoothing methods, Dirichlet smoothing should be the best method. Compared with the Laplace smoothing and Lidstone smoothing, which are the discounting methods, the interpolation method would consider the relative frequency of a word in a large collection instead of treating unseen words equally.

Which language models are expected to be more similar and why? Laplace smoothing and Lidstone smoothing should be more similar and they are different in the weights to unseen terms but these two methods use the same mechanism.

Comment on the value of ϵ = 0.1 in the Lidstone correction The chosen value of ϵ is suitable in this case. In the given dataset, the document lengths are not too large and the ϵ value should not be too large so that the unseen word would not be shared with too much weight.

Would $\mu = 5000$ be a more appropriate value the 5000 for μ is not a good choice. Because the $\lambda = \frac{N}{N+\mu}$, if the value of μ in this case is too large, the value of λ tend to close to 0, which means the background probability would dominate the results of the estimate of the word. As a result, the $\mu = 5000$ should not a good choice in this case.