Parametric Classification using Zipf's law

Janani Kalyanam
University of Wisconsin, Madison
kalyanam@cae.wisc.edu

Abstract

The project explores the possibility of classification using Zipf's law. Zipf's law states that, given some corpus of natural language utterances, frequency of occurrence (f) of any word is inversely proportional to its rank (r) in the frequency table. $f \propto \frac{1}{r}$. Hence Zipf's law gives some insight about the frequency distribution of the words in a document. Our goal was to use this information to see if similar documents indeed form clusters. We represent each document with 3 parameters. Results have produced clearly discernable clusters of different document classes.

1 Introduction

Document classification techniques are used to enhance information retrieval. Information retrieval deals with searching for documents, text or data within a document etc. Documents with similar content are classified into a single group in order to easen the process of information retrieval. There exist different kinds of classification techniques: supervised classification where an external system (humans) helps classify each document, unsupervised classification where an automated algorithm classifies each document, and semi-supervised classification which is a mixture of both supervised and unsupervised classification. Generally, in semi-supervised classification, a training set (called labeled data) is used to produce classification boundaries/thresholds using 'some heuristics'. These classification boundaries/thresholds are in turn used on the unlabeled data which are hence classified. The contribution of this project is the proposal a possible 'heuristic' that could be used to classify documents.

2 Zipf's Law and Mandelbrot's Law

Given a document, there are various basic parameters that characterize it. Some immediate questions like - total number of words in the document, number of distinct words (formally known as word types), and the frequency of each word (formally known as word token) arise. Linguist George Kingsley Zipf first proposed Zipf's law which relates the frequency of occurrence of each word, to its rank in the rank-frequency table. That is,

$$f = k\frac{1}{r} \tag{1}$$

The significance of Zipf's law is that, the document is extremely sparse for most of the word types. There exist very less number of word types that occur frequently. Zipf claims that this law exhibits the *Principle of Least Effort* which argues that humans will act in such way as to minimize the average effort.

Mandelbrot's law characterizes the rank-frequency relationship in a document by a richer choice of parameters. It expresses frequency as a function of rank with the following equation:

$$\begin{aligned} f &= \theta_1(r+\theta_2)^{\theta_3} && (2) \\ \log f &= \log\theta_1 + \theta_3 \log(r+\theta_2) && (3) \end{aligned}$$

It is noted that Eq[2] reduces to Zipf's law (Eq[1]) for $\theta_2 = 0$ and $\theta_3 = -1$. From this equation, it can be seen that:

- θ_3 will always be < 0. Its stands for the rate of decay of f.
- θ_1 provides a constant shift in the rank-frequency curve and its value is representative of the size of the corpus. For example, if two documents created from same vocabulary have the same θ_3 (assume $\theta_2 = 0$), but different θ_1, then it means, for the every rank, one of the documents has higher frequency then the other.
- θ_2 is like an extra leeway term which is added to the original rank. Generally large corpora, have

extremely long tails. (Most of the words occur only once). In such cases, θ_2 expected to be large to serve the purpose of better curve fitting.

We parametrize each document from various classes using $\theta_1, \theta_2, \theta_3$.

3 Simulation and Results

Four classes of documents, ~ 100 movie reviews, ~ 100 news articles, ~ 10 debates and ~ 10 novels were collected.

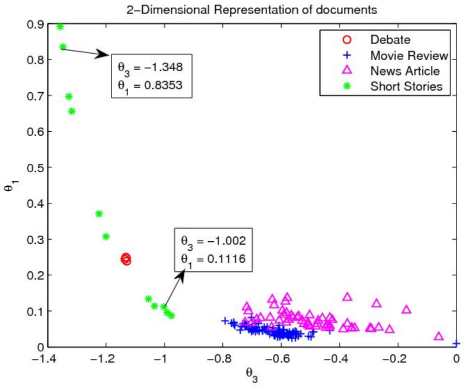

Figure 1: 2D representation of document

Each document was stemmed and tokenized following which normalized frequency and rank for each document was found. In Eq[3], θ_2 was set to zero, and a straight-line fit for each document was obtained. Using the parameters of the straight-line fit, θ_3 (which is the slope of the line) and θ_1 ($\log \theta_1$ is the y-intercept) was obtained, and each document was represented using these two dimensions. It can be seen in fig[1] that movie reviews, news article and debates form tight clusters, while there is a lot of spread with the short stories (novels). We explore this a little more by taking 2 samples of short stories marked in fig[1], one of which has $\theta_1 = 0.1116$, and the second $\theta_1 = 0.8353 \approx 8(0.1116)$. We remind ourselves about the argument made in section[2] that θ_1 is representative of the length of the document. Conforming with the arguement in section[2], the size of the first corpus was found to be 20319, and that of the second was 146828, which is $\approx 8(20319)$.

Figure[2] incorporates θ_2 as well in the plots. The approximate position of the same two points from fig[1] are indicated. Arguement about θ_2 made in section[2] can again be confirmed by the fact that larger θ_2 corresponds to the larger corpus. It can also be seen that there is not much spread along the θ_2 axis which is again attributed to the fact that except for short stories, most other documents were short in length.

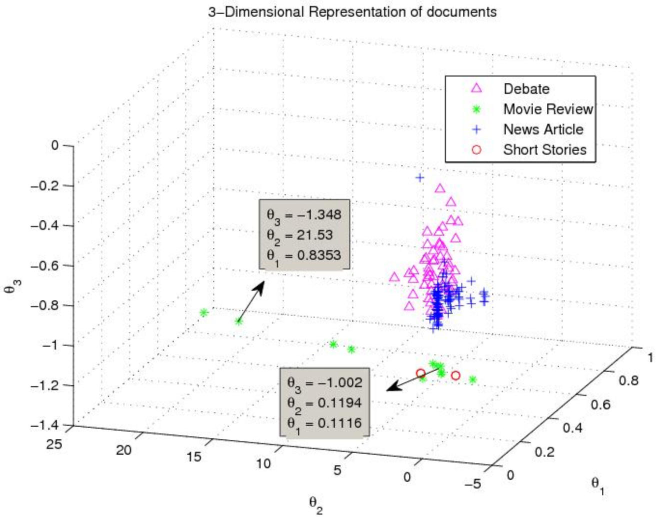

Figure 2: 3D representation of document

4 Conclusion

It can be seen that the documents from the same label are indeed clustered together. One advantage of the process is that it requires very less pre-processing work. The flip side of the coin is in the fact that the semantic information from each document is lost.

References

[1] C.Manning, H.Schutze, *Foundations of Statistical Natural Languege Processing*, MIT Press, Cambridge MA, 1999

[2] J.Zhu, *Class Notes*, Madison, WI, 2009.

[3] Movie Reviews `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

[4] Debates `http://www.cnn.com/`

[5] New Articles `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[6] Short Stories `http://www.gutenberg.org/`