Identification of Lost or Deserted Written Texts Using Zipf's Law with NLTK

Devanshi Gupta, Priyank Singh Hada, Deepankar Mitra, and Niket Sharma

Manipal University Jaipur, ComputerScience Department, Jaipur, India
{devanshigupta06,pshada1,deepankar.mitra4ever,
niket.sharma1408}@gmail.com

Abstract. Sometimes it becomes very difficult to identify the valuable text written by some great personalities; especially when the text is not having a signature or the author is anonymous. Deserted manuscripts or documents without a title or heading can be an additional pain. It might happen that the work of dignitaries are lost or only some part of their valuable piece of work is found available in the libraries or with other storage media's. By deploying Zipf's law with the NLTK module available in python, this problem can be solved to a great extent, helping save the originality of the valuable texts and not leaving them unidentified. This can also be helpful in some real time data analysis where frequency plays an important role; plagiarism detection in written texts is one such example. NLTK is a strong toolkit which helps in extracting, segmenting, parsing, tagging and searching etc. of many natural languages with the help of python modules. In this paper it has been tried to combine Zipf's Law with NLTK to come up with a tool to identify the anonymous or deserted valuable texts.

Keywords: Keywords: NLTK, Zipf's law, NLP.

1 Introduction

Dealing with Natural Language Processing (NLP) is a fascinating and active research field. With the help of Natural Language processing Toolkit (NLTK) module available in Python, graphical analysis of natural languages has also become possible. The graphical user interface present in NLTK helps to plot and study graphs to understand the results of natural language processing in a much better an efficient way. NLTK suits both linguists and researchers as it has enough theory as well as practical practice examples in the NLTK book itself [1], [2]. NLTK is an extensive collection of documentation, corpora and few hundreds exercises making it a huge framework providing a better understanding of the natural language's and their processing. NLTK is entirely self-contained providing raw as well as annotated text versions and also simple functions to access these [3], [4].

Moving to Zipf's law; it is based on power law distribution which emphasizes that if given a corpus containing utterances of a natural language, then the frequency of any word in the given corpus is inversely proportional to the rank in frequency table so formed. It implies that the word which is occurring most frequently will occur

M.K. Kundu et al. (eds.), *Advanced Computing, Networking and Informatics - Volume 1*,
Smart Innovation, Systems and Technologies 27,
DOI: 10.1007/978-3-319-07353-8_59,

approximately twice as the frequency of the second most frequently occurring word in the same corpus and so on. The general Zipf's law studies the influence relation that frequency is suffered by rank where rank, is equivalent to the independent variable and frequency can be looked upon as the dependent variable. It is observed that the randomly generated distribution for frequency of words in texts is quite similar to Zipf's law. Occurrence frequency of any word is similar to the inverse power law taking into consideration its word rank and the exponent of the inverse power law is found almost very close to 1, because the transformation from the word's length to its corresponding rank stretches or expands an exponential behavior to power law function behavior [5].

In this paper Zipf's law has been deployed along with the available corpus in NLTK module in python in order to generate plots. Further, analysis of these plots have been done to show how a particular style of writing hampers a writer in his writing and usage of words and grammar in the related context. The frequency of words in various texts has been calculated and the results have been obtained in the form of graphical plots which has helped to show the existence of similarity between various texts by the same writer/ author.

2 Background Theory

Many independent works have already been done using either NLTK or Zipf's law. But, a combination of two has made no considerable contribution in the present influenced world of natural language processing.

Several works has been done in NLP wherein two parsers have been fused together [6]. In this paper the author has described how in present scenario where natural language processing applications and its implementations are equally existent for the programmers having null linguistic knowledge were being able to build some specific NLP linguistic systems. Two parsers were fused together to achieve more accuracy and generality were tested on a corpus showing a higher level of accuracy in the results.

Other previous works in NLTK include a focus on computational semantics software and how it is easy to do computational semantics with NLTK [7]. This paper shows how Python gives an advantage to those students who are not exposed to any programming language in the merits of Prolog and Python when compared relatively. Zipf's law was introduced by George Kingsley Zipf and proposed it in Zipf 1935, 1949. It is an empirical law which is formed using the mathematical statistics, and it is referred that various kinds of data types have been studied in the fields of physics as well as social sciences which can be aptly approximated in accordance with Zipfian distribution, which is very much related to discrete mathematics probability distributions of power law family.

On the other hand Zipf's law like distribution have been implemented in various real time and internet analysis, where an analysis has been done to check the revenue

of 500 top firms' of China and its rank and frequency following Zipf's like distribution with inverse power law where the slope is found to be close to 1 [8].

3 Motivation

Significant work has been done independently using NLTK or Zipf's Law. This work has tried to combine the above two where NLTK has been used to get some of the texts of the Project Gutenberg to do the analysis and Zipf's law has helped to plot and study those analyses.

Project Gutenberg is an electronic text archive which preserves cultural text, which is freely available to all and a small section of texts have been included in NLTK creating a Gutenberg Corpus, which is a large body of text [9]. To see the distribution of words and texts and the frequency of words used Zipf's law has been implemented on some of the texts of the Gutenberg Corpus in NLTK. The Gutenberg Corpus of NLTK contains three texts by Chesterton, namely; Chesterton-ball, Chesterton-brown and Chesterton-Thursday and three by Shakespeare, namely; Shakespeare-hamlet, Shakespeare-Macbeth and Shakespeare-Caesar, amongst other texts in the corpus. These are the various texts written and composed by the same author/ writer. The word type and its frequency of occurring in the text singly and comparatively has been checked using NLTK module and using Zipf's law, comparative graph has been plot to analyze the results of the research.

Here it is tried to see how much the writing is hampered, influenced and varies according to the speech and types of words used by a particular author/writer. Writing, usage of words, frequency of using words, adjectives, presenting those phrases and everything differs and varies person to person. Even a common man has a different way of writing essays and texts than another common man.

Often it is found that the problem of text kept or preserved anonymously, sometimes it is done deliberately on author's request but sometimes it's not. This valuable text or cultural texts' identification becomes important so that its master's originality can be kept intact. This can be done by employing Zipf's law on the text and counting the words and its corresponding frequency in the respective text. Here after plotting a comparative graph of the same with some identified text which is thought to be written by the same author as this unidentified one.

The results can then show that whether the text has been identified is according to one which is assumed or not.

A number of corpora from NLTK have been studied and various comparison results have been plotted to study the Zipf's law and its scope and slope over natural languages and words from daily usage of various language but mainly English. As a first step the frequency distribution is calculated and stored in a file for some nine corpuses from a suite of corpuses in NLTK namely Reuters Corpus, Project Gutenberg, Movie Reviews Corpus, State Union Corpus, Treebank, and Inaugural Corpus from USA, Webtext Corpus, Nps Chat Corpus, and Cess Corpus. Frequency occurrence of words, word, and its rank are also calculated in a file and finally a comparison plot is generated among various corpuses and texts of NLTK.

4 Getting Started

Zipf's law is explained as let f (w) be the frequency of a word w in any free text. It is supposed that the most frequently occurring word is ranked as one, i.e. all words in a text are ranked according to the frequency of occurrence in the text. According to Zipf's law; frequency of occurrence of a word of any type is inversely proportional to its rank.

$$f \alpha 1/r$$
$$f = k/r$$
$$f * r = k$$

where k is a constant.

Power law is when the frequency of any event varies as the power of some attributes of that event the frequency is said to follow a power law. Significance of power law here is taken in sense of logarithmic scale because values here are very large. So to denote these values on scale, a power law is used in form of log scale to demonstrate these values on x-y axes.

At first it is started with counting words in the text and calculating how many times each word is appearing in the given text. Each word in each line is read by stripping of the front and back whitespaces around the word, converting each word into lowercase to make the things a little more manageable. A dictionary is maintained to store all the words with their corresponding frequency.

Next step is assigning the rank to each word. Rank of a value is one plus the number of higher values. This implies that if an item x is taken wherein it is said that x is the fourth highest value in the list means its rank is four then it is said that there are three other items whose values are higher than x. Therefore, to assign a rank to each word by its corresponding frequency, it should be known that how many words have higher frequency than that word. So grouping is to be done first i.e. group the words according to their frequency. Certain boxes are taken and each box is labeled according to the frequency of the word i.e. putting all words appearing three times in a corpus, in a box labeled as three or 3 and so on. To check how many words have higher frequency than a given word, the label of the box is checked to which a selected word belongs to and then adding up the number of words in the boxes with larger labels.

Now putting up all the words having the same frequency into a box representing that box with the corresponding frequency, then to determine the rank, see which box it belongs to, identify all boxes that keep words with higher frequency then add number of words that belong to the box identified in the previous step and lastly add one to the resulting sum. To put all words with the same frequency into a box with that frequency, a dictionary is used and the dictionary created earlier is used to lookup for the frequency of the word and a rank function is created with three parameters as word, frequency of word and group of words in box dictionary.

It is also observed that many real systems don't show true power law behavior because they are either incomplete or possibly inconsistent with the undertaken conditions under which it is expected for power laws to emerge. In general Zipf's law doesn't hold for subsets of objects or events or a union or combination of Zipfian sets. Some missing elements produce deviations from a pure Zipfian or Zipf's law in the

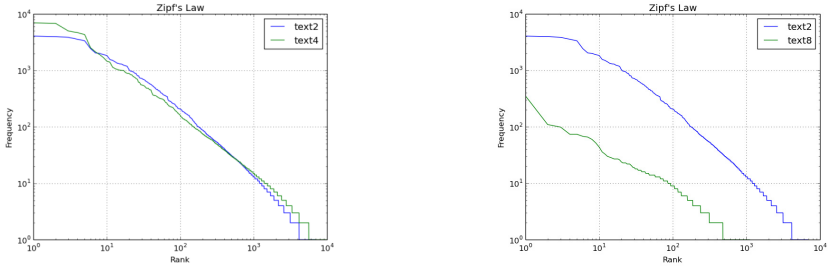
subset. The line is sometimes not well fit, for highest rank words and lowest rank words which are overestimated and underestimated respectively.

5 Result Analysis

At first Zipf's law is implemented on text 2: Sense and Sensibility by Jane Austen 1811 and text 4: Inaugural Address Corpus of NLTK and found that Zipf's law is obeyed with some deviations at the extremes and it is also seen that there are quite some similarities in these two texts (Fig. 1(a)), i.e. the way in which the two texts are written and the usage of words and its frequencies is quite similar and in fact the graph intersects at a place near the lower mid signifying that some of the words usage is same.

Similarly implementing Zipf's law on text 2: Sense and Sensibility by Jane Austen 1811 and text 8: Personals Corpus of NLTK, it is found that there are no similarities between words of these two corpuses (Fig. 1(b)) i.e. the frequencies of using words is not similar and a considerable gap is maintained among the two texts.

With the result analyses and comparison on these three texts of NLTK it is found that how one text i.e. text 2 is quite similar to text 4 but the same text i.e. text 2 has absolutely no similarities with text 8; while all the texts deviate at the extremities due to underestimated or overestimated values at the two extremes respectively.



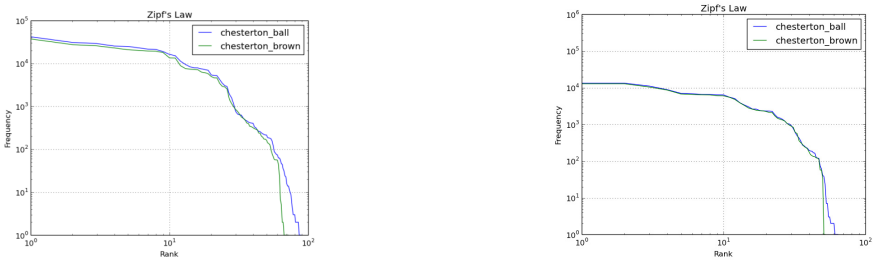
(a) (b)

Fig. 1. (a) Comparision results of text 2: Sense and Sensibility by Jane Austen 1811 and text 4: Inaugural Address Corpus of NLTK, (b) Comaprision results of text 2: Sense and Sensibility by Jane Austen 1811 and text 8: Personals Corpus of NLTK

Now analyzing the results of the main objective i.e. to see how well it works for the texts (taken as sample) by Shakespeare and Chesterton, two great authors and writers and how important it is too see a similarity in their existing texts so that the analyses of these results can be used to test some valuable unidentified or anonymous text; where these two tests and these two authors are just taken to check the results.

When simply the words of the two texts i.e. Chesterton Ball and Chesterton Brown, composed by Chesterton are compared (no frequency count is considered), a graph with almost similar and overlapping pattern is generated till the lower mid and a little deviating in the later part (Fig. 2(a)), which shows that the words used are quite similar.

And when the comparison of these two texts according to the frequency of occurrence of words is done it is found that the graph is completely overlapping except for a few words at the lower extremity. This means that the frequency of using words is very similar in these two texts by the same writer/ author (Fig. 2(b)).



(a) (b)

Fig. 2. (a) Comparision of two different texts written by Chesterton (Brown and Ball), **(b)** Comparison according to the frequency of words in the text Ball and Brown by Chesterton

When the comparison for all the three texts available by Chesterton in Project Gutenberg namely; Ball, Brown and Thursday is done, it is seen that in the composition Thursday the frequency of words used is a little less though the pattern followed is similar but at lower extremity, the frequency of words of text Thursday matches with that of Brown. Here also the frequency of occurrence of words is compared (Fig. 3).

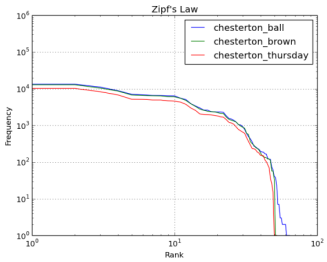
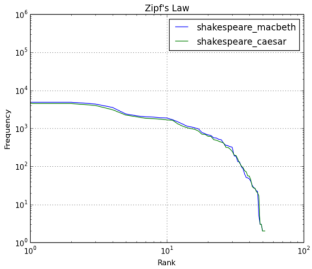


Fig. 3. Comparison according to the frequency of words in the text Ball, Brown and Thursday by Chesterton

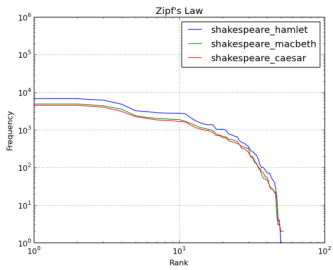
In the figure below, the results say that when the frequency of occurrence of words in both the compositions by Shakespeare are compared i.e. Shakespeare_Macbeth and Shakespeare_caesar, it is noticed that the words usage is almost similar in both the texts and also the graph does not touch the axes at the lower end, which means that there are no words in the text that are used almost negligibly or with very less frequency and hence the usage of words is distributed in a certain pattern (Fig. 4(a)), the words and their frequency count is used for the analyses.

When all the three plays by Shakespeare namely; Caesar, Macbeth and Hamlet, available in NLTK's Project Gutenberg, when included to test the results on the basis of the frequency of occurrence of words, it was noticed that the graph for text Hamlet touches the axis at the lower end signifying that words usage in play Hamlet is different and is not used very frequently making it touch the lower axis. It is also noticed that the graph for Hamlet follows the similar pattern as that of the other two

but the words used in Hamlet are with a higher frequency (Fig. 4(b)), this means the pattern followed by a particular writer is similar though the frequency of using words may vary.



(a)



(b)

Fig. 4. (a) Comparison according to the frequency ofoccurence of words with their corresponding ranks in the text Caesar and Macbeth by Shakespeare, (b) Comparison according to the frequency of occurrence words with their corresponding ranks in the text Caesar, Hamlet and Macbeth by Shakespeare

When the texts by Chesterton are compared, here considering Ball and Brown and the compositions by Shakespeare considering here Caesar and Macbeth are compared with each other, it is found that a considerable gap in the usage of words and the frequency with which they appear and are used by the respective authors and writer (Fig. 5), it is seen that a constituent gap is maintained between the writing patterns of various writers and the frequency with which they tend to use words in their texts and compositions follows a certain kind of pattern.

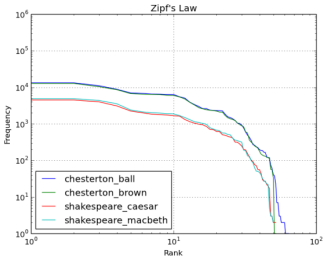


Fig. 5. Comparison according to the frequency of occurrence of words with their corresponding ranks in the text Caesar and Macbeth by Shakespeare and Ball and Brown by Chesterton

6 Conclusion and Future Work

After studying the above results it is concluded that the writing pattern and the usage of words differ from person to person and can help in some major result findings and research. After seeing the result in Fig. 5, it is seen that how the compositions by two people vary and the frequency of usage of words differs producing a considerable gap in the graph.

These results can also help in understanding this kind of pattern which is being followed in many cases and places like the population in largest cities in World, or increase in revenue of an organization over a period of time, to check the progress of students in a year etc., with the help of Zipf's law results generation and its analyses is done to see whether the result is achieved or not. It may also help in some real time aspects around us and their results play a significant role like the progress of students for a school is very important.

Similarly it is very important sometimes to judge the anonymity of a text or composition which can be done through this measure and find the existence. though it is also seen that exact Zipf's law is not followed in here and hence can be said a failure to Zipf's law but the concept can be used to to determine and identify the anonymity of not only texts but THIS can be used in various other aspects as in Intrusion Detection as a future application, where a set pattern of Intrusion can be detected and saved and if any deviation from the existing pattern is found, it can be said that there might be an Intrusion in the organization.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media (2009)
2. Lobur, M., Romanyuk, A., Romanyshyn, M.: Using NLTK for Educational and Scientific Purposes. In: 11th International Conference on The Experience of Designing and Application of CAD Systems in Microelectronics, pp. 426–428 (2011)
3. Abney, S., Bird, S.: The Human Language Project: Building a Universal Corpus of the World's Languages. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 88–97 (2010)
4. Li, W.: Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. IEEE Transactions on Information Theory 38(6), 1842–1845 (1992)
5. Rahman, A., Alam, H., Cheng, H., Llido, P., Tarnikova, Y., Kumar, A., Tjahjadi, T., Wilcox, C., Nakatsu, C., Hartono, R.: Fusion of Two Parsers for a Natural Language Processing Toolkit. In: Proceedings of the Fifth International Conference on Information Fusion, pp. 228–234 (2002)
6. Garrette, D., Klein, E.: An Extensible Toolkit for Computational Semantics. In: Proceedings of the 8th International Conference on Computational Semantics, pp. 116–127 (2009)
7. Chen, Q., Zhang, J., Wang, Y.: The Zipf's Law in the Revenue of Top 500 Chinese Companies. In: 4th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2008)
8. Shan, G., Hui-xia, W., Jun, W.: Research and application of Web caching workload characteristics model. In: 2nd IEEE International Conference on Information Management and Engineering (ICIME), pp. 105–109 (2010)
9. Project Gutenberg Archive, https://archive.org/details/gutenberg