

PAPER • OPEN ACCESS

Towards a graph model application for automatic text processing in data management

To cite this article: E.G. Grigoryeva *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **483** 012077

View the [article online](#) for updates and enhancements.

Towards a graph model application for automatic text processing in data management

E.G. Grigoryeva, L A Kochetova, Y V Pomelnikov, V V Popov, T V Shtelmakh

Volgograd State University, 100 Universitetsky Prosp., Volgograd, 400062, Russia

E-mail: kochetova@volsu.ru

Abstract.Based on the two models, “bag-of-words” and graph model, the paper deals with the development of methods for automated text analysis with the purpose to classify natural language texts and randomly generated documents. Within “bag-of-words” model, the authors have found that the primary Zipf’s law, which states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table, does not hold continuously true. Modifications to this law have been proposed that enable us to classify texts more efficiently. Using the graph model of the text, which takes into account the occurrence of two random words in a sentence, and the median degree of the vertices of the graph, the authors demonstrate that it can be applied to differentiate meaningless texts from meaningful ones even though the word lists of the two texts are identical.

Introduction

Automated text analysis and text mining methods have received a great deal of attention because of the remarkable increase of digital documents. Typical tasks involved in these two areas include text classification, information extraction, document summarization, text pattern mining etc. Most of them are based on text representation models which are used to represent text content. A common and standard approach to model text document is bag-of-words. This model is suitable for capturing word frequency, however structural and semantic information is ignored. Most existing text classification methods (and text mining methods at large) are based on representing the documents using the traditional vector space model [1]. Recently, graph-based models have emerged as alternatives to text representation model. As compared to the traditional models, graph representation is mathematical constructs that represents not only the content of a document but also the relationship among the keywords, therefore, it can model relationship and structural information effectively. A text can appropriately be represented as Graph using vertex (node) as feature term and edge relation can be significant relation between the feature terms [2]. Text representation using Graph model provides computations related to various operations like term weight, ranking which is helpful in many applications in information retrieval.

The application of Graph based representation of text elements provides processing of the information in various areas like document clustering, document classification, word sense disambiguation, prepositional phrase attachment. However, Graph algorithms or techniques need to be extended in order to capture the requirement and complexities of the applications.

As data can be represented by an array of texts, which requires structured representation, it is required to construct a model of ‘useless’ texts in order to solve the “garbage” filtering problem. In the article, we focus on two simple models. The first model A is based on randomly selected wordlist and punctuation marks. The second model B is randomly selected wordlist and punctuation marks, but for each element of the text probability of sampling is given. The purpose of the study is to define text characteristics that identify a text as “garbage” by applying the first or second model. Mathematically, this means building a function that calculates the value that is used to discriminate a text. We



investigate modifications of the Zipf's laws (G.K. Zipf). In the first part of the article, the "bag of words" text model is used, and in the second part of the article a graph model is proposed.

1. Study based on the model "bag-of-words"

1.1. The primary Zipf's Law "rank-frequency"

Let us have a document D . For every word w from the document we can count the frequency of hits $k(w)$, i.e. the number of occurrences of this word in the document. The probability $p(w)$ of the word w occurrence in a random sentence in the document is calculated as the ratio of the number $k(w)$ to the total number of words N_w in the document.

The words that occur in the document are ranked as their frequencies decrease and are numbered 1,2,3. The frequency sequence number is called the frequency rank. The most frequently occurring words will have a rank of 1, followed by 2 and so on. If two words have the same frequency, then their ranks are equal. The text word list will be compiled after lemmatization has been conducted.

The first Zipf's law is usually formulated as a hyperbolic dependence between the rank $r(w)$ of the word w and its frequency $k(w)$. This dependence can be presented as follows [3, p. 378]:

$$k(w) = A \cdot (r(w))^{-\gamma}, \quad (1)$$

where A and γ —some constant values. The parameter γ is usually close to 1 and there are good reasons for which it is convenient to assume that for many texts $\gamma = 1$ [3, p. 386]. If equality (1) is divided by the number N_w of all the words in the analyzed text and put $\gamma = 1$, then we arrive at the equality:

$$p(w) \cdot r(w) = C, \quad (2)$$

Where $C = A/N_w$ —constant value. The value of the constant C is dependent on the text topic and genre: stylistic features, language variance [3] and other features. Thus, it has been found that for the Russian language texts on electronics $C = 0.145$ [5, p. 105]. The rationale for the formation of laws of this type can also be found in [6].

We will study the frequency characteristics of the text based on Zipf's law in the form (2), allowing us to reveal any properties of the natural text represented by the "bag of words" model. The experiments performed show that both for the individual texts and for the collection of texts significant deviations from this law are observed.

For all the words w of the rank r , the value obtained from the formula $C(w) = p(w) \cdot r(w)$ is the same and depends only on the value of the rank $r = r(w)$. Figures 1 and 2 show the dependence of $C(w)$ on the rank r graphs for 8 texts of classical Russian fiction. In the first case, word lists of full and short forms of adjectives were compiled, in the second case, word lists of all the used words were built. Figure 1 shows that most authors use adjectives from the first half of the range of ranks. In Figure 2, an increase in the value of the parameter N_w (the total number of words in the collection) and the number of ranks r leads to a decrease in $C(w)$.

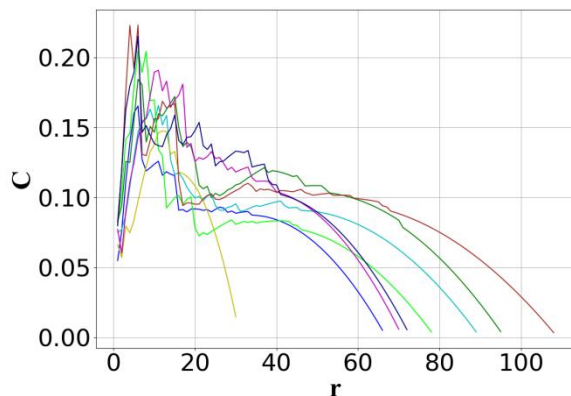


Figure 1. The graph of the function $C=p \cdot r$ for the collection of eight files and their mean value (0.088), the wordlist — adjectives in full and short forms.

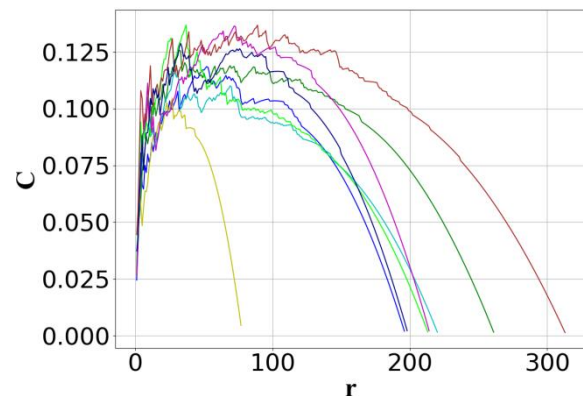


Figure2. The graph of the function $C=p \cdot r$ for 8 text files, the wordlist — all the words from the files.

The graphs of the sample files with the respective samples of words are shown below:

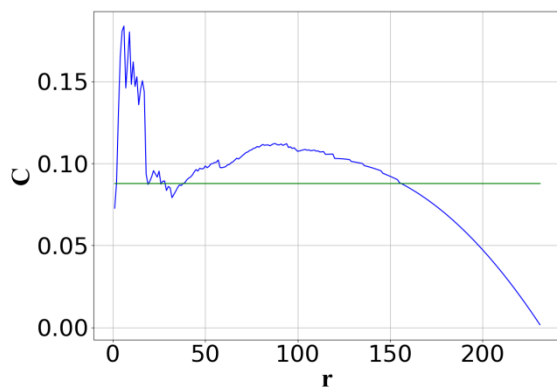


Figure3. The graph of the function $C=p \cdot r$ for the eight text files and the mean (0.088), the wordlist — adjectives in full and short forms.

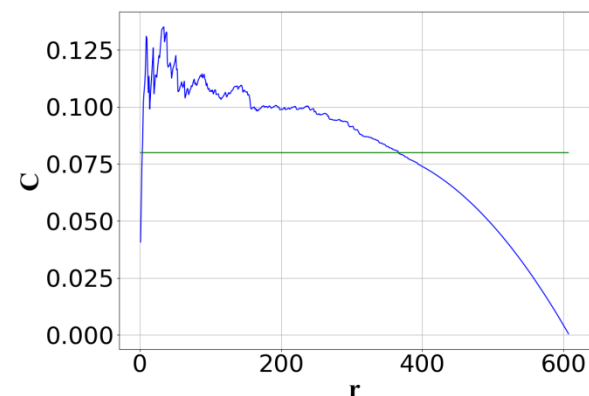


Figure4. The graph of the function $C=p \cdot r$ for the eight text files and the mean (0.08), the wordlist — all the words from the files.

We assumed that a significant increase in the number of ranks for the texts represented in Figure 3 indicates significant differences in the sets of adjectives used. Drawing up a collection (Figure 3 and Figure 4) leads to a more uniform distribution of frequencies according to the ranks and “blurring” of copyright differences in the texts. That is, the selection of certain language features from the collection of texts can serve to identify some of their individual characteristics. Rewriting of individual text files with the help of a synonymizer leads to very similar graphs of the function $C(r)$, i.e. the frequency dependence under study does not allow to identify the original text.

We also analyzed a collection of sixty Russian language text files written by Russian and foreign authors of approximately 36Mb. We imposed restrictions on the type of words - only strings of more than two characters were counted. The change in the value of C depending on r will have a character similar to the graphs presented for the samples in Figure 2.

We point out some features of the collection that are found when analyzing it by applying the first Zipf's law. Let us use for analysis $C(r)$ the second Zipf's law, which shows the hyperbolic distribution of the number of words of a certain frequency. The most salient part of the curve is formed by the words that significantly change the frequency characteristics of the text. Formula (2) is scaled by the value of K , equal to the ratio of the total number of words in the text to the one in the dictionary. A

comparison of the values considered for the word list of the collection showed that the value $C(r) \cdot K$ would then be in a decreasing range between the mean value and the value of approximately 0.06–0.07.

1.2. Modification of the first Zipf's law parameters

Let us try to estimate numerically the deviation of a document from Zipf's law in the form (2). If we assume that this law is perfectly satisfied for the document D , then $C(w) = p(w) \cdot r(w) = C$ for any word w from D , where C is a constant. In this case, the mean MC of the values of $C(w)$ for all the words w of D is also equal to C . If there are deviations from the law, then the differences $C(w) - MC$ are not always equal to zero. Therefore, you introduce at least three different characteristics of the deviation from the law:

$$\text{Norm0} = \max \{|C(w) - MC| : w \in D\}, \quad (3)$$

$$\text{Norm1} = \sum \{|C(w) - MC| : w \in D\}, \quad (4)$$

$$\text{Norm3} = \{\sum |C(w) - MC|^2\}^{1/2} : w \in D\}. \quad (5)$$

For all the nouns from the novel by F.M. Dostoevsky *The Brothers Karamazov* the following values were obtained by formulas (3) - (5): Norm0 = 0.11558, Norm1 = 94.8598, Norm2 = 1.94546.

We introduce other ways to determine the rank of a word, which reduces the total number of ranks for this document and, mostly important, deviations from Zipf's law are reduced.

1.2.1. Method 1. Let us choose an integer $m \geq 1$ and assume that the words w_1 and w_2 have the same rank if the numbers $k(w_1)$ and $k(w_2)$ if divided by m give the same partial numbers (that is, the integral parts of $k(w_1)$ and $k(w_2) / m$ are equal).

We can give another description. We divide the set of non-negative real numbers into half-intervals $[0, m)$, $[m, 2m)$, $[2m, 3m)$, and so on. Then, the words w_1 and w_2 will have the same rank if the numbers $k(w_1)$ and $k(w_2)$ lie in the same half-interval.

If $m = 1$, then the ranks of the two words are equal only if $k(w_1) = k(w_2)$, that is, we get the usual definition of rank. As m grows, Norm0, Norm1, Norm2 values usually decrease. So, for the above text file, we have:

Table1. Deviations from Zipf's law in relation to the parameters.

	m=1	m=2	m=3	m=10
Norm0	0.11558	0.11105	0.10315	0.09822
Norm1	94.8598	74.6352	61.7326	37.8910
Norm2	1.94546	1.63000	1.38936	0.96081

Method 1 is used when we want to reduce the number of word groups with the same ranks. The way in which the words w_1 and w_2 are assigned the same rank is not quite correct, if the difference in the numbers $k(w_1)$ and $k(w_2)$ in absolute values does not exceed the previously fixed number m . Let, for example, the frequencies of the words w_1 , w_2 , w_3 and w_4 are equal to 70, 69, 67 and 66, respectively, and $m = 2$. Then the adjacent words in this chain, namely the pairs (w_1, w_2) , (w_2, w_3) , and (w_3, w_4) must have the same rank, therefore the words at the edge of the chain (w_1, w_4) must have the same rank, however the frequency difference $70 - 66$ is greater than m and, therefore, their ranks must be different.

Let us consider another way of introducing a rank that considers words close, the frequency ratio of which differs little from unity.

1.2.2. Method 2. Let us choose a real number λ greater than one (but close to it) and divide the set of non-negative real numbers into half intervals $[1, \lambda)$, $[\lambda, \lambda^2)$, $[\lambda^2, \lambda^3)$, etc.

We assign the same rank to the words w_1 and w_2 if the numbers $k(w_1)$ and $k(w_2)$ lie in the same half-interval. Let us $\lambda = 1.02$. Then, if the frequencies $k(w_1)$ and $k(w_2)$ are not greater than 70, then the same

ranks are assigned to the words w_1, w_2 only if their frequencies coincide. Simultaneously, the same ranks will be assigned to the words, the frequencies of which fall into one of the half-intervals [380, 386), [511, 521], [4431, 4519).

If $\lambda = 1.02$, then for the above-mentioned novel by F.M. Dostoevsky, we have: Norm0 = 0.10098, Norm1 = 75.3674, Norm2 = 1.54230. The deviations obtained are smaller as compared with the deviations found by the traditional method of determining the word rank.

2. Study based on graph text model

To construct a graph model, we introduce the graph $G=(V,E)$, in which the set of vertices is formed by the set of all the words in the text, and the edges indicate the fact that a pair of words is found in one sentence. Moreover, each edge is assigned its weight b - the number of sentences in the text containing this pair of words. Some properties of such a graph were studied in [7]. Note that when the words in the text are mixed, its graph changes its structure. In accordance with the model B, the “garbage” text has the same frequency distribution of words as the source text. Therefore, we set the task of finding the characteristics of a text graph that would not be invariant, regardless of mixing. Among the available characteristics of the graph, we consider:

- average value of the graph edges weights,
- median value of the graph edge weights,
- the average value of the graph vertices degree,
- the median value of the graph vertices degree.

Note that the distribution of the degrees of the vertices of a random graph [8–10] is the subject of a theoretical study and has practical application, for example, when analyzing the graph structure of the Internet. In this case, it is assumed that the degrees have a density described by Pareto distribution: $p(x)=x^{-t}$ [11,12].

To put our model to the test, we took fiction texts with a restricted file size. We considered all the files from the specified directory and its subdirectories with a size limit of 60Kb. The results of the calculations are shown in the figures below.

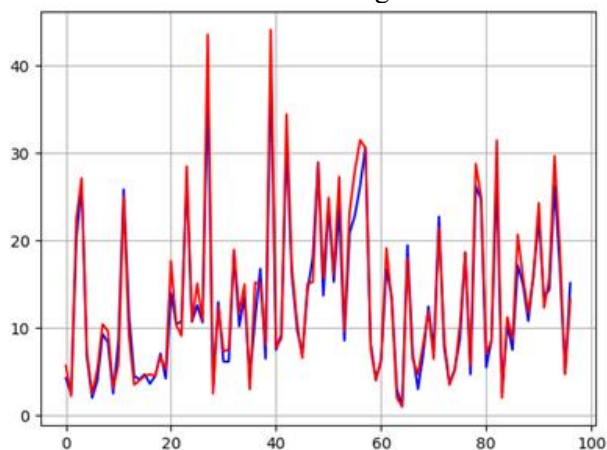


Figure 5. Average weight of the graph edges

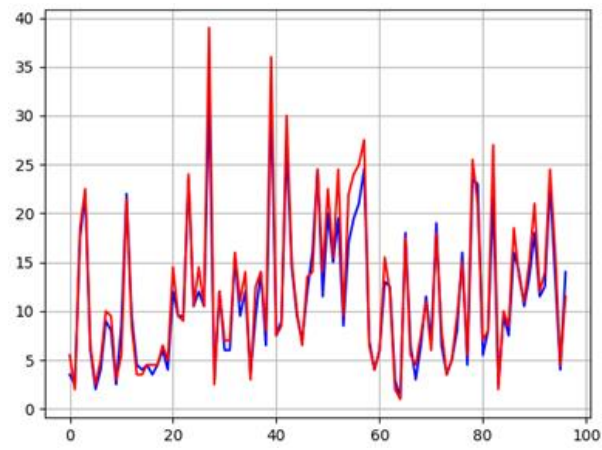


Figure 6. Median weight of the graph edges

In Figure 5, the number of texts in the corpus is plotted on the abscissa axis, and the average weight of the edges is plotted on the ordinate axis. The graph for the source texts is marked blue, the source texts obtained by random mixing are marked red. Calculations show that in 39 out of 97 texts, the first value does not exceed the second.

Figure 6 shows the result of calculating the median value of the edge weights for the same corpus of texts. In this case, in 40 out of 97 texts, the red graph is larger than the blue graph. In both cases

considered, the value of the average or median value of the weights of the edges of the graph cannot be an example of the discriminant function of the text.

Figures 7 and 8 show calculations of the average values of the degrees of the vertices of the graph and the median of the degrees of the vertices of the graph.

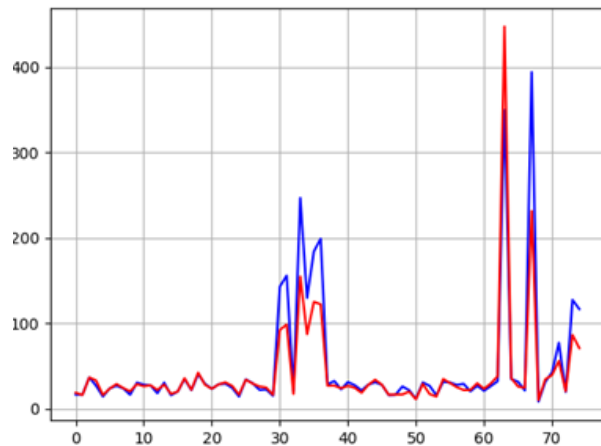


Figure7. The average value of the graph vertices degree.

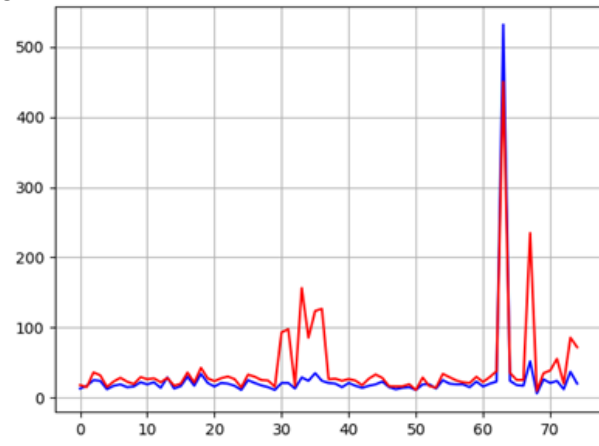


Figure8. The median value of the graph vertices degree.

Calculations show in 41 out of 97 texts, the average value of the graph vertices for the source text is less than the same value for the mixed text. As in the case with the previous two characteristics, this value cannot be used as a criterion to classify the text as “garbage”.

In Figure 8 the red graph mainly lies above the blue. Calculations show that this does not hold true for only 7 texts out of 97. In other words, the median value of the vertices of the text graph can be used to identify “garbage”. In this case, the identification scheme may be as follows. For the source text, the vertices of the graph are calculated. Then, their median value is calculated. Further, the source text is subjected to multiple mixing, and after each such operation the median values of the new graph vertices are calculated. After that, the mean value of the calculated medians is taken and compared with the median value of the source text. If the resulting value is greater, the text is assumed to be real, otherwise it is classified as “garbage” [13].

Thus, the last value is a reliable parameter with which you can separate meaningful texts from the “garbage”.

Conclusion

To sum up, computer experiments were conducted to find out the applicability of Zipf's laws to the bag-of-words text model. Using different wordlists based on certain type of words as well as modifying the parameters of Zipf's laws, the study identifies distinctive characteristics of meaningful texts. The constructed graph model of the text, in which links are established for all pairs of words in one piece of text, in this particular case, in one sentence, allowed to effectively investigate the parameters of the model and obtain meaningful information.

Acknowledgement

The reported research was funded by Russian Foundation for Basic Research and the government of Volgograd region, grant № 18-412-340007

References

- [1] Mikolov T et. al. Efficient Estimation of Word Representations in Vector Space, arxiv.org. URL: <http://arxiv.org/pdf/1301.3781.pdf>

- [2] Sonawane S S, Kulkarni P A 2014 *International Journal of Computer Applications* **96** (19) 1-7
- [3] Arapov M V 1981 *Systematic Research* (Moscow: Nauka)
- [4] Bakulina M P 2007 *Descree Analysis and Study of Operations* **14** (2) 3-13
- [5] Piotrovskiy R G 1975 *Text, Machine, Human* (Leningrad.: Nauka)
- [6] Haitun S D 1983 *Naukometria. Sostoyanie I Perspektivi* (Moscow: Nauka)
- [7] Grigoryeva E G., Klyachin V A., Pomelnikov Yu V and Popov V V *Vestnik Volgogradskogo gosudarstvennogo universiteta Seriya 2 Yazykoznanie* **16**(2) 58-67
- [8] Fedotov N G, Moiseev A V, Syemov A A, Lizunkov V G, Kindaev A Y 2017 New Methods of Three-Dimensional Images Recognition Based on Stochastic Geometry and Functional Analysis *IOP Conference Series: Materials Science and Engineering* **177** (1) 012047
- [9] Raygorodskiy A M 2009 *Matematika v zadachakh* (Moscow: MTsNMO)
- [10] Erdős P, Rényi A. 1959 *Publ. Math. Debrecen* **6** 290- 97
- [11] Newman M E, Strogatz S H and Watts D J *Phys. Rev.* **64** 26-118
- [12] Pavlov V Yu and Cheplukova I A 2008 *Descree Math* **20** (3) 3–18
- [13] Pavlov Yu L 2009 *Descree Math* .**21** (3) 14–23