American Journal of Computing Research Repository, 2013, Vol. 1, No. 1, 1-5
Available online at http://pubs.sciepub.com/ajcrr/1/1/1

DOI:10.12691/ajcrr-1-1-1

Investigating the Distribution of Arabic and English Keywords and Their Progress Over Different Text File Formats

Boumedyen Shannaq*

Computer science and Information Technology Department, Mazoon University College, Muscat, Sultanate of Oman
*Corresponding author: aboumedyen@gmail.com

Received August 11, 2013; Revised November 01, 2013; Accepted November 13, 2013

Abstract This paper explicates a systematic approach of implementing text format categorization. It also emphasizes defined corpus linguistics and accordingly demonstrates how various Text files Html, Pdf, Doc and Txt format respectively could be analyzed. This work concentrates on comparing Arabic text format with English text format, for which various text formats have been considered. Hence the idea is implemented by calculating a distributed factor for the keywords distribution with respect to Arabic and English text documentation. All the text selected is from the Computer Technology domain. The text categorization process is implemented on the text collection and consists of two main corpus namely, Arabic and English text respectively. The obtained results show that the Arabic text format document is well distributed in Doc files compared to the English text document which is well distributed in Xml files. These results shall contribute in handling and building an effective Electronic Learning System for Arabic and English Texts. The results and conclusions are presented here with various graphical outputs for better understanding.

Keywords: *information retrieval, text categorization, distributing factor, natural language processing, future trends*

Cite This Article: Boumedyen Shannaq, "Investigating the Distribution of Arabic and English Keywords and Their Progress Over Different Text File Formats." *American Journal of Computing Research Repository* 1, no. 1 (2013): 1-5. doi: 10.12691/ajcrr-1-1-1.

1. Introduction

Internet Technology has flooded the world with online information. The organization and control over this abundant information has become a major challenge to the world. In today's world; children, students, schools, universities, colleges companies, government etc, all rely on information retrieved from internet. The dependence on the internet increases the traffic on the net. Thus there is a challenge in front of one's search engines to find new and effective ways to deal with this enormous volume of information and flood of internet users. Search engines play an important role in the information age. Since we live in an information age, most of the information, even related with product and services are derived from the internet. The government as well as private institutions mainly rely on the internet, it is thus necessary to design or make a search engine which can effectively index or classify the web pages in a manner that help its users to derive the exact information required by them. But despite of companies claiming their success in producing a search engine which will satisfy the internet users still the user complain of the lack of accuracy and relevance of the information desired [1]. Further the Internet Technology need to look for new techniques to organize this volume of information over the internet, and this could be done through analyzing the text format that available on the internet to generate useful text analysis and categorization over different file formats.

1.1. Text Collection

The text file formats were collected from various media, such as, weekly, internet sites, Computer Technology e-books, Computer Technology encyclopedias, and Computer Technology electronic publications, etc. In this connection, the text categorization is an emerging trend in the field of research and it is very useful and necessary to analyze the corpus data. This text categorization has been performed in a strategic way since text exists in, and is exemplified by, huge volumes of data. Therefore the fruit of such a strategy is that this kind of application may help and support researchers in selecting the best text collection for the purpose, for example, building a knowledge base; an ontology; Thesaurus, a glossary [2]. Furthermore, it may be useful as a learning tool for instructors and students. That means the students and teachers may find rich and useful information in these formats. In general Arabic users prefer to use the Doc files as distinct from other text formats.

Table 1 illustrates the statistics of the collected text file formats.

Table 1. Statistics of the collected text file formats

Text Format	Files Number	Files Size(Bytes)	Words Number	Characters Number
DOC	2907	1421551518798	337361903586	1355209056522
PDF	1903	954590637310	238703952108	1861085152429
HTML	1069	9771755378	2442938845	18562424994
XML	2327	62410325121	60093154016	309156254031
TXT	2480	47503451097	86785783828	418737168295

2. Background

The process of analyzing text collection is an essential task for assessing the corpus for completeness and representativeness. Linguistics traditionally considers the following measurement to evaluate corpus and collections [17], such as:

• Average word length, measured in characters per word.

• Average sentence length, measured in words per sentence.

• Token-to-type ratio, the ratio between the length of the text and different word forms used (and thus sensitive to the text length).

• Relative perplexity, based on unigram entropy.

• Collocation richness, which measures how many typical collocations (derived from the Co build Bank of English) are used in the environment of a word. Video and audio represent a sizable segment of online content, but search and linguistic analysis such as part-of-speech (PoS), tagging and concordance require machine-readable written text. In [3], the rating factor have been developed to find the distribution of the words over different text types i.e. Classical text, Stories, IS and others and in their work they found that classical text are well distributed over other text types [3]. More information can also be found in [4,5].

3. Analysis of Text Files

The data has been collected from various magazines and electronic publications are not having all words equally significant for representing the semantics of a document. In written language, some words carry more meaning than others. Usually, noun words (or groups of noun words) are the ones which are most representative of document content. Therefore, it is usually considered worthwhile to preprocess the text of the documents in the collection to determine the terms to be used as index terms. Table 2 shows a sample of stop words in English and Arabic.

Table 2. Sample of stop words

Arabic Stop words	English Stop words
من إلى و هذا لكن على	On once one only Onto or the

During this preprocessing phase other useful text operations can be performed such as elimination of stop words. The document representation is by sets of index terms, which lead to a rather imprecise representation of the semantics of the documents in the collection. For instance, a term like 'the', 'that', 'from' etc. has no meaning by itself and might lead to the retrieval of various documents which are unrelated to the recent user query in information retrieval system [4,6]. However, using the set of all words in a collection to index its documents generates too much noise for the retrieval task. One way to reduce this noise is to reduce the set of words which can be used to refer to (index) documents. Thus, the preprocessing of the documents in the collection might be viewed simply as a process of controlling the size of the vocabulary (the number of distinct words used as an index terms). It is expected that the use of a controlled vocabulary leads to an improvement in retrieval performance. While controlling the size of the vocabulary is a common technique with commercial systems, it does introduce an additional step in the indexing process which is frequently not easily perceived by the users [7]. The raw data collected needs to be processed before carrying out other tasks; the processing includes reformatting the text to a unified format and the same code set. The data collected have to be indexed into one database. The purpose of indexing the data is to recognize every token in the collection. The indexing step produces different sets of data, word lists, and frequency lists. We summarized various statistical information to describe documents files in each file format. For example by removing HTML tags, stop words, aligning numbers and punctuations and finally removing extra spaces between adjacent words. The same process was applied to other file formats in Arabic and English corpuses respectively.

3.1. Data Set Assessments

The necessity of a suitable Text collection with a wider coverage considered as samples of the languages is a key for this research. Information Retrieval (IR) and other Natural Language processing (NLP) disciplines need a text collection which contains useful information for the experimentation. The performance of IR and NLP techniques that uses corpus as a resource or dataset in this research. Therefore linguists carry out a variety of tests to evaluate the correctness of the data. These measures and evaluations vary with the task, the language, and the techniques. The assessment tools can rearrange such a corpus store so that various observations can be made. The corpus assessment tools are used to validate the collections by applying statistical and probability tests, such as Zipf's law and the Mandelbrot formula. These tests are useful for describing the frequency distribution of the words in the corpus dataset. They provide evidence of any inequality of the dataset for more information see [8,9,10,11,12,13].

3.2. Zipf's Law

According to Zipf's law [15], if we count up how often each word occurs in a corpus and then list these words in the order of their frequency of occurrence, then the relationship between the frequency of a given word *f* and its position in the list (its rank *r*) will be a constant *k* such that: f.r = k. An enhanced theory of Zipf's law is the Mandelbrot distribution. Figure 1 and Figure 2 demonstrate the development of Zipf's law using our text collection, for more information see [14,15,16].

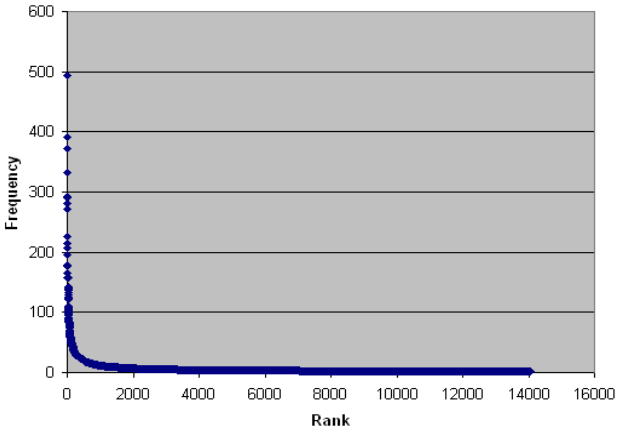

Figure 1. Real word frequency vs. their ranks and fitted curve with Zipf's law for documents files

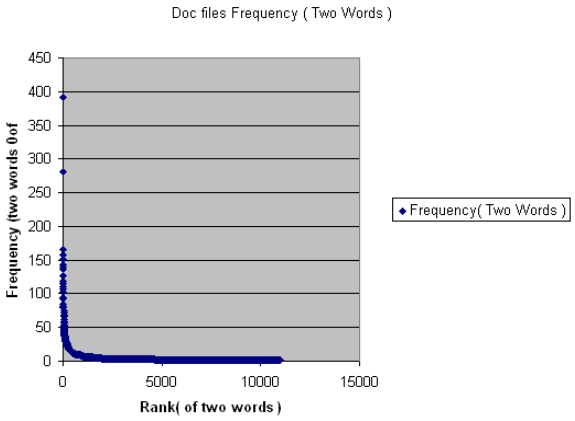

Figure 2. Real Two word frequency vs. their ranks and fitted curve with Zipf's law for doc Files

Figure 1 illustrates the rank/frequency profile of the text collection with .DOC format (frequency of one word). Frequency (on the y axis) is plotted on a logarithmic scale, the frequency of the most frequent words is much higher than the frequency of the long end of rare words that a figure of this size without a logarithmic transformation would look like the hyperbolic graph. The scheme illustrates, the frequency curve decreases very steeply from the extremely high values corresponding to the most frequent words, and it becomes progressively flatter, until it reaches a very wide area of stability in correspondence to the ranks assigned to the end of words occurring once, the same illustrated in Figure 2 but for combinations of frequencies of two words i.e. occurring twice. The interesting and values of information obtained from these figures, means that only first percentage words of text collection (text with .doc format) offered the semantic meaning of text collection (text with .doc format).

4. Results and Discussion

This work analyses the text document of various formats using rating factor to discover the richness of the information inside the files taken for analysis. Some Arabic and English text files have been taken for analysis and the rating factor (Ψ) has been calculated as follows,

$$\Psi = N(R) / N(O)$$

Where, N(R) is the number of words in the Normalized document (avoiding repeating and stop words)

N(O) is the total number of words in the text file.

For Ex, N(O) is given as :

Information System helps information systems students to compete in a global environment. Many students confused with information technology and Information Systems.

N(R) is calculated as:

Information System helps students compete global environment. Confused technology.

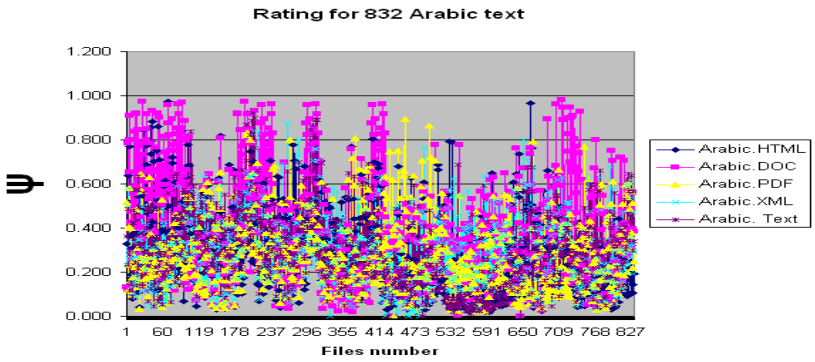

Figure 3. Rating for 832 Arabic texts

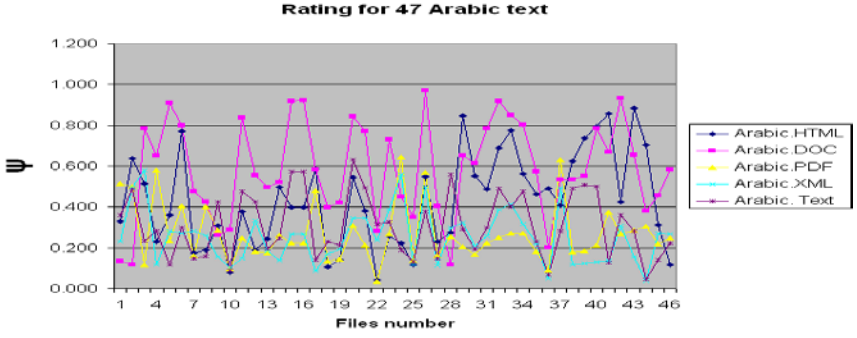

Figure 4. Rating for 47 Arabic texts

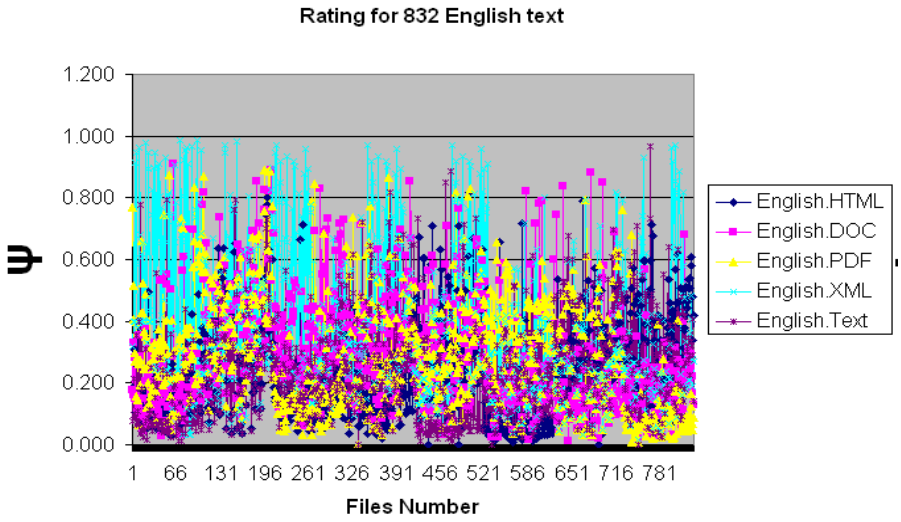

Figure 5. Rating for 832 English texts

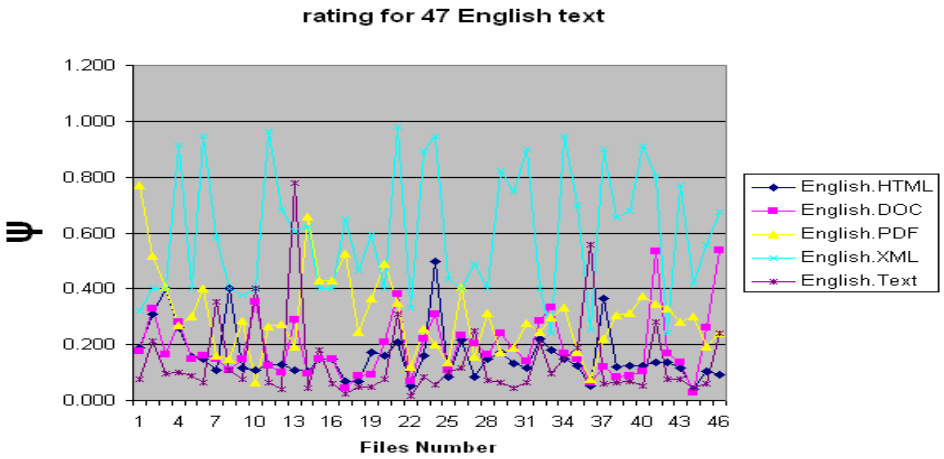

Figure 6. Rating for 47 English texts

As it is shown in the table above, the rating factor (Ψ) is calculated as follows, Ψ = 9/21 Where, N(R) is 9 and N(O) is 21. The experimental results are illustrated in Figure 3, Figure 4, Figure 5 and Figure 6. These figures, illustrate obtained rating factor (Ψ) for investigated Arabic text collection in different format i.e. HTML, DOC, PDF, XML and TXT. The blue plot in Figure 3 represents distribution of rating factor for Arabic text with .HTML format, Pink plot for .DOC, Yellow plot for .PDF, green plot for .XML and light purple plot for .TXT files. Rating factor is on the y axis and file numbers are on the X axis. This Figure 3 illustrates results for 832 Arabic text files and Figure 4 illustrates results of 47 Arabic text files, just for butter presentation. The Figure 5 illustrates results for 832 English text files and Figure 6 illustrates results of 47 English text files, just for butter presentation.

The obtained results shown an the above figures could help the internet user for decreasing the number of retrieved documents and for their relevancy. Assume that an Arabic user for "حاسوب" 'computer in English', in this connection Google search engine retrieved 7420000 Arabic documents, as shown in Figure 7. However Google search engine retrieved 19900 Arabic documents , if the user limit the query to ".Doc" files such as : " حاسوب Filetype:doc" , Figure 8 shows this output.

Figure 7. ""حاسوب" " query

Figure 8. حاسوبFiletype:doc

5. Conclusion

This paper explicates a systematic approach of implementing Text format categorization. It also emphasizes defined corpus linguistics and accordingly demonstrates how various Text files Html, Pdf, Doc and Txt format could be analyzed respectively. The implementation results have been presented in Figure 3, Figure 4, Figure 5, and Figure 6 respectively. These results indicates that those Keywords in the Arabic Doc files are showing its rich and fastness in their growth, while keywords in the English specific to Xml files indicate its rich and fastness in their growth. This work has been considered for the application of the Strategic Rating factor and shows how it could be used in investigating the distribution of keywords and their progress over different Text files formats. These results shall contribute in handling and building an effective Electronic Learning System for Arabic and English Texts.

References

[1] Boumedyen, "The New Arabic Document Summarization techniques (NADST)", MECIT, Oman, 2011.

[2] P.P.Kokorin, Boumedyen.Shannag, E. V. ShChelkunova, “Algorithm of normalization and ontological Clusters texts” information-measuring and operating systems Journal, 2010. http://www.radiotec.ru/catalog.php?cat=jr.

[3] Aksenov A.Y., Zaytseva A. A., Boumedyen Shannaq. "The rank method of text data regions localization", information-measuring and operating systems Journa, 211l.

[4] P.P.Kokorin, B.Shannag, E.V.ShChelkunova, “Algorithm of normalization and ontological Clusters texts” information-measuring and operating systems Journal, 2010. http://www.radiotec.ru/catalog.php?cat=jr.

[5] Kokorin P. P. Kolesnikov R. A., Andreeva N. A, Frolov K. V, Boumedyen Shannaq, “The info logical approach to develop edutainment systems”. St. Petersburg institute for Informatics and Automation of Russian RAS, Academy of Sciences, 199178, Russia, VAX UDC 004.9, Information-measuring and operating systems Journal, 2009. http://www.radiotec.ru/catalog.php?cat.

[6] Boumedyen Shannaq, S.V. Kuleshov,” Super Arabic morphological analyzer (SAMA1) “ St. Petersburg institute for Informatics and Automation of Russian RAS ,Academy of Sciences, 199178, Russia ,VAX UDC 003.9, *information-measuring and operating systems Journal*, 2009.

[7] Alekcandov V.V, Kuleshov S.V, Boumedyen Shannaq, “ Phenomenon of identification” *information-measuring and operating systems Journal*, 2010. http://www.radiotec.ru/catalog.php?cat=jr

[8] Baayen, Harald (2001), Word frequency distributions. Dordrecht: Kluwer.

[9] Baldi, Pierre/Frasconi, Paolo/Smyth, Padhraic (2003), Modeling the Internet and the web. Chichester: Wiley.

[10] Ha, Le Quan, Sicilia-Garcia, E. I., Ming, Ji, and Smith, F. J. (2002), extension of Zipf’s law to words and phrases, in Proceedings of COLING 2002, Taipei, Taiwan.

[11] Heaps, H. S. (1978) Information Retrieval – Computational and Theoretical Aspects, Academic Press.

[12] Li, Wentian (1992). Random texts exhibit zipf’s-law-like word frequency distribution. IEEE Transactions on Information Theory, 38(6), 1842–1845.

[13] Sichel, H. S. On a distribution law for word frequencies. *Journal of the American Statistical Association*, vol: 70, 542-547.

[14] H. Guiter and M. V. Arapov, editors. Studies on Zipf's Law, Wissens chaftlicher Verlag Trier, 1982.

[15] R. G□unther, L. Levitin, B. Schapiro, and P. Wagner, Zipf's law and the act of ranking on probability distributions. *International Journal of Theoretical Physics*, 15:395, 1996.

[16] Wentian Li. References on Zipf's law. URL: http://linkage.rockefeller.edu/wli/zipf/.

[17] Mason, Oliver; Berglund, Ylva, Low-level parameters reflecting the naturalness of texts. Proceedings of JADT2002, 6th *International Conference on Textual Data Statistical Analysis*, Saint Malo, March 13-15 2002. Vol.2, p.507-516. ISBN: 2-7261-1198X.