"Improving Document Processing and Indexing by Preprocessing and Tokenization"

Alharith Alkafije , George Ajam
Department of Computer Science University of Babylon, Babylon-Iraq.

1- Abstract

Information Retrieval is the science of searching information within documents. Documents are in huge quantity and still growing. It is very difficult to find the information according to requirements of user. So different algorithms are being proposed based on long research in information retrieval and data mining.Where in this paper we analyze the documents in the collection Sense and Sensibility available on the web page under the subheading "data files" we download these files and build programs that are able to index a collection of documents and calculate text statistics across the corpus. Text processing (or document processing) includes tokenization, preprocessing (converting upper case letters to lower, Unicode conversion, and removing diacritics from letters, punctuations, or numbers), stop words removal, and stemming. These steps save indexing time and space, especially for a huge set of data. Also, the experiment results at the end of this paper approve the reliability and efficiency of the algorithms.

الخلاصة

إسترجاع المعلومات هو العلم الخاص بالبحث في المستندات. المستندات التي تكون كبيرة كما ومتزايدة بالحجم. حيث انه من الصعب إيجاد المعلومات حسب متطلبات المستخدم. لذا فإن خوارزميات مختلفة يتم إقتراحها إعتمادا على بحوث في إسترجاع المعلومات والتنقيب عن البيانات.أما في هذا البحث فإننا نحلل المستندات من وجهة نظر توافر البيانات على صفحات الويب في تجمعات حسية وإن نوع تلك الملفات من خلال مقدمتها الجزئية هي "ملفات بيانات". نقوم بتحميل الملفات ونبني البرامج التي تكون قادرة على فهرسة مجاميع المستندات لحساب إحصائيات النصوص الرئيسية.معالجة النص (أو معالجة المستند) تتضمن معالجة اولية للكلمات tokenization (حيث يتم تحويل الحروف الكبيرة الى الحالة الصغيرة، ويتم تحويل صيغة الترميز الى صيغة موحدة Unicode ويتم إزالة الحركات من الحروف وكذلك يتم إزالة التنقيط او الارقام). هذه الخطوات تخزن فهرسة حسب الوقت والمجال خصوصا لمجموعة ضخمة من البيانات وكذلك فإن نتائج التجربة في هذا البحث تثبت الاعتمادية والكفاءة الخاصة بالخوارزمية.

2- Introduction

Information Retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics, and physics. (wikipedia)

Web search engines are the most visible IR applications. Web search engines implementation of many features formerly found only in experimental IR systems. Search engines become the most common and maybe best instantiation of IR models, research, and implementation. In this study, search engine approaches for ranking and classify documents are studied and improvements are being done by novel approach for classification.

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query. Similarity scores are given to each query after calculating it within corpus of documents. (David Grossman)

Many different measures for evaluating the performance of information retrieval systems have been proposed. The measures require a collection of documents and a query. Every document is known to be either relevant or non-relevant to a particular query. (Manning 2009)

In this paper we analyze the documents by using tokenization, preprocessing (converting upper case letters to lower, Unicode conversion, removing diacritics from letters, punctuations, or numbers), stop words removal, and stemming.

In the end of the paper an improvement document processing and indexing the steps that proposed save indexing time and space, especially for a huge set of data.

3- Implementation descriptions

Programming language

We decided to use MATLAB because the paper was ideally suited to be implemented as a Matlab application.

Proposed programs steps

Download the data set file (sense.txt)

We are preprocessed the text file (sense.txt), tokenized it and convert to each word to lower case (ws_wno_withspecial.txt).

After that we remove special character and numbers we get the file (without_special.txt).

Then we remove stop words according to the set of from the web site and we get (without_stw.txt).(Gerard Salton , Chris Buckley)

After that we do the stemming for the data set, according to the porter stemming algorithm available on the website to get the file (all_stemmed.txt). (Martin Porter)

The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.

We find the number of the document in the collection where when we made the token we will see that each document starts with <P ID=xxxx> and ends with </P>, we find 1862 document.

We find the total number of words in the corpus, the total number of distinct words, corpus frequency(corpus_freq.txt), document frequency (unique_freq.txt), most frequent 50 words in the corpus and the 500th, 1000th, and 2000th most frequent word and their collection frequencies. After that we compare these collection frequencies with those that are suggested by Zipf 's law. We do that before and after stemming. Zipf 's law is an empirical law formulated using mathematical statistics, refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution.(Manning 1999)

Also we find the number of words which occur only in one document and the percent of all words in our data set.

We repeat the entire above request on our data set after we passed through a lemmatizer and we compared between the results where we noticed.

4- Test results

This item contains the test results of our programs, statistics are shown in each table below:

Table 1: Main statistics results with and without Lemmatizer.

Item	With Stem.	Without Stem.
No. of the document in the collection	1862	1862
No. of the word in the collection	40507	40507
No. of the unique words in the collection	6885 (17%)	4760 (11.7%)
No of words repeated one time in just one document	3321 (8.2%)	2185 (5.3%)

Table 2: Statistics result for 50 most frequency word without stemming

Term	Term Frequency	REL. FREQ. (%)	Term	Term Frequency	REL. FREQ. (%)
'elinor'	[618]	'1.5257%'	'man'	[114]	'0.2814%'
'mrs'	[526]	'1.2985%'	'ferrars'	[109]	'0.2691%'
'marianne'	[490]	'1.2097%'	'felt'	[105]	'0.2592%'
'time'	[237]	'0.5851%'	'young'	[103]	'0.2543%'
'dashwood'	[224]	'0.5530%'	'long'	[102]	'0.2518%'
'sister'	[214]	'0.5283%'	'replied'	[99]	'0.2444%'
'edward'	[210]	'0.5184%'	'left'	[98]	'0.2419%'
'miss'	[209]	'0.5160%'	'happy'	[97]	'0.2395%'
'jennings'	[203]	'0.5011%'	'kind'	[94]	'0.2321%'
'mother'	[200]	'0.4937%'	'world'	[90]	'0.2222%'
'thing'	[184]	'0.4542%'	'barton'	[89]	'0.2197%'
'mr'	[178]	'0.4394%'	'middleton'	[87]	'0.2148%'
'willoughby'	[176]	'0.4345%'	'hope'	[86]	'0.2123%'
'colonel'	[163]	'0.4024%'	'cried'	[85]	'0.2098%'
'lucy'	[157]	'0.3876%'	'town'	[84]	'0.2074%'
'house'	[148]	'0.3654%'	'present'	[83]	'0.2049%'
'great'	[147]	'0.3629%'	'family'	[82]	'0.2024%'
'good'	[145]	'0.3580%'	'morning'	[81]	'0.2000%'
'day'	[139]	'0.3432%'	'place'	[79]	'0.1950%'
'lady'	[137]	'0.3382%'	'affection'	[78]	'0.1926%'
'give'	[126]	'0.3111%'	'love'	[76]	'0.1876%'
'heart'	[123]	'0.3037%'	'letter'	[74]	'0.1827%'
'sir'	[119]	'0.2938%'	'feelings'	[72]	'0.1777%'
'brandon'	[116]	'0.2864%'	'found'	[71]	'0.1753%'
'dear'	[115]	'0.2839%'	'brother'	[70]	'0.1728%'

Table 3: statistics results for the maximum frequencies in the collection and all words not appear are 1 freq (2000) without stemming.

Term	Corpus Frequency	REL. FREQ. (%)	Term	Corpus Frequency	REL. FREQ. (%)
'elinor'	[618]	'1.5257%'	'comfort'	[60]	'0.1481%'
'mrs'	[526]	'1.2985%'	'brought'	[58]	'0.1432%'
'marianne'	[490]	'1.2097%'	'attention'	[57]	'0.1407%'
'time'	[237]	'0.5851%'	'continued'	[56]	'0.1382%'
'dashwood'	[224]	'0.5530%'	'find'	[55]	'0.1358%'
'sister'	[214]	'0.5283%'	'side'	[54]	'0.1333%'
'edward'	[210]	'0.5184%'	'end'	[53]	'0.1308%'
'miss'	[209]	'0.5160%'	'returned'	[52]	'0.1284%'
'jennings'	[203]	'0.5011%'	'friend'	[51]	'0.1259%'
'mother'	[200]	'0.4937%'	'chapter'	[50]	'0.1234%'
'thing'	[184]	'0.4542%'	'body'	[49]	'0.1210%'
'mr'	[178]	'0.4394%'	'evening'	[48]	'0.1185%'
'willoughby'	[176]	'0.4345%'	'days'	[47]	'0.1160%'
'colonel'	[163]	'0.4024%'	'determined'	[46]	'0.1136%'
'lucy'	[157]	'0.3876%'	'business'	[45]	'0.1111%'
'house'	[148]	'0.3654%'	'account'	[44]	'0.1086%'
'great'	[147]	'0.3629%'	'called'	[43]	'0.1062%'
'good'	[145]	'0.3580%'	'directly'	[42]	'0.1037%'
'day'	[139]	'0.3432%'	'added'	[41]	'0.1012%'
'lady'	[137]	'0.3382%'	'carriage'	[40]	'0.0987%'
'give'	[126]	'0.3111%'	'back'	[39]	'0.0963%'
'heart'	[123]	'0.3037%'	'child'	[38]	'0.0938%'
'sir'	[119]	'0.2938%'	'coming'	[37]	'0.0913%'
'brandon'	[116]	'0.2864%'	'conduct'	[36]	'0.0889%'
'dear'	[115]	'0.2839%'	'appeared'	[35]	'0.0864%'
'man'	[114]	'0.2814%'	'girl'	[34]	'0.0839%'
'ferrars'	[109]	'0.2691%'	'answer'	[33]	'0.0815%'
'felt'	[105]	'0.2592%'	'assure'	[32]	'0.0790%'
'young'	[103]	'0.2543%'	'common'	[31]	'0.0765%'
'long'	[102]	'0.2518%'	'believed'	[30]	'0.0741%'
'replied'	[99]	'0.2444%'	'anxious'	[29]	'0.0716%'
'left'	[98]	'0.2419%'	'appearance'	[28]	'0.0691%'
'happy'	[97]	'0.2395%'	'affair'	[27]	'0.0667%'
'kind'	[94]	'0.2321%'	'acquainted'	[26]	'0.0642%'
'world'	[90]	'0.2222%'	'advantage'	[25]	'0.0617%'
'replied'	[99]	'0.2444%'	'agreeable'	[24]	'0.0592%'
'barton'	[89]	'0.2197%'	'afraid'	[23]	'0.0568%'
'middleton'	[87]	'0.2148%'	'affectionate'	[22]	'0.0543%'
'hope'	[86]	'0.2123%'	'address'	[21]	'0.0518%'
'cried'	[85]	'0.2098%'	'admiration'	[20]	'0.0494%'
'town'	[84]	'0.2074%'	'age'	[19]	'0.0469%'
'present'	[83]	'0.2049%'	'alarm'	[18]	'0.0444%'
'family'	[82]	'0.2024%'	'agitation'	[17]	'0.0420%'
'morning'	[81]	'0.2000%'	'aware'	[16]	'0.0395%'
'place'	[79]	'0.1950%'	'affliction'	[15]	'0.0370%'
'affection'	[78]	'0.1926%'	'ago'	[14]	'0.0346%'
'love'	[76]	'0.1876%'	'approbation'	[13]	'0.0321%'
'letter'	[74]	'0.1827%'	'acknowledge'	[12]	'0.0296%'

'feelings'	[72]	'0.1777%'	'attempted'	[11]	'0.0272%'
'found'	[71]	'0.1753%'	'absence'	[10]	'0.0247%'
'brother'	[70]	'0.1728%'	'abilities'	[9]	'0.0222%'
'hear'	[69]	'0.1703%'	'advise'	[8]	'0.0197%'
'spirits'	[68]	'0.1679%'	'accepted'	[7]	'0.0173%'
'person'	[67]	'0.1654%'	'accepting'	[6]	'0.0148%'
'pleasure'	[66]	'0.1629%'	'abode'	[5]	'0.0123%'
'feel'	[65]	'0.1605%'	'abhorrence'	[4]	'0.0099%'
'poor'	[64]	'0.1580%'	'ability'	[3]	'0.0074%'
'behaviour'	[63]	'0.1555%'	'abhorred'	[2]	'0.0049%'
'acquaintance'	[62]	'0.1531%'	'abandoned'	[1]	'0.0025%'
'elinors'	[61]	'0.1506%'	So on	So on	So on

Table 4: Zipf's law

Collection Frequency (Cfi)	With Stem.	Actual	Without Stem	Actual
1	618	618	704	704
500	1.236	1	1.76	1
1000	0.618	1	0.704	1
2000	0.309	1	0.352	1

Table 5: Statistics result for 50 most frequency word with stemming

Term	Corpus Frequency	REL. FREQ. (%)	Term	Corpus Frequency	REL. FREQ. (%)
'mr'	[704]	'0%'	'kind'	[136]	'0.3357%'
'elinor'	[679]	'1.6763%'	'hope'	[127]	'0.3135%'
'mariann'	[561]	'1.3849%'	'heart'	[126]	'0.3111%'
'sister'	[313]	'0.7727%'	'middleton'	[120]	'0.2962%'
'dashwood'	[280]	'0.6912%'	'sir'	[119]	'0.2938%'
'time'	[258]	'0.6369%'	'dear'	[115]	'0.2839%'
'edward'	[253]	'0.6246%'	'friend'	[114]	'0.2814%'
'mother'	[249]	'0.6147%'	'moment'	[113]	'0.2790%'
'miss'	[216]	'0.5332%'	'repli'	[110]	'0.2716%'
'willoughbi'	[214]	'0.5283%'	'affect'	[109]	'0.2691%'
'thing'	[205]	'0.5061%'	'live'	[108]	'0.2666%'
'jen'	[203]	'0.5011%'	'return'	[106]	'0.2617%'
'dai'	[186]	'0.4592%'	'felt'	[105]	'0.2592%'
'luci'	[183]	'0.4518%'	'expect'	[103]	'0.2543%'
'colonel'	[174]	'0.4296%'	'love'	[102]	'0.2518%'
'make'	[168]	'0.4147%'	'comfort'	[101]	'0.2493%'
'feel'	[163]	'0.4024%'	'mind'	[100]	'0.2469%'
'happi'	[160]	'0.3950%'	'left'	[98]	'0.2419%'
'hous'	[153]	'0.3777%'	'hear'	[97]	'0.2395%'
good'	[152]	'0.3752%'	'letter'	[94]	'0.2321%'
'great''	[150]	'0.3703%'	'call'	[93]	'0.2296%'
'thought'	[149]	'0.3678%'	'continu'	[91]	'0.2247%'
'brandon'	[140]	'0.3456%'	'manner'	[90]	'0.2222%'
'engag'	[138]	'0.3407%'	'acquaint'	[89]	'0.2197%'
'made'	[137]	'0.3382%'	'spirit'	[88]	'0.2172%'

Table 6: statistics results for the maximum frequencies in the collection and all words not appear are 1 freq (2000) with stemming.

Term	Corpus Frequency	REL. FREQ. (%)	Term	Corpus Frequency	REL. FREQ. (%)
'mr'	[704]	'0%'	'pleasur'	[68]	'0.1679%'
'elinor'	[679]	'1.6763%'	'daughter'	[67]	'0.1654%'
'mariann'	[561]	'1.3849%'	'natur'	[66]	'0.1629%'
'sister'	[313]	'0.7727%'	'end'	[65]	'0.1605%'
'dashwood'	[280]	'0.6912%'	'doubt'	[64]	'0.1580%'
'time'	[258]	'0.6369%'	'behaviour'	[63]	'0.1555%'
'edward'	[253]	'0.6246%'	'answer'	[62]	'0.1531%'
'mother'	[249]	'0.6147%'	'delight'	[61]	'0.1506%'
'miss'	[216]	'0.5332%'	'arriv'	[58]	'0.1432%'
'willoughbi'	[214]	'0.5283%'	'hand'	[57]	'0.1407%'
'thing'	[205]	'0.5061%'	'attach'	[56]	'0.1382%'
'jen'	[203]	'0.5011%'	'express'	[55]	'0.1358%'
'dai'	[186]	'0.4592%'	'peopl'	[54]	'0.1333%'
'luci'	[183]	'0.4518%'	'invit'	[53]	'0.1308%'
'colonel'	[174]	'0.4296%'	'busi'	[52]	'0.1284%'
'make'	[168]	'0.4147%'	'bodi'	[51]	'0.1259%'
'feel'	[163]	'0.4024%'	'account'	[50]	'0.1234%'
'happi'	[160]	'0.3950%'	'farther'	[49]	'0.1210%'

'hous'	[48]	'circumst'	'0.3777%'	[153]	'0.1185%'
'good'	[47]	'disappoint'	'0.3752%'	[152]	'0.1160%'
'great'	[46]	'attend'	'0.3703%'	[150]	'0.1136%'
'thought'	[45]	'care'	'0.3678%'	[149]	'0.1111%'
'brandon'	[44]	'carriag'	'0.3456%'	[140]	'0.1086%'
'engag'	[43]	'admir'	'0.3407%'	[138]	'0.1062%'
'made'	[42]	'ad'	'0.3382%'	[137]	'0.1037%'
'kind'	[41]	'began'	'0.3357%'	[136]	'0.1012%'
'hope'	[40]	'believ'	'0.3135%'	[127]	'0.0987%'
'heart'	[39]	'attempt'	'0.3111%'	[126]	'0.0963%'
'middleton'	[38]	'come'	'0.2962%'	[120]	'0.0938%'
'sir'	[37]	'beauti'	'0.2938%'	[119]	'0.0913%'
'dear'	[36]	'acknowledg'	'0.2839%'	[115]	'0.0889%'
'friend'	[35]	'civil'	'0.2814%'	[114]	'0.0864%'
'moment'	[34]	'case'	'0.2790%'	[113]	'0.0839%'
'repli'	[33]	'assist'	'0.2716%'	[110]	'0.0815%'
'affect'	[32]	'astonish'	'0.2691%'	[109]	'0.0790%'
'live'	[31]	'affair'	'0.2666%'	[108]	'0.0765%'
'return'	[30]	'advantag'	'0.2617%'	[106]	'0.0741%'
'felt'	[29]	'accept'	'0.2592%'	[105]	'0.0716%'
'expect'	[28]	'charm'	'0.2543%'	[103]	'0.0691%'
'love'	[27]	'allow'	'0.2518%'	[102]	'0.0667%'
'comfort'	[26]	'bed'	'0.2493%'	[101]	'0.0642%'
'mind'	[25]	'afford'	'0.2469%'	[100]	'0.0617%'
'left'	[24]	'affection'	'0.2419%'	[98]	'0.0592%'
'hear'	[23]	'act'	'0.2395%'	[97]	'0.0568%'
'letter'	[22]	'ask'	'0.2321%'	[94]	'0.0543%'
'call'	[21]	'constant'	'0.2296%'	[93]	'0.0518%'
'continu'	[20]	'ag'	'0.2247%'	[91]	'0.0494%'
'manner'	[19]	'afflict'	'0.2222%'	[90]	'0.0469%'
'acquaint'	[18]	'advanc'	'0.2197%'	[89]	'0.0444%'
'spirit'	[17]	'advis'	'0.2172%'	[88]	'0.0420%'
'cri'	[16]	'agre'	'0.2123%'	[86]	'0.0395%'
'famili'	[15]	'action'	'0.2098%'	[85]	'0.0370%'
'town'	[14]	'admit'	'0.2074%'	[84]	'0.0346%'
'assur'	[13]	'absolut'	'0.2049%'	[83]	'0.0321%'
leav'	[12]	'abil'	'0.2024%'	[82]	'0.0296%'
'brother'	[11]	'approv'	'0.2000%'	[81]	'0.0272%'
'speak'	[10]	'absenc'	'0.1975%'	[80]	'0.0247%'
'found'	[9]	'abroad'	'0.1926%'	[78]	'0.0222%'
'opinion'	[8]	'ah'	'0.1901%'	[77]	'0.0197%'
'ey'	[7]	'acquit'	'0.1876%'	[76]	'0.0173%'
'attent'	[6]	'absurd'	'0.1852%'	[75]	'0.0148%'
'subject'	[5]	'abod'	'0.1827%'	[74]	'0.0123%'
'hour'	[4]	'abhorr'	'0.1802%'	[73]	'0.0099%'
'gener'	[3]	'abhor'	'0.1777%'	[72]	'0.0074%'

5- Conclusions

The main steps in our paper was Text processing (or document processing) it includes tokenization, preprocessing (converting upper case letters to lower, Unicode conversion, and removing diacritics from letters, punctuations, or numbers), stop words removal, and stemming. These steps save indexing time and space, especially for a huge set of data.

We can see from our data that the number of the unique words is decreased after stemming with a percent of nearly 30%, also we find that the words appears just in one document is also decreased in a percent of 35%. From that we can conclude that the stemming improves the performance. Also the results of the zipf's law are improved after stemming as we can see in the figures 1 and 2.

Figure 1: Before Zip's Law

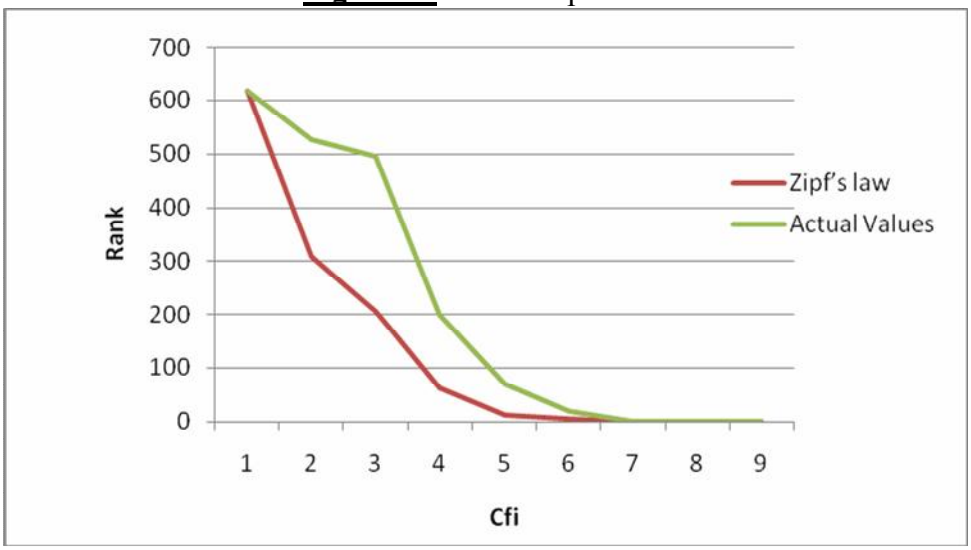
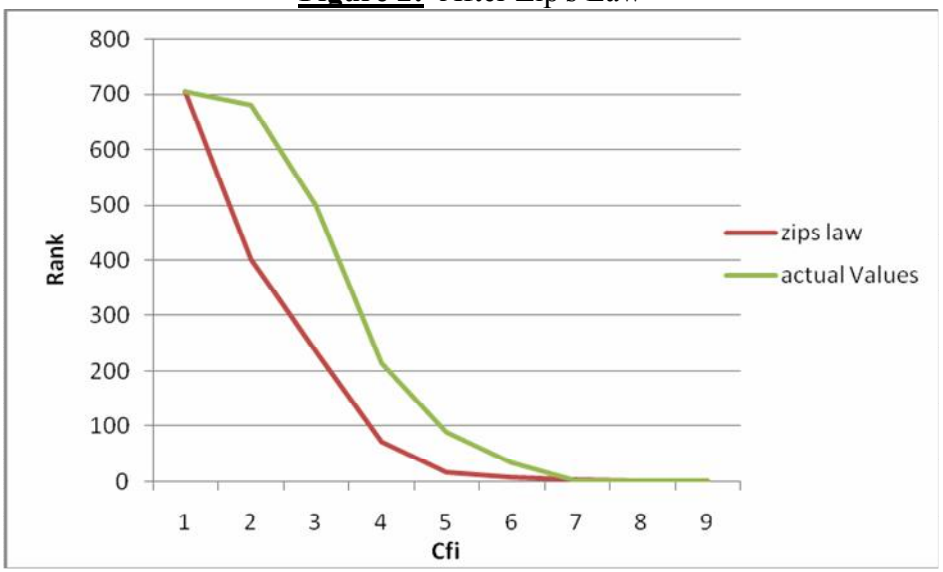


Figure 2: After Zip's Law



6- References

Wikipedia–Information Retrieval, retrieved in 2011, June. http://www.wikipedia.org/wiki/Information_retrival.

David Grossman, Ophir Frieder, "Vector Space Implementation", Illinois Institute of Technology,2010.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze "An Introduction to Information Retrieval", Cambridge University Press Cambridge, England, 2009.

Gerard Salton , Chris Buckley , Onix Text Retrieval Toolkit API Reference http://www.libertypages.com/manuals/onix/stopwords2.html

Martin Porter, "The Porter Stemming Algorithm http://tartarus.org/~martin/PorterStemmer/

Christopher D. Manning, Hinrich Schütze Foundations of Statistical Natural Language Processing, MIT Press (1999), ISBN 978-0262133609, p. 24