PAPER REVIEW

Presented By: Udochukwu Nweke

CS734: Introduction to Information Retrieval

Dr. Michael Nelson

October 12, 2017

Paper 1

Extension of Zipf’s Law to Words and Phrases

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, F. J. Smith
School Computer Science
Queen’s University of Belfast Belfast
BT7 1NN, Northern Ireland

Date: 2002

Paper 2

A Comparison of Document, Sentence, and Term Event Spaces

Catherine Blake
School of Information and Library Science
University of North Carolina at Chapel Hill
North Carolina, NC 27599-3360

Date : 2006

Zipf’s Law

- Observation: frequent words and rare words
- In English, in a random piece of text: “of” and “the” make up 10% of all occurrences
- A Word like “aardvark” is extremely rare
- Zipf’s law states that the frequency of word tokens in a large corpus of natural language is inversely proportional to the rank.

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

Where f is the frequency of word in a corpus, r is the rank, and k is the constant for the corpus

Zipf’s Curve

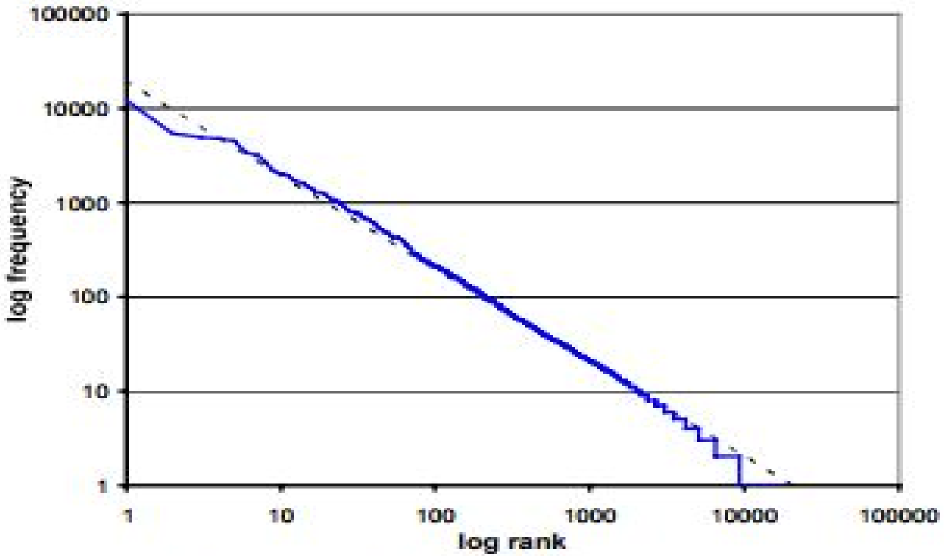
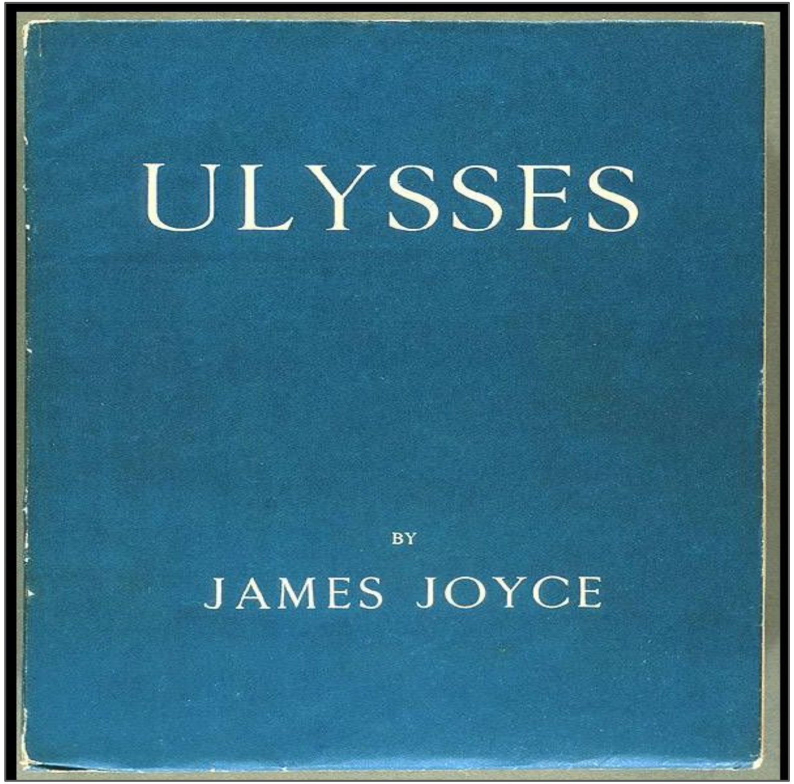


Figure 1 Zipf curve for the unigrams extracted from a 250,000 word tokens corpus

- After computing the frequencies and sorting the frequencies by rank

- Zipf’s curve is a straight line with a slope of -1.

Zipf’s Law Discovery



- Zipf discovered the law by manually analyzing the frequency of words in the novel “Ulysses”
- The novel contains 29, 899 different word types associated with 260,430 word tokens

Zipf’s Law For Larger Corpora

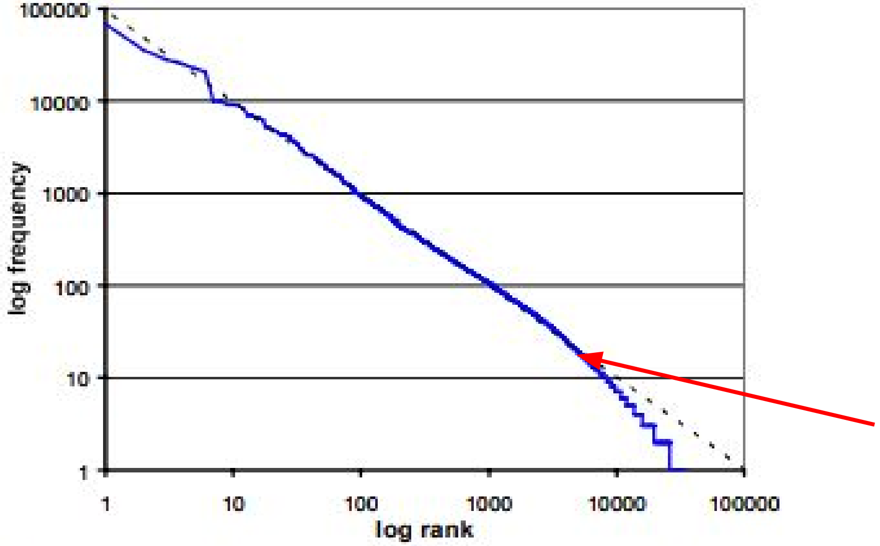
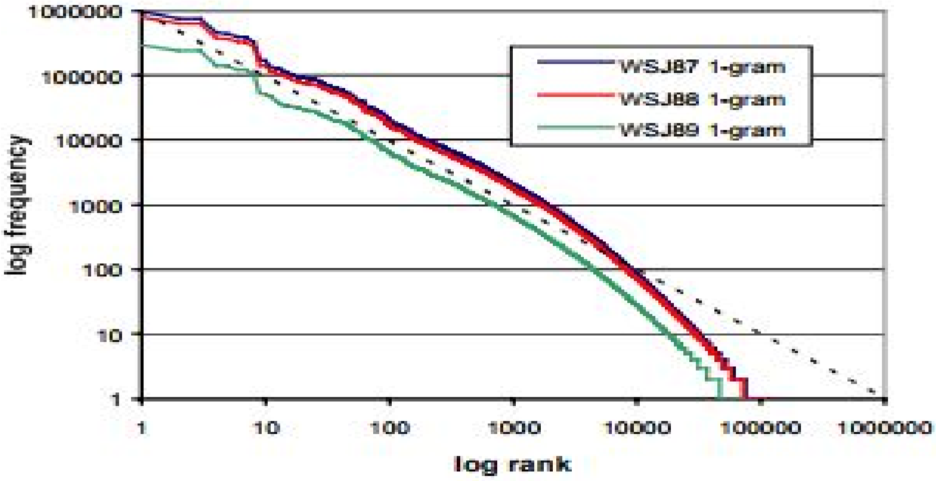


Figure 2 Zipf curve for the unigrams extracted from the 1 million words of the Brown corpus

After the development of more advanced PCs in 1980, a larger corpora of over a million word was processed and when Zipf’s curve was drawn, it was found to drop below the zipf straight line with a slope of -1 for r>5000.

Exploring the Invalidity of Zipf’s Law for Larger Corpora In Two Languages (Paper 1)



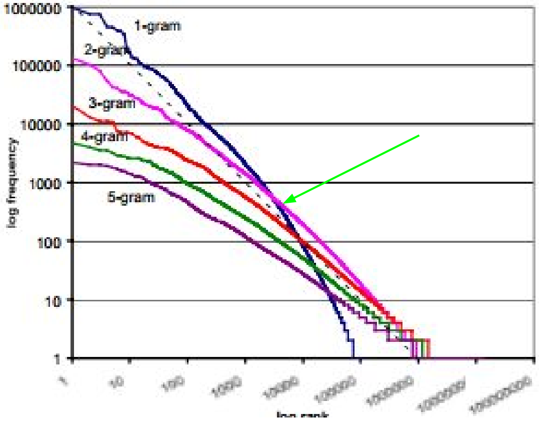
The English corpora used for these experiments were taken from the Wall Street journal (Paul & Baker, 1992) for 1987, 1988, 1989, with sizes approximately 19 million, 16 million and 6 million tokens respectively.

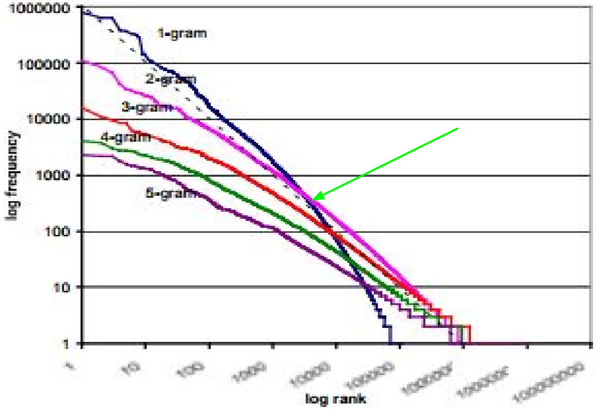
English Corpora

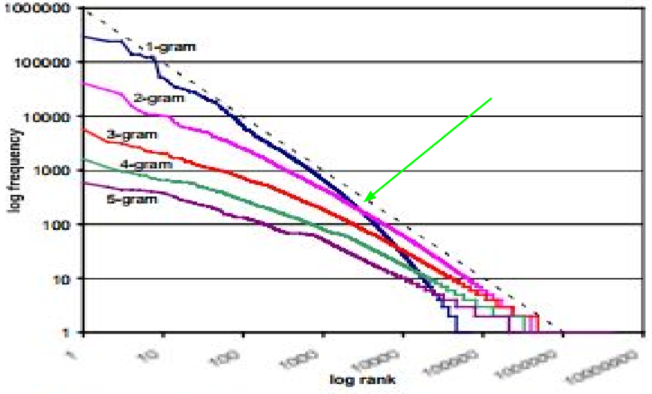
- Documents are not made of individual words but also consist of phrases of 2, 3 and more word, usually called n grams.

- In order to draw Zipf’s curve for n = 2 to 5 grams, the frequence of n-gram in each corpus is computed and placed in a rank order.

Zipf’s curve for WSJ Corpus (n=1, 2, 3, 4 and 5 grams)







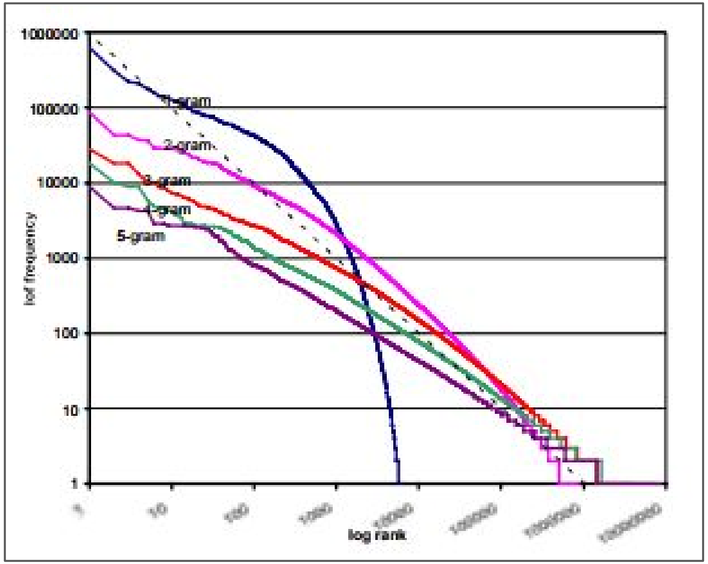
Although the Zipf’s curves for the WSJ corpora follow a straight line, the unigram curves crosses the bigram when the rank is approximately 3,000 in all 3 cases

Mandarin Corpora

The Mandarin corpus used in for experiments is the TREC Corpus. It was obtained from the People’s Daily Newspaper from 01/1991 to 12/1993 and from the Xinhua News Agency for 04/1994 to 09/1995 from the Linguistic Data Consortium (http://www.ldc.upenn.edu).

Note: The Mandarin language is a syllable-class language in which each syllable is the same time a word and a chinese character.

Zipf’s Curve For TREC Mandarin Corpus

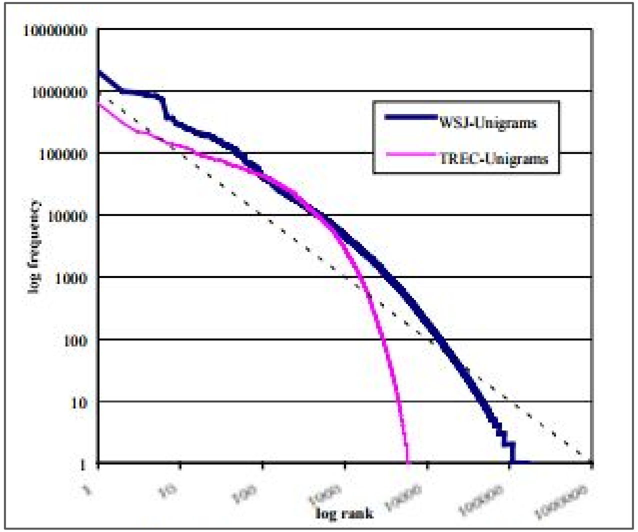


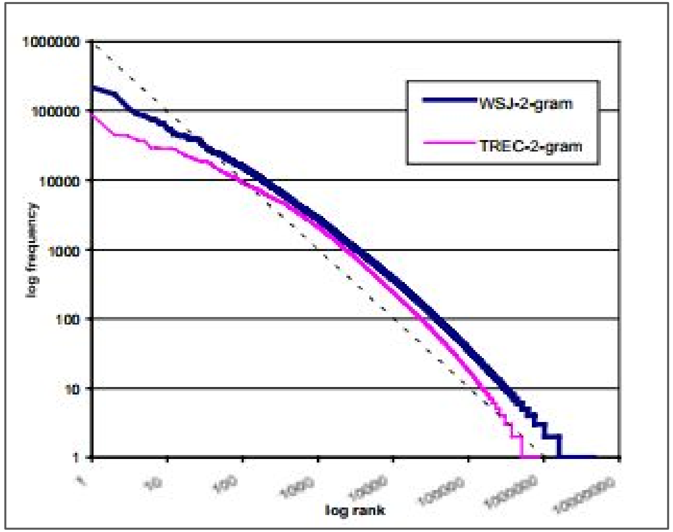
Observations:

- The bigram curve in the mandarin is more curved than the bigram curve in English Corpora.
- The shapes of the other TREC n-gram Zipfian curves are similar to but not the same for those in the English Corpora.

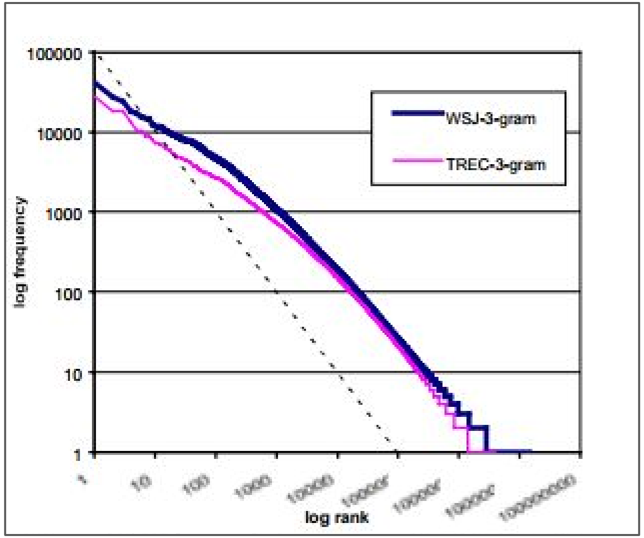
Comparing Unigrams and 2-grams For The WSJ English And TREC Mandarin Corpora

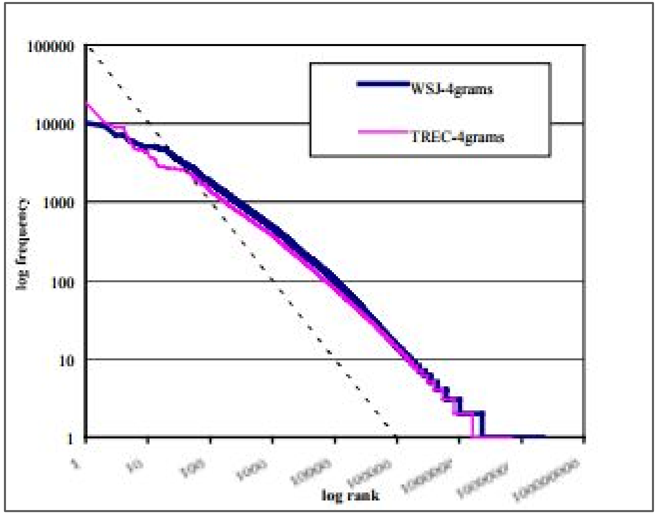
The English curves are for the 3 WSJ corpora joined together making 40 million word corpus.





Comparing Trigram and 4-grams For The WSJ English And TREC Mandarin Corpora

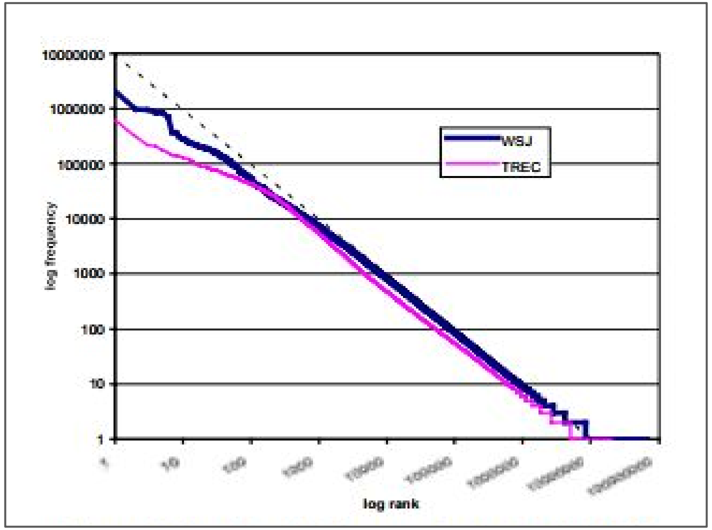




Why Zipf’s Law Is Invalid for Large Corpora

- Zipf’s law derivation was solely based on single words and it failed for English when the number of word types was greater than 5000 words
- In Mandarin, it failed almost immediately for unigrams
- Probably combining Mandarin compound words in bigram, trigram and higher n-gram statistics wouldn’t have failed Zipf’s law

Combined n-grams Will Give A Better Zipf Curve



Zipf’s curve for all unigram and n-gram together with their frequencies, sorted on frequency and put in a rank.

The combined n-gram curve follows Zipf’s straight line with slope close to -1

Findings

- The combined curve for both languages are straight line with slope close to -1 for all ranks > 100.
- The result has been found to be remarkable for 3 other natural languages: Irish, Latin and Vietnamese, in preliminary experiment.

Paper 2

A Comparison of Document, Sentence, and Term Event Spaces

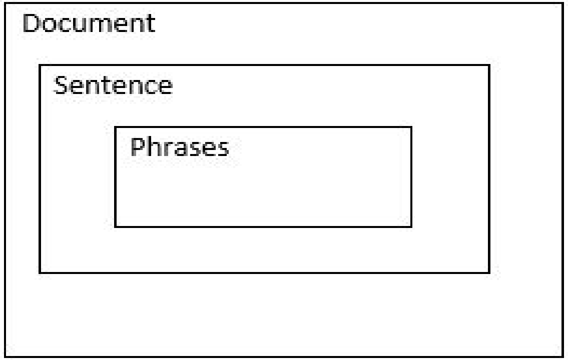
Catherine Blake
School of Information and Library Science
University of North Carolina at Chapel Hill
North Carolina, NC 27599-3360

Date: 2006

Identifying Relevant Documents In Terms of IDF

- In Information retrieval, relevant documents are identified by comparing query terms with terms from a document corpus
- The most common corpus weighting scheme is term frequency (TF) x inverse document frequency (IDF)
- $$\text{IDF} = \frac{\text{Number of Document}(N)}{\text{Number of Document}(n_i)\ \ \text{with term}(t_i)}$$
- N is the total number of corpus documents; n_i is the number of documents that contain at least one occurrence of the term t_i; and t_i is a term, which is typically stemmed.

IDF Should Not Be The Only Corpus Weighting Scheme



- Information retrieval systems trends from document to sub-document retrieval
- Sentence for summarization
- Words or Phrases for question answering system
- IDF should be replaced with ISF and ITF when sub-documents are retrieved

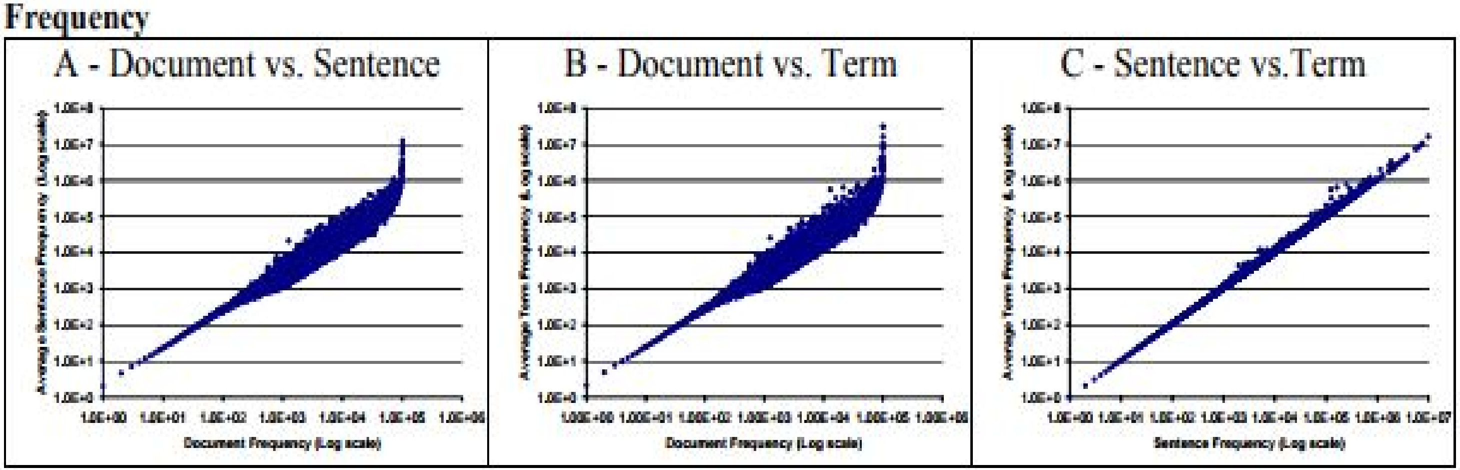
$$\text{ISF} = \frac{\text{Number of Sentence(N)}}{\text{Number of Sentence}(n_i)\text{ with term}(t_i)}$$

ISF and ITF Challenges

- The challenge is that although document language models (IDF) has had unprecedented empirical success, language models based on a sentence or term do not appear to work well (Robertson, 2004).

- The goal is to uncover why IDF is the best weighting scheme using different metrics

Raw Term Comparison

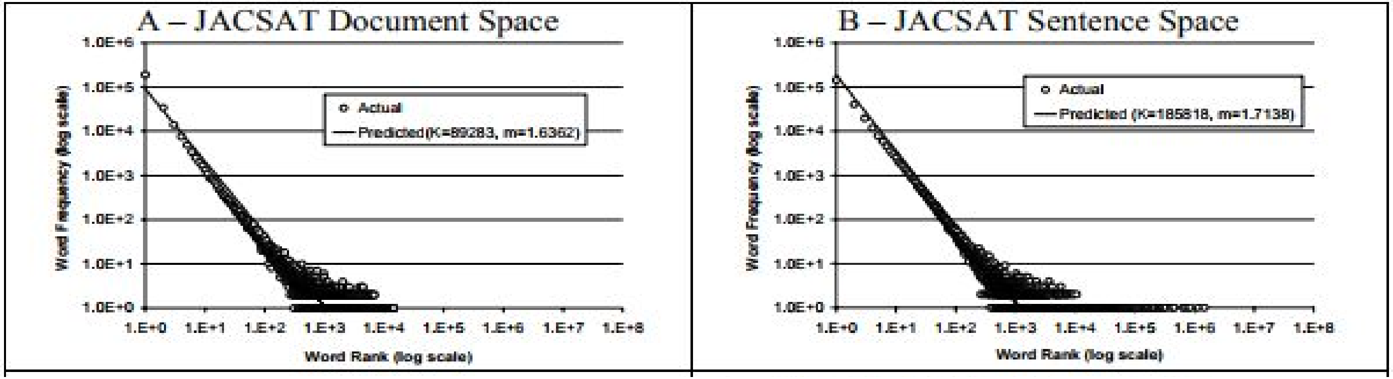


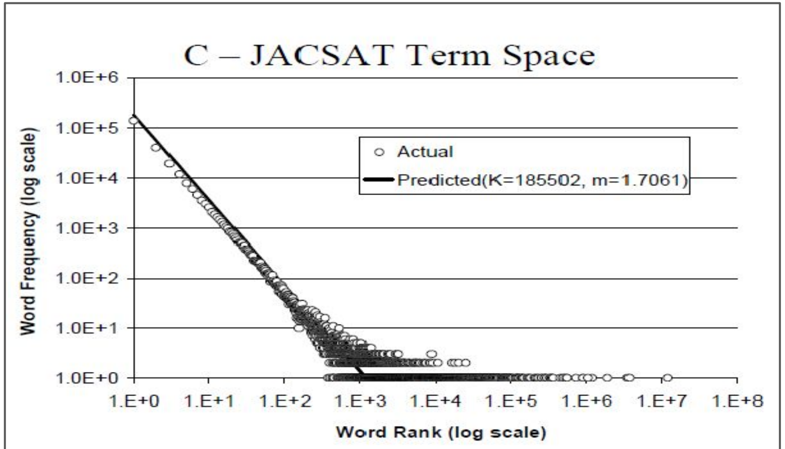
- A compares the document and the average term frequency

- B compares the the document and the average term frequency

- C shows the sentence frequency and average term frequency, demonstrating that sentence and term are highly correlated, unlike A and B

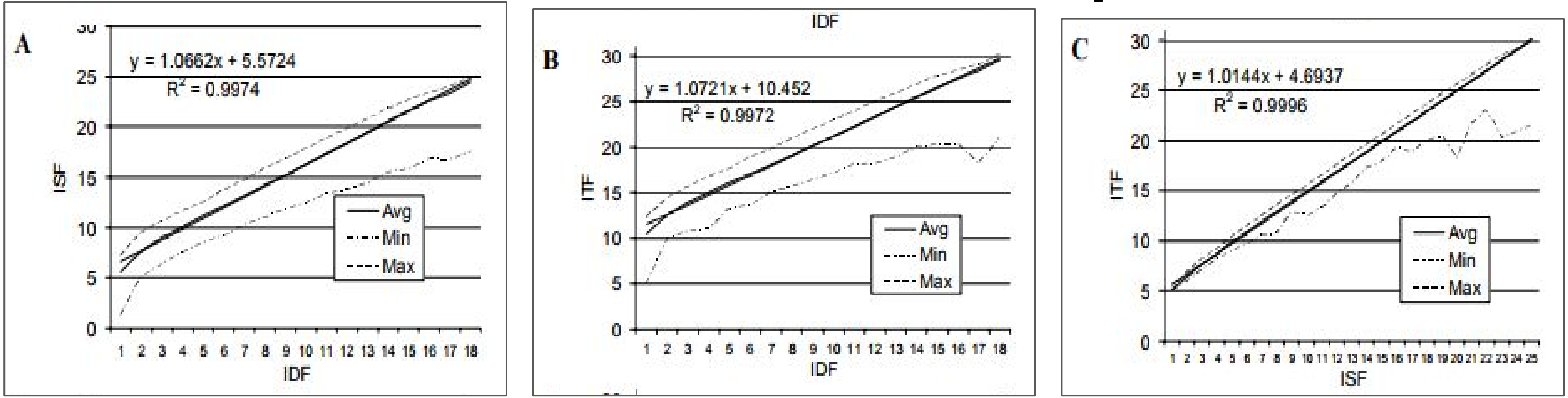
Zipf Law Holds For All Event Spaces





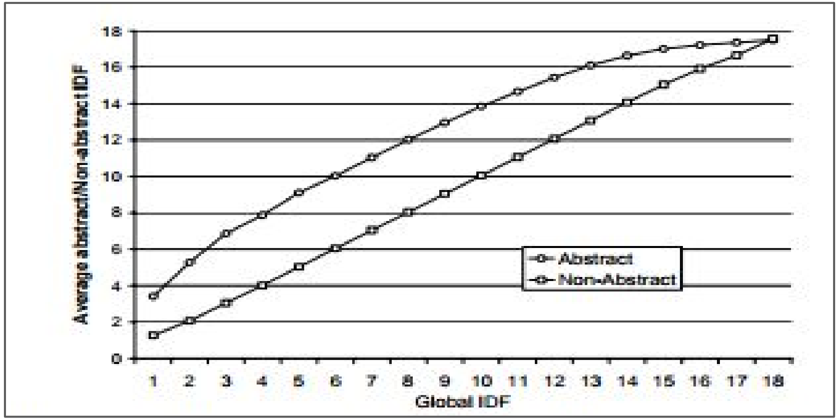
Zipf’s curve showing word frequency plot against rank.

Direct IDF, ISF, and ITF Comparisons



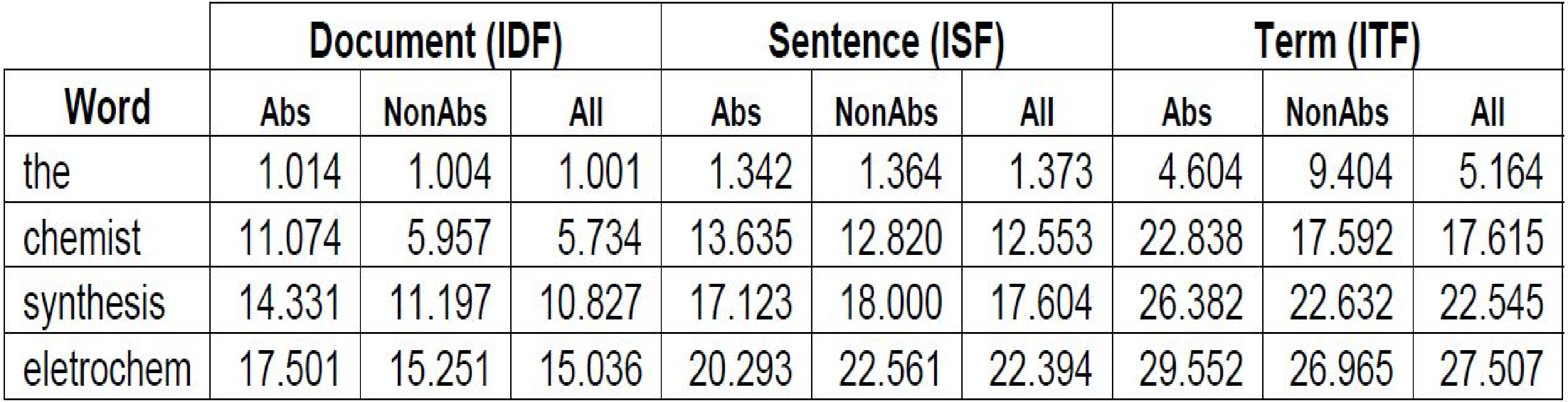
- A shows the average, minimum and maximum ISF for each rounded IDF value. Although the graph shows a correlation for A and B, the average ISF and ITF values are 5.57 and 10.45 times greater than the corresponding IDF respectively.
- Although C shows a stronger correlation between ITF and ISF, the ITF values are higher than the equivalent ISF values

Abstract Versus Full-Text Comparison



The graph explains that the weight assigned to stemmed terms in abstract were higher than than the weight assigned to terms in the body of a document.

IDF, ISF And ITF Values of Abstract and Main Text



	Document (IDF)			Sentence (ISF)			Term (ITF)		
Word	Abs	NonAbs	All	Abs	NonAbs	All	Abs	NonAbs	All
the	1.014	1.004	1.001	1.342	1.364	1.373	4.604	9.404	5.164
chemist	11.074	5.957	5.734	13.635	12.820	12.553	22.838	17.592	17.615
synthesis	14.331	11.197	10.827	17.123	18.000	17.604	26.382	22.632	22.545
eletrochem	17.501	15.251	15.036	20.293	22.561	22.394	29.552	26.965	27.507

The Table shows IDF, ISF and ITF values for abstract and non abstract documents. ITF assigned the highest weights followed by ISF and IDF has the least value.

Findings

- A linear transformation between document to sub document will be difficult because there is no direct correlation between document and sub documents.
- Although IDF, ISF and ITF are highly correlated, replacing IDF with ISF or ITF will result in a weighting scheme where corpus weight dominated the weight assigned to a query and document terms.

Relationship Between Paper 1 and Paper 2

Zipf’s law based solely on single term failed for large corpora when the number of word type approached 5000.

Both papers use Zipf’s law to understand the distribution of words in a corpus. Paper 1 shows that different corpora follow the Zipf’s law, but when the size of a corpus grows, some of the curves deviated from Zipf’s straight line curve.

Paper 2 shows that Zipf’s law is applicable to multiple event spaces such as Document (IDF), Sentence (ISF), and term (ITF).

Zipf’s law was a better fit when term space comprised of both word and phrases