

Searchable Words on the Web

Hugh E. Williams Justin Zobel
Department of Computer Science, RMIT University,
GPO Box 2476V, Melbourne 3001, Australia
{hugh,jz}@cs.rmit.edu.au

Abstract

In designing data structures for text databases, it is valuable to know how many different words are likely to be encountered in a particular collection. For example, vocabulary accumulation is central to index construction for text database systems; it is useful to be able to estimate the space requirements and performance characteristics of the main-memory data structures used for this task. However, it is not clear how many distinct words will be found in a text collection or whether new words will continue to appear after inspecting large volumes of data. We propose practical definitions of a word, and investigate new word occurrences under these models in a large text collection. We inspected around two billion word occurrences in 45 gigabytes of world-wide web documents, and found just over 9.74 million different words in 5.5 million documents; overall, 1 word in 200 was new. We observe that new words continue to occur, even in very large data sets, and that choosing stricter definitions of what constitutes a word has only limited impact on the number of new words found.

1 Introduction

The World-Wide Web is growing exponentially, with at least two billion searchable pages in some single repositories.¹ We can only speculate as to whether the new pages appearing every day cover new topic material, but the scope of the current web is far greater than that of any traditional encyclopedia. The pages indexed by the major search engines include not just material such as news articles, literature, technical material, legislation, advertising, corporate data, and personal information, but are written in a multitude of languages—some of them, such as Klingon and Furbish, imaginary.

Web search engines use the words occurring in web pages to locate particular pages in response to user queries. The words are extracted from the pages during construction of inverted indexes [12]. Wide experience with text querying has shown that effectiveness (the ability to find pages that satisfy users' needs) is greatest when all words are indexed, other than perhaps a limited number of function words such as “the” and “moreover” [12]. However, complete indexing presents significant practical problems, one of which is vocabulary size: the smaller the vocabulary, the greater the likelihood it can be maintained in memory, and the simpler it is to support query modes such as vocabulary browsing.

Perhaps surprisingly, the number of new words observed in web data does not appear to taper off as the volume of data increases. That is, even after many gigabytes of text have

¹For example, see Google, at <http://www.google.com/>.

been processed, the vocabulary continues to grow [17]. There are several likely causes for such ongoing growth: neologisms; first occurrences of rare personal names and place names; first occurrences of complex chemical names and material such as DNA strings; words in documents written in unusual languages; URLs; and typographical errors. The space of all strings grows very rapidly with string length: there are for example over 8 million 7-character strings of lower-case letters.

We have investigated the rate at which new words are observed and whether new words that might reasonably occur in a query are observed after processing of large volumes of text. To some extent these questions depend on how the parser defines “word”; using different definitions of word—such as whether the document appears to be in English, whether words are allowed to contain digits, whether the word appears to derive from a small dictionary, and by inspection—we conclude that potentially valuable new words do continue to occur.

2 Defining “word”

In English, a simple definition of what constitutes a “word” is any sequence of alphanumeric characters bounded by non-alphanumerics. If this is extended to include single occurrences of the quote or hyphen characters within a word (thus including “don’t” and “right-handed” but not “students’ ”), but to exclude strings with more than two digits, it covers almost every string that in English might reasonably be regarded as a word and can be used in practice for vocabulary accumulation tasks [4, 10, 16]. We refer to this as the ALNUM class.

However, the ALNUM class clearly includes many strings that would not be generally regarded as words; the goal of an ideal—but implausible—parser would be to eliminate all such non-words from ALNUM. We have explored several different more restrictive classes.

One of these classes, ALPHA, is the words containing only alphabetic characters; to further restrict this class we converted upper-case characters to lower-case, so that “the” and “The” were regarded as identical and “don’t” was treated as two words, “don” and “t”. This approach is probably overly restrictive for most tasks, eliminating strings that would generally be regarded as words—and certainly as useful search terms—such as “3M”, “MP3”, “747”, “R2D2”, and “ISO9001”.

Another approach is to only consider the words occurring in English-language documents, thus eliminating—hopefully—not just documents in other languages but documents containing no text at all; this would also permit categorisation of documents by language, providing additional filtering in searching tasks or vocabulary browsing. However, there is no simple, reliable way of identifying documents as being in English; we used heuristics that, by inspection, appear to be reasonably reliable. These heuristics were based on the words occurring in a dictionary of 126,001 words distributed with a version of the publicly-available `ispell` spell-checker utility. First, we required that at least 60% of the words parsed from the document occur in the dictionary; second, we required that the words constitute at least 40% of the page (other than material in HTML markup tags). Testing these heuristics with a sample of English pages and other pages, we found that these thresholds were well below the percentages found in all the English pages and well above those found in the others. Use of this filter gave us two further classes, ENGLISH-ALNUM and ENGLISH-ALPHA, the set of words occurring in “English” pages.

The final approach we considered was the class DICT, of ENGLISH-ALPHA words that are in the `ispell` dictionary. This class is extremely restrictive: the dictionary does not include personal names, chemical names, place names, and so on, and is even deficient in technical

Table 1: *Occurrence statistics for different definitions of “word”.*

| | ALNUM | ALPHA | ENGLISH -ALNUM | ENGLISH -ALPHA | DICT | STEM -DICT |
|------------------------------------|-------|-------|-------------------|-------------------|-------|---------------|
| Pages ($\times 10^6$) | 5.5 | 5.5 | 4.1 | 4.1 | 4.1 | 4.1 |
| Distinct words ($\times 10^6$) | 9.74 | 6.36 | 4.79 | 2.96 | 0.10 | 0.20 |
| Word occurrences ($\times 10^6$) | 1,915 | 1,897 | 1,586 | 1,576 | 1,576 | 1,576 |
| New per last 1,000,000 | 247 | 144 | 221 | 114 | 1 | 8 |

terminology. This is an impractical definition of a word, but a useful comparison point for the experiments described in the next section.

A slightly broader class is STEM-DICT, of ENGLISH-ALPHA words that stemmed to the same root as a word in the `ispell` dictionary. Stemming is the process of removing variant suffixes from words, such as “-tion”, “-ed”, or “-s”; we used a publicly-available version of the Porter stemmer [8]² that, while possibly being inferior to other stemmers [13], is used widely in practice. The utility of this class of words is also unclear, but it again provides a useful comparison point that is less restrictive than DICT.

In all classes we truncated strings at 32 characters.

3 Data

To examine trends in occurrences of new words we used the data gathered for the Web track of TREC, an international collaborative evaluation of information retrieval techniques [2]. We used approximately 45 Gb of the Web data, which derives from a crawl of the Web undertaken in 1997.

Before extracting words from the data we preprocessed it. We eliminated material in tags, excepting comments, because such material includes long non-word strings generated for purposes such as security.

We also eliminated material that appeared to be binary data, again through simple heuristics. These were based on the assumption that a word should be in the neighbourhood of a white space character, and that legitimate text should include some white space. Long strings without white space but including punctuation were assumed not to be text, thus, for example, eliminating most `http` addresses and directory path names—both of which are common sources of distinct strings that might be treated as words. Additionally, we processed only unencoded ASCII text and did not process encoded data that may be a source of words. For example, MIME-encoded postscript using the “application/postscript” content type was not decoded to ASCII text.

4 Results

The number of pages, distinct words, and word occurrences for each class is shown in Table 1. The last line shows the number of new words found in the last million word occurrences

²Many public-domain implementations of the Porter stemmer are available. Stuart J. Barr’s implementation in the C programming language works well and is available from Mark Sanderson’s IR resources at http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/

inspected. Using our English filtering scheme, we found that around three-quarters of the pages appear to be in English; therefore, while all experiments process around 45 gigabytes of data, the experiments with the ALNUM and ALPHA classes are on around 5.5 million pages, while the other classes are tested on 4.1 million pages.

Table 1 shows that, for the non-dictionary classes, even after almost two billion words observed, new words occur on average at a rate of more than 1 in 10,000 (for the most restrictive ENGLISH-ALPHA class). The rate is more than 1 in 4000 for the least restrictive ALNUM class. The dictionary classes are impractically restricted, but serve a useful comparison: over the whole collection, new words in these classes occur, in DICT at a rate over 1 in 16,000 and in STEM-DICT at just over 1 in 8,000. Indeed, a new word was found in the last million word occurrences in even the DICT class (“glitteringly”).

Observing these words in more detail, from the beginning to the end of our collection we found documents with unique words, that is, words found in one document only. As discussed below, many of these are errors but some are not. Thus the phenomenon is not a consequence of document order.

Interestingly, for the dictionary-based schemes using the dictionary of 126,001 words, less than 80% of the words are found in the English documents inspected. (We wonder at the utility or correctness of the remaining words.) Correctness in general of new words occurring after inspecting large text collections is also of interest, and we discuss a simple evaluation of the correctness of new word appearances later in this section.

In developing the practical restrictions of the ALNUM class, we found only a tiny fraction of word occurrences contain more than two digits. The reduction in vocabulary size from 9.74 in ALNUM to 6.36 million words in ALPHA—around a third of the distinct words disappear—is largely due to the standardised case in the ALPHA class; the result is similar for the English-filtered ENGLISH-ALNUM and ENGLISH-ALPHA.

For the non-dictionary classes, we have seen that on average over one word occurrence in 10,000 is new. However, Figure 1 shows the decreasing trend in the number of new words seen under each definition of a word with the word occurrences inspected. We show a coarse value, where each point represents the total new words seen in the last 100 million word occurrences. While the trend is towards fewer new words, the rate of arrival of new words varies: for the ALNUM class, around 395,000 new words were seen in the 100 million word occurrences before the 900th million word, while over 470,000—around 19% more—were seen in the 100 million before the 1,100th million word occurrence. After inspecting another 800 million words and adding another 2.9 million new words to the vocabulary, almost 440,000 more new words were seen in the 100 million words before the 1,900th word occurrence. Not surprisingly, with stricter definitions of “word” the variation is reduced.

The variation in frequency of new words is emphasised further in Figure 2. We show here the total new words seen with each 10 million word occurrences; we show only three representative schemes for readability. We have observed, in general, that peak regions of new word occurrence result from encountering usually only a few documents that cover new topics. For example, a page reporting research in chemistry or physics containing new terminology—but meeting thresholds for the ENGLISH-ALNUM class—can be a source of many new words; encountering a site of such pages results in a peak. Another common source of words are lists of rare place names or other proper nouns. This occurrence of peak regions continues to occur despite English filtering as can be seen in the ENGLISH-ALNUM class.

To investigate the utility of new words, we carried out a simple experiment. We produced two largely distinct lists of new words occurring in the inspection of the collection using the

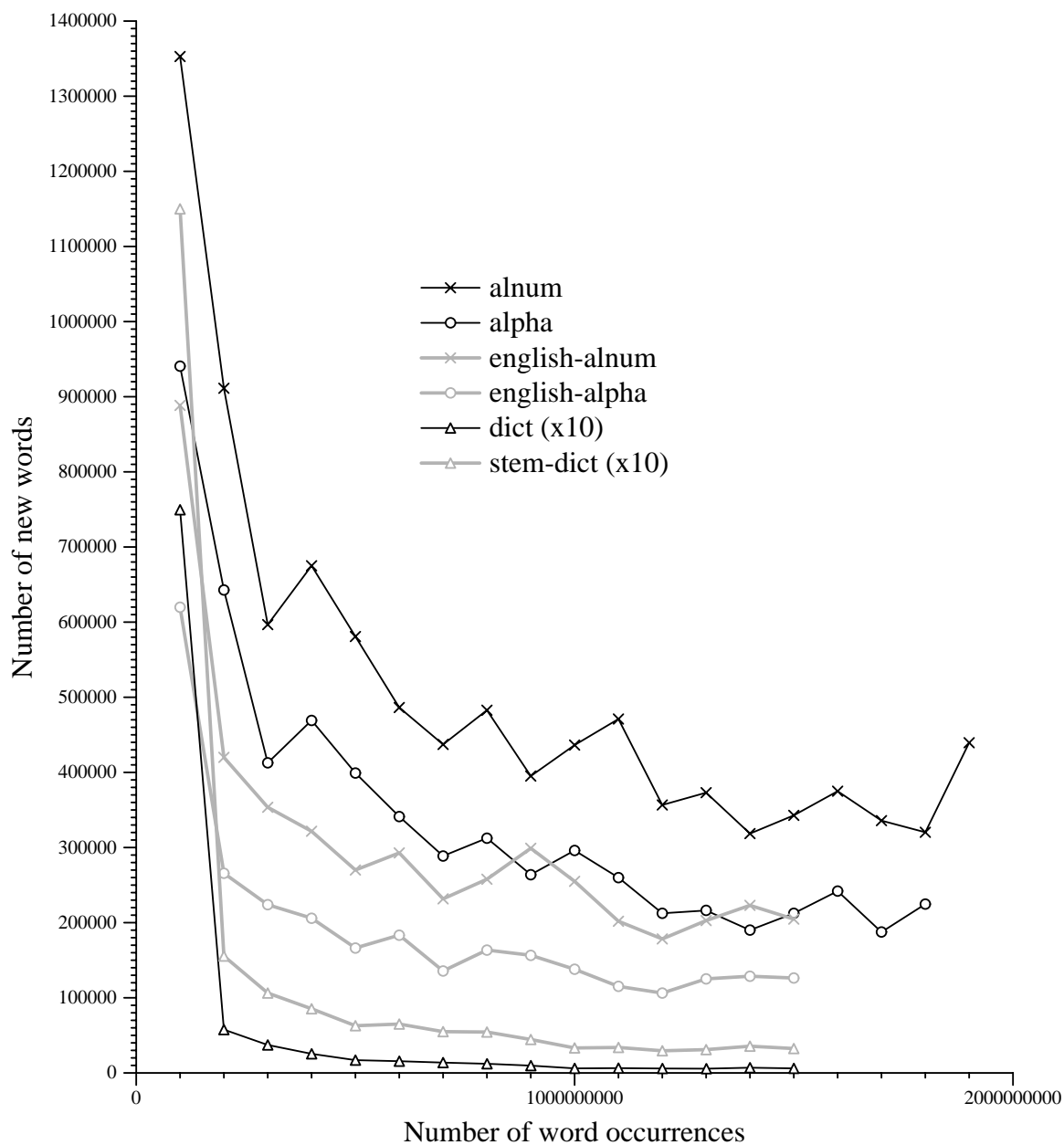


Figure 1: *New word occurrences versus words inspected in a collection of 45 gigabytes of world-wide web documents. The points shown are the total new words seen at each 100 million word-occurrence interval. Six schemes derived from different definitions of a word are shown. The ALNUM and ALPHA schemes show new words seen in processing 5.5 million unfiltered pages; the other schemes show new words in 4.1 million pages that were filtered as being written in English.*

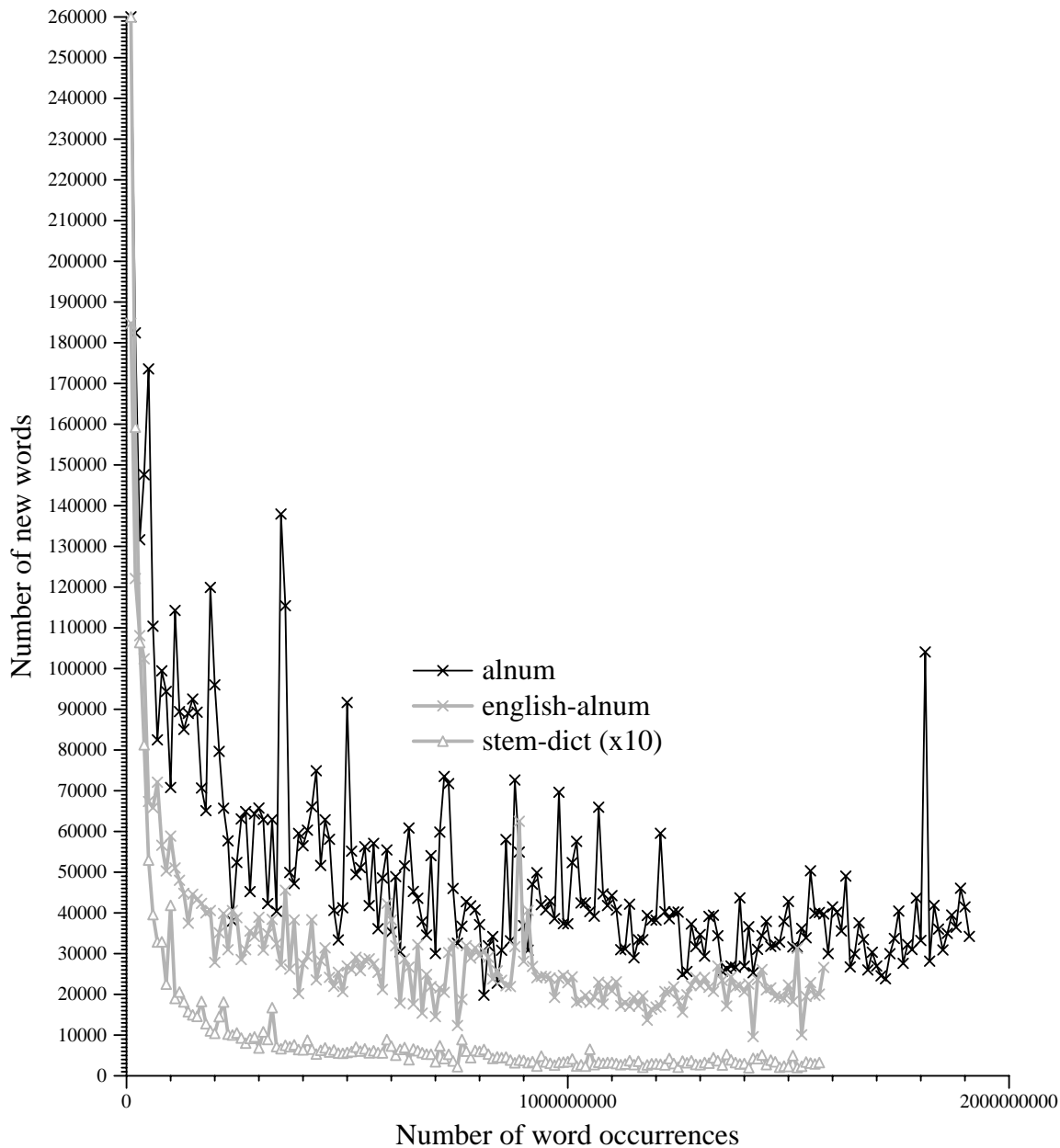


Figure 2: *New word occurrences versus words inspected in a collection of 45 gigabytes of world-wide web documents. The points shown are the total new words seen at each 10 million word-occurrence interval. Three schemes from Figure 1 are shown for readability.*

Table 2: *New words in the ALPHA class. The table shows four different localities of word occurrences from 1×10^3 to 1×10^9 and the percentage of 200 words judged as correct new words. Examples of correct and incorrect words are shown on the right of the table: the first line for each locality shows the first three words judged correct and incorrect by the first author; the second line shows the first three words judged correct and incorrect by the second author. The first author judged all 100 words near the first 1,000 occurrences as correct words.*

| Word Occurrence | Correct | Acceptable Examples | Unacceptable Examples |
|-------------------------|---------|--|---|
| $\approx 1,000$ | 95% | jul, gnu, make results, search, mailed | — dev, dec, internationaltesting |
| $\approx 100,000$ | 84% | protecting, denounced, clawback caps, unreliable, dimple | universitetsadjunkt, lindex, alot ctron, mis, cogentdata |
| $\approx 10,000,000$ | 68% | stings, pender, honker fastings, gunpowder, sobered | vilain, intensit, essouffl lacomb, stensas, umhoefer |
| $\approx 1,000,000,000$ | 39% | alapatt, myeloschisis, rainsong rudsdale, khamar, berrybank | nephroma, theprairies, maalstom oakv, sevron, mmys |

ALPHA class. The lists contained 100 new words from near each of the 1,000th, 100,000th, 10,000,000th, and 1,000,000,000th word occurrence localities. We extracted every new word, every 5th new word, every 25th new word, and every 125th new word respectively until a list of 100 words for each locality was produced. After this, we sorted both lists of 400 words, and independently judged each word as either a reasonable new word, or an erroneous word choice.

Table 2 shows the results of our word-judgement experiment. As expected, the average percentage of correct new words falls as more words are inspected. However, even after inspecting one billion words, more than a third of the words are still correct, suggesting more than 40 correct new words per million word occurrences are likely to be encountered if new pages are added to our collection. Twelve selected words from each locality, six of which we judged to be acceptable and six that were unacceptable words, are shown at the right of the table; the first line for each locality are the first six words judged correct and incorrect by the first author, while the second line shows words judged by the second author.

Many of the words that first occur after a billion word occurrences appear to be spelling or typographic errors. These fall into several categories; for example, a considerable proportion of them are concatenations, and others are transpositions. It is attractive to consider automatic correction of these words. However, automatic correction has significant pitfalls. One, it is necessary to have a list of words that are known to be correctly spelt. In the context of the web—which has many languages, unusual place names, technical literature, and tens of millions of personal home pages—no dictionary can possibly fulfil this role. Two, it is not easy to identify misspelt words, even given an “oracle” of correct spellings. As a simple experiment, for each word that occurs less than ten times we tried to find a near neighbour (as measured by an edit distance [5, 15]) amongst the more common words. This was a complete failure: with a reasonable definition of neighbourhood, most of the rare words did not have a near neighbour, while the other rare words typically had several. Many words that were obviously misspellings did not, surprisingly, have a neighbour at all. Three, rare words are often not misspellings. For example, “villein” and “serf” are rare words but are correctly

spelt. Emerging terms, such as the use of place names in newswires and of chemical names in technical literature, are rare but valid. A recent thesis has explored these issues further [3].

The value of internet search services is that they are able to find matches to a wide range of queries. While many of the rare words are clearly errors, there is no obvious method for distinguishing the mistakes from words that are potential query terms. While most rare terms are unlikely to be used in queries, the options for a search engine are to keep them all or discard them all; the latter means that some queries cannot be correctly resolved. As a simple example, URLs are commonly posed as queries; yet many of the words in URLs occur in one document only.

We counted the occurrence of rare words in a large query log [9]. Among 1.7 million queries, we found approximately 250,000 distinct query words after casefolding and removing non-alphabetic characters. In the class of words ALPHA, there were 2,000,000 words first seen after the billionth word occurrence. Of these, 8000 (one in 250) were in the query log. That is, one in 30 of the words in the query log were extremely rare in the data.

Overall, the most significant result is that new, correct words continue to occur under all practical definitions of a word.

5 Models of Word Occurrences

When discussing the occurrence of words, reference is often made to the so-called “Zipf’s law” (in fact, it is a conjecture) of George Zipf. His observation was that the frequency f of occurrence of an event, as a function of the rank r (when the rank r is determined by the frequency f of occurrence), is an inverse power-law function $f_r = 1/(r \times \alpha)$, where α is a constant. Zipf [14] observed—most famously—this phenomena in the frequency of use of words in English text, but also in the size of cities (the largest city has twice the population of the second-largest city, and so on), and concerning the profitability of companies.³ Indeed, the phenomena is one of the most widely studied, and has been explained adequately by statistics, while also faithfully viewed by others as illustration of Zipf’s principle of least-human effort [6].

Consider the frequency of distinct words in English text as an example. Using text from the Wall Street Journal (WSJ) distributed as part of the TREC project [2], the constant α in $f_r = 1/(r \times \alpha)$ has been observed to be 0.1 for the frequency of words. In WSJ, the most frequently-occurring word (“the”) occurs just over twice as often as the second most-frequent word (“of”), that in turn occurs 1.1 times more frequently than the next word (“to”), and so on. However, for the Web data, the law does not hold; the relationship is approximately $f_r = 1/(r^{1.5} \times \alpha)$, and the original formulation overestimates the frequencies of the rare words by a factor of 50 or so.

It has also been observed (and assumed) that the occurrence of new words in English text follows a Zipf-like distribution [1, 11]. Other rank-frequency models have also been proposed to predict the number of distinct words in English text collections. For example, Baeza-Yates and Ribeiro-Neto report that it has been shown on medium-size text collections that the occurrence of distinct words grows sublinearly with the collection size, in a proportion close

³This observation was in fact made much earlier by Pareto in his economics and finance work “Cours d’économie politique” (Rouge, Lausanne et Paris, 1897). Indeed, similar observations have been earlier made in various areas including the “Lotka’s (inverse-square) Law” that the number of authors making n contributions is about $1/n^2$ of those making one contribution (Alfred J. Lotka, “The Frequency Distribution of Scientific Productivity”, Journal of the Washington Academy of Sciences, 16(12):317–323, 1926).

to the square root [1]. More precisely, they state Heaps' Law that vocabulary size is $O(n^\beta)$, where n is the size of the text and β is a positive value less than one; they report a typical value of β as between 0.4 and 0.6.

Heap's Law works reasonably well as a predictor of overall collection size given a good estimate of the constant β . The relatively small collection of around 500 Mb of the WSJ, with the ALNUM word model, has a vocabulary of around 1.4×10^6 words in 7.9×10^6 word occurrences. Therefore, for this collection, $\beta = 0.16$. But different text sources display strikingly different properties: for the much larger collection used in our experiments $\beta = 0.59$ for the ALNUM class.

The number of new words in newly inspected text cannot be accurately predicted using models, and constants cannot be used on text drawn from different sources. As we have shown in the last section, the frequency of arrival of new words varies unpredictably. However, Heap's Law works well as a predictor of the overall vocabulary size using a sample of a fraction of the collection. For example, after sampling the first ten million word occurrences in our collection β can be accurately determined as 0.59 for the ALNUM class.

6 Conclusion

New words continue to occur in large text collections as more data is inspected. We have developed several definitions of words, the most practical being any sequence of alphanumeric characters bounded by non-alphanumerics but permitting a single quote or hyphen, and not permitting more than two digits. With this definition of a word, we have found that over one in each 4000 occurrences is a new word, even after inspecting 45 gigabytes of data. Restricting the definition of a word so that strings containing numbers are not included, or processing only pages that are likely to be in English, reduces the frequency of new words somewhat, but, even after almost two billion occurrences, new words still occur at a rate of more than 1 in 10,000.

More restrictive definitions of a word have limited utility, as they seem unlikely to result in significant overall space or time savings in main-memory vocabulary accumulation or searching tasks on modern hardware. For example, only 39 Mb of main-memory might be saved after processing 45 gigabytes of data if only alphabetic strings are permitted (assuming an average of eight characters per additional word, and a low overhead of four bytes per word). However, we offer a caution: while the mean number of new words falls as more data is inspected, the arrival of new words does not occur at a constant or predictable rate. Indeed, models such as Heap's law are only useful for approximating the overall vocabulary size, not the rate of new word occurrence, and Zipf's conjecture does not hold for larger collections. Moreover, the more flexible the definition of a word, the more apparent the unpredictability.

Our results show that there is no simple bound to the number of words that are likely to be found in a text collection: as the web grows, so does its vocabulary, apparently without limit. This finding contradicts common wisdom on this topic [1]. At a practical level, this has implications for technologies such as internet search engines. For example, one of the preferred methods for constructing an index for a search engine assumes that the complete vocabulary of the text can be held in memory [7]. Our results show that this assumption is incorrect: the vocabulary will grow linearly with data volume.

Acknowledgments

This work was supported by the Australian Research Council.

References

- [1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, May 1999.
- [2] D. Harman. Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, 31(3):271–289, 1995.
- [3] J. Hasan. Automatic dictionary construction from large collections of text. Master’s thesis, School of Computer Science and Information Technology, RMIT University, 2001. RT-35.
- [4] S. Heinz, J. Zobel, and H.E. Williams. Burst tries: A fast, efficient data structure for string keys. *ACM Transactions on Information Systems*, 20(2):192–223, 2002.
- [5] K. Kukich. Techniques for automatically correcting words in text. *Computing Surveys*, 24(4):377–440, 1992.
- [6] W. Li. Comments on Zipf’s law and the structures and evolution of natural language. *Complexity*, 3(5):9–10, 1998.
- [7] A. Moffat and T.A.H. Bell. In-situ generation of compressed inverted files. *Journal of the American Society for Information Science*, 46(7):537–550, 1995. To appear.
- [8] M.F. Porter. An algorithm for suffix stripping. *Program*, 13(3):130–137, 1980.
- [9] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 52(3):226–234, 2001.
- [10] H.E. Williams, J. Zobel, and S. Heinz. Self-adjusting trees in practice for large text collections. *Software Practice and Experience*, 31(10):925–939, 2001.
- [11] I.H. Witten and T.C. Bell. Source models for natural language text. *International Journal on Man Machine Studies*, 32:545–579, 1990.
- [12] I.H. Witten, A. Moffat, and T.C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, second edition, 1999.
- [13] J. Xu and W.B. Croft. Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, 16(1):61–81, 1998.
- [14] G.K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading, Mass., 1949.
- [15] J. Zobel and P. Dart. Finding approximate matches in large lexicons. *Software Practice and Experience*, 25(3):331–345, March 1995.
- [16] J. Zobel, S. Heinz, and H.E. Williams. In-memory hash tables for accumulating text vocabularies. *Information Processing Letters*, 80(6):271–277, 2001.
- [17] J. Zobel and H.E. Williams. Combined models for high-performance compression of large text collections. In *String Processing and Information Retrieval (SPIRE)*, pages 224–231, Cancun, Mexico, 1999. IEEE Computer Society Press.