Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping

Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila

Department of Information and Computer Science, Aalto University,
Helsinki Institute for Information Technology (HIIT), Finland
{firstname.lastname}@aalto.fi

Abstract. Comparing frequency counts over texts or corpora is an important task in many applications and scientific disciplines. Given a text corpus, we want to test a hypothesis, such as "word X is frequent", "word X has become more frequent over time", or "word X is more frequent in male than in female speech". For this purpose we need a null model of word frequencies. The commonly used bag-of-words model, which corresponds to a Bernoulli process with fixed parameter, does not account for any structure present in natural languages. Using this model for word frequencies results in large numbers of words being reported as unexpectedly frequent. We address how to take into account the inherent occurrence patterns of words in significance testing of word frequencies. Based on studies of words in two large corpora, we propose two methods for modeling word frequencies that both take into account the occurrence patterns of words and go beyond the bag-of-words assumption. The first method models word frequencies based on the spatial distribution of individual words in the language. The second method is based on bootstrapping and takes into account only word frequency at the text level. The proposed methods are compared to the current gold standard in a series of experiments on both corpora. We find that words obey different spatial patterns in the language, ranging from bursty to non-bursty/uniform, independent of their frequency, showing that the traditional approach leads to many false positives.

Keywords: burstiness, sequence analysis, natural language modeling.

1 Introduction

Analyzing word frequencies is important in many application domains, such as data mining and corpus linguistics. Suppose we have a set of texts and we want to test a hypothesis, such as "word X is frequent", "word X has become more frequent over time", or "word X is more frequent in male than in female speech". For such tasks, we need to have a null model of word frequencies. The standard for statistical testing of such hypothesis is based on the bag-of-words assumption, i.e., every word can occur at any position in a text with equal probability.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part II, LNAI 6912, pp. 341–357, 2011.

This assumption has been pervasively used by both data mining [15] and linguistics communities [5] for finding words with significantly elevated occurrences in a text. We show in this paper that for almost no word, its frequency distribution observed in text corpora corresponds to a binomial distribution. Thus, the binomial distribution is almost always an inappropriate null model for word frequency distribution.

In linguistics, frequencies of words and other phenomena such as proverbs, semantic tags, n-grams, etc., are widely used to study how people communicate. It is well known that the bag-of-words model is a poor descriptor of word occurrences. Linguists have gone as far as claiming that hypothesis testing of word frequencies is rarely useful to finding associations, and often leads to misleading results [14]. Others have noted that a measure of *dispersion* is necessary to improve significance testing [21], or that each significant result should be checked using an effect size measure [9] or manual investigation [20].

In information retrieval, the fraction of documents where a word occurs is used to detect content-related words. The *inverse document frequency*, used in the classic *tf-idf*, or more recent approaches such as *Okapi BM25* [2,22], is useful because content-related words are less dispersed than words with a grammatical function. Such models implicitly assume the bag-of-words setting. Usually the statistical significance of word frequencies is of no interest, because the task is not to find or study individual words that describe the documents but to rank documents according to their relevance to a given set of query words. Our problem setting, however, is very different and thus not directly comparable.

Our Approach. Comparing frequency counts over texts or corpora is an important task in many applications and scientific disciplines. The commonly used bag-of-words model, which can be described as a Bernoulli process with fixed rate, does not account for any structure present in natural languages. It can be easily shown that words have very different behavior in language, even at the word frequency level. In Figure 1, we illustrate the frequency histograms of the words *for* and *i* (lowercase *I*) in the British National Corpus [24]. These words are both very frequent, and approximately equally frequent. Yet, their frequency distribution is very different, thus employing the bag-of-words model in this example would be misleading.

Contextual behavior of words varies in language and is affected by several factors, such as genre, topic, author (gender, age, social class) etc. For example, in written language, especially in newspaper texts, there is avoidance of repeating a word, due to stylistic ideals, whereas in conversation, priming of words and syntactic structures plays an important role [10,23]. Hence, it is evident that natural language is non-homogeneous. There is great variance in word frequencies which depends on the specific word.

To model the natural behavior of words, we study their distribution throughout texts. The essential unit here is the interval between two occurrences of a word. We refer to this interval as the *inter-arrival time* between two instances. A recent study suggests that *inter-arrival times* in natural language can be modeled to a good accuracy using a Weibull distribution [1]. This parametric

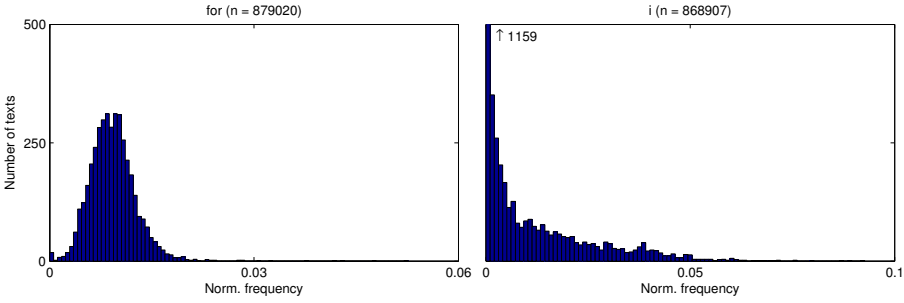

Fig. 1. Histogram of normalized frequencies vs. number of texts for the words *for* and i in the British National Corpus

distribution gives rise to a parameter β that can be interpreted as the *burstiness* of a word; we show this has a direct effect on the word frequency distribution. Bursty words tend to exhibit long inter-arrival times followed by short inter-arrival times, while the inter-arrival times for non-bursty words have smaller variance. The lower the burstiness parameter, the burstier the word: for example, $\beta_{for} = 0.93$ and $\beta_i = 0.57$.

Our Contributions. We propose two methods for modeling word frequencies that both take into account the behavior patterns of words. The first method is based on the inter-arrival time distribution of individual words. The second model is based on bootstrapping and takes into account only word frequency at the text level. We compare these methods to the current gold standard in a series of experiments on two large corpora: the British National Corpus (BNC) [24] and the San Francisco Call Newspaper Corpus (SFCNC) [17]. These corpora contain about 100 million and 63 million words, respectively. The experiments are based on comparing word frequencies over writing styles in the BNC and over time in SFCNC.

We show that taking the behavior of individual terms into account matters: in many cases it increases the frequency thresholds for the word to be reported as significantly frequent and therefore reduces the number of reported words. In addition, we find that the inter-arrival distribution can be used to give good predictions for the word-frequency distribution and that the inter-arrival and bootstrap methods give similar results.

2 Related Work

Word frequencies have been studied and analyzed in several domains. Research on graphs and networks has shown that many natural phenomena and patterns in human activity exhibit power-law behavior [3,7,16,18]. The discovery of power-law statistics occurred in the study of natural language; Zipf's law [26], relating the rank of words and their frequencies, describes the oldest known example of a power-law. It is surprising that for word frequencies in text documents, no such heavy-tailed modeling has been attempted.

The Bernoulli model has been widely used in modeling text in both data mining and linguistics. Dunning et al. [5] adopts the bag-of-words model to assess the statistical significance of word frequencies in text, assuming a Multinomial distribution, while Kleinberg [15] assumes multiple levels of frequency rates in text, where bursts of low frequencies may contain bursts of higher frequencies.

A significant amount of work has focused on detection of bursty structure in text, where bursty words are clustered to represent topics [8,12], or they are classified based on their frequency trajectories [11]. Additional work includes burstiness detection methods for query logs [25] or streams [13]. A burstiness-aware search framework has been introduced by Lappas et al. [17] which is fully non-parametric. All these methods, however, do not perform any significance assessment of word frequencies, thus they are orthogonal to the work presented in this paper.

Several effects of contextual behavior of words have been addressed in linguistics, such as text genre differences [4], word priming in conversation [10,23], differences in language use between males and females, age groups, and social classes [21]. Recent work by Altmann et al. [1] has shown that the distribution of successive occurrences of words can be modeled by the Weibull distribution, which is used in this paper.

3 Problem Setting

Let $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$ be a corpus, i.e., a set of n texts, defined over a lexicon $\Sigma = \{q_1, \ldots, q_{|\Sigma|}\}$. Each text $S_i = w_1 \ldots w_{|S_i|}$ is a sequence of words, with $w_j \in \Sigma$ for $j = 1, \ldots, |S_i|$.

The frequency $\text{freq}(q, S_i)$ of a word q in a document S_i is the number of occurrences of q in S_i; the frequency of a word q in a corpus $\mathcal{S}$ is the total number of occurrences of q in $\mathcal{S}$, i.e., $\text{freq}(q, \mathcal{S}) = \sum_{i=1}^{n} \text{freq}(q, S_i)$. The size $\text{size}(\mathcal{S})$ of a corpus $\mathcal{S}$, is the total number of words in it, which is the sum of the lengths of all texts, i.e., $\text{size}(\mathcal{S}) = \sum_{i=1}^{n} |S_i|$.

We focus on assessing the statistical significance of word frequencies in texts. Given a word q and a corpus $\mathcal{S}$, we would like to decide whether the frequency of q is significantly higher in $\mathcal{S}$ than in some given corpus $\mathcal{T}$ that conveys background knowledge on the word frequency distribution. For this purpose, we define an appropriate model for probability and use the one-tailed p-value:

$$p(q, \mathcal{S}, \mathcal{T}) = Pr\left(\frac{\text{freq}(q, \mathcal{S})}{\text{size}(\mathcal{S})} \leq \frac{\text{freq}(q, \mathcal{T})}{\text{size}(\mathcal{T})}\right). \tag{1}$$

We are interested in words for which this p-value is less than a user-defined significance threshold $\alpha \in [0, 1]$:

Definition 1 *(Dominant word). Given a word q, a corpus $\mathcal{S}$, a background corpus $\mathcal{T}$, a p-value function p, and a significance threshold α, q is a* dominant word *in $\mathcal{S}$ if and only if*

$$p(q, \mathcal{S}, \mathcal{T}) \leq \alpha. \tag{2}$$

We consider the following two problems:

Problem 1. *Given a word q and two corpora $\mathcal{S}$, $\mathcal{T}$, decide whether q is a dominant word in $\mathcal{S}$, given $\mathcal{T}$.*

For example, $\mathcal{S}$ can include articles written by male authors and $\mathcal{T}$ articles written by female authors. Given a word q, we would like to assess the significance of the frequency of q in $\mathcal{S}$ compared to the frequency in $\mathcal{T}$. In other words, we would like to determine whether q is used by males at a significantly higher rate than by females.

Problem 2. *Given two corpora $\mathcal{S}$ and $\mathcal{T}$, find the set of words $\mathcal{Q} \subseteq \Sigma$, such that each $q_j \in \mathcal{Q}$ is a dominant word in $\mathcal{S}$, given $\mathcal{T}$.*

For example, $\mathcal{S}$ may include newspaper articles written, e.g., in one year, while $\mathcal{T}$ may include newspapers written over some previous years. Our task then is to detect all dominant words for that year (set $\mathcal{S}$) compared to the previous years (set $\mathcal{T}$). Using this set of words we may infer the most important topics during that year and also observe gradual change of the language.

Note that we allow the case where both $\mathcal{S}$ and $\mathcal{T}$ contain only a single text. In the experiments in Section 5 we show that even when $\mathcal{S}$ consist of only one text, taking into account the structure of the language is meaningful and our approach gives results that differ substantially from the bag-of-words model.

4 Methods

In Section 4.1 we briefly discuss the baseline method, whereas in Section 4.2 we introduce the method based on inter-arrival times. This method comes in two flavors: fully non-parametric or using the Weibull distribution to model inter-arrivals. In Section 4.3 we introduce the bootstrapping method.

4.1 Method 1: Bernoulli Trials

A popular method for significance testing in frequency comparison is based on the assumption that a word occurs at any position in a text with equal probability. This setting is modeled by a repetition of Bernoulli trials, and the frequencies then follow a binomial distribution. The binomial distribution can be accurately approximated with the Poisson distribution for faster computation. Computational methods for these distributions are available in any modern statistical software package. This method serves as the baseline for comparison.

Let $p_{q,\mathcal{S}}$ denote the probability of observing q at any position in $\mathcal{S}$. The probability of observing q exactly k times after n trials is given by the probability mass function of the binomial distribution:

$$f(k; n, p_{q,\mathcal{S}}) = \binom{n}{k} {p_{q,\mathcal{S}}}^k (1 - p_{q,\mathcal{S}})^{n-k}. \tag{3}$$

Let $\hat{p}_{q,\mathcal{T}}$ denote the empirical probability of observing q at any position in $\mathcal{T}$: $\hat{p}_{q,\mathcal{T}} = \text{freq}(q, \mathcal{T})/\,\text{size}(\mathcal{T})$. Since the null hypothesis is that $p_{q,\mathcal{S}} = p_{q,\mathcal{T}}$, we can

use $\hat{p}_{q,\mathcal{T}}$ as an estimate for $p_{q,\mathcal{S}}$. The p-value for the Bernoulli model is then given by setting $n = \text{size}(\mathcal{S})$, $p = \hat{p}_{q,\mathcal{T}}$ and summing over $k \in [freq(q,\mathcal{S}), n]$:

$$p_1(q,\mathcal{S},\mathcal{T}) = \sum_{k=\text{freq}(q,\mathcal{S})}^{\text{size}(\mathcal{S})} \binom{\text{size}(\mathcal{S})}{k} \hat{p}_{q,\mathcal{T}}^{k} (1-\hat{p}_{q,\mathcal{T}})^{\text{size}(\mathcal{S})-k}. \tag{4}$$

Function $p_1(q,\mathcal{S},\mathcal{T})$ gives the one-tailed p-value of observing a frequency at least as high as $\text{freq}(q,\mathcal{S})$, given the size of $\mathcal{S}$ and the estimate $\hat{p}_{q,\mathcal{T}}$. Its value can be computed using the cumulative distribution function of Equation (4). The computational complexity of this approach is $\mathcal{O}(\text{size}(\mathcal{S}) + \text{size}(\mathcal{T}))$. For the remainder of this paper, this method will be denoted as `Bin`.

4.2 Method 2: Inter-arrival Times

This approach takes into account the natural behavior of words as expressed by inter-arrival times between words. The method is again based on computing a one-tailed p-value for observing a certain frequency or higher in $\mathcal{S}$, similar to the Bernoulli method. However, we do not assume that a word can occur at any position with fixed and equal probability. Instead, the probability of a word occurrence depends on the distance to the previous occurrence. Two null models are considered: the first is non-parametric and is based directly on the observed inter-arrival times, whereas the second is based on the Weibull distribution. First, we define inter-arrival times.

Inter-arrival Times. Let $\mathcal{S}$ be an ordered set of texts, which we concatenate to produce one long text $S = w_1 \dots w_{\text{size}(\mathcal{S})}$. For each word $q_i \in \Sigma$ with $n = freq(q_i,\mathcal{S})$, let $q_i^1, \dots, q_i^n$ denote the *positions* where q_i occurs in S, i.e., $q_i^l = j$ if and only if w_j is the l^{th} occurrence of q_i in S. The j-th *inter-arrival time* of word q_i, denoted as $a_{i,j}$, is given by

$$a_{i,j} = q_i^{j+1} - q_i^j \text{ , for } j = 1, \dots, n-1. \tag{5}$$

We take the inter-arrival time before the first occurrence of the word and after the last occurrence by considering $\mathcal{S}$ to be a ring. For simplicity, we define:

$$a_{i,n} = q_i^1 + |S| - q_i^n. \tag{6}$$

This ensures that the probability of observing the word is computed properly. Note there are as many inter-arrival times as words.

Empirical p-value. To obtain a null model and associated p-values, we use Monte Carlo simulation to create randomized texts, and compare the frequencies in the randomized texts against the observed frequency $\text{freq}(q,\mathcal{S})$.

Consider N random texts $\mathcal{R}_1, \dots, \mathcal{R}_N$, which have a size equal to $\mathcal{S}$: $\text{size}(\mathcal{R}_i) = \text{size}(\mathcal{S})$ for $i = 1, \dots, N$. We produce the random texts using a probability distribution for inter-arrival times learned from the background corpus $\mathcal{T}$. That

is, we construct a sequence of occurrences of word q by repeatedly sampling randomly from the set of inter-arrival times of q. We approximate the one-tailed p-value using the empirical p-value [19]:

$$\hat{p}_2(q, \mathcal{S}, \mathcal{T}) = \frac{1 + \sum_{i=1}^{N} I(\text{freq}(q, \mathcal{S}) \leq \text{freq}(q, \mathcal{R}_i))}{1 + N}, \tag{7}$$

where $I(\cdot)$ is the indicator function. We do not have to normalize the frequencies, since $\mathcal{S}$ and $\mathcal{R}_i$ are by definition of the same size.

Empirical Inter-arrival Time Distribution. The main step of the algorithm is to sample an inter-arrival time from the inter-arrival distribution, which we denote as $f(x)$. In the non-parametric case, we sample an inter-arrival time uniformly at random from the observed inter-arrival times, i.e., each observed inter-arrival time has equal chance of being chosen. This can be implemented by keeping a vector of inter-arrival times.

The first occurrence is treated separately. We can be at any point in any possible inter-arrival time at the beginning of the text. However, it is more likely we are at some point in a long inter-arrival than in a short one. To be precise, this is proportional to the length of the inter-arrival and thus we should sample uniformly from $g(x) = C \cdot x \cdot f(x)$, where C is a normalization constant such that $\sum_x C \cdot x \cdot f(x) = 1$. This gives us the current inter-arrival time, and within this inter-arrival time, any position is equally likely. Fast sampling from this distribution can be implemented by associating a normalized probability with each unique element in $f(x)$.

Random corpora produced using this Monte Carlo sampling procedure can be used to compute estimates for the one-tailed p-value. For the remainder of this paper, this method will be denoted as $\texttt{IA}_E$.

Weibull Inter-arrival Time Distribution. Recent work suggests that inter-arrival times between words can be modeled well using the Weibull (or stretched exponential) distribution [1]. It is shown that for almost any word the Weibull model fits the data much better than a Poisson distribution, as measured by the explained variance (R^2). Nonetheless, this recent study is mostly based on one data source: discussions on Google groups [1]. As far as we know, this result has not been validated by any other study. The comparison of the Monte Carlo simulation between the Weibull distribution and the empirical inter-arrival distribution will be the first evaluation of this result.

The probability density function for the Weibull distribution is given by

$$f(x) = \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta - 1} e^{-(x/\alpha)^{\beta}}, \tag{8}$$

where $\alpha, \beta > 0$ are the *scale* and the *shape* parameters, respectively. If $\beta = 1$, the Weibull distribution equals the Poisson distribution, and if $\beta \to 0$ it approaches a power-law and becomes heavy-tailed. We fit the parameters using the maximum-likelihood estimation. Implementations for fitting and sampling

for Weibull distributions are available in software packages for statistical analysis, such as R and Matlab.

The Monte Carlo simulation using the Weibull distribution is implemented as follows: sampling from $f(x)$ can be accomplished using standard statistical software, while the distribution $g(x) = C \cdot x \cdot f(x)$,

$$g(x) = C \cdot x \cdot f(x) = \frac{k}{\alpha \Gamma(1 + \frac{1}{\beta})} \left(\frac{x}{\alpha}\right)^{\beta} e^{-(x/\alpha)^{\beta}} \tag{9}$$

can be sampled by using the cumulative distribution function for $g(x)$,

$$G(x) = \int_0^a dx \cdot g(x) = 1 - \frac{\Gamma\left(1 + \frac{1}{\beta}, \left(\frac{a}{\alpha}\right)^{\beta}\right)}{\Gamma\left(1 + \frac{1}{\beta}\right)}. \tag{10}$$

Now, if we substitute $y = \left(\frac{a}{\alpha}\right)^{\beta}$, then $G(x)$ becomes the Gamma distribution with shape $k = 1 + 1/\beta$ and scale $\theta = 1$. We can sample a random number y from the Gamma distribution and obtain an inter-arrival time by $r = \lceil \alpha \cdot y^{1/\beta} \rceil$. The rounding is necessary because the Weibull distribution is continuous and > 0, while inter-arrival times are discrete and ≥ 1.

Again, using Equation (7) and the random samples obtained from this Monte Carlo simulation, we can compute an estimate for the one-tailed p-value. The computational complexity of this method is $\mathcal{O}(N \operatorname{size}(\mathcal{S}) + \operatorname{size}(\mathcal{T}))$ and the memory requirement is $\mathcal{O}(\operatorname{size}(\mathcal{T}))$. Since computations are done word-per-word, the memory cost can be reduced by storing only one vector of inter-arrival times at a time. For the remainder of this paper, this method will be denoted as $\texttt{IA}_W$.

4.3 Method 3: Bootstrapping

Instead of using inter-arrival times we can use a non-parametric approach to model the word frequency distribution directly.

Let $\mathcal{S}$ contain only one text, i.e., $\mathcal{S} = \{S\}$, and let $\mathcal{T}$ be a corpus with many texts. A straightforward approach to compute a p-value for the observed word frequency in $\mathcal{S}$ is to count the fraction of texts in $\mathcal{T}$ where the normalized frequency is larger. However, if $\mathcal{S}$ contains multiple texts, we would like to take into account heterogeneity between texts in both $\mathcal{S}$ and $\mathcal{T}$. We use bootstrapping [6] to approximate the p-value, although for this purpose also analytical estimates might be used.

The procedure is as follows: we take N random sets of texts $\mathcal{R}_1, \ldots, \mathcal{R}_N$, from the background corpus $\mathcal{T}$, each set having the same number of texts as $\mathcal{S}$: $|\mathcal{R}_i| = |\mathcal{S}|$. This leads to the problem that $\operatorname{size}(\mathcal{R}_i)$ is not necessarily equal to $\operatorname{size}(\mathcal{S})$, thus we should use normalized frequencies. We use the pooled frequency, divided by the pooled text size. Alternatively, one could use averages of frequencies that are normalized per text. Now, the empirical p-value (similar to Equation (7)) is

$$\hat{p}_3(q, \mathcal{S}, \mathcal{T}) = \frac{1 + \sum_{i=1}^{N} I\left(\frac{\operatorname{freq}(q,\mathcal{S})}{\operatorname{size}(\mathcal{S})} \leq \frac{\operatorname{freq}(q,\mathcal{R}_i)}{\operatorname{size}(\mathcal{R}_i)}\right)}{1 + N}. \tag{11}$$

The computational complexity of this method is $\mathcal{O}(N|\mathcal{S}| + \text{size}(\mathcal{S}) + \text{size}(\mathcal{T}))$ and the memory requirement is $\mathcal{O}(|\mathcal{T}|)$. For the remainder of this paper, this method will be denoted as `Boot`.

5 Experiments

The performance of our methods has been benchmarked on two large corpora:

The *British National Corpus* (BNC) [24] is the largest linguistically annotated corpus that is available in full-text format. It contains almost 100 million words of British English, spread over 4,049 texts, which are classified in text genres, such as fiction, academic prose, newspaper articles, transcribed conversation and more. The corpus is a result of careful digitization and has been annotated with meta information such as author gender, age, etc. and has been part-of-speech-tagged automatically with manual validation. We preprocess the data by removing all capitalization.

The *San Francisco Call Newspaper Corpus* (SFCNC) contains tokenized and stemmed newspaper articles published in the San Francisco Call, a daily newspaper, between 1900 and 1909, with stopwords removed. The SFCNC has been constructed and used by Lappas et al. [17]. The corpus consists of three periods:

- Period I: 110,387 articles published from 01/01/1900 to 31/12/1901.
- Period II: 133,684 articles published from 01/01/1903 to 31/12/1904.
- Period III: 129,732 articles published from 01/01/1908 to 31/12/1909.

The experiments are based on comparing word frequencies over writing styles in the BNC and over time in SFCNC. In Section 5.1, we present a simple proof of concept benchmark to show that taking into account individual behavior of words matters. We discuss the differences between male/female authors and four text-genres in the BNC in Sections 5.2 and 5.3. Significant language changes over time in the SFCNC are illustrated in Section 5.4 and the proposed methods also managed to detect dates of significant events.

5.1 BNC: A Simple Benchmark

We performed a simple benchmark on the BNC to show that *burstiness* matters when assessing the statistical significance of word frequencies. For simplicity, we used in this experiment a fixed text length of 2,000 words both for $\mathcal{S}$ and $\mathcal{T}$, which leaves us with 3,676 texts. We compared the significance thresholds for the most frequent words in the BNC and words with frequency just below 100,000. In detail, for each of the 30 words, we computed the word frequency that is required to make that word significant at the level $\alpha \leq 0.01$. Because the texts in the BNC are not ordered, we order them randomly.

In Table 1 we show the results of the proposed methods, $\texttt{IA}_\texttt{E}$, $\texttt{IA}_\texttt{W}$, and `Boot`, compared to `Bin`. Also, $\beta_{q,T}$ indicates the value of the shape parameter β of the Weibull distribution for each word q in $\mathcal{T}$. $\beta = 1$ corresponds to an exponential distribution, which we consider to be *non-bursty*. The lower the β the *burstier* the word. If $\beta > 1$, then the distribution is more regular than exponential, which we shall also consider to be *non-bursty*.

Table 1. Actual frequencies, parameters $\beta_{q,\mathcal{T}}$, and significance thresholds for the most frequent words in the BNC and words with frequency just below 100,000. Thresholds are computed for a text of length 2,000 words and $\alpha \leq 0.01$. $freq(q, \mathcal{T}$ is the frequency of the Word in the BNC. $\beta_{q,\mathcal{T}}$ is the burstiness of the word, given by the Weibull distribution. `Bin`, is the binomial model. $\texttt{IA}_E$ and $\texttt{IA}_W$ are the inter-arrival methods using empirical and Weibull distribution. `Boot` is the bootstrapping method. Largest differences occur when $\beta_{q,\mathcal{T}}$ is lowest.

Word	freq$(q,\mathcal{T})$	$\beta_{q,\mathcal{T}}$	Bin	$\texttt{IA}_E$	$\texttt{IA}_W$	Boot	Word	freq$(q,\mathcal{T})$	$\beta_{q,\mathcal{T}}$	Bin	$\texttt{IA}_E$	$\texttt{IA}_W$	Boot
the	6043900	1.10	149	152	143	197	how	99022	0.65	7	9	9	10
of	3043549	1.02	82	85	80	116	most	98051	0.77	7	8	8	7
and	2617864	1.08	72	72	70	95	back	96978	0.66	7	9	9	11
to	2594656	1.05	71	72	70	82	get	96000	0.60	7	10	10	20
a	2165365	1.01	61	63	61	72	way	95763	0.78	7	8	8	7
in	1938440	1.01	56	57	55	73	our	93272	0.53	7	11	10	17
that	1119422	0.87	35	40	38	69	down	92084	0.67	7	9	9	10
it	1054755	0.79	34	39	37	79	made	91383	0.80	7	8	7	8
is	990747	0.77	32	40	37	54	right	90533	0.57	7	10	9	38
was	881620	0.72	29	39	35	53	between	90519	0.70	7	8	8	8
for	879020	0.93	29	31	30	37	got	90165	0.51	7	12	12	20
i	868907	0.57	29	57	48	110	er	89845	0.43	7	28	26	54
's	784709	0.75	27	33	31	70	much	89842	0.79	7	7	8	7
on	729923	0.91	25	27	27	37	work	89344	0.61	7	9	9	11
you	667623	0.56	24	49	42	100	think	88665	0.56	7	11	10	17

The first observation we make concerns the six most frequent words (*the* – *in*), which have $\beta \geq 1.00$ and are thus non-bursty. The inter-arrival methods give similar frequency thresholds as the binomial model, although the bootstrapping method suggests that even for these words we should be more conservative in estimating p-values.

On the left side of the table are the words *for* and *i*, used previously in the example of Figure 1. The binomial model does not distinguish between the two words, while the three proposed methods do, by requiring a much higher frequency for *i* to be considered significant. The words on the right side of the table suggest there is difference between various words, regardless of frequency. Words such as *right* and *er*, but also *get*, *got*, and *think* are bursty and all three methods suggest we should assess the significance much more conservatively. Regarding the rest of the words in the table, we can conclude that both inter-arrival based methods perform similarly, with $\texttt{IA}_W$ consistently requiring slightly lower frequencies than $\texttt{IA}_E$. `Boot` often gives the highest threshold, but for few words (*most, way, much*) the results are similar to the binomial model.

5.2 BNC: Differences between Male and Female Authors

Next, we studied text variation between male and female authors in the BNC. For this experiment, we selected all fiction texts from BNC and split them into two groups: those written by males BNC_{male} and those written by females

Table 2. Number of dominant words for written fiction by male or female authors at various significance thresholds α. `Bin`, is the binomial model. $\texttt{IA}_E$ and $\texttt{IA}_W$ are the inter-arrival methods using empirical and Weibull distribution. `Boot` is the bootstrapping method. *Any* is the number of words reported as dominant by any of the methods. The inter-arrival and bootstrap methods show many of the words reported as significantly frequent by the binomial method are not significant. The inter-arrival method using Weibull distribution is most conservative.

Gender	α	`Bin`	$\texttt{IA}_E$	$\texttt{IA}_W$	`Boot`	*Any*	Gender	α	`Bin`	$\texttt{IA}_E$	$\texttt{IA}_W$	`Boot`	*Any*
Male	0.1%	183	133	110	119	185	Female	0.1%	202	147	123	131	202
Male	1.0%	264	210	182	186	266	Female	1.0%	290	222	195	210	290
Male	10.0%	417	375	359	366	417	Female	10.0%	470	420	400	405	471

BNC_{female}. We conducted two experiments: in the first we searched for dominant words in BNC_{male}, thus we set $\mathcal{S} = BNC_{male}$ and $\mathcal{T} = BNC_{female}$, and secondly we performed the reverse experiment. The performance of the proposed methods was compared to that of `Bin` for different significance thresholds α.

In Table 2, we can see the number of dominant words produced by each method for $\alpha = 0.1\%, 1.0\%, 10\%$. We also recorded the number of words detected as dominant by at least one of the methods, which is denoted as *Any* in the table. We can conclude that the number of dominant words detected by the three proposed methods are always less than those detected by `Bin` for both genders. For example, a significance threshold of 0.1%, the number of dominant words detected by `Bin` are approximately 1.7 times as many as those detected by $\texttt{IA}_W$, 1.4 times as many as those detected by $\texttt{IA}_E$ and 1.5 times as many as those detected by `Boot`. Also, $\texttt{IA}_W$ consistently detects the smallest number of dominant words. We also observed that dominant words detected by the proposed methods were nearly always flagged as dominant by `Bin`. Further investigation showed these words were reported by one of the inter-arrival methods and have p-values just above α for all other methods.

An overview of all p-values resulting from this experiment is given in Figure 2. The six displays compare all methods pairwise to each other. The in-sets enlarge the view at small p-values. We found that the inter-arrival time methods and `Boot` report *smoothed* p-values in many cases, i.e., p-values below 0.5 are higher and p-values above 0.5 become lower, in comparison to the binomial model. We find also that there is much agreement between $\texttt{IA}_E$ and `Boot`. The Weibull distribution appeared to give more variable results and larger differences compared to the binomial model than the other two methods. In general, the inter-arrival time method and `Boot` have greater agreement with each other than with the binomial model, as is clearly shown in the in-sets in all six figures.

5.3 BNC: Differences between the Main Genres

We studied text variation between the four main genres in BNC, i.e., *conversation*, *fiction prose*, *newspaper articles*, *academic prose*. Texts were split into four groups, one group per genre. Then for each genre, we set $\mathcal{S}$ to contain all

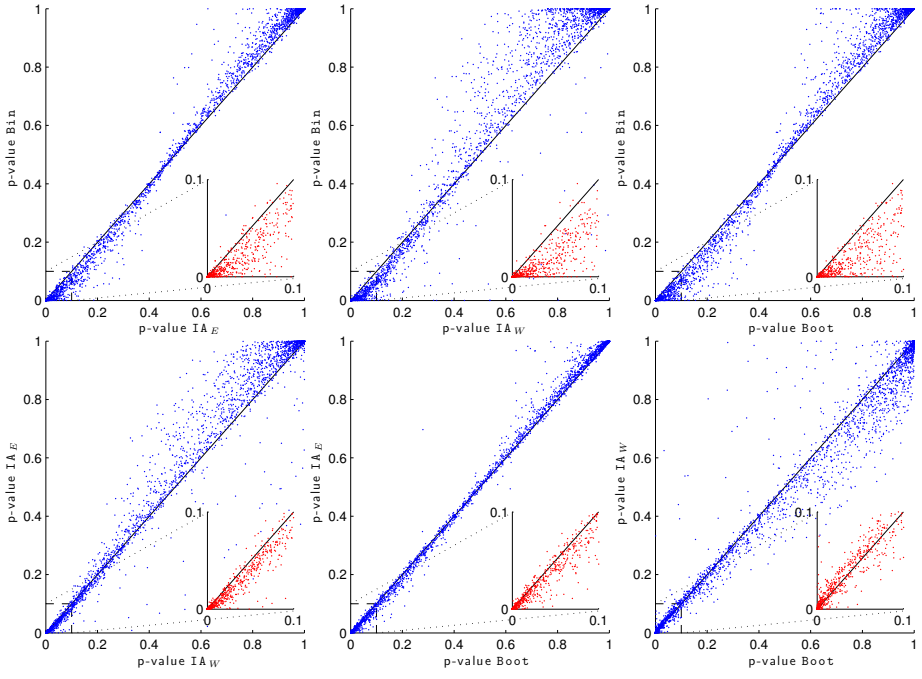

Fig. 2. Comparison of p-values between the four methods for male and female authors in the BNC. Each figure gives p-values from one method, against p-values in another method. Each point corresponds to a word. For explanation of labels see Table 2. The p-values from both experiments (male vs. female and vice versa) have been aggregated. The in-sets show more detail for the lower p-values < 0.1. We found that the binomial model gives very different results from all three other methods (top figures). The inter-arrival methods using empirical distribution and the bootstrap method show great agreement (bottom-centre figure). The inter-arrival method using Weibull distribution shows greater variance (bottom-right, bottom-left, and top-centre figure).

texts of that genre and $\mathcal{T}$ to contain the rest of the corpus. The performance of the proposed methods was compared to that of the binomial model for different significance thresholds α.

In Table 3, we can see the number of dominant words produced by each method and for each genre, for $\alpha = 0.1\%, 1.0\%, 10\%$. The behavior is the same as that observed for the male vs. female experiment. Again, we observed that nearly all dominant words detected by the proposed methods were also flagged as dominant by `Bin`. A figure illustrating the comparison of p-values is omitted due to space limitations. The results support the observations made in Figure 2.

5.4 SFCNC: Language Change over Time

We studied language variation between the three periods in SFCNC. For each period, we set $\mathcal{S}$ to contain all texts of that period and $\mathcal{T}$ to contain all texts from

Table 3. Number of words marked as dominant for each genre at various significance thresholds α. For explanation of labels see Table 2. The inter-arrival and bootstrap methods show many of the words reported as significantly frequent by the binomial method are not significant. The inter-arrival method using Weibull distribution and bootstrapping are most conservative. For *conversation* the differences between binomial and the other methods are smallest and for *news* they are greatest.

Genre	α	Bin	IA_E	IA_W	Boot	*Any*	Genre	α	Bin	IA_E	IA_W	Boot	*Any*
Conv	0.1%	381	328	308	314	381	News	0.1%	532	363	315	316	532
Conv	1.0%	412	384	363	367	412	News	1.0%	634	488	420	434	634
Conv	10.0%	473	453	447	446	474	News	10.0%	796	717	670	668	796
Fict	0.1%	505	388	339	352	507	Acad	0.1%	680	600	552	562	681
Fict	1.0%	573	496	446	464	573	Acad	1.0%	746	677	644	653	746
Fict	10.0%	682	629	619	610	682	Acad	10.0%	842	811	787	787	844

Table 4. Number of words marked as dominant for each news period at various significance thresholds α. For explanation of labels see Table 2. The inter-arrival and bootstrap methods show many of the words reported as significantly frequent by the binomial method are not significant. The inter-arrival method using Weibull distribution is most conservative. The differences between IA_E and Boot are small.

Period	α	Bin	IA_E	IA_W	Boot	*Any*	Period	α	Bin	IA_E	IA_W	Boot	*Any*
I	0.1%	141	73	50	73	141	II	10.0%	334	269	268	279	337
I	1.0%	229	144	113	134	231	III	0.1%	119	65	46	66	119
I	10.0%	423	339	346	340	428	III	1.0%	172	112	96	117	173
II	0.1%	113	41	19	46	113	III	10.0%	305	250	254	266	305
II	1.0%	182	99	74	98	182							

the other two periods. The performance of the proposed methods was compared to that of the binomial model for different significance thresholds α.

In Table 4, we can see the number of dominant words produced by each method and for each period, for $\alpha = 0.1\%, 1.0\%, 10\%$. The results at large are the same is in the experiment on the BNC. The differences between the proposed methods and the binomial model are even larger than before, especially at $\alpha = 0.1\%$. About half of the words marked as dominant by the binomial model, are false-positives according to the inter-arrival method using empirical distribution or bootstrap method. Using the Weibull distribution suggests even fewer truly significant words.

A full comparison of the p-values computed by all methods, aggregated over the three news periods, is shown in Figure 3. The in-sets show more detail for the lower p-values. Again, as in the BNC, the three proposed methods give higher p-values than Bin when $\alpha \leq 0.1$. In addition, for the same significance level, IA_W is clearly more conservative than the other methods. Also, the agreement between IA_E and Boot has decreased slightly, where IA_E gives slightly more conservative estimates.

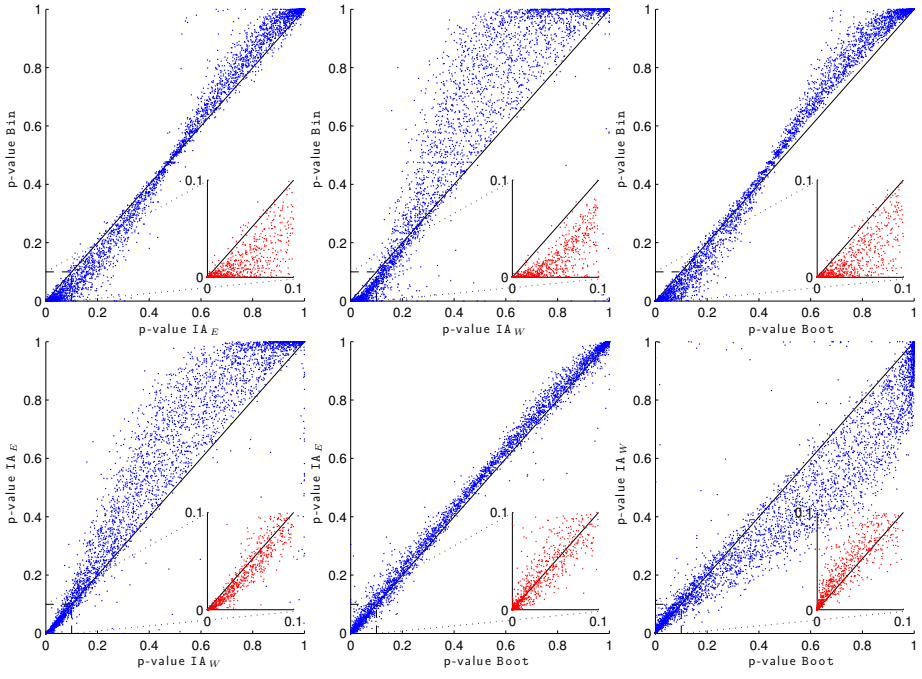

Fig. 3. Comparison of p-values between the four methods for the three periods in the SFCNC. Each figure gives p-values from one method, against p-values in another method. Each point corresponds to a word. For explanation of labels see Table 2. The p-values from all experiments have been aggregated. The in-sets show more detail for the lower p-values < 0.1. The figures confirm the findings of Figure 2. All three methods give more conservative p-values than binomial, and the pairwise differences between the inter-arrival time methods and the bootstrap method are similar to the genre experiment.

5.5 SFCNC: Locating Dates of Important Events

As a final test, we studied the intervals of word bursts reported in Lappas et al. [17]. These intervals correspond to bursts of some word after or around a significant historical event. We computed for each of the query words the days where this word is dominant, using $\alpha = 1\%$, and compare these to the corresponding intervals given by the search framework presented in their paper.

In Table 5 we find the results for one such query: *Jacksonville*. This interval (27 Apr–20 May) corresponds to the great fire at Jacksonville, Florida that occurred at May 3rd, 1901. We find that using any of the methods discussed in this paper find a shorter interval (5 May–8 May), and significant discussion one week later. The inter-arrival and bootstrap methods restrict the set of days even further. Due to lack of space the other results are omitted. In most cases the results were similar to the finding above, and for certain words, the intervals corresponded to those found in Lappas et al.

Table 5. Dates where the word *Jacksonville* occurs significantly frequent. `Lappas` is the method used in Lappas et al. [17]. `Bin`, is the binomial model. $\texttt{IA}_E$ and $\texttt{IA}_W$ are the inter-arrival methods using empirical and Weibull distribution. `Boot` is the bootstrapping method. An "x" corresponds to the word being dominant in the SFCNC at that day. All methods suggests stricter intervals than `Lappas` and the inter-arrival and boostrap methods flag the smallest sets of days.

	0427	0428	0429	0430	0501	0502	0503	0504	0505	0506	0507	0508	0509	0510	0511	0512	0513	0514	0515	0516	0517	0518	0519	0520
1901																								
`Boot`	.	.	.	.	.	.	.	.	x	x	.	x	.	.	.	.	.	x	.	.	x	.	.	.
$\texttt{IA}_W$	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.
$\texttt{IA}_E$	.	.	.	.	.	.	.	.	x	x	.	x	.	.	.	.	.	x	.	.	.	.	.	.
`Bin`	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	x	x	.	.	x	.	.	.
`Lappas`	x	.	x	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	.	.	x	x	x	x

6 Conclusion

Models based on the bag-of-words assumption have been prevalent in text analysis and have been proven to perform well in a wide variety of contexts. The bag-of-words assumption provides a good estimate of the expected number of word occurrences in text. However, the variance—or more generally, the shape of the word frequency distribution—is seriously misestimated. We have introduced a method for assessing the significance of word frequencies that is based on the inter-arrival times. The method can use either the empirical distribution or a parametric distribution such as Weibull. By comparing the sets of dominant words given by the binomial model, the inter-arrival based method and the bootstrap-based method, we have shown that any statistical significance test on word occurrences that is based on the bag-of-words assumption tends to overestimate the significance of the observed word frequencies and hence result to false positives. Thus, bag-of-words based methods should not be used to asses the significance of word frequencies. One should either use an empirical method such as the bootstrap model presented in the paper, or the inter-arrival time based method.

An interesting direction for future work is to use the idea of inter-arrival times instead of bag-of-words in other scenarios, such as information retrieval, and to study test statistics other than word frequencies, which could be based on inter-arrival times directly. Also, further research on parametric distributions for inter-arrival times of words is warranted by the significant differences in the experimental results between the empirical and Weibull distribution.

Acknowledgements. This work was supported by the Finnish Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN) and the Academy of Finland (Project 1129300). We thank Terttu Nevalainen, Tanja Säily, and Turo Vartiainen for their helpful comments and discussions.

References

1. Altmann, E.G., Pierrehumbert, J.B., Motter, A.E.: Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS ONE 4(11), e7678 (2009)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman, Amsterdam (1999)
3. Barabási, A.-L.: The origin of bursts and heavy tails in human dynamics. Nature 435, 207–211 (2005)
4. Biber, D.: Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press, Cambridge (1995)
5. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19, 61–74 (1993)
6. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC (1994)
7. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: ACM SIGCOMM, pp. 251–262 (1999)
8. Fung, G.P.C., Pui, G., Fung, C., Yu, J.X., Yu, P.S., Yu, S., Lu, H.: Parameter free bursty events detection in text streams. In: VLDB, pp. 181–192 (2005)
9. Gries, S.T.: Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. Corpus Linguistics and Linguistic Theory 12, 277–294 (2005)
10. Gries, S.T.: Syntactic priming: A corpus-based approach. Journal of Psycholinguistic Research 34(4), 365–399 (2005)
11. He, Q., Chang, K., Lim, E.-P.: Analyzing feature trajectories for event detection. In: ACM SIGIR, pp. 207–214 (2007)
12. He, Q., Chang, K., Lim, E.-P.: Using burstiness to improve clustering of topics in news streams. In: IEEE ICDM, pp. 493–498 (2007)
13. He, Q., Chang, K., Lim, E.-P., Zhang, J.: Bursty Feature Representation for Clustering Text Streams. In: SIAM SDM, pp. 491–496 (2007)
14. Kilgarriff, A.: Language is never, ever, ever, random. Corpus Linguistics and Linguistic Theory 1-2, 263–275 (2005)
15. Kleinberg, J.: Bursty and hierarchical structure in streams. DMKD 7, 373–397 (2003)
16. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. World Wide Web 8(2), 159–178 (2005)
17. Lappas, T., Arai, B., Platakis, M., Kotsakos, D., Gunopulos, D.: On burstiness-aware search for document sequences. In: ACM SIGKDD, pp. 477–486 (2009)
18. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM TKDD 1(1) (2007)
19. North, B.V., Curtis, D., Sham, P.C.: A note on the calculation of empirical p-values from Monte Carlo procedures. The American Journal of Human Genetics 71(2), 439–441 (2002)
20. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: 38th ACL Workshop on Comparing Corpora, pp. 1–6 (2000)
21. Rayson, P., Leech, G., Hodges, M.: Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. International Journal of Corpus Linguistics 2(1), 133–152 (1997)
22. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: ACM SIGIR, pp. 232–241 (1994)

23. Szmrecsanyi, B.: Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. Corpus Linguistics and Linguistic Theory 1(1), 113–149 (2005)
24. The British National Corpus, version 3, BNC XML edn. (2007)
25. Vlachos, M.: Identifying similarities, periodicities and bursts for online search queries. In: ACM SIGMOD, pp. 131–142 (2004)
26. Zipf, G.K.: Human behavior and the principle of least effort. Addison-Wesley, Reading (1949)