Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping

Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila
Aalto University, Finland

Problem setting
Given two collections of texts (corpora) S and T and a significance threshold α, find all words that are *significantly* more frequent in S than in T, that is with p-value $p \leq \alpha$.

Hypothesis
- Previously used methods are based on the bag-of-words model
  - This model ignores any structure in texts and corpora
- Figure 1 illustrates the effect of *burstiness*
- The distribution of bursty words is poorly predicted by the bag-of-words model

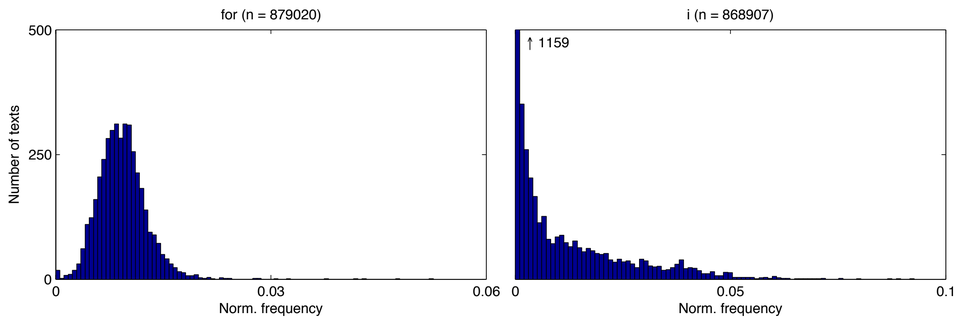

Figure 1: frequency histograms for the words *for* and *I* in the British National Corpus [1]. The total frequency of the words in the corpus is similar, but the frequency distributions are very different.

Solution
- We propose two novel methods that take into account the *burstiness* of words
- The first method is based on the distribution of *inter-arrival times*
  - An *inter-arrival time* is the number of words between two consecutive occurrences of a word
- The second method is based on bootstrapping of the frequency distribution

Conclusion
- The bag-of-words model underestimates the variance of the frequency distribution
- The two introduced methods give better estimates of this shape
- These tests provide more accurate results for computing the statistical significance of word frequencies

Results: A simple benchmark

Word	Freq (10^6)	β	*Bin*	*IA-W*	*Boot*
the	6.0	1.10	149	143	197
of	3.0	1.02	82	80	116
and	2.6	1.08	72	70	95
to	2.6	1.05	71	70	82
a	2.2	1.01	61	61	72
in	1.9	1.01	56	55	73
that	1.1	0.87	35	38	69
it	1.1	0.79	34	37	79
is	1.0	0.77	32	37	54
was	0.9	0.72	29	35	53
for	0.9	0.93	29	30	37
i	0.9	0.57	29	48	110
‘s	0.8	0.75	27	31	70
on	0.7	0.91	25	27	37
you	0.7	0.56	24	42	100

Table 1: Frequency thresholds for a text with 2,000 words and $\alpha = 0.01$, using the British National Corpus [1] as corpus T. We found there is a clear correspondence between the burstiness β and the given significance thresholds for the inter-arrival time and bootstrap test. For bursty words ($\beta < 1.00$) there is a large difference between the previously used binomial test and the proposed methods. Bootstrapping always gives the highest, most conservative threshold.

Results: Comparing p-values between all methods

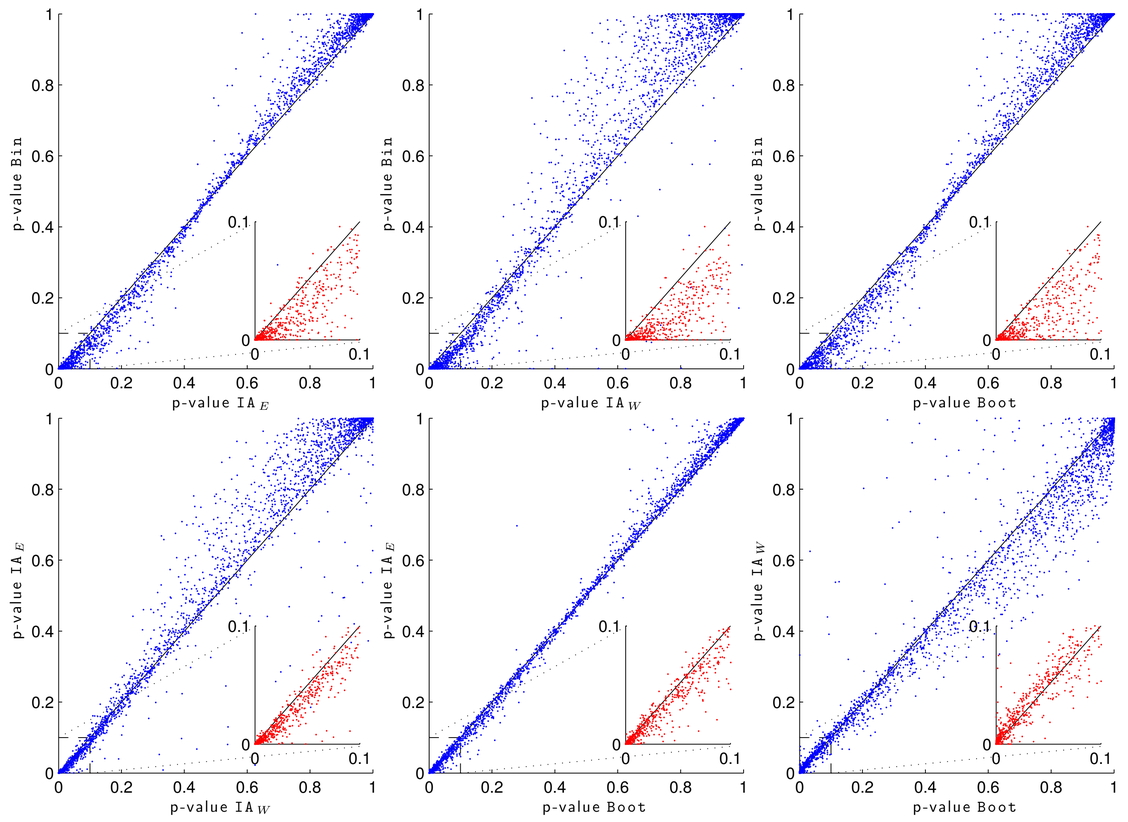

Figure 2: Comparison of p-values between the four methods for male and female authors of fiction texts in the British National Corpus [1]. Each figure gives p-values from one method, against p-values in another method. Each point corresponds to a word. The in-sets show more detail for p-values < 0.1. We found that the binomial test gives very different results from all three other methods (top figures). The inter-arrival test using empirical distribution and the bootstrap test show great agreement (bottom-centre figure). The inter-arrival test using Weibull distribution shows greater variance (bottom-right, bottom-left and top-centre figures). More details and results for the San Francisco Call Newspaper Corpus [2] can be found in the paper.

Input
- Two sets of texts: S and T
- Each text is a sequence of words

Binomial test
- An exact test, similar to χ^2-test and log-likelihood ratio test [2]
- Based on assumption that all words in a corpus are independent
- $p = \sum_{k=freq(q,S)}^{size(S)} \binom{size(S)}{k} p_{q,T}{}^{k} (1 - p_{q,T})^{size(S)-k}$
- $p_{q,T}$ is the probability of word q in corpus T

Burstiness and inter-arrival times
- An *inter-arrival time* is the number of words between two consecutive occurrences of a word
- The distribution can be described well by a Weibull distribution [3]
- The shape parameter of the distribution β, can be interpreted as the *burstiness* of a word
- $\beta = 1$ corresponds to an exponential distribution
- The lower the β, the burstier the word

Inter-arrival time test
- Based on the distribution of *inter-arrival times*
- The significance test is done by creating N random corpora R_1 to R_N, using the inter-arrival time distribution learned from corpus T
- We use the empirical p-value:
- $\hat{p} = \frac{1 + \sum_{i=1}^{N} I\left(freq(q,S) \leq freq(q,R_i)\right)}{1 + N}$
- The inter-arrival time test can use the empirical distribution, or a parametric distribution, such as Weibull

Bootstrap test
- Based on the word frequency distribution
- Create N random corpora by repeatedly sampling $|S|$ texts from corpus T
- Again the empirical p-value:
- $\hat{p} = \frac{1 + \sum_{i=1}^{N} I\left(\frac{freq(q,S)}{size(S)} \leq \frac{freq(q,R_i)}{size(R_i)}\right)}{1 + N}$

Session: Text Mining & Recommender Systems
Thursday September 8, 14:00-15:50 Attica Hall

References
[1] The British National Corpus, version 3 (BNC XML Edition), 2007. http://www.natcorp.ox.ac.uk/
[2] T. Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19:61–74, 1993.
[3] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS ONE, 4(11):e7678, 2009.
[4] T. Lappas, B. Arai, M. Platakis, D. Kotsakos, and D. Gunopulos. On burstiness-aware search for document sequences. In ACM SIGKDD, pages 477–486, 2009.

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.