Physica A 445 (2016) 328–334

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

On variation of word frequencies in Russian literary texts

Vladislav Kargin

Binghamton University, United States

HIGHLIGHTS

- We examine a large online library of Russian literary texts.
- The variation in the word frequencies across texts is related to the average word frequency by a non-linear power law.
- The finding is consistent with "burstiness" (increased relative variation) of rare words.
- A latent Dirichlet allocation (LDA) model is estimated.
- The non-linearity result can be explained by asymmetry in the distribution of latent factors.

ARTICLE INFO

Article history:
Received 22 June 2015
Received in revised form 4 October 2015
Available online 19 November 2015

Keywords:
Burstiness
Word frequency variation
Latent Dirichlet allocation

ABSTRACT

We study the variation of word frequencies in Russian literary texts. Our findings indicate that the standard deviation of a word's frequency across texts depends on its average frequency according to a power law with exponent $\frac{1}{2} < \alpha < 1$, which shows that the rarer words have a relatively larger degree of frequency volatility (that is, higher "burstiness").

A latent factor model has been estimated to investigate the structure of the word frequency distribution. The findings suggest that the dependence of a word's frequency volatility on its average frequency can be explained by the asymmetry in the distribution of latent factors.

1. Introduction

The study of word frequency variation in different texts arose first in the problem of author attribution [1–3]. Recently, the explosive growth in the computing power and in the text data volume led to many new applications. For example, the text indexing problem asks to associate documents with queries for fast retrieval; the authorship profiling problem requires to describe features of the author (sex, age, religious and political beliefs, etc.) based on texts that the author produced. In addition, the classic authorship attribution problem found new applications in security and forensics (see surveys by Holmes [4], Juola [5], Koppel, Schler, and Argamon [6] and Stamatatos [7]).

For all these applications, the fundamental statistical issue is the distribution of word frequencies[1] in different texts. For example, if a word in a query has its frequency in a document higher than its average frequency, then this document can be regarded as more relevant to the query.

Some properties of the word frequency distribution were noticed a long time ago. For example, Zipf's law [1] describes the distribution of word frequencies in a particular text, and Heaps' law (p. 207 in Ref. [8], p. 75 in Ref. [9]) relates the number of distinct words in a text to its length. Some new research on these laws was done in Refs. [10–12], and [13]. See also surveys

E-mail address: vladislav.kargin@gmail.com.

[1] In this paper we use the term "frequency" as usual in statistics, that is, the number of the word occurrences in a document divided by the document's total number of words.

http://dx.doi.org/10.1016/j.physa.2015.11.014
0378-4371/

in Refs. [14,15]. This paper focuses on a different set of properties and investigates the variation of word frequencies across documents.

One has to understand the structure of the word-document frequency matrix for applications in the information retrieval, in order to handle the problems of word synonymity and polysemy. For this purpose, there have been recently developed tools such as LSA ("latent semantic analysis", Deerwester et al. [16]), pLSA ("probabilistic latent semantic analysis", Hofmann [17]), and LDA ("latent Dirichlet allocation", Blei, Ng, and Jordan [18]). The main idea of these methods is the dimension reduction. The variation of word frequencies across texts is assumed to stem mainly from the variation in relatively small amount of factors (or "topics") across texts.

The goal of this study is to establish basic facts about the fluctuations of word frequencies across documents such as the dependence of the fluctuation size on the average word frequency. In order to clarify this dependence, we will apply a latent factor technique, the LDA.

The paper is organized as follows. First, in Section 2 we describe the data. Then, in Section 3 we study how the size of frequency fluctuations across texts depends on the word's average frequency. Next, in Section 4 we apply a latent factor model to analyze the variation of vocabulary across texts in more detail. Finally, Section 5 concludes.

2. A preliminary look at the data

We use data from Flibusta, a Russian online library. It covers Russian and translated fiction works from many historical periods and literary genres. The data is freely available either via the torrent network or by an automatic download. To the author's best knowledge, it have not been used previously for linguistic research.

Currently, it has between 200,000 and 300,000 texts by about 85,000 authors, where the author is understood to include translators and sometimes organizations that published a particular text. Our analysis uses only a part of this dataset (around 25,000 books). In particular, we use only books which are available in a text format (more precisely, in the "FB2" book format) and we exclude the documents that are available only as pdf, djvu, doc, and other binary formats.

The library works using the wiki principle and the texts are uploaded by users, therefore the number of texts depends both on how many texts were written by the author and on how many of them were uploaded by users.

To illustrate the content of the library, the two authors with the largest number of texts are the American and Russian science fiction writers Ray Bradbury and Kir Bulychev, with 550 and 540 texts, respectively. Many of the other top authors are authors and translators of books in popular genres such as science fiction, mystery, romance, action, historical fiction, sensational and how-to literature.

If the authors working in the genres associated with popular culture are excluded, then we find many well-known classic authors, most of whom are short story writers. To illustrate, for the first 25 of these authors the number of texts in the online library ranges from 446 for Anton Chekhov to 144 for Franz Kafka.

3. Variation of word frequencies across texts

In this section, as a first step we establish that there is significant variation in word frequencies across different texts. Then we connect the size of the variation to the average frequency of the word in a given text. We find a power function dependence between these two quantities.

Let $\xi_{b,w}^{(t)}$ be an indicator variable which equals 1 if the word at place t in book b equals w. Then, the frequency of word w in book b can be written as

$$x_{b,w} = \frac{1}{T_b}\sum_{t=1}^{T_b}\xi_{b,w}^{(t)}, \tag{1}$$

where T_b is the length of the book b.

Suppose that for a given w the random variables $\xi_{b,w}^{(t)}$ are independent and identically distributed with the expectation parameter p_w, which does not depend on b. Then $\mathbb{E}x_{b,w} = p_w$, and

$$\mathbb{V}\left(x_{b,w}\right) = \frac{p_w(1-p_w)}{T_b}. \tag{2}$$

To test this hypothesis, we estimate p_w by using the whole sample:

$$\widehat{p}_w := \frac{1}{T}\sum_{b=1}^{B}\sum_{t=1}^{T_b}\xi_{b,w}^{(t)}, \tag{3}$$

where T is the total number of words in the data and B is the number of texts. Then we compute the normalized variance of $x_{b,w}$ across books.

$$V_w = \frac{1}{\widehat{p}_w(1-\widehat{p}_w)}\frac{1}{B}\sum_{b=1}^{B}T_b(x_{b,w}-\widehat{p}_w)^2. \tag{4}$$

This statistic should be compared with 1.