DOI 10.32460/ishmu-2021-9-0030
УДК 004.415:81-139
ББК 81.11

Программа для вычисления статистических индексов, используемых в стилевой классификации русскоязычных текстов

Филимонов В. В.[1], старший преподаватель,
Живодеров А. А.[2], канд. физ.-мат. наук, ст. науч. сотр.,
Терентьев А. С.[1], студент,
Дерябина Е. И.[1], студент,
[1]УрФУ, Екатеринбург
[2]ЦНБ УрО РАН, Екатеринбург

Ключевые слова: программа, статистические индексы, доверительные интервалы

В статье описывается программа PROC, предназначенная для вычисления различных статистических индексов, статистического анализа полученных значений в больших массивах текстов.

A program for calculating statistical indexes used in the stylistic classification of Russian-language texts

Filimonov V. V.[1], senior lecturer,
Zhivodyorov A. A.[2], PhD in Physico-Mathematical sciences, senior researcher,
Terentev A. S.[1], student,
Deryabina E. I.[1], student,
[1]UrFU, Ekaterinburg
[2]CSL UB RAS, Ekaterinburg

Keywords: program, statistical indices, confidence intervals

The article describes the PRO program designed for calculating various statistical indices in the statistical analysis of the obtained estimates in large arrays of texts.

Введение

С каждым годом объем текстовой информации растет и увеличивается потребность в эффективных инструментах для обработки и анализа документов, поскольку продуктивность выполняемой работы тесно связана с используемыми средствами.

Одним из современных инструментов является информационная система Corpus, разработанная сотрудником ЦНБ УрО РАН Горбичем Л. Г., которая позволяет рассчитывать статистические индексы, такие как: индекс триграмм, индекс биграмм, индексы сжимаемости текстов, информационные индексы (функция соотношения порядка и хаоса в системе и «функция развития»).

Индекс триграмм и биграмм: значение статистики χ^2 для троек и пар букв соответственно. В различных исследованиях использовались тройки и двойки, как всех букв алфавита, так и только гласных. Методика вычисления для биграмма и триграмма описана в работах [1-3].

Индексы сжимаемости текстов: отношение размера файла в байтах после сжатия стандартными алгоритмами к исходному размеру файла.

Информационные индексы были предложены старшим научным сотрудником ЦНБ УрО РАН В. Б. Вяткиным в работах [4-6], как R и D функции, они рассчитываются, как отношение и произведение аддитивной синтропии (I∑) кэнтропии отражения (S)

$$Rf = \frac{I_\Sigma}{S}, \; Df = I_\Sigma S \quad \text{, где} \qquad (1)$$

$$I_\Sigma = \sum_{i=1}^{N} \frac{m_i}{M} log_2 m_i \qquad (2)$$

$$S = -\sum_{i=1}^{N} \frac{m_i}{M} log_2 \frac{m_i}{M} \qquad (3)$$

M – общее количество элементов в составе системы, N – число частей системы, mi – количество элементов в i-й части.

Программа ЦНБ УрО РАН была написана для работы с конкретным текстом, и в случае, когда необходимо исследовать статистически значимую выборку, работа с программой требует значительного времени, так

как большую часть операций необходимо проводить вручную. В связи с этим появилась задача разработки инструмента для обработки больших корпусов текстов. Что и было сделано, результаты работы представлены в настоящей статье.

Была написана программа PROC, которая рассчитывает любые индексы, не только те, что были описаны выше, а также ряд статистических индексов, описанных в статьях В. В. Филимонова, А. А. Живодерова [7; 8]. Кроме самих индексов была поставлена задача расчета доверительных интервалов и центральных значений доверительных интервалов индексов для неограниченного количества текстов без долговременной необходимости проводить через инструмент каждый документ лично.

Программа для вычисления статистических индексов PROC:

Программа написана на языке программирования Golang, это относительно молодой язык, набирающий популярность, благодаря своей простоте. Он прекрасно сочетает в себе лаконичность и хорошую производительность, а это позволяет создавать высоконагруженные приложения в меньшие сроки. Также нужно отметить понятный синтаксис, что облегчает работу с этим языком.

Требования к входным данным таковы, что на вход должны подаваться текстовые файлы в формате txt, содержащие текст, написанный на русском языке. Кодировка UTF-8.

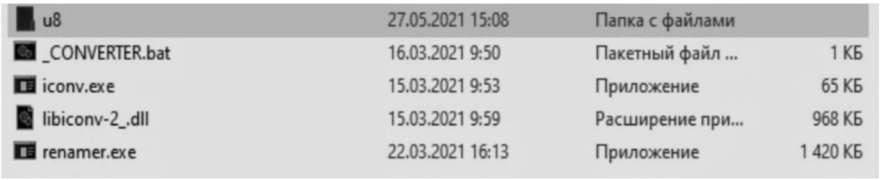
Работа с программой PROC

Для начала нужно убедиться, что все документы в нужной кодировке. Для этого файлы в формате txt конвертируются в UTF-8 конвертором, но прежде необходимо занести нужные файлы в папку с программой (рис. 1, 2).

После запускается конвертер, программа меняет кодировку у файлов, отличных от UTF-8.

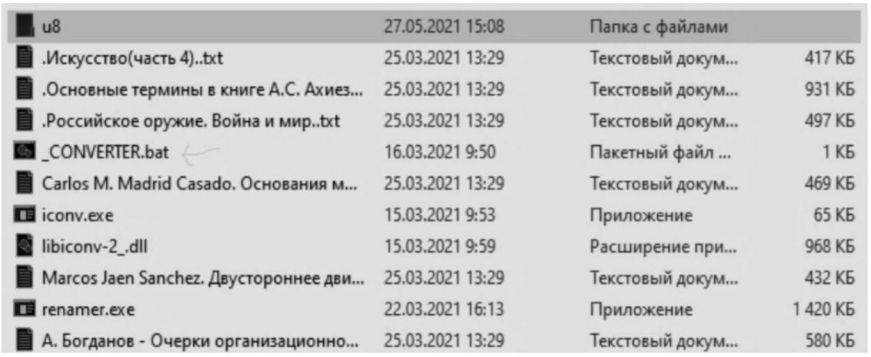
Если файл уже находится в нужной кодировке, программа его удалит, поэтому необходимо занести документ в папку с PROC вручную. Этот список программа выводит на экран (рис. 3).

После всех необходимых преобразований можно перейти к расчету параметров. Для этого запускаем PROC. Открывается небольшое окно программы, где автоматически рассчитываются индексы, и программа сохраняет csv файл со списком текстов и результатов расчетов по ним. Итоговый файл называется PROC_RESULTS.csv и сохраняется в той же папке, где находится сама программа.



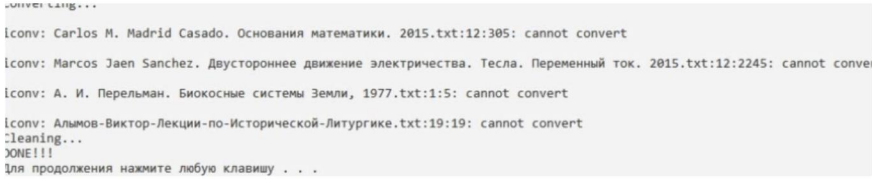
u8	27.05.2021 15:08	Папка с файлами	
_CONVERTER.bat	16.03.2021 9:50	Пакетный файл ...	1 КБ
iconv.exe	15.03.2021 9:53	Приложение	65 КБ
libiconv-2_.dll	15.03.2021 9:59	Расширение при...	968 КБ
renamer.exe	22.03.2021 16:13	Приложение	1 420 КБ

Рис. 1. Конвертер



u8	27.05.2021 15:08	Папка с файлами	
.Искусство(часть 4)..txt	25.03.2021 13:29	Текстовый докум...	417 КБ
.Основные термины в книге А.С. Ахиез...	25.03.2021 13:29	Текстовый докум...	931 КБ
.Российское оружие. Война и мир..txt	25.03.2021 13:29	Текстовый докум...	497 КБ
_CONVERTER.bat	16.03.2021 9:50	Пакетный файл ...	1 КБ
Carlos M. Madrid Casado. Основания м...	25.03.2021 13:29	Текстовый докум...	469 КБ
iconv.exe	15.03.2021 9:53	Приложение	65 КБ
libiconv-2_.dll	15.03.2021 9:59	Расширение при...	968 КБ
Marcos Jaen Sanchez. Двустороннее дви...	25.03.2021 13:29	Текстовый докум...	432 КБ
renamer.exe	22.03.2021 16:13	Приложение	1 420 КБ
А. Богданов - Очерки организационно...	25.03.2021 13:29	Текстовый докум...	580 КБ

Рис. 2. Конвертер с нужными файлами



```
onverting...

iconv: Carlos M. Madrid Casado. Основания математики. 2015.txt:12:305: cannot convert

iconv: Marcos Jaen Sanchez. Двустороннее движение электричества. Тесла. Переменный ток. 2015.txt:12:2245: cannot conve

iconv: А. И. Перельман. Биокосные системы Земли, 1977.txt:1:5: cannot convert

iconv: Алымов-Виктор-Лекции-по-Исторической-Литургике.txt:19:19: cannot convert
Cleaning...
DONE!!!
Для продолжения нажмите любую клавишу . . .
```

Рис. 3. Результат работы конвертера

В качестве примера приведем расчет ряда статистических индексов, а также границ доверительных интервалов и центральных значений индексов.

В одном из наших исследований, результаты которого находятся в стадии публикации была поставлена задача рассчитать значения, доверительные интервалы и центры для индексов триграмм, биграмм, их отношения, индекса сжимаемости текстов, и информационных индексов: соотношения порядка и хаоса в системе, и «функции развития».

В качестве материала для исследования было подобрано 1077 текстов на русском языке, из них административных – 112 (АД), научно-популярных – 75 (НП), поэзии – 139 (ПО), прозы – 387 (ПР), публицистики – 232 (ПУ), религиозных – 97 (РЕ), социально-политических – 35 (СП).

Научные тексты не были рассмотрены, потому что им посвящено отдельное исследование.

В таблице 1 приведены результаты вычислений центров индексов, в таблице 2 - границ интервалов, выполненные с помощью программы PROC:

Таблица 1

Центры индексов

Жанр	Центр χ^2	Центр Deflate	ЦетрRf	Центр Df
АД	4,472	0,1556	1,0092	0,01818
НП	3,122	0,2786	1,0555	0,01923
ПР	3,013	0,2962	1,0711	0,01971
ПО	3,91	0,2995	1,1067	0,02055
ПУ	3,38	0,2877	1,0064	0,01854
РЕ	3,41	0,2685	1,0164	0,01877
СП	3,32	0,2723	1,0074	0,01863

Таблица 2

Границы интервалов индексов

Жанр	Границы χ^2	Границы Deflate	Границы Df	Границы Rf
АД	4.28775; 4.65552	0.14917; 0.16212	0.01792; 0.01844	0.82170; 1.19672
НП	2.86739; 3.37754	0.27411; 0.28316	0.01905; 0.01941	1.04132; 1.06984
ПР	2.87604; 3.07859	0.29487; 0.29797	0.01959; 0.01973	1.06296; 1.07374
ПО	3.74559; 4.07459	0.29498; 0.30400	0.02017; 0.02093	1.02129; 1.19223
ПУ	3.26621; 3.49529	0.28396; 0.29154	0.01849; 0.01859	1.00350; 1.00921
РЕ	3.23865; 3.58746	0.26338; 0.27356	0.01863; 0.01891	1.00089; 1.03192
СП	3.01556; 3.61780	0.26503; 0.27964	0.01847; 0.01878	0.99658; 1.01828

По результатам вычисления в среде MATLAB были построены визуальные трехмерные модели центров и доверительных интервалов. По осям отложены значения отношения хи2 для триграмм к хи2 для биграмм, R-функции и индекса сжимаемости текстов (рис. 4). Цвет элементов выбран таким только для наглядности и не несет смысловой нагрузки.

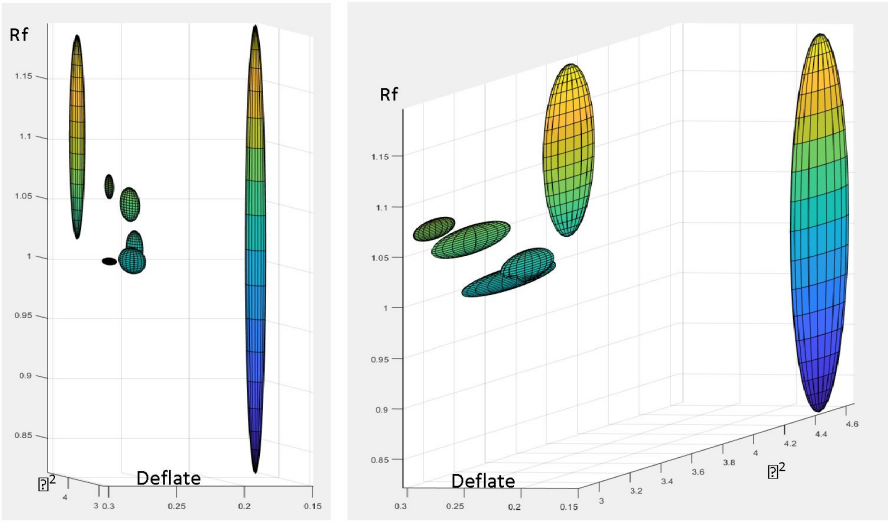


Рис. 4. Визуальные трехмерные модели центров и доверительных интервалов индексов

Заключение

Была написана программа PROG, позволяющая вычислять значения статистических индексов, а также доверительные интервалы и центральные значения указанных индексов для различных жанров текстов на русском языке.

Список источников

1. Филимонов В. В. Экспрессия и упорядоченность в письменной речи / В. В. Филимонов, А. А. Живодеров, Л. Г. Горбич // Известия Уральского Федерального Университета. Серия 1. Проблемы образования, науки и культуры. 2012. № 3(104) С. 313-319.

2. Filimonov V. V. Clustering of Russian-language texts using χ^2 statistics / V.V. Filimonov, A. M. Amieva, A. P. Sergeev // Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12-13, 2016). Information: transmission, processing, perception. Ekaterinburg : UrFU named after the first President of Russia B. N. Yeltsin, 2016. P. 164-174.

3. Горбич Л. Г. Опыт различения поэтических и прозаических текстов на основе сравнения распределений биграмм гласных букв / Л. Г. Горбич, В. В. Филимонов, А. А. Живодеров // Количественные методы в искусствознании: сборник материалов конференции. Екатеринбург, 2013. С. 163-166.

4. Вяткин В. Б. Хаос и порядок дискретных систем в свете синергетической теории информации // Научный журнал КубГАУ. 2009. № 47. URL: http://ej.kubagro.ru/2009/03/pdf/08.pdf (дата обращения: 10.06.2021).

5. Vyatkin V. B. A synergetic theory of information / V. B. Vyatkin. DOI: 10.3390/info10040142 // Information. 2019. V. 10, № 4. P. 142.

6. Vyatkin V. B. "Characteristic text length" in Informatics: problems, methodology, technology / V. B. Vyatkin // Characteristic text length. Proc. XIV Intern. Sci. and Method. Conf. of Informatics: Problems, Methodology, Technology. 2014. V. 1.P. 263-266.

7. Application of the random walk model for describing Russian-language texts / A. A. Kramarenko, V. V. Filimonov, A. A. Zhivodyorov, A. M. Amieva // Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12-13, 2017). Information: transmission, processing, perception. Ekaterinburg : UrFU named after the first President of Russia B.N. Yeltsin, 2017. P. 138-164.

8. Attribution of Russian-language texts using the law of large numbers / V. V. Filimonov, A. M. Amieva, A. A. Zhivodyorov, A. A. Kramarenko // Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12-13, 2017). Information: transmission, processing, perception. Ekaterinburg : UrFU named after the first President of Russia B.N. Yeltsin. 2017. P. 10-18.

References

1. Filimonov V. V., Zhivoderov A. A., Gorbich L. G. Ekspressiya i uporyadochennost' v pis'mennoj rechi [Expression and orderliness in written speech]. Izvestiya Ural'skogo Federal'nogo Universiteta. Seriya 1. Problemy obrazovaniya, nauki i kul'tury, 2012, no. 3(104), pp. 313-319. (In Russ.).

2. Filimonov V. V., Amieva A. M., Sergeev A. P. Clustering of Russian-language texts using χ2 statistics. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2016). Information: transmission, processing, perception. Ekaterinburg: UrFU named after the first President of Russia B.N. Yeltsin, 2016, pp. 164–174.

3. Gorbich L. G., Filimonov V. V., Zhivoderov A. A. Opyt razlicheniya poeticheskih i prozaicheskih tekstov na osnove sravneniya raspredelenij bigramm glasnyh bukv [The experience of distinguishing poetic and prose texts based on comparing the distributions of bigrams of vowel letters]. Kolichestvennye metody v iskusstvoznanii: sbornik materialov konferencii. Ekaterinburg, 2013, pp. 163-166. (In Russ.).

4. Vyatkin V. B. Haos i poryadok diskretnyh sistem v svete sinergeticheskoj teorii informacii [Chaos and order of discrete systems in the light of synergetic information theory]. Nauchnyj zhurnal KubGAU, 2009, no. 47. URL: http://ej.kubagro.ru/2009/03/pdf/08.pdf. (In Russ.).

5. Vyatkin V. B. A synergetic theory of information. DOI: 10.3390/info10040142. Information, 2019, vol. 10, no. 4. p. 142.

6. Vyatkin V. B. "Characteristic text length" in Informatics: problems, methodology, technology. Characteristic text length. Proceedings of the XIV International Scientific and Methodological Conference. Voronezh: VSU, 2014, vol. 1, pp. 263-266.

7. Kramarenko A. A., Filimonov V. V., Zhivodyorov A. A., Amieva A. M. Application of the random walk model for describing Russian-language texts. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2017). Information: transmission, processing, perception. Ekaterinburg: UrFU named after the first President of Russia B.N. Yeltsin, 2017, pp. 138–164.

8. Filimonov V. V., Amieva A. M., Zhivodyorov A. A., Kramarenko A. A. Attribution of Russian-language texts using the law of large numbers. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2017). Information: transmission, processing, perception. Ekaterinburg: UrFU named after the first President of Russia B.N. Yeltsin, 2017, pp. 10–18.