АЛГОРИТМ АНАЛИЗА СРЕДНЕЙ ДЛИНЫ СЛОВА В СЛОВАРЕ В.И. ДАЛЯ

Кузнецов Е. В., Приходько Н. Н., Чехомов Е. Г., Косарева Е. М.
Факультет компьютерного проектирования, Кафедра ПИКС, БГУИР
Минск, Республика Беларусь
E-mail: {yauheni.kuznetsou, nikitaprix50, genchek1, kksrvvv }@gmail.com

В работе представлен алгоритм статистического анализа лексического состава русского языка на основе Толкового словаря живого великорусского языка В. И. Даля. С использованием языка программирования Python и библиотек pandas, matplotlib реализована обработка текстовых данных словаря, включающая извлечение словарных единиц, вычисление средней длины слова и построение распределения длин слов. Результаты анализа показывают, что средняя длина слова в словаре Даля составляет 8,49 символа, что превышает средние показатели для современного русского языка. Исследование демонстрирует применение методов Data Science для лингвистического анализа.

Введение

Толковый словарь живого великорусского языка В. И. Даля – один из крупнейших словарей русского языка, содержащий около 200 тысяч слов и 30 тысяч пословиц, поговорок и загадок [1, 2]. Словарь был создан в середине XIX века и отражает состояние русского языка того периода, включая народную речь, областные диалекты и профессиональную лексику [3].

Статистический анализ лексического состава языка позволяет выявить закономерности в структуре словарного запаса и провести сравнительные исследования [4]. Современные методы обработки данных с использованием языка программирования Python и специализированных библиотек открывают новые возможности для количественного анализа лингвистических материалов.

Цель данного исследования – разработка и реализация алгоритма анализа средней длины слова в словаре Даля с применением инструментов Data Science.

I. Описание исходных данных

Исходным материалом для исследования послужил электронный текстовый файл, содержащий словарные статьи из словаря Даля. Данные представлены в текстовом формате с кодировкой Windows-1251 (CP1251), что характерно для русскоязычных текстов. Общий объём исходного файла составил более 500 тысяч символов.

Структура словаря предполагает, что каждая строка файла начинается с заглавного слова словарной статьи, после которого следует толкование, примеры употребления и другая лингвистическая информация. Для статистического анализа требовалось извлечь только заглавные слова – основные словарные единицы.

Предварительный анализ данных показал необходимость преобразования кодировки из CP1251 в UTF-8 для корректной обработки кириллических символов в современных версиях Python.

II. Методика обработки данных

Алгоритм обработки данных реализован на языке программирования Python с использованием библиотек re (регулярные выражения), pandas (анализ данных) и matplotlib, seaborn (визуализация) [5].

III. Извлечение словарных единиц

Первый этап обработки данных – извлечение заглавных слов из каждой строки исходного файла. Для этого применяются регулярные выражения, позволяющие находить последовательности буквенных символов русского и латинского алфавитов. Используется шаблон \b[а-яА-Яа-zA-Z]+\b, где \b обозначает границу слова.

Алгоритм последовательно обрабатывает каждую строку исходного файла, извлекает из неё все слова с помощью функции re.findall() и сохраняет первое слово (заглавное слово словарной статьи) в выходной файл. Результатом данного этапа является текстовый файл, содержащий список всех заглавных слов словаря – по одному на каждой строке.

IV. Статистический анализ

На втором этапе выполняется статистический анализ полученного списка слов. Для этого все слова загружаются в память и преобразуются в структуру данных pandas.DataFrame – табличное представление, удобное для анализа.

Для каждого слова вычисляется его длина – количество символов. Создаётся новый столбец таблицы, содержащий длины всех слов. На основе этих данных вычисляются следующие статистические показатели:

- средняя длина слова (среднее арифметическое);
- минимальная и максимальная длина слова;
- распределение частот для каждой длины слова;
- медианная длина и мода распределения.

V. Визуализация результатов

Третий этап – визуализация полученных результатов. С использованием библиотеки pyplot строится гистограмма распределения длин слов. Гистограмма дополняется линией ядерной оценки плотности (KDE), которая позволяет визуализировать распределение в виде непрерывной кривой (см. рис. 1).

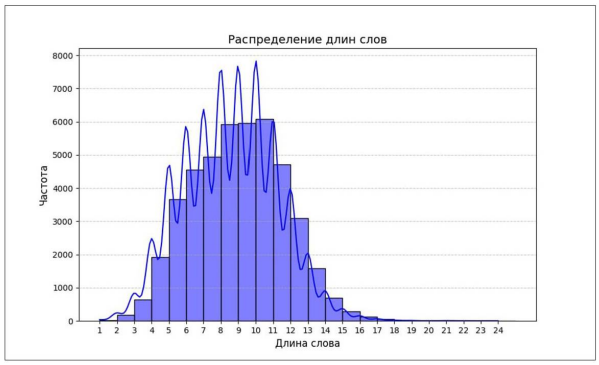

Рис. 1 – Гистограмма распределения длин слов

VI. Результаты исследования

В результате обработки словаря Даля было извлечено 44 691 заглавное слово. Статистический анализ показал следующие результаты:

- средняя длина слова составила 8,49 символа;
- минимальная длина – 1 символ;
- максимальная длина – 24 символа;
- наиболее частая длина слова (мода) – 10 символов (13,61% всех слов);
- медианная длина – 9 символов.

Распределение слов по диапазонам длин представлено в таблице 1.

Таблица 1 – Распределение слов по длине

Диапазон длин	Количество слов	Доля, %
1–3 символа	1 054	2,36
4–6 символов	10 154	22,72
7–9 символов	16 827	37,65
10–12 символов	13 888	31,08
13–15 символов	2 566	5,74
16 и более	202	0,45

Анализ показывает, что наибольшую долю (37,65%) составляют слова длиной от 7 до 9 символов. Слова длиной 10–12 символов занимают второе место (31,08%). Таким образом, около 69% всех слов словаря имеют длину от 7 до 12 символов.

Полученное значение средней длины слова (8,49 символа) существенно превышает среднюю длину слова в современном русском языке, которая, по различным оценкам, составляет от 5,28 до 6 символов [6, 7]. Это объясняется спецификой словаря Даля, который включает большое количество диалектных, устаревших и сложных составных слов.

Среди самых длинных слов в словаре – «понедоброжелательствовал» (24 символа), «беспокровительственный» (22 символа), «переосвидетельствовать» (22 символа).

Распределение слов по первым буквам алфавита показало, что наибольшую долю (24,78%) составляют слова, начинающиеся на букву «П», что связано с большим количеством приставочных образований в русском языке.

Заключение

В работе представлен эффективный алгоритм статистического анализа лексического состава исторического словаря с использованием современных инструментов Data Science. Разработанный метод позволяет извлекать количественные характеристики словарного состава и визуализировать результаты анализа.

Результаты исследования показывают, что словарь Даля характеризуется относительно большой средней длиной слова (8,49 символа), что отражает богатство словообразовательных возможностей русского языка XIX века и широкое представление диалектной и архаичной лексики.

Предложенная методика может быть применена для сравнительного анализа различных словарей и текстовых корпусов, а также для исследования эволюции лексического состава языка.

1. Толковый словарь живого великорусского языка / В. И. Даль. – М.: Русский язык, 1863–1866. – 4 т.
2. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа: `https://ruscorpora.ru`. – Дата доступа: 17.10.2025.
3. Бодуэн де Куртенэ, И. А. О словаре Даля / И. А. Бодуэн де Куртенэ // Толковый словарь живого великорусского языка В. И. Даля. – СПб., 1903. – Т. 1. – С. I–XLII.
4. McKinney, W. Python for Data Analysis / W. McKinney. – O'Reilly Media, 2017. – 544 p.
5. Лутц, М. Изучаем Python / М. Лутц. – СПб.: Символ-Плюс, 2019. – 1280 с.
6. Пиотровский, Р. Г. Математическая лингвистика / Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская. – М.: Высшая школа, 1977. – 383 с.
7. Гусейн-Заде, С. М. О средней длине слова в русском языке / С. М. Гусейн-Заде // Проблемы структурной лингвистики. – М.: Наука, 1982. – С. 82–88.