УДК 681.3 DOI: 10.26102/2310-6018/2021.32.1.025

Тематический анализ текстовой информации на основе частотных характеристик

Д.В. Меняйлов, А.П. Преображенский, Е.И. Чопорова
Воронежский институт высоких технологий,
Воронеж, Российская Федерация

Резюме. В настоящее время происходит активное развитие методов, связанных с исследованием текстовых массивов. При этом подобные подходы нацелены либо на то, чтобы измерять пространственные характеристики в текстовых массивах, таких как длины строк, размеры шрифтов и т. п., либо на рассмотрение общелингвистических задач, в которых изучаются смыслонесущие единицы, такие как предложения, фразы и др. Во втором классе задач перспективным можно считать использование частотного анализа. В работе дан анализ подходов, которые могут при этом использоваться. Авторами составлен алгоритм обработки текста на естественном языке. Созданный в работе программным образом алгоритм реализуется с помощью Python, Jupyter Notebook, WordCloud, NLTK. При обработке текстовый массив разбивается на слова, после чего происходит формирование списка токенов. Даны рекомендации по удалению союзов, предлогов и других частей речи, чтобы осуществлять полноценный анализ тематики. Показаны основные этапы алгоритма частотного анализа текста, которые заключаются в том, что выгружаются данные, производится первичная обработка текстовых массивов, осуществляется процесс замены слов, проводится оценка статистических данных, убираются лишние слова, осуществляется визуальное представление. В статье продемонстрирован пример фрагментов программного кода, описывающих работу ключевых этапов алгоритма.

Ключевые слова: текстовая информация, модель, частотный анализ, программа, слово, язык.

Для цитирования: Меняйлов Д.В., Преображенский А.П., Чопорова Е.И. Тематический анализ текстовой информации на основе частотных характеристик. https://moitvivt.ru/ru/journal/pdf?id=944 DOI: 10.26102/2310-6018/2021.32.1.025.

Thematic Analysis of Text Information Based on Frequency Characteristics

D.V. Menyaylov, A.P. Preobrazhenskiy, E.I. Choporova
Voronezh Institute of High Technologies
Voronezh, Russian Federation

Abstract: Currently, there is a development of methods related to the study of text arrays. In doing so, they aim to either measure their spatial characteristics, such as line lengths, font sizes, and more, or for consideration of general linguistic problems, in which the study of meaning-bearing units, such as sentences, phrases, and others, is carried out. In the second class of tasks, the use of frequency analysis can be considered promising. The paper analyzes the approaches that can be used in this case. The authors in the article developed an algorithm for processing text in a natural language. The algorithm created in the work is programmatically implemented using Python, Jupyter Notebook, WordCloud, NLTK. During processing, the text array is split into words, after which a list of tokens is formed. Recommendations are given for removing conjunctions, prepositions, and other parts of speech to carry out a comprehensive analysis of the topic. The main stages of the text frequency analysis algorithm are demonstrated. The data are unloaded, the primary processing of text arrays is carried out, after which the process of replacing words is carried out, the statistical data is evaluated, unnecessary words are

removed, and a visual presentation is carried out. The main stages of the algorithm have also been demonstrated based on fragments of the program code.

Keywords: text information, model, frequency analysis, program, word, language.

For citation: Menyaylov D.V., Preobrazhenskiy A.P., Choporova E.I. Thematic Analysis of Text Information Based on Frequency Characteristics. *Modeling, Optimization and Information Technology*. 2021;9(1). Available from: https://moitvivt.ru/ru/journal/pdf?id=944 DOI: 10.26102/2310-6018/2021.32.1.025 (In Russ).

Введение

В современных условиях весьма актуальным является вопрос, связанный с возможностями анализа и применения знаний, которые накоплены в достаточно больших объемах [1,2]. Среди них следует отметить подходы к решению задач, в которых осуществляется автоматическая обработка, распознание и классификация текстовой информации. Перспективным является метод частотного анализа текстов. Данный метод базируется на том, что существует нестандартное статистическое распределение символов внутри текстовых массивов. Практическое применение данного подхода может быть самым разным, например - выявление антиплагиата. Данной проблеме посвящено большое число работ. Например, в [3] научная деятельность в области классификации данных на естественном языке исследуется на основе ежегодной публикации научных работ в этой области. Авторы предлагают к рассмотрению метод классификации русскоязычных текстов, сочетающий алгоритмы частотного, морфологического и интеллектуального анализа. Также проблемы частотного анализа встречаются, когда необходим процесс дешифровки, выделение необходимого множества данных в больших массивах, анализ текстов, которые были написаны на древних языках, проведение процессов категоризации. Реализация частотного анализа может найти применение в рекомендательных системах [4]. Например, в работе [5] показано, каким образом можно разработать нейросетевой метод, связанный с автоматическим извлечением знаний из массивов полнотекстовых документов. Тексты в нем представляются как совокупность ключевых термов. В [5] показано, каким образом анализируется тематический состав информационных библиотек на базе такого способа. При этом частотная составляющая лежит в основе меры близости текстов. В исследовании [6] рассмотрены алгоритмы, которые связаны с построением рефератов с привлечением методик частотного анализа текстов. Проведен сравнительный анализ алгоритмов и предложены методы повышения качества рефератов, реализована методика составления сводного реферата, программный модуль создан для решения задачи автореферента на основе разработанных методик. В работе [7] предлагается использовать программные продукты частотного анализа текстовых массивов, которые размещены в Интернете. Автоматические частотные анализаторы лежат в основе подобных программных продуктов. Для того, чтобы быстро получать информацию относительно структурных характеристик, лингвистических свойств, доступности для аудитории, а также разных текстовых параметров возникают соответствующие возможности. На базе программных продуктов можно проводить оптимизацию коммуникации.

В [8] проводится анализ, обсуждается эффективность алгоритма, связанного с частотным анализом текстовых данных. При этом осуществляется статическая балансировка нагрузок. Сама реализация алгоритма осуществлена на базе параллельного подхода.

Приложение, разработанное авторами, является многопоточным. Это предоставляет возможности для того, чтобы осуществлять сравнение того, как вычисления будут распределены среди потоков - с частотностью и без нее.

Цель данной работы состоит в разработке алгоритма обработки текстовых данных. Сами данные рассматриваются как предложения на естественном языке.

Были обозначены такие ключевые задачи:

1. Рассмотреть особенности разных систем анализа текстовой информации. Важно было определить главные параметры, которые оказывают влияние на извлечение информации из текстовых массивов.

2. Провести разработку алгоритма обработки текста на естественном языке. При этом используются методики, связанные с частотным анализом текстов [9].

3. Рассмотреть особенности экспериментальной оценки эффективности предлагаемого подхода. Рассматривается пример текстовых данных на русском языке.

Особенности алгоритма частотного анализа текста

Частотность – это простой метод обработки текста на естественном языке:

$$Freq_x = \frac{Q_x}{Q_{all}}, \quad (1)$$

здесь Freqx - частотность слов "x", Qx – количество словоупотреблений слов "x", Q_{all} – количество словоупотреблений.

Исследователями показывается, что как правило частотность может выражаться в процентах. Для словарей характеристики частотности слов могут быть отражены при помощи пометок - употребительное, малоупотребительное и т.д.

Но в результате мы получаем список частым образом встречающихся слов. Также используем тематику и основные понятия.

С целью анализа данных – текста на русском языке, обработку целесообразно произвести с помощью Python, Jupyter Notebook, WordCloud, NLTK. Это и было осуществлено в данной работе.

При анализе русскоязычного текста необходимо учитывать некоторые его особенности. Весь анализ можно представить следующим образом (Рисунок 1).

Для начала мы выгрузим данные и произведем первичную обработку текста. Затем ведется процесс замены, статистическая обработка, мы убираем лишние слова. После этого визуальным образом представляем информацию.

1) Прочитаем содержимое файла, в итоге получим t, это текстовый массив. Таким образом, получаем длину текста с помощью len и первые 250 символов текста (Рисунок 1):

```
f = open ('File_medicina_01.txt', "w")
t = f.read ()
len (t)
t [:250]
```

Рисунок 1 - Пример считывания первых 250 символов текста

Figure 1 - An example of reading the first 250 characters of text

2) Очистим текст от лишних знаков, пробелов и цифр. Приведем все символы к нижнему регистру. Совокупность символов пунктуации стандартная из string. Совокупность специальных символов для удаления может расширяться.

Таким образом, необходим анализ исходного текста на предмет лишних символов. Дополнительно, к знакам пунктуации необходимо проводить анализ по символам переноса строк и табуляции, а также другие символы , при наличии их в исходном тексте.

Разделим t на символы и оставим не входящие в набор spec_chars и затем объединим список символов. Объявим удаление совокупности символов из исходного текста, и затем удаление специальных символов и цифр из исходного текста (Рисунок 2):

```
t = t.lower ()
import string
print (string.punctuation)
spec_chars = string.punctuation + '\n\0«»\t-…'
t = "".join([ch for ch in t if ch not in spec_chars])
def remove_chars_from_t(t, chars):
   return "".join([ch for ch in t if ch not in chars])
t = remove_chars_from_t (t, spec_chars)
t = remove_chars_from_t (t, string.digits)
```

Рисунок 2 - Пример удаления специальных символов и цифр из исходного текста
Figure 2 - An example of removing special characters and numbers from the original text

3) После первичной обработки необходимо разбить текст на части, для частотности разделим текст на слова. С помощью метода NLTK. t_tok - список токенов. Длина списка токенов: len(t_tok). Выведем первые 20 слов.

Преобразуем к классу T. Выведем тип переменной t. Выведем 20 первых токенов из текста (Рисунок 3):

```
from nltk import word_tokenize
t_tok = word_tokenize(t)
t_tok[:20]
import nltk
t = nltk.T(t_tok)
print(type(t))
t[:20]
```

Рисунок 3 - Пример вывода первых 20 токенов из текста
Figure 3 - An example of the output of the first 20 tokens from the text

4) Для подсчета статистики необходимо применить класс FreqDist. При выводе переменной fdist отобразится словарь с токенами и их частоты, то количество, сколько эти слова встречались в тексте. most_common для получения списка действий с токенами, которые часто встречаются (Рисунок 4).

Дадим визуальное представление с помощью графика (Рисунок 5). Класс FreqDist содержит plot.

Указываем количество токенов. Если мы упорядочим по убыванию частоты использования слов, то частота n слова будет обратно пропорциональной его порядковому номеру n:

```
from nltk.probability import FreqDist
fdist = FreqDist(t)
FreqDist({'он': 22, 'и': 126, 'не': 49, 'в': 81, 'что': 34, 'она': 19, 'с': 24, 'ее':
19, 'было': 17, 'на': 11, ...})
fdist.most_common(5)
[('и', 126), ('в', 81), ('не', 49), ('что', 34), ('с', 24)]
fdist.plot(30,cumulative=False)
```

Рисунок 4 - Пример определения частоты слов
Figure 4 - Example of determining the frequency of words

Рисунок 5 - Результаты убывания частоты использования слов
Figure 5 - Results of decreasing word usage

Таким образом, чаще всего встречались союзы, предлоги и другие части речи, смысловая нагрузка отсутствует, присутствуют только семантико-синтаксические связи. Поэтому, чтобы в итоге частотного анализа просматривалась тематика, нужно удалить данные слова.

5) Так как предлоги, союзы, междометия, частицы не имеют смысловой нагрузки, значит они избыточны.

Для русского языка предусмотрены шумовые слова. Для текстов с разной тематикой шумовые слова могут отличаться.

То есть нужно проанализировать текст и исключить шумовые слова, которые не вошли в стандартный набор.

Список может расширяться с помощью extend. После удаления данных слов частота токенов выглядит по другому.

Результаты стали точнее отражать тематику текста. Но есть, например, токены "регистрационный" и "регистрационного", это слово представлено в разных формах. Поэтому нужно для слов исходного текста найти основу (Рисунок 6):

```
from nltk.corpus import stopwords
russian_stopwords = stopwords.words ("russian")
russian_stopwords.extend11 (['нею', 'это'])
fdist_sw.most_common(10)
[('печать', 18),
('номер', 15),
('имя', 12),
('регистрационного', 4),
('было', 4),
('поминутно', 3),
('карта', 2),
('несколько', 3),
('регистрационный', 3),
('время', 4)]
```

Рисунок 6 - Пример определения основы для слов
Figure 6 - An example of defining a stem for words

6) Визуализация происходит с помощью WordCloud и matplotlib. Нужно передать строку. Это можно сделать с помощью join, пробел в качестве разделителя. В итоге мы получим визуализацию для исходного текста, можно извлечь основную информацию:

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
%matplotlib inline
t_raw = " ".join (t)
wordcloud = WordCloud().generate (t_raw)
```

Рисунок 7 - Пример извлечения основной информации
Figure 7 - An example of extracting basic information

На Рисунке 7 приведены основные этапы алгоритма частотного анализа текста. На Рисунке 8 показано объединение фрагментов программного кода, которые применялись в алгоритме.

Заключение

Таким образом, в работе дан анализ подходов, связанных с частотным анализом текстовой информации. Рассмотрены возможности оценок применяемых в них параметров. Разработан алгоритм обработки текста на естественном языке на базе методик частотного анализа текста. Осуществлена экспериментальная оценка относительно характеристик эффективности предлагаемого способа. На примере текста на русском языке было рассмотрено два варианта. В первом варианте чаще всего встречались союзы, предлоги и другие части речи, смысловая нагрузка отсутствовала, присутствовали только семантико-синтаксические связи. Поэтому, чтобы в итоге частотного анализа просматривалась тематика, необходимо было удалять данные слова. Во втором варианте результаты стали точнее отражать тематику текста. Была получена визуализация для исходного текста, из которой можно извлечь основную информацию.

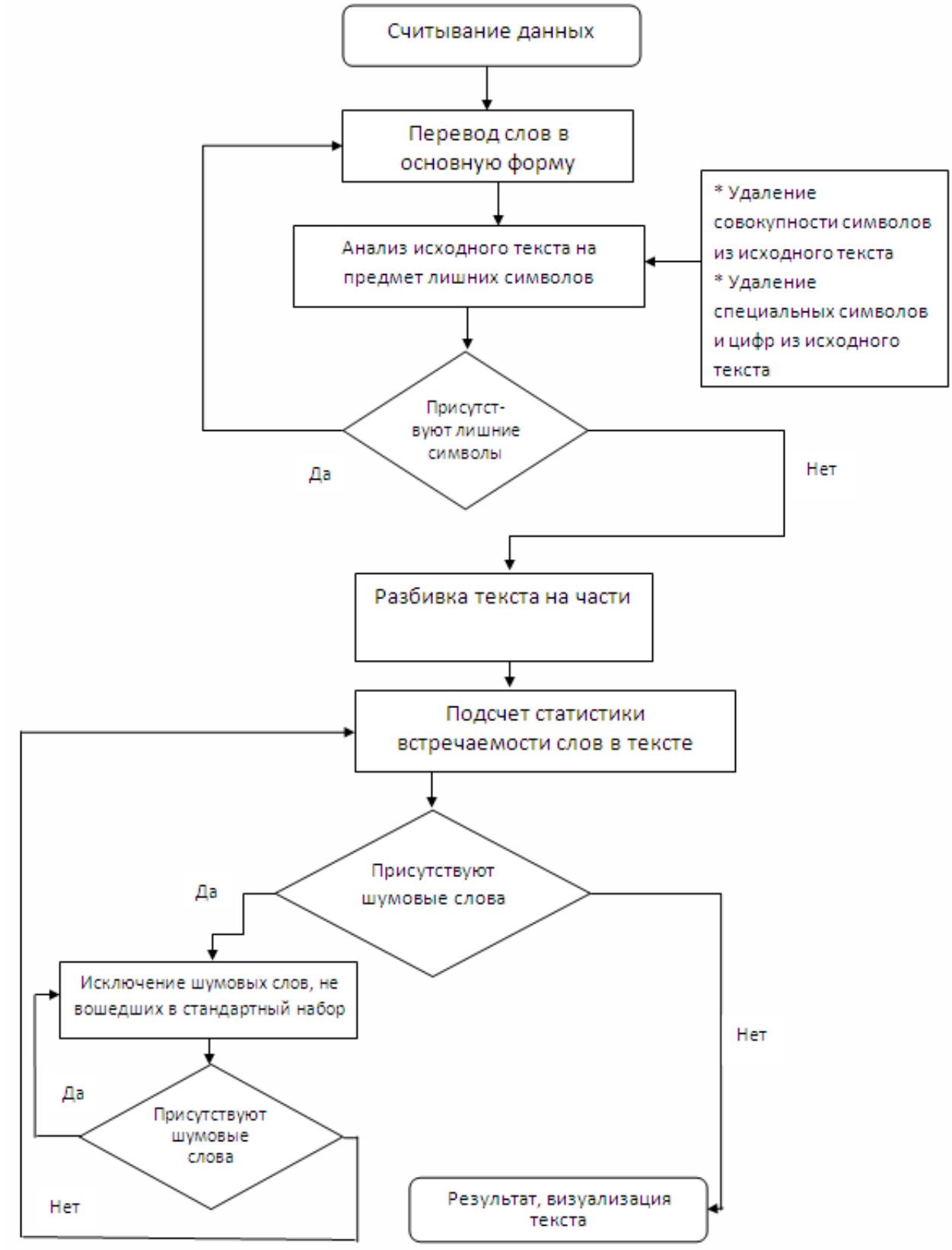

Рисунок 8 - Основные этапы алгоритма частотного анализа текста
Figure 8 - The main stages of the text frequency analysis algorithm

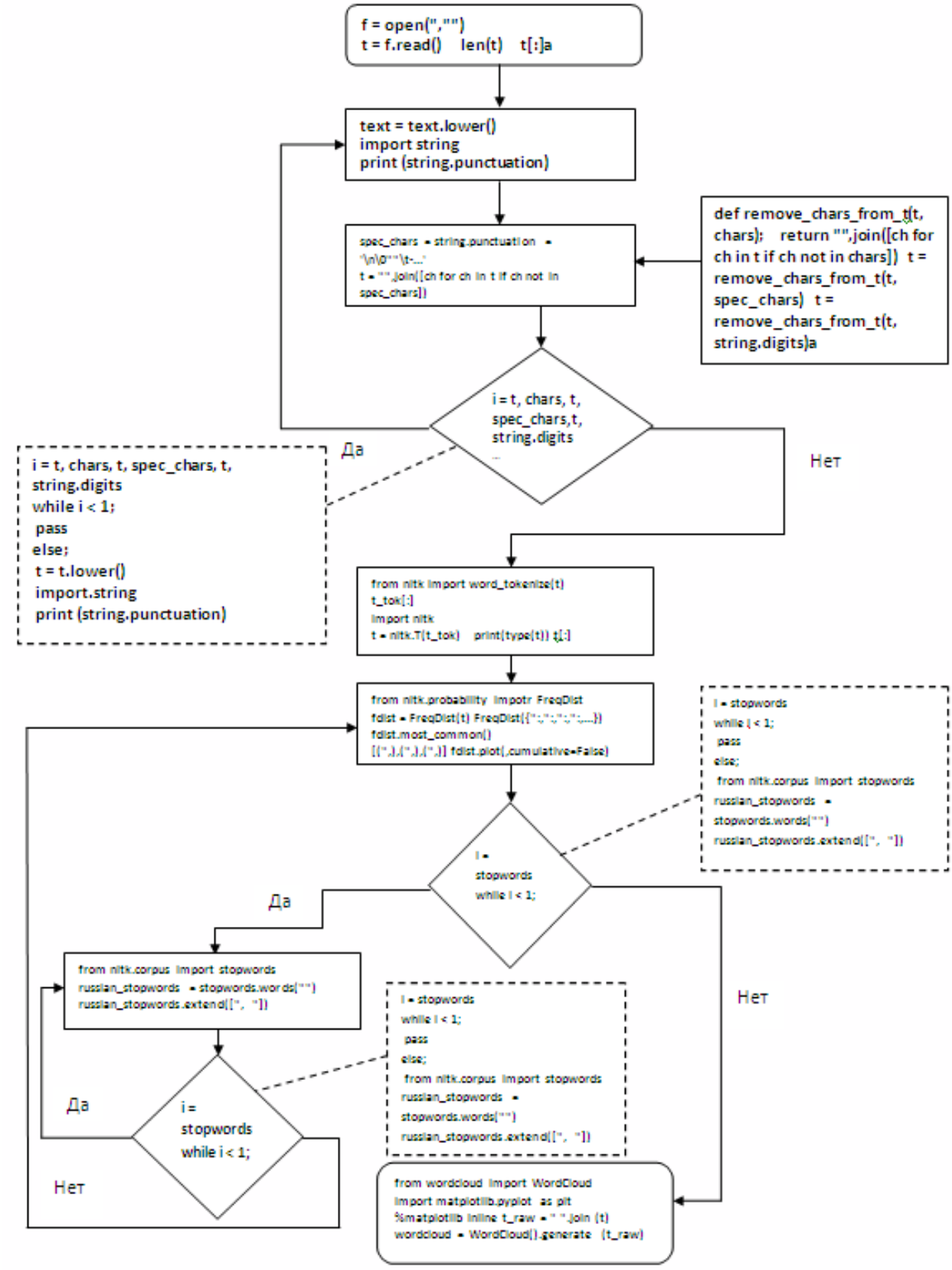

Рисунок 9 - Основные этапы алгоритма частотного анализа текста, представленные на основе фрагментов программного кода

Figure 9 - The main stages of the text frequency analysis algorithm, presented on the basis of fragments of the program code

ЛИТЕРАТУРА

1. Свиридов В.И., Чопорова Е.И., Свиридова Е.В. Лингвистическое обеспечение автоматизированных систем управления и взаимодействие пользователя с компьютером *Моделирование, оптимизация и информационные технологии.* 2019;1(24):430-438.
2. Цепковская Т.А., Чопорова Е.И. Проблемы построения автоматизированных обучающих систем *Моделирование, оптимизация и информационные технологии.* 2017;1(16):20.
3. Осочкин А.А., Фомин В.В., Флегонтов А.В. Метод частотно-морфологической классификации текстов. *Программные продукты и системы*. 2017;3(30):478–486.
4. Смирнова И.Г., Чопорова Е.И., Серостанова Н.Н. Особенности разработки профильных учебных пособий по иностранному языку с учетом формирования информационно-коммуникативной компетенции обучающихся. *Вестник Воронежского института высоких технологий*. 2017;3(22):64-68.
5. Шеменков П.С. Нейросетевой метод извлечения знаний на основе совместной встречаемости ключевых термов. *Сборник материалов 61 научно-технической конференции профессорско-преподавательского состава*, СПб ГУТ.2009:42–43.
6. Третьяков Ф.И., Серебряная Л.В. Методы автоматического построения рефератов на основе частотного анализа текстов. *Доклады Белорусского государственного университета информатики и радиоэлектроники*. 2014;3(81):40–44.
7. Шумилина Т.В. Применение частотного анализа текстов СМИ для оптимизации процесса коммуникации. *Вестник Московского Университета. Сер. 10. Журналистика.* 2017;(2):67–79.
8. Тхан Б. Х., Лупин С.А., Тайк А. М., Тун Х. Статическая балансировка нагрузки в параллельной реализации алгоритма частотного анализа текстовой информации. *International Journal of Open Information Technologies*. 2016;4(11):27-33.
9. Ляшевская О. Н., Шаров С. А. *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка).* М: Азбуковник, 2009.

REFERENCES

1. Sviridov V. I., Choporova E. I., Sviridova E.V. Linguistic support of automated control systems and user-computer interaction *Modeling, optimization and information technology.* 2019;1(24):430-438.
2. Tsepkovskaya T.A., Choporova E.I. Problems of building automated training systems *Modeling, optimization and information technology*. 2017;1(16):20.
3. Osochkin A.A., Fomin V.V., Flegontov A.V. Method of frequency-morphological classification of texts. *Software products and systems*. 2017;3(30):478–486.
4. Smirnova I.G., Choporova E.I., Serostanova N.N. Features of the development of specialized teaching aids in a foreign language, taking into account the formation of information and communication competence of students. *Vestnik Voronezhskogo institute vysokih tekhnologij.* 2017;3(22):64-68.
5. Shemenkov P.S. Neural network method of knowledge extraction based on the co-occurrence of key terms. *Proceedings of 61st scientific and technical conference of the teaching staff*, SPb GUT. 2009:42–43.
6. Tretyakov F.I., Serebryanaya L.V. Methods for automatic construction of abstracts based on the frequency analysis of texts. *Reports of the Belarusian State University of Informatics and Radioelectronics.* 2014;3(81):40–44.

7. Shumilina T.V. Application of frequency analysis of media texts to optimize the communication process. *Vestnik Moskovskogo Universiteta. Ser. 10. Journalism*. 2017;(2):67–79.
8. Than B.H., Lupin S.A., Taik A.M., Tun H. Static load balancing in parallel implementation of the algorithm for frequency analysis of text information. *International Journal of Open Information Technologies*. 2016;4(11):27-33.
9. Lyashevskaya O.N., Sharov S.A. *Frequency dictionary of the modern Russian language (on the materials of the National corpus of the Russian language)*. M .: Azbukovnik, 2009.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Меняйлов Дмитрий Владимирович, аспирант, Воронежский институт высоких технологий, Воронеж, Российская Федерация.
e-mail: dmitriy.menyaylov111@yandex.ru

Dmitriy V. Menyaylov, postgraduate student, Voronezh Institute of High Technologies, Voronezh, Russian Federation.

Преображенский Андрей Петрович, профессор, Воронежский институт высоких технологий, Воронеж, Российская Федерация.
e-mail: app@vivt.ru

Andrey P. Preobrazhenskiy, professor, Voronezh Institute of High Technologies, Voronezh, Russian Federation.

Чопорова Екатерина Ивановна, доцент, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: choporov_oleg@mail.ru

Ejkaterina I. Choporova, assistant professor, Voronezh State Technical University, Voronezh, Voronezh, Russian Federation.