УДК 004.8

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ЗАДАЧИ ОЦЕНКИ ПРИНАДЛЕЖНОСТИ ТЕКСТОВ ОДНОЙ ТЕМАТИКЕ

[1]Каспранская А.И., [2]Сметанина О.Н.

[1]ИП Гимальдинова О.Н., Уфа, e-mail: annakaspranskaya@gmail.com;
[2]ФГБОУ ВО «Уфимский государственный авиационный технический университет», Уфа, e-mail: smoljushka@mail.ru

Данная статья посвящена программной реализации задачи определения принадлежности двух текстов одной теме. В ходе работы был проведен анализ современного состояния проблемы, показавший актуальность работы и необходимость написания собственного программного решения для задачи, поставленной перед авторами, так как в данной задаче нет заранее известных тематик для текстов, они могут меняться в процессе работы, и отсутствует обучающая выборка. Был проведен обзор существующих библиотек для построения эмбеддингов, на основе которого выбрана библиотека Natasha. Данная библиотека имеет более подробную документацию, малый вес, высокую скорость работы, менее требовательна к аппаратному обеспечению. Natasha обучена на большом наборе текстов русской художественной литературы, включающем в себя более 300 тыс. текстов, с размером словаря 5×105 элементов. Результат работы программного обеспечения был показан на задаче распределения обращений от граждан между различными министерствами. После предварительной обработки текстов программа показала 97,9 % качество определения назначения для обращения. Данное значение точности показало работоспособность предлагаемого авторами решения проблемы определения принадлежности текстов общей теме. Возможны и другие области применения разработанного решения.

Ключевые слова: обработка текстов на естественном языке, эмбеддинг, принадлежность текстов, семантическая близость текстов, модель машинного обучения, методы классификации текстов

SOFTWARE IMPLEMENTATION OF THE PROBLEM OF EVALUATION OF BELONGING TO TEXTS OF THE SAME SUBJECT

[1]Kaspranskaya A.I., [2]Smetanina O.N.

[1]Individual Entrepreneur Gimaldinova O.N., Ufa, e-mail: annakaspranskaya@gmail.com;
[2]Ufa State Aviation Technical University, Ufa, e-mail: smoljushka@mail.ru

This article is devoted to the software implementation of the problem of determining whether two texts belong to the same topic. In the course of the work, an analysis of the current state of the problem was carried out, which showed the relevance of the work and the need to write your own software solution for the task assigned to the authors, because in this problem there are no previously known topics for texts, they can change in the course of work, and there is no training sample. A review of existing libraries for building embeddings was carried out, on the basis of which the Natasha library was selected. This library has more detailed documentation, light weight, high speed, less demanding on hardware. Natasha is trained on a large set of Russian fiction texts, which includes more than 300 thousand texts, with a dictionary size of 5×105 elements. The result of the software was shown on the task of distributing applications from citizens between different ministries. After pre-processing the texts, the program showed 97.9% quality in determining the destination for the appeal. This value of accuracy showed the operability of the solution proposed by the authors of the problem of determining whether texts belong to a common theme. Other areas of application of the developed solution are also possible.

Keywords: natural language processing, embedding, text ownership, semantic similarity of texts, machine learning model, text classification methods

Определение семантической близости текстов является одной из важнейших задач области компьютерной лингвистики. Ее решение может быть использовано при классификации текстов, автоматизации информационного поиска и пр.

Вопросам классификации текстов посвятили свои исследования многие специалисты в России и за рубежом: Т. Батура [1], Д.О. Долбин, В.И. Адамчук [2], А.М. Федотова, С.Е. Шаньшин, А.В. Куртукова, А.С. Романов [3], Х.Т. Максудов, Б.Б. Иномов [4], И.А. Батраева, А.Д. Нарцев, А.С. Лезгян [5], Yilin Niu, Chao Qiao, Hang Li, Minlie Huang [6], Omid Shahmirzadi, Adam Lugowski, Kenneth Younge [7].

Анализ существующих решений позволил сделать вывод об их научной и практической значимости. Однако существующие решения не могут быть использованы для задачи, поставленной перед авторами данной статьи. Особенностью решаемой авторами задачи является то, что тематики текстов заранее не определены, они могут меняться в процессе работы, и отсутствует обучающая выборка.

Современное состояние проблемы

Вопросы автоматизации извлечения информации из текстов широко применяются для решения ряда прикладных задач, среди которых можно выделить задачу

тематической классификации текстов; анализа тональности, выявления спама и др. Как правило, классификация может быть точной или пороговой (используется мера подобия). Для классификации текстов часто используются ряд методов на основе машинного обучения (рис. 1). Подробный обзор методов, оценку их достоинств и недостатков рассматривает в своей работе Т. Батура [1].

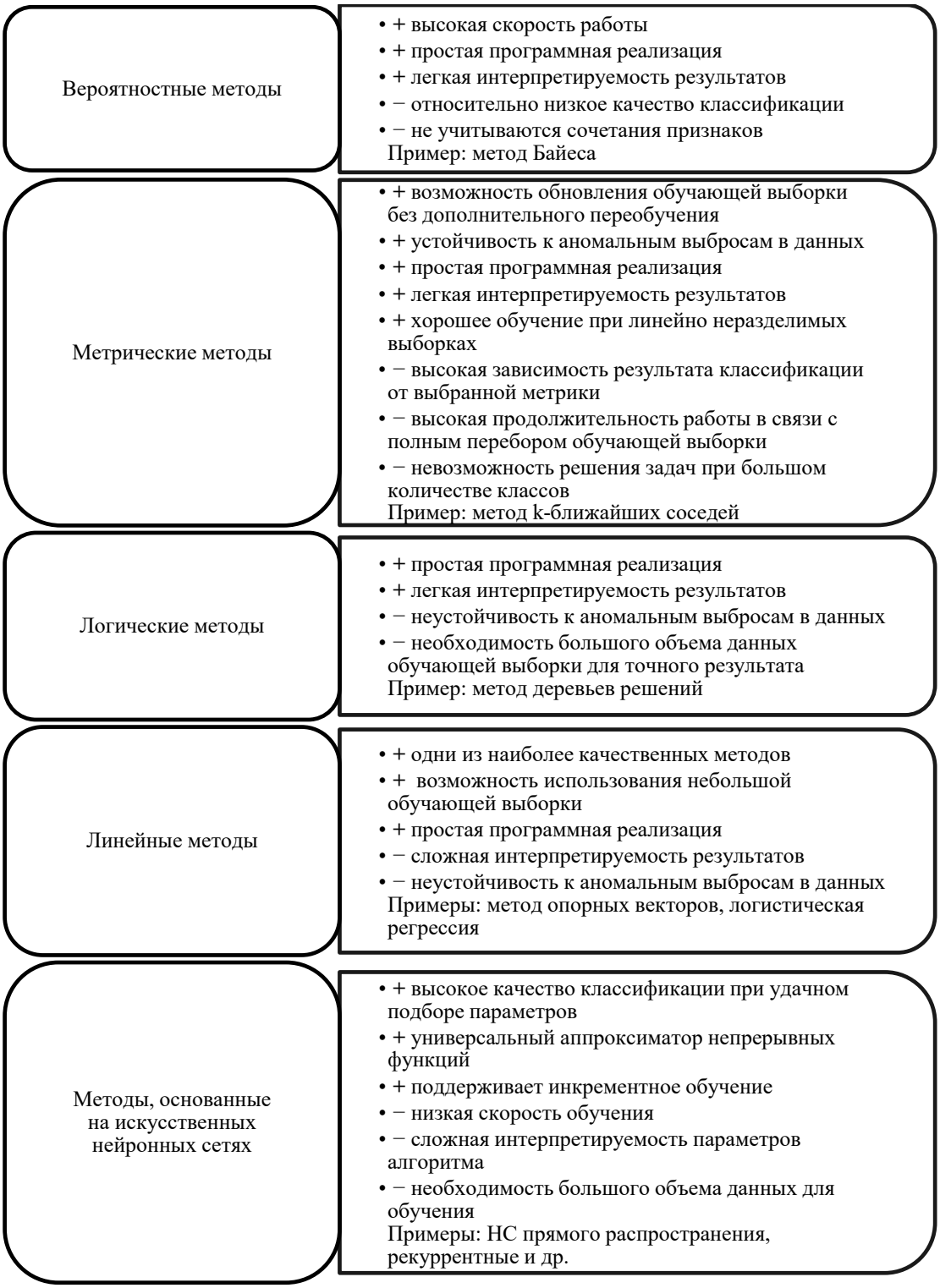

Рис. 1. Краткий обзор методов классификации текстов

Вопросу классификации текстов посвящены исследования как российских, так и зарубежных авторов, что подтверждает актуальность создания программного продукта, реализующего функцию вычисления принадлежности текстов одной тематике. Д.О. Долбин и В.И. Адамчук для классификации текста по темам используют нейронную сеть с моделью многослойного персептрона [2]. А.М. Федотова, С.Е. Шаньшин, А.В. Куртукова и А.С. Романов рассматривают применение моделей RuBert, MultiBert, SVM и MLP для задачи определения автора текстов, заранее обучая модели на четырёхстах художественных текстах пятидесяти авторов [3]. Для определения специализации научных текстов Х.Т. Максудов и Б.Б. Иномов рассматривают методы k-ближайших соседей и логистической регрессии [4]. И.А. Батраева, А.Д. Нарцев, А.С. Лезгян при решении задачи определения жанровой принадлежности текстов используют векторное представление слов с помощью модели word2vec и подают его сверточной нейронной сети [5].

Теме статьи посвящены работы и зарубежных авторов. Так, YilinNiu, Chao Qiao, Hang Li, Minlie Huang для определения близости текстов используют пословные эмбеддинги [6], а Omid Shahmirzadi, Adam Lugowski, Kenneth Younge для решения задачи сходства используют меры близости на основе TF IDF и эмбеддингов [7].

Существующие решения имеют научную и практическую значимость, но не подходят для решения задачи, поставленной перед авторами, так как в данной задаче нет заранее известных тематик для текстов, они могут меняться в процессе работы, и отсутствует обучающая выборка. Вследствие чего возникла необходимость в написании собственной программной реализации определения принадлежности текстов одной тематике.

Постановка задачи оценки принадлежности текстов одной тематике

Для оценки семантического сходства заранее не известных текстов на русском языке на произвольные темы используется функция определения принадлежности текстов одной теме. Математическая постановка задачи заключается в оценке функции принадлежности двух текстов одной теме, а именно определению $sim(text_1, text_2)$, где $text_1$ и $text_2$ – два произвольных текста, для которых определяется принадлежность одной теме, A – некоторая тема, которой могут принадлежать тексты. Учитывается то, что тексты $text_1$ и $text_2$ (в общем виде – $text_j$, где j=1,2, состоят из отдельных слов $text_j = (t_{ij}, .., t_{nj})$, $text_j$ – j-й текст, t_{ij} – i-е слово в j-м тексте, n – количество уникальных слов) – произвольные. Функция принадлежности определяется как

$$sim(text_1, text_2) = \begin{cases} 1, |(text_1 \in A) \wedge (text_2 \in A) \\ 0, |(text_1 \notin A) \vee (text_2 \notin A) \end{cases}.$$

Разработанное программное решение должно учитывать функцию принадлежности и предложенный ранее авторами [8] алгоритм решения задачи.

Таким образом, необходимо разработать программное решение для определения степени семантической близости между произвольными текстами, приложение должно иметь простой и удобный интерфейс.

Программный комплекс для реализации информационных процессов

Программный комплекс состоит из нескольких запускаемых файлов, связанных между собой по типу клиент-сервер.

Клиентская часть программы имеет дружественный интерфейс и реализована на языке C#. Интерфейс программы (рис. 2) создавался для прикладного решения задачи оценки семантического сходства заранее не известных текстов на русском языке – определение направления обращений граждан среди различных ведомств.

Задача клиентской части программного решения – предоставление пользователю удобного формата работы с текстами, а именно загрузка множества текстов разом, наглядный вывод степеней близости текстов.

Рис. 2. Интерфейс разработанного программного решения

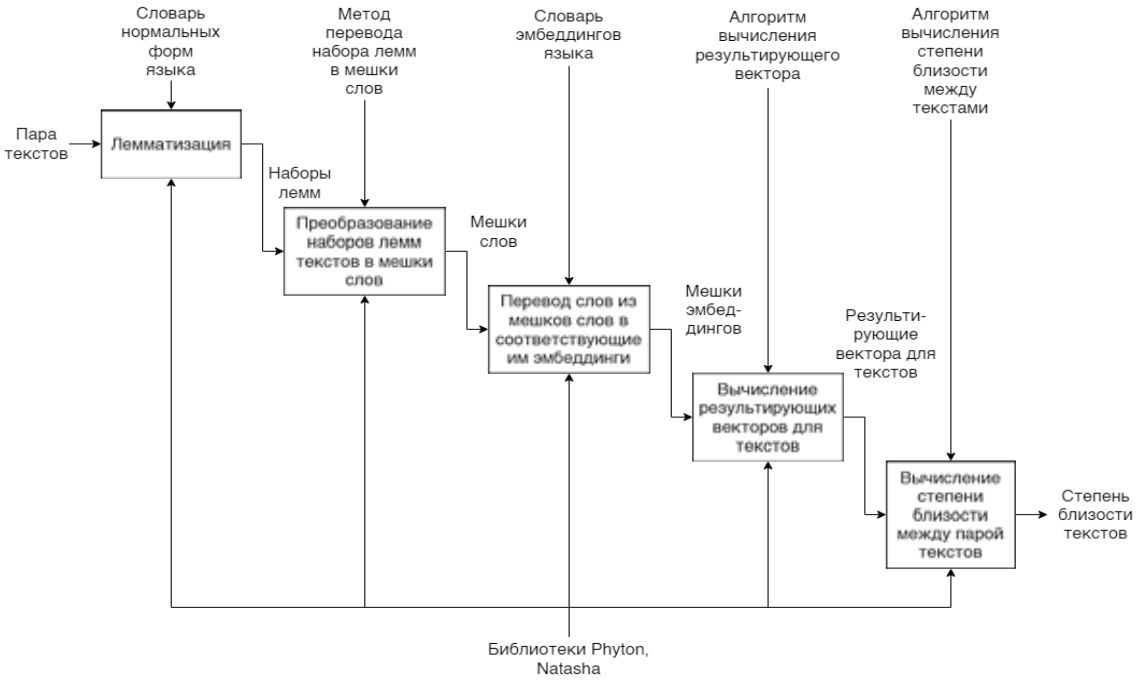

Рис. 3. Обобщенная функциональная модель серверной части программного решения

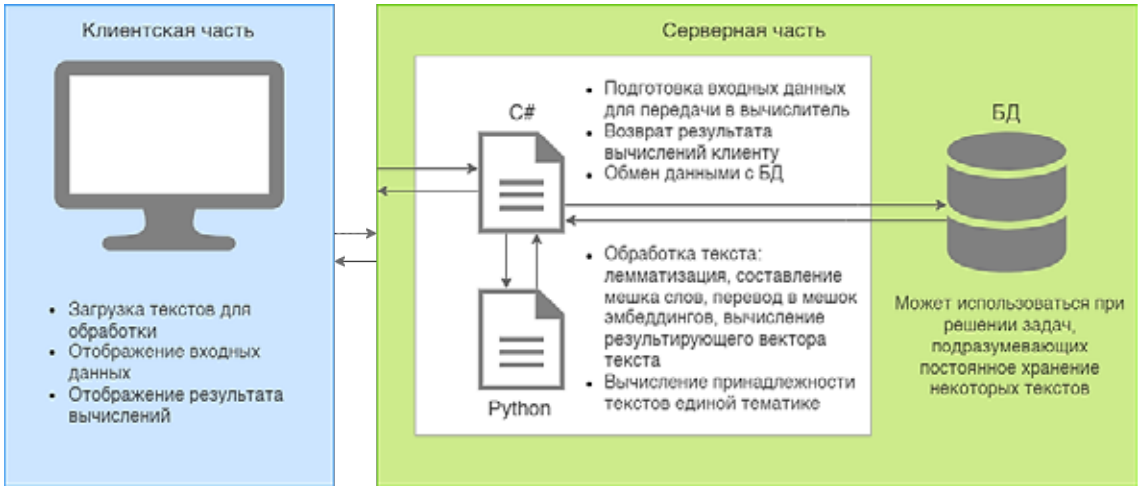

Рис. 4. Архитектура программного комплекса

Серверная часть программного решения предназначена для обработки текстов и вычисления степени близости между ними и представлена в обобщённом виде на следующей функциональной модели (рис. 3).

Основная часть логики программы написана на языке Python. Разумность установки этой части на сервер обусловлена повышенными требованиями языка к окружению. Для работы программы необходима установка python 3.9.9 и следующих библиотек: Xmlrpc.server, Natasha, Navec, Pymorhy2, Scipy, Numpy.

На рис. 4 представлена архитектура программного комплекса, взаимодействие между серверной и клиентской частями, их задачи и логические компоненты.

Обзор библиотек для реализации основных этапов задачи. Для того чтобы машина научилась понимать человека, в области компьютерных наук выделилось направление технологий искусственного интеллекта – обработка естественного языка, или Natural Language Processing (NLP). Эта группа технологий занимается проблемами компьютерного анализа и синтеза текстов на человекочитаемых

языках, позволяет распознавать тексты, классифицировать документы, выполнять машинный перевод, определять спам-письма, создавать чат-боты и виртуальных помощников.

Основными библиотеками, которые включают в себя эмбеддинги для русского языка, можно считать RusVectores, DeepPavlov и Natasha (рис. 5).

Для решения задачи было решено использовать библиотеки проекта Natasha по ряду критериев: а) более понятная и подробная документация к проекту, простой и удобный для использования интерфейс, прозрачная обработка текста (явная инициализация компонент, загрузка эмбеддингов, вызов необходимых методов – разбиения на токены, анализа морфологии и прочего); б) на основе результатов сравнения Natasha с моделью ruBertot DeepPavlov в задаче выделения именованных сущностей (табл. 1) и в) на основе результатов сравнения Natasha с инструментом RusVectores при оценке качества эмбеддингов на задаче семантической близости (табл. 2).

Рис. 5. Основные библиотеки, включающие в себя эмбеддинги для русского языка

Таблица 1

Сравнение DeepPavlov и Natasha

	Natasha, Slovnet NER	DeepPavlov BERT NER
PER/LOC/ORG F1 по токенам, среднее по Collection5, factRuEval-2016, BSNLP-2019, Gareev	0.97/0.91/0.85	0.98/0.92/0.86
Размер модели	27 МБ	2 ГБ
Потребление памяти	205 МБ	6 ГБ (GPU)
Производительность, новостных статей в секунду (1 статья ≈ 1КБ)	25 на CPU (Core i5)	13 на GPU (RTX 2080 Ti), 1 на CPU
Время инициализации, с	1	35
Библиотека поддерживает	Python 3.5+, PyPy3	Python 3.6+
Зависимости	NumPy	TensorFlow

Таблица 2

Сравнение RusVectores и Navec

		Среднее качество на 6 датасетах	Время загрузки, секунды	Размер модели, МБ	Размер словаря, x10^3
Navec	hudlit_12B_500K_300d_100q	0,719	1,0	50,6	500
	news_1B_250K_300d_100q	0,653	0,5	25,4	250
RusVectores	ruscorpora_upos_cbow_300_20_2019	0,692	3,3	220,6	189
	ruwikiruscorpora_upos_skipgram_300_2_2019	0,691	5,0	290,0	248
	tayga_upos_skipgram_300_2_2019	0,726	5,2	290,7	249

Несмотря на то, что Natasha на 1% показала качество ниже, чем DeepPavlov, размер ее модели меньше в 75 раз, потребление памяти меньше в 30 раз, а скорость работы на CPU выше в 2 раза, что, несомненно, делает библиотеки Natasha более приемлемым вариантом для решения поставленной задачи.

По таблице видно, что качество модели hudlit_12B_500K_300d_100q лучше, чем у моделей от RusVectores, при этом словарь больше в 2–3 раза, а размер модели меньше в 5–6 раз.

Используемая модель для построения эмбеддингов обучена на большом наборе текстов русской художественной литературы (более 300 тыс. текстов с размером словаря 5×105 элементов).

Пример использования программного решения для практической задачи

Данный программный комплекс был разработан для решения задачи автоматизации распределения обращений от граждан между различными министерствами. Семантическую близость необходимо было определить между входящим обращением и «функцией» министерства.

Важной особенностью решаемой задачи стало содержание большого количества вводных и общих фраз как для обращений граждан, таких как «Добрый день», «Прошу обратить внимание» и пр., так и в положениях о министерствах, таких как «участвует в разработке», «обеспечивает работу» и пр. Это значительно затруднило определение назначения обращения, поэтому привело к решению задачи о поиске значимых слов, только тех, которые точно передают функцию ведомства или суть обращения, и отбрасывании лишних.

После предварительной обработки программа показала высокую долю качественно распределенных обращений – 97,9%.

Заключение

Особенностью решаемой авторами задачи по оценке семантического сходства текстов является то, что тематики текстов заранее не определены, они могут меняться в процессе работы, а также отсутствует обучающая выборка. Результаты анализа готовых программных решений задачи позволили сделать вывод о необходимости разработки собственного программного решения, в основу которого положены предложенные авторами функция принадлежности и алгоритм решения задачи.

Программный комплекс состоит из нескольких запускаемых файлов, связанных между собой по типу клиент-сервер. Клиентская часть программы имеет дружественный интерфейс и реализована на языке C#. Основная часть логики программы написана на языке Python. Для решения задачи следует использовать библиотеки проекта Natasha.

Программный комплекс апробирован для задачи автоматизации распределения текстовых обращений граждан между различными министерствами.

Программное решение может быть использовано при оценке сходства новых версий клинических рекомендаций в медицине с текущими.

Результаты исследований, приведенные в статье, частично поддержаны грантом РНФ 22-19-00471.

Список литературы

1. Батура Т. Методы автоматической классификации текстов // Международный журнал «Программные продукты и системы». 2017. С. 85–99.

2. Долбин Д.О., Адамчук В.И. Практическое применение нейронных сетей для классификации текстов по темам // Научно-практические исследования. 2021. С. 25–28.

3. Федотова А.М., Шаньшин С.Е., Куртукова А.В., Романов А.С. Модели RUBERT, SVM и MLP в задаче определения автора текста // Сборник избранных статей научной сессии ТУСУР. 2021. С. 203–206.

4. Максудов Х.Т., Иномов Б.Б. Оценка эффективности методов k-ближайших соседей и логистической регрессии при определении специальности научных текстов // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2019. № 4. С. 34–38.

5. Батраева И.А., Нарцев А.Д., Лезгян А.С. Использование анализа семантической близости слов при решении задачи определения жанровой принадлежности текстов методами глубокого обучения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2019. С. 14–22.

6. Yilin Niu, Chao Qiao, Hang Li, MinlieHuang. Word Embedding based Edit Distance. [Электронный ресурс]. URL: https://arxiv.org/abs/1810.10752 (дата обращения: 14.06.2022).

7. Omid Shahmirzadi, Adam Lugowski and Kenneth Younge. Text Similarity in Vector Space Models: A Comparative Study. [Электронный ресурс]. URL: https://arxiv.org/abs/1810.00664 (дата обращения: 14.06.2022).

8. Каспранская А.И., Сметанина О.Н. Подход к оценке принадлежности текстов одной тематике // Современные наукоемкие технологии. 2022. № 5. С. 43–47.