2. Программа развития университета 2030 : сайт / Тюменский государственный университет. — URL: https://www.utmn.ru/priority 2030/ (дата обращения: 25.05.2022). — Текст : электронный.
3. Sergeev P. P. The use of digital footprints to create psychographic portraitsfor increased efficiency in advertising messages / P. P. Sergeev, D. A. Samylina. — Текст : непосредственный // Communications. Media. Design. — 2021. — Vol. 6, № 3. — С. 115-128.
4. Широбокова С. Н. Аспекты разработки инструментария для поиска аудитории потенциальных абитуриентов с помощью интерфейсов программирования приложений социальной сети «Вконтакте» / С. Н. Широбокова, В. С. Холодков, А. М. Бейбалаев. — Текст : непосредственный // Инновационная наука. — 2017. — № 1-2. — С. 101-103.
5. Рашка С. Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2 / С. Рашка, В. Мирджалили. — 3-е изд. — Санкт-Петербург : ООО «Диалектика», 2020. — 848 с. : ил. — Текст : непосредственный.
6. Де Берг, М. Computational Geometry: Algorithms and Applications / М. Де Берг, О. Чион, М. Кревелд, М. Овермарс. — 3rd Edition. — Springer, 2008. — 398 с.: ил. — Текст : непосредственный.
7. Стаффер, М. Laravel. Полное руководство / М. Стаффер. — Санкт-Петербург : Питер, 2020. — 512 с. : ил. — Текст : непосредственный.

А. И. Дадашзаде, Е. В. Егорова, М. С. Воробьева
Тюменский государственный университет, г. Тюмень

УДК 004.912

РАЗРАБОТКА СЕРВИСА ОБЗОРА И АНАЛИЗА НОВОСТЕЙ ДЛЯ СТУДЕНТОВ ТюмГУ

Аннотация. *В статье рассматривается разработка сервиса для обзора и анализа публикаций на основе данных, полученных из социальной сети «ВКонтакте». Реализован метод для обработки данных, а именно для определения категории для каждой публикации. В результате работы разработан пользовательский сервис для отображения обработанных данных — отсортированных публикаций.*

Ключевые слова: *сбор данных, обработка данных, ключевые слова, пользовательский сервис, Python, ВКонтакте.*

Введение. В наше время мы сталкиваемся с огромным количеством информации. На сегодняшний день почти все студенты являются пользователями более чем одной социальной сети, и каждый день они просматривают множество новостей. Поскольку объем производимого контента растет экспоненциально, пользователям необходимы инструменты, которые классифицировали бы и сортировали информацию. Чтобы найти нужную публикацию, связанную с университетом, студенты тратят немало время на поиск, например, если новость была опубликована давно. Поэтому вопрос о категоризации публикаций становится очень важным.

Данная проблема является актуальной. Фильтрация новостей решает большинство проблем: поиск нужной информации, возможность просматривать новости на определенную тематику, удобная навигация между публикациями, опубликованными в разное время.

Для нахождения подходящих решений необходимо применять различные подходы для обработки неструктурированные тексты, и этому вопросу посвящены современные исследования. Например, В. С. Мальчиц в своих статьях исследует методы машинного обучения и опорных векторов для классификации новостей [1-2]. В книге [3] рассматриваются основные задачи и универсальные методы анализа текстовых данных. В работе [4] автором А. А. Романенко предлагается использовать метод ближайшего соседа для решения задачи и методы отбора признаков для улучшения категоризации.

Постановка задачи. Необходимо разработать пользовательский сервис, который позволит студентам быстро искать и просматривать публикации официального сообщества Тюменского государственного университета (ТюмГУ) в социальной сети «ВКонтакте». Для разработки сервиса необходимо собрать и проанализировать данные публикаций сообщества.

Архитектура сервиса (рис. 1) содержит 5 модулей: Data Loader (Сбор данных), Algorithms (Реализация алгоритмов), Data Structures (Необходимые структуры данных), Analyzer (Обработка данных), User Interface (Интерфейс пользовательского сервиса).

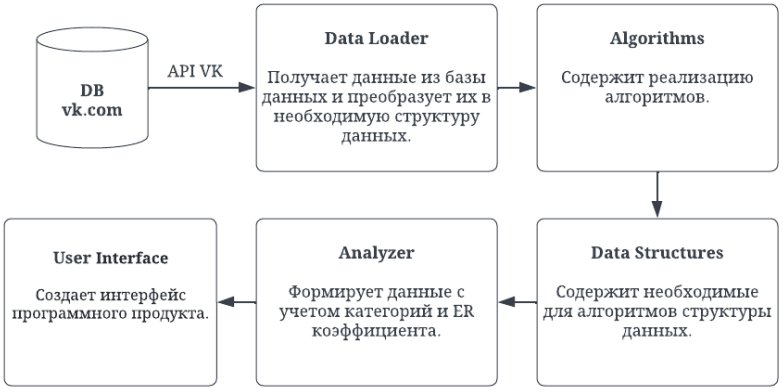

Рис. 1. Архитектура сервиса

Разработка модуля для сбора данных

Данные публикаций (дата, текст, фото, реакции пользователей) официального сообщества ТюмГУ в социальной сети «ВКонтакте» были получены с использованием интерфейса API VK, количество полученных публикаций — 5000 (период с апреля 2016 по май 2022 г.). Для точного решения о тематиках категорий был проведен опрос, состоящий из 10 вопросов, о категориях, интересующих новостей, в итоге было получено мнение 50 студентов 2 курса направления МОиАИС.

После проведения предварительного просмотра собранных данных (текстов публикаций группы) и анализа результатов опроса студентов об интересующих их тематиках новостей было решено выделить 9 категорий публикаций: мероприятия, олимпиады, конкурсы, достижения, мнения студентов, стипендия, работа, практики, другие новости.

Разработка модуля для обработки данных

Рассмотрим алгоритм обработки данных публикаций официального сообщества ТюмГУ в социальной сети «ВКонтакте» (рис. 2).

Этап 1. Нормализация полученных данных, токенизация и фильтрация стоп-слов текста, эмодзи и ссылок с использованием регулярных выражений, лемматизация. Подготовка данных реализована с помощью библиотек языка Python (nltk, pymorphy2, re).

Этап 2. Определение ключевых слов категорий с помощью статистической меры TF-IDF, которая применяется с целью оценки важности слова. Значимость слова прямо пропорциональна частоте использования слова в тексте публикации и обратно пропорциональна частоте использования слова во всех новостях.

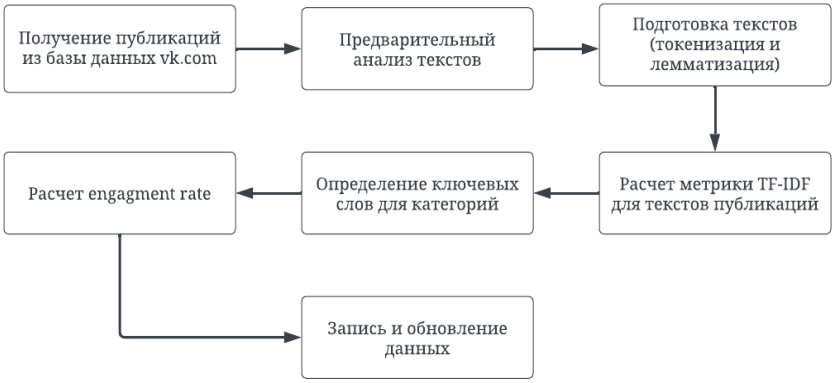

Рис. 2. Схема работы модулей сбора и обработки данных

Суть заключается в том, чтобы сделать меньшей значимость слов, используемых повсеместно [5]. TF — отношение количества вхождений некоторого слова к общему количеству слов текста публикации по формуле (1). Другими словами, оценивается значимость слова в границах одной новости.

$$tf(t,d) = \frac{n_t}{\sum_k n_k}. \quad (1)$$

IDF — значение инверсии частоты, отношение количества публикаций к количеству новостей, в которых содержится некоторое слово. Это значение уменьшает значимость широкоупотребительных слов. Имеется лишь одно значение IDF для каждого неповторимого слова в пределах текстов новостей, и вычисление осуществляется по формуле (2).

$$idf(t,D) = \log\frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}. \quad (2)$$

Статистическая мера TF-IDF считается по формуле (3), где множителями являются значения TF и IDF.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D). \quad (3)$$

В результате вычислений слова с высокой частотой в границах определенной публикации и с невысокой частотой использования в других новостях приобретут более высокое значение TF-IDF. Принимая во внимание значение метрики TF-IDF, составляется словарь ключевых слов категорий.

Для определения категории публикации реализован поиск ключевых слов в тексте новости.

Этап 3. Подсчет значения Engagement rate (ER), которое является коэффициентом вовлеченности аудитории в активности сообщества. Автор Е. В. Бахчева в статье [6] отмечает, что коэффициент вовлеченности позволяет судить о качестве и эффективности размещаемых материалов, а также результативности инструментов продвижения. Для определения популярных публикаций по формуле (4) используются значения количества реакций (likesCount), репостов (repostsCount), комментариев (commentsCount) и охвата публикации, т. е. количества просмотров (viewsCount).

$$ER = \frac{likesCount + repostsCount + commentsCount}{viewsCount}. \quad (4)$$

Коэффициент вовлеченности является динамическим параметром, потому что используемые для расчета ER данные зависят от активности студентов на публикации официального сообщества университета.

В результате разработки алгоритма были получены следующие результаты. Для каждой категории выделены ключевые слова (табл. 1).

Для каждой новости определена категория, соотношение количества публикаций по категориям, можно увидеть на диаграмме (рис. 3).

Таблица 1

Таблица категорий и количествово ключевых слов

Id категории	Название	Количество ключевых слов
1	Мероприятия	34
2	Олимпиады	16
3	Конкурсы	12
4	Достижения	29
5	Мнения студентов	4
6	Стипендии	6
7	Работы	8
8	Практики	6

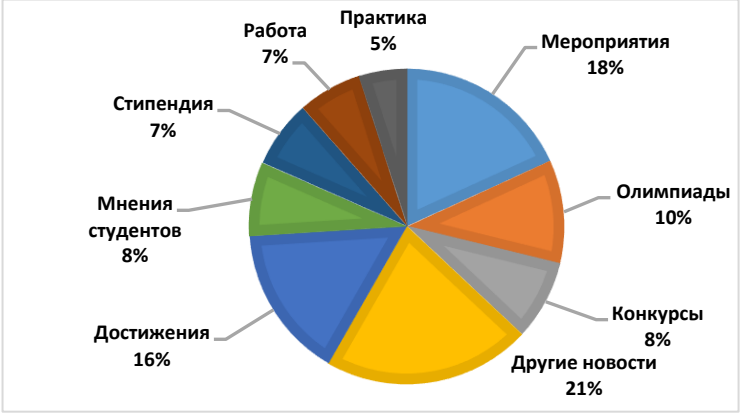

Рис. 3. Распределение публикаций по категориям

Сводные данные, представленные на диаграмме, о соотношении количества публикаций по определенной категории показывают следующее:

1. Категории — *мероприятия (900 новостей)*, *достижения (800 новостей)* и *олимпиады (500 новостей)* набрали большее количество

публикаций, в сумме эти категории занимают около 44% производимого контента.

2. Новости, которые не попали ни в одну из категорий, принадлежат к категории *другие новости*, к ней относятся публикации с объявлениями, видеороликами и т. д.

Для динамического обновления данных реализована метод, в котором коэффициент вовлеченности для новости пересчитывался через 3, 5, 7 и 14 дней после публикации.

Разработка пользовательского сервиса

После получения данных заполняется пользовательский сервис для отображения обработанных новостных публикаций, размеченных по 9 категориям с применением сортировок: по дате и популярности (рис. 4).

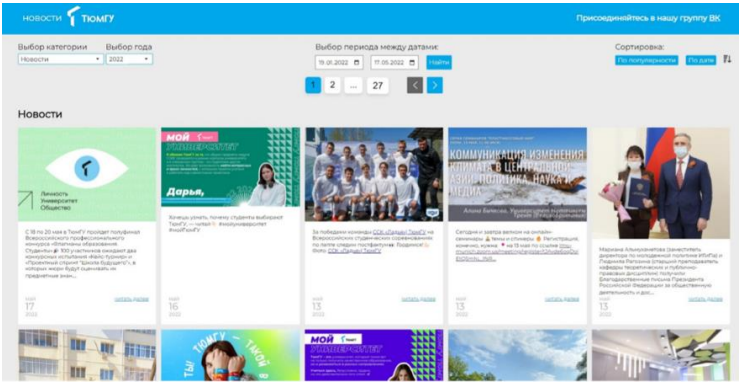

Рис. 4. Новостные публикации за 1-й квартал 2022 г.

В ходе выполнения пользовательского сервиса реализованы функции:

1) сортировка, предназначенная для отображения новостных событий по популярности, дате и периоду;

2) фильтрация для отображения публикаций по категориям.

Например, за 1-й квартал 2022 г. было выгружено 261 новостей официального сообщества ТюмГУ социальной сети «ВКонтакте».

Заключение. Разработанный пользовательский сервис позволяет студентам быстро находить нужную новость или изучать

прошлые публикации официального сообщества ВКонтакте «Тюменский государственный университет | ТюмГУ», применяя фильтрацию по категориям и времени.

В дальнейшем планируется расширение проекта:

- добавление сообществ для обзора;
- добавление методов определения категорий.

СПИСОК ЛИТЕРАТУРЫ

1. Мальчиц В. С. Применение методов машинного обучения для классификации новостей / В. С. Мальчиц. — Текст : электронный // Молодежь XXI ВЕКА: шаг в будущее. — 2019. — С. 208-209. — URL: https://www.elibrary.ru/item.asp?id=39379772 (дата обращения: 20.04.2022).
2. Мальчиц В. С. Обработка данных для машинного обучения и применение метода опорных векторов для реализации классификатора новостей / В. С. Мальчиц, А. Н. Гетман. — Текст : электронный // Вестник Амурского государственного университета. Серия: Естественные и экономические науки. — 2019. — № 87. — С. 8-13. — URL: https://www.elibrary.ru/item.asp?id=41456363 (дата обращения: 22.04.2022).
3. Ломакина Л. С. Информационные технологии анализа и моделирования текстовых данных / Л. С. Ломакина, В. Б. Родионов, А. С. Суркова. — Текст : непосредственный // Системы управления и информационные технологии. — 2012. — № 2. — С. 39-44.
4. Романенко А. А. Категоризация текстов на основе монотонного классификатора ближайшего соседа / А. А. Романенко. — Текст : непосредственный // Математические и информационные технологии. — 2011.
5. Скотт У. TF-IDF с нуля в python на реальном наборе данных / У. Скотт. — Текст : электронный // towardsdatascience.com : [сайт]. — 2019. — 15 февр. — URL: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089 (дата обращения: 29.04.2022).
6. Бахчева Е. В. Коэффициент вовлеченности как инструмент оценки коммуникации / Е. В. Бахчева. — Текст : непосредственный // Язык. Культура. Медиакоммуникация. — 2021. — Т. 1, № 1.