

ПЕРМСКИЙ
ГОСУДАРСТВЕННЫЙ
НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

ТЕКСТОВАЯ АНАЛИТИКА В ОБЩЕСТВЕННОЙ ГЕОГРАФИИ

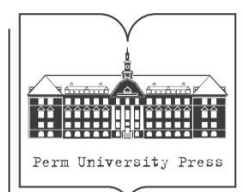


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное
образовательное учреждение высшего образования
«ПЕРМСКИЙ ГОСУДАРСТВЕННЫЙ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»

ТЕКСТОВАЯ АНАЛИТИКА В ОБЩЕСТВЕННОЙ ГЕОГРАФИИ

*Допущено методическим советом
Пермского государственного национального
исследовательского университета в качестве
учебно-методического пособия для студентов,
обучающихся по направлению подготовки магистров
«География»*



Пермь 2025

УДК 911.3:3(075.8)

ББК 65.04я73

Т307

Текстовая аналитика в общественной географии [Электронный ресурс] : учебно-методическое пособие / автор-составитель Е. В. Коньшев ; Пермский государственный национальный исследовательский университет. – Электронные данные. – Пермь, 2025. – 3,63 Мб ; 101 с. – Режим доступа: <http://www.psu.ru/files/docs/science/books/uchebnie-posobiya/Konyshhev-Tekstovaya-analitika-v-obshchestvennoj-geografii.pdf>. – Заглавие с экрана.

ISBN 978-5-7944-4238-0

Учебный курс «Текстовая аналитика в общественной географии» относится к комплексным и междисциплинарным и направлен на формирование навыков обработки больших данных методами текстового анализа. В учебно-методическом пособии представлены программа курса, теоретические материалы по разделам дисциплины, перечень практических работ и рекомендации по их выполнению, список вопросов для самоконтроля и для подготовки к итоговому контрольному мероприятию, основная и дополнительная литература по курсу, а также цифровые образовательные ресурсы. Предназначено для студентов, обучающихся по программе 04.05.02 «География», в качестве учебно-методического пособия.

УДК 911.3:3(075.8)

ББК 65.04я73

*Издается по решению ученого совета географического факультета
Пермского государственного национального исследовательского университета*

Рецензенты: кафедра туризма, геоурбанистики и экономической географии
Уфимского университета науки и технологий (зав. кафедрой, канд.
геогр. наук, доцент **И. В. Закиров**);

проректор по развитию на основе анализа данных Вятского государственного университета, д-р социол. наук, профессор **Е. В. Митягина**

ISBN 978-5-7944-4238-0

© ПГНИУ, 2025

© Коньшев Е. В. (составление), 2025

ОГЛАВЛЕНИЕ

Введение.....	
1. Программа курса.....	
2. Материалы по курсу.....	
Тема 1. Введение в анализ больших данных.....	
Тема 2. Методика текстового анализа в общественной географии...	
Тема 3. Сбор данных из открытых источников и разметка текста....	
Тема 4. Обработка данных low-code и no-code инструментами.....	
Тема 5. Текстовая аналитика на платформе PolyAnalyst.....	
Тема 6. Инструменты создания пользовательских отчетов.....	
3. Методические указания для обучающихся по освоению дисциплины.....	
4. Вопросы для подготовки к итоговому контролю.....	
5. Список рекомендуемых источников.....	
6. Приложения.....	

ВВЕДЕНИЕ

Учебный курс «Текстовая аналитика в общественной географии» находится на стыке общественной географии, социологии, информатики, экономики и психологии. В рамках данного курса обучающиеся приобретут знания, умения и навыки работы с текстовыми данными и их применения в общественно-географических исследованиях. Обучающиеся получат представления о направлениях и принципах текстового анализа, правовых и этических нормах применения данных, познакомятся с реализованными проектами, приобретут умения по сбору данных из открытых источников и их обработке, научатся работать с low-code инструментами, обрабатывать текстовую аналитику на платформе PolyAnalyst, визуализировать данные и презентовать их. Акцент в подготовке планируется сделать не только на технических аспектах и алгоритмах обработки данных, но и их интерпретации, пространственному анализу и выявлению закономерностей, применению полученных результатов для целей территориального планирования и управления.

Цель учебно-методического пособия – повышение эффективности учебного процесса и самостоятельной работы студентов по освоению дисциплины, приобретение обучающимися знаний, умений и навыков работы с текстовыми данными для проведения исследований в общественной географии.

Задачами являются:

- 1) Формирование знаний об особенностях применения методов текстового анализа в общественной географии.
- 2) Приобретение умений, необходимых для работы с инструментами, применяемых в текстовом анализе данных.
- 3) Формирование практических навыков по обработке текстовых данных, их анализу и визуализации.

Для выполнения практических и лабораторных работ требуется специализированное программное обеспечение (Orange, Gephi и PolyAnalyst).

1. ПРОГРАММА УЧЕБНОГО КУРСА

Введение. Цель, задачи и структура курса. Место общественной географии в системе наук и ее роль в решении общегосударственных задач. Актуальность и конструктивность географических исследований.

Тема 1. Введение в анализ больших данных. Общее понятие о больших данных. Источники данных. Правовые и этические аспекты использования данных. Принципы текстовой аналитики и разработка концепции исследования. Обзор направлений применения методов текстового анализа в общественных науках.

Тема 2. Методика текстового анализа в общественной географии. Направления текстового анализа в общественной географии. Выявление проблемы и разработка концепции исследования. Обзор реализованных кейсов. Оценка восприятия туристско-рекреационного пространства. Анализ социального самочувствия населения в муниципальных образованиях.

Тема 3. Сбор данных из открытых источников и разметка текста. Принципы составления технического задания. Процедура сбора данных. Разметка данных.

Составление технического задания. Составление таблицы по источникам данных. Сбор данных. Аннотирование данных. Контроль качества. Обучение и тестирование модели. Разметка текста для задач обработки естественного языка (классификация текста, аннотирование эмоциональной наполненности, категоризация тем, категоризация языка). Аннотирование сущностей. Извлечение и тегирование ключевых фраз или слов. Тегирование частей речи.

Тема 4. Обработка данных low-code и no-code инструментами. Основные функциональные возможности Orange и Gephi. Загрузка и обработка данных в Orange. Анализ социальных сетей с помощью Gephi

Тема 5. Текстовая аналитика на платформе PolyAnalyst. Обзор возможностей платформы PolyAnalyst. Функциональные узлы платформы PolyAnalyst для работы с текстовыми данными. Применение инструментов: модификация колонок, фильтрация строк, поисковые запросы, индексирование, проверка орфографии, выявление уникальных записей и текстов, извлечение ключевых слов, таксономия, кластеризация, создание облака ключевых слов, связь терминов, извлечение сущностей, анализ тональности. Интерпретация результатов.

Тема 6. Инструменты создания пользовательских отчетов. Шаблоны визуализаций. Построение и настройка дашборда. Инструменты визуализации PolyAnalyst и Visiology. Приемы визуализации данных. Процедура выбора и настройки цветовой палитры, типа визуализации, области детализации, включения графических объектов. Построение и настройка дашборда. Публикация отчетов.

2. МАТЕРИАЛЫ ПО КУРСУ

ТЕМА 1. Введение в анализ больших данных

Лекция 1. Общее понятие о больших данных. Источники данных. Правовые и этические аспекты использования данных

В 2017 году в Российской Федерации была принята Программа «Цифровая экономика», направленная на развитие инновационного потенциала страны. Большие данные стоят на первом месте в качестве «сквозных технологий», то есть в качестве прорывных цифровых технологий. Более того, большие данные упоминались в принятой в 2013 году. Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы.

Большие данные (англ. Big Data) – это термин, обозначающий огромные неупорядоченные объемы данных, которые постоянно с огромной скоростью генерируются из множества источников [1].

Свойства и особенности. Big Data кардинально отличаются от традиционных структурированных данных. Во-первых, большие объемы данных измеряются в терабайтах, петабайтах или даже экзбайтах. Это связано с постоянным ростом данных, генерируемых пользователями, датчиками и другими источниками. Во-вторых, необходима высокая скорость обработки данных в реальном времени или с минимальными задержками: процесс обработки и анализа должен быть масштабируемым и способным обрабатывать данные на высоких скоростях. В-третьих, разнообразие поступающих данных происходит из-за того, что они загружаются с различных источников, таких как социальные сети, сенсоры интернета вещей и веб-сайты. Они могут иметь различную природу – тексты, изображения, аудио, видео или иные показатели. Также они бывают структурированными, полуструктурированными и неструктурированными. В-четвертых, необходима проверка подлинности данных для обеспечения их целостности и точности, а также избежания подделки или изменения данных. В-пятых, требуются сложные алгоритмы и инструменты, такие как машинное обучение и искусственный интеллект для обнаружения и анализа сложных связей, шаблонов или трендов, так как традиционным методам это не под силу.

В качестве основных **источников больших данных** выступают:

- Интернет: глобальная сеть содержит огромный объем информации в виде сайтов, блогов, почты, чатов, видео и другого контента.
- Мобильные устройства: смартфоны и планшеты генерируют данные о звонках, сообщениях, фото, видео, геолокации, использовании приложений.
- Различные приборы: данные поступают от многочисленных датчиков и сенсоров, которые с высокой точностью измеряют температуру, давление, влажность, освещенность, звук и другие физические параметры.
- Медицинская техника, предоставляющая данные анализов, снимков, электрокардиограмм, томографии для диагностики и лечения.
- Спутники: они делают снимки и карты Земли, а также собирают данные о других космических объектах.
- Социальные сети, видеорекамеры, умные дома и автомобили.

Для сбора информации из этих источников используются специальные программы, которые называются краулерами (crawlers), парсерами (parsers), скраперами (scrapers) или коллекторами (collectors). Эти программы способны автоматически обходить веб-страницы, извлекать нужную информацию, преобразовывать ее в нужный формат и отправлять в систему хранения.

При извлечении данных важно учитывать:

- Тип и формат данных: текст, изображение, видео или что-то другое. Для каждого типа подходят свои методы сбора и обработки.
- Доступность данных: открытые или закрытые. Нужно соблюдать правила и учитывать ограничения на их использование.
- Разрешение: выбирать оптимальное качество данных для анализа.
- Частоту и объем поступления данных: подбор подходящих протоколов и форматов передачи данных.

Учет всех этих факторов позволит максимально эффективно извлечь данные из первоисточников для дальнейшего анализа и использования. Это критически важный этап работы с Big Data.

Хранение больших объемов данных – сложная задача, требующая тщательно продуманного подхода. Традиционные реляционные базы данных не подходят для работы с огромными массивами информации, поскольку не масштабируются и не отвечают требованиям по продуктивности

и гибкости. Чтобы решить данную проблему, применяют специализированные системы, такие как: NoSQL – базы данных и распределенные файловые системы. Также для хранения Big Data активно используются облачные решения, такие как Amazon S3, Google Cloud Storage, Azure Blob Storage. Они предоставляют практически неограниченное хранилище данных и избавляют от необходимости закупать и обслуживать дорогостоящее оборудование.

Перед проведением анализа данных необходимо выполнить **процесс подготовки данных** для того, чтобы устранить шум, исправить ошибки и привести данные в удобный для анализа формат. Обработка больших данных включает в себя следующие этапы:

- Очистка – удаление лишних, ошибочных или некорректных данных. Устранение дубликатов, пропусков, шумов.
- Интеграция – объединение данных из разных источников в единый набор. Соединение таблиц, приведение к одному формату.
- Трансформация – изменение формы или содержания данных. Происходит фильтрация, сортировка, группировка, вычисление новых значений.
- Редукция – сокращение объема данных за счет удаления или сжатия избыточной информации.
- Агрегация – обобщение данных путем вычисления итогов по группам. Подсчет средних, минимумов, максимумов, сумм.

Для реализации этих этапов применяются специальные технологии и инструменты: Hadoop, Spark, Hive и Pig, а также параллельные вычисления, распределенные вычисления, потоковая обработка, облачные вычисления и другие технологии.

Аналитика больших данных (Big Data Analytics) – это процесс извлечения полезных знаний из огромных массивов информации при помощи специальных технологий. Цель – выявить закономерности, тенденции, корреляции, прогнозы.

Существует четыре типа анализа:

- Описательная аналитика показывает общую картину данных. Использует статистику для подсчета, измерения, визуализации. Помогает увидеть основные характеристики наборов данных.

- Диагностическая аналитика выявляет причины явлений в данных. Определяет факторы, влияющие на результаты или поведение систем. Помогает найти корень проблем.

- Предиктивная аналитика прогнозирует будущее на основе прошлого и настоящего. Использует машинное обучение для построения моделей и выработки прогнозов по новым данным.

- Прескриптивная аналитика предлагает оптимальные решения и действия для достижения нужного результата. Применяет методы оптимизации и поиска лучших вариантов.

При использовании больших данных возникают следующие **проблемы**:

- нарушение приватности;
- проблемы кибербезопасности;
- проблемы этики;
- низкое качество данных из-за шумов, неполноты, противоречий.

Для снижения рисков необходимы меры защиты данных: шифрование, аудит, анонимизация; соблюдение этических и правовых норм при работе с Big Data. Ключевыми принципами являются прозрачность, ответственность, справедливость, согласие пользователя.

Правовые и этические аспекты использования данных. Среди ключевых проблем использования больших данных – правовые и этические. С одной стороны люди, используя социальные сети, добровольно публикуют личную информацию, делятся фотографиями, новостями. С другой стороны, при заполнении профиля в социальных сетях некоторые поля могут по ошибке или преднамеренно оставаться незаполненными или содержать ложную информацию о фактах биографии, интересах и предпочтениях [3]. С правовой точки зрения сбор информации о респонденте без прямого обращения к нему считается нереактивным методом исследования, примером которого являются «большие данные» (Big Data). В России официально действует федеральный закон «О персональных данных» от 27.07.2006, регулирующий отношения, связанные с обработкой персональных данных юридическими и физическими лицами с использованием средств автоматизации, в том числе в информационно-телекоммуникационных сетях.

С точки зрения этики сбор данных зачастую нарушает принцип добровольности участия в исследовании. Также возникают опасения, связанные с возможным причинением вреда здоровью испытуемых и вторжением в их личную жизнь в случае деанонимизации персональных данных, собранных в ходе эксперимента. Поэтому для обеспечения этических требований исследователи предлагают выполнять следующие шаги:

- предоставить пользователям подробную информацию о том, какие данные будут использоваться для исследования и кто будет иметь к ним доступ;
- предоставить пользователям информацию о текущих результатах исследования;
- сообщить пользователям о публикации результатов исследования.

Большинство социальных сетей в правилах указывают, что пользователи, создавая аккаунт, автоматически соглашаются на доступность информации о них широкой аудитории. Однако, учитывая этические нормы, при работе с текстовыми данными особое внимание необходимо уделить следующим условиям:

- а) полная анонимизация данных и отсутствие возможности их деанонимизировать;
- б) отсутствие взаимодействия между участниками выборки;
- в) опубликованные результаты не должны давать возможность раскрыть участников исследования [2].

Подробнее о правовых и этических нормах можно узнать здесь: <http://ipcmagazine.ru/legal-issues/legal-regulation-of-big-data-foreign-and-domestic-experience>

Список использованных источников

1. Big Data: основные понятия. URL: <https://elbrusboot.camp/blog/big-data-osnovnyie-poniatiia/>
2. Соснин К. А. Правовое регулирование Больших данных: зарубежный и отечественный опыт // Журнал Суда по интеллектуальным правам. 2019. № 25. С. 30–42. URL: <http://ipcmagazine.ru/legal-issues/legal-regulation-of-big-data-foreign-and-domestic-experience>). EDN: YVVHEA
3. Щеглова И. А. Этические и правовые аспекты использования данных из социальных медиа // Вестник Томского государственного университета. 2018. № 431. С. 81-87. DOI: 10.17223/15617793/431/10, EDN: UYPPEI

Задание для практической работы

Напишите эссе о роли больших данных в развитии современного общества.

Общая оценка состоит из 10 баллов, выставляемым по следующим параметрам:

- соответствие текста эссе выбранной теме;
- связь с географическими науками;
- убедительность приводимых аргументов;
- согласованность, последовательность, логичность изложения;
- оригинальность эссе.

Самостоятельная работа

Изучите статью о правовых и этических нормах использования больших данных (<http://ipcmagazine.ru/legal-issues/legal-regulation-of-big-data-foreign-and-domestic-experience>).

Сделайте краткий конспект.

Вопросы для самопроверки

1. Назовите свойства и источники больших данных.
2. Из каких этапов состоит процесс обработки данных?
3. Дайте характеристику правовых и этических проблем применения больших данных.

Лекция 2. Принципы текстовой аналитики

Лавинообразный рост объема текстовых данных, которые могут быть использованы в пространственном анализе, связывают с распространением сети Интернет и ростом популярности социальных сетей.

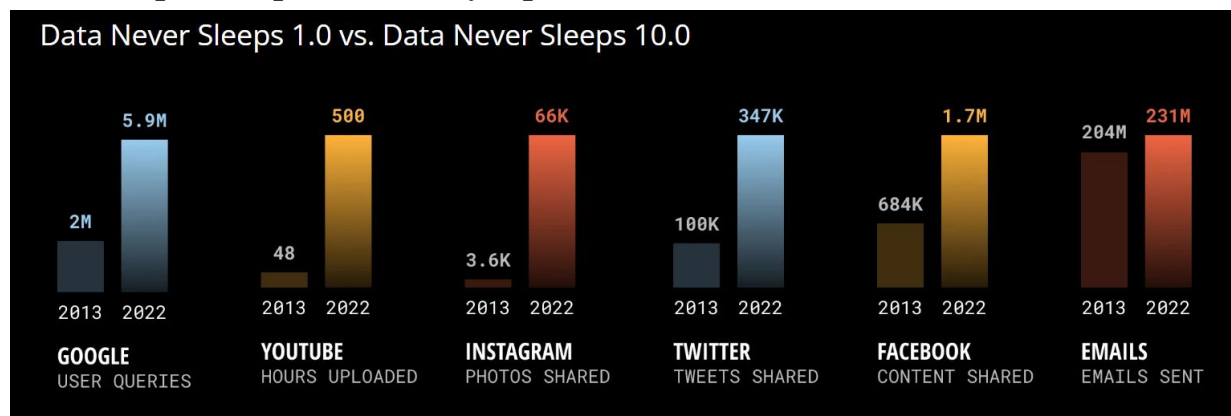


Рис. 1. Объем данных, которые ежеминутно генерируется в Интернете (сравнение 2013 и 2022 годов, по данным Data Never Sleeps [1])¹.

По состоянию на апрель 2022 года Интернетом пользуются 63% населения мира, что составляет примерно 5 миллиардов человек. Из этого общего числа 4,65 миллиарда, более 93%, были пользователями социальных сетей (рис. 1). По данным Statista, общий объем данных, который, по прогнозам, будет создан, захвачен, скопирован и использован во всем мире в 2022 году, составит 97 зеттабайт, а к 2025 году это число вырастет до 181 зеттабайта [1].

Весь объем данных условно можно разделить на **структурированные, частично структурированные** и **неструктурированные**. Следует учитывать, что чем менее структурированы данные, тем сложнее с ними работать и применять в исследовательских моделях. Для пространственного анализа, который является одним из наиболее применяемых методов в общественной географии, более всего подходят именно структурированные количественные данные. Однако примерно 80% всех цифровых данных относятся к неструктурированным.

К **неструктурированным** данным относятся те, что не соответствуют заранее определенной модели данных с произвольным расположением в

¹ Социальные сети Facebook и Instagram принадлежат компании Meta, признанной экстремистской и запрещенная на территории Российской Федерации. Доступ к Twitter также заблокирован.

тексте необходим для исследования элементов. Неструктурированные данные являются частью более обширного понятия Big Data (в него входят структурированные и частично структурированные данные). Массив неструктурированных данных может формироваться из выгрузки постов, комментариев, профилей пользователей социальных сетей, из множества текстовых документов (например, техническое описание изделий, товарные накладные и пр.), из электронных писем и даже из совокупности изображений, видео, показаний датчиков. Можно предположить, что весь этот объем неструктурированных данных содержит в себе нужные для аналитики компоненты, но без предварительной обработки извлечь их крайне трудно.

Выделяется два основных типа неструктурированных данных [2]:

- *генерируемые людьми* – к ним относятся различные виды создаваемого людьми контента (текстовые документы, электронные письма, посты в соцсетях, изображения, видео);

- *машиногенерируемые* – она создаются устройствами и датчиками (это файлы журналов, данные GPS, результаты работы Internet of Things (IoT) и другая телеметрическая информация).

Наиболее востребованными в общественно-географических исследованиях являются данные текстовых документов, выгрузки из социальных сетей и специализированных сайтов, изображения, данные датчиков.

В некоторых случаях для анализа текстовых данных не требуется применения специальных программных продуктов и методов. Достаточно прочитать книгу, несколько статей или постов на стене сообщества в социальной сети. Но если требуется изучить десятки, сотни тысяч постов из социальных сетей, чтобы выявить наиболее популярные темы и оценить тональность их восприятия, то необходимо применить автоматический анализ текстов с применением методов машинного обучения. Автоматическая обработка естественного языка (Natural Language Processing или NLP) – это отрасль компьютерных наук и искусственного интеллекта, основной целью которой является разработка технологий для взаимодействия человека и компьютера с использованием естественных языков [3].

Основные возможности, предоставляемые текстовыми данными в области получения данных:

- Извлечение данных из множества различных источников. Смещение акцента с поиска на обработку извлеченных данных.

- Поиск и извлечение информации в режиме реального времени.
- Определение типов данных. Выбор обработчика в зависимости от типа данных. Обработка аудио- и видеоинформации.

- Широкие возможности по извлечению данных из социальных сетей, публичных источников данных, наиболее посещаемых сайтов и специализированных источников.

Основные преимущества в области категоризации текстовых данных:

- Кластеризация документов, создание и наполнение категорий в документах. Выделение параграфов и разделов документов с использованием алгоритмов кластеризации.

- Группировка и категоризация документов и частей документов в соответствии с заданными сценариями.

- Использование самообучаемых (статистических) алгоритмов и алгоритмов обучения с учителем.

- Использование категорий для последующего извлечения информации из неструктурированного текста.

Основные возможности применения методов текстового анализа для извлечения данных:

- Извлечение значимой информации из текста.

- Выделение собственных имен, людей, организаций, географических мест, наименований, ссылок на сайты, e-mail, сообщений на форумах, сообщений в соцсетях и пр.

- Выделение продуктов/услуг и их характеристик, в том числе характеристик, связанных с обслуживанием клиентов.

- Наполнение данными CRM-систем. Поиск и наполнение данными о профилях клиентов, отношениях между клиентами, выявление родственных связей, выявление степени влияния между клиентами.

Основные возможности применения методов текстового анализа для анализа мнений:

- Анализ мнений используется для дополнительного извлечения скрытой информации, которую сложно извлечь, используя классические методы.

- Анализ мнений включает выявление тональности сообщений, определение эмоциональной составляющей высказывания, выделение суждений, в которых проявляются мечты и намерения клиента.

- Использование результатов оценки тональности для разработки сценариев снижения негатива со стороны клиентов.

- Анализ желаний и мечтаний клиентов для разработки новых продуктов и новых маркетинговых стратегий.

Основные возможности применения машинного обучения для анализа текстовых данных:

- Использование математических алгоритмов для обучения компьютера анализу текстовой информации.

- Использование комплексных методов обучения: статистических методов, нейронных сетей, метода энтропии, дерева решений и пр.

- Использование машинного обучения вместе с разбором вручную для повышения качества работы ТА.

- Разработка новых алгоритмов машинного обучения, направленных на подражание человеку для автоматизации диалога с клиентом.

Этапы текстового анализа данных:

1. Предобработка текстовых данных. Тексты не унифицированы сами по себе: они могут быть написаны на разных языках, состоять из эмодзи, иметь ошибки, сокращения, цифры и знаки препинания. Разнообразие элементов делает текст более сложным для анализа. Чтобы работать с текстовыми данными, их необходимо упростить и привести к стандартной форме, которая подходит для используемого алгоритма. Вот примерный список того, что нужно или можно (в зависимости от алгоритма) сделать:

- удалить знаки препинания;
- заменить заглавные буквы на строчные;
- удалить лишние переносы, табуляцию и знаки переноса;
- удалить стоп-слова;
- лемматизировать слова;
- токенизировать текст.

Стоп-слово – обычно это служебная часть речи, которая употребляется так часто, что отдельное ее употребление не вносит в текст какой-то отдельный особенный смысл.

Токенизация – это разделение текста на единицы, которые будут анализироваться в ходе текстового анализа.

Лемматизация помогает сократить количество токенов в документе. Идея заключается в том, чтобы привести разные уникальные формы слов к начальной, исходной форме.

2. Извлечение именованных сущностей. *Именованные сущности* (named entities) – уникальные идентификаторы сущностей, упомянутых в тексте (даты, имена, названия, адреса). Извлечение именованных сущностей (named entity recognition, NER) – это извлечение конкретной информации из текстовых данных. Дальше именованные сущности могут, например, использоваться как категориальные переменные в исследованиях, где единица наблюдения – текст, а именованные сущности – одна из возможных характеристик текста. Бывают случаи, когда данные должны остаться обезличенными, тогда извлечение сущностей используется для их удаления из текста. Например, имена из медицинских карт, даже если они попали к исследователю легально, стоит извлечь и сразу удалить их из текста. Выделение и распределение обычно происходит с помощью предварительно составленного словаря, который исследователь дополняет необходимыми ему сочетаниями. Альтернативный способ – это использование правил, позволяющих находить именованные сущности в соответствии с их положением. Также можно задать пользовательский набор именованных сущностей и классов в дополнение к тому, на который опирается исследователь.

3. Анализ тональности. Анализ тональности (sentiment analysis) – выявление эмоциональной оценки авторов текстов (тональности) или эмоционально-окрашенных слов по отношению к объектам, предметам или явлениям, о которых идет речь в текстах. Сформулировать алгоритм, который позволил бы делать аналогичную оценку, не так просто. Анализ тональности – одна из задач обработки естественного языка. Базовая задача определения тональности состоит в том, чтобы понять окраску текста: отрицательную или положительную. Как и любая другая переменная, тональность сообщения должна быть измерена с помощью какой-то шкалы. Самый простой вариант – это бинарный классификатор, состоящий из двух типов тональности: положительной и отрицательной.

Но это не единственный вариант. Например, мы можем кодировать тональность пятью категориями по степени согласия (ликертова шкала):

- крайне положительная;
- положительная;
- нейтральная;
- отрицательная;
- крайне отрицательная.

Более совершенные методы определяют шкалы, связанные с эмоциональным состоянием: радость, грусть, злость и так далее. Однако все усложнения бинарного классификатора делают алгоритм сложнее и менее надежным.

При анализе тональностей нужно определить единицу, тональность которой мы оцениваем: предложение, абзац, полный текст. Для анализа тональностей рекомендуется разбивать текст на отдельные предложения и определять тональность каждого предложения. Например, комментарий из социальных сетей делят на отдельные предложения и определяют тональность каждого предложения, после высчитывается оценка тональности всего комментария.

Для анализа тональностей хорошо подойдут сообщения из социальных сетей, отзывы на продукты и сервисы, новости из медиа, тексты произведений и тексты с веб-страниц. С помощью анализа тональности в социальных сетях описывают реакции пользователей на определенные события. Не советуем проводить анализ тональности в текстах, содержащих формальную лексику. Например, юридические документы или отчеты о проделанной работе предполагают небольшое количество эмоционально-окрашенной лексики, поэтому в них сложно вычленить тональность.

Перед анализом тональности рекомендуется не удалять стоп-слова в ваших данных. Наличие или отсутствие частицы может полностью изменить тональность единицы текстового анализа.

Список использованных источников

1. Data Never Sleeps (ежегодная инфографика) //URL: <https://www.domo.com/data-never-sleeps>
2. Неструктурированные данные: примеры, инструменты, методики и рекомендации // URL: <https://habr.com/ru/articles/756454/>
3. Сосновская Я. Прикладной анализ данных в социальных науках / электронный учебник. URL: <https://education.yandex.ru/handbook/data-analysis/article/tekstovye-dannye-i-sposoby-ih-obrabotki>

Самостоятельная работа

Изучите отчет Data Never Sleeps по использованию данных (<https://www.domo.com/data-never-sleeps>). Сделайте краткое сообщение на семинаре.

Вопросы для самопроверки

1. Какие основные возможности предоставляются текстовыми данными в области получения данных?
2. Дайте характеристику этапа «предобработка текстовых данных».
3. Назовите основные рекомендации по оптимальной обработке неструктурированных данных.

Лекция 3. Обзор направлений применения методов текстового анализа в общественных науках

Методы текстового анализа с использованием различных аналитических платформ широко применяются для изучения общественных явлений и процессов. Примеры реализованных кейсов представлены в галерее дашбордов (<https://gallery.pa6.megaputer.ru>). Все примеры визуализаций построены на данных из открытых источников, на которые указывают ссылки. Подробнее о функциональных возможностях для визуализации результатов исследования можно узнать на платформе PolyAnalyst (<https://gallery.pa6.megaputer.ru/polyanalyst/static/paclient/publication-view.html?reportUUID=2dd651ba-0837-42f9-bb54-1ff3daf4d60c&locale=rus&guest=1#sheetId=98>).

Также рекомендуется ознакомиться с работой помощника выбора визуализаций (https://disk.yandex.ru/i/jJ_ovdiAkGLxjQ) и с аннотациями статей из научных журналов, посвященных изучению общественных и гуманитарных проблем методами текстового анализа.

Текстовый анализ данных в социологии. Применение текстовых данных в социологии основывается на стремлении изучать поведение людей в цифровом пространстве. Так, в некоторых статьях обсуждается возможность использования нового источника данных – цифровых следов пользователей – для изучения качества жизни и благополучия населения. Представлен обзор актуального состояния исследований качества жизни и благополучия населения, которые используют в качестве источника данных цифровые следы (социальные медиа, поисковые запросы, сервис Google Books и т. д.) и методы цифровой социологии (большие данные, машинное обучение) для анализа и обработки этих данных [8].

Текстовый анализ данных в образовании. В исследованиях акцент сделан на выявлении отношения целевых групп к различным процессам в

образовании и их поведения в цифровом образовательном пространстве. Так, в одной из статей рассматриваются трудности российского студенчества в условиях перехода на дистанционное обучение в связи с пандемией новой коронавирусной инфекции COVID-19 на примере четырех вузов Приморского края. На основе предварительно выгруженных с использованием специализированного программного обеспечения (Polyanalyst, библиотек машинного обучения) сообщений сообществ вузов региона была проведена разметка всех сообщений по релевантности изучаемой темы, что позволило в дальнейшем систематизировать сообщения по тематике и тональности. В тематике сообщений были выделены такие группы, как: организационно-методические вопросы, вопросы социально-финансовой сферы, самочувствие и личные качества, другие. В тональности сообщений были выделены группы сообщений негативной тональности, нейтральной и позитивной тональности. Лемматизация слов в сообщениях позволила определить топ-150 слов. На основе проведенного анализа содержания, тональности сообщений студентов, опубликованных в социальных сетях вузов, были выявлены основные трудности, с которыми столкнулись студенты в связи с переходом на дистанционное взаимодействие в образовательном процессе в условиях карантина [3].

В другой статье представлены результаты исследования субъективного мнения населения о качестве образования в Республике Саха (Якутия) (рис. 2).



Рис. 2. Структура тональности высказываний по сюжетам в категории «Школа» [5].

Данное исследование проведено на основе методов текстового анализа контента социальных сетей (цифровых следов) с использованием Big Data и информационно-аналитической системы PolyAnalyst.

Исследование предполагало осуществление сбора данных в социальных сетях (постов и комментариев, релевантных цели исследования), предобработку данных (очистку от «мусорных» сообщений) и непосредственно анализ. На этапе анализа сообщения были категоризованы. Как наиболее крупная определилась категория «школа». Во всех категориях, в том числе рассматриваемой, были выделены ключевые слова, соответствующие основным сюжетам, сделаны оценка тональности сообщений и статистический анализ. В результате были выявлены проблемные и актуальные, с точки зрения населения, вопросы школьной системы образования и дана их оценка [5].

Текстовый анализ данных в лингвистике. В лингвистике текстовые данные рассматриваются как источник для моделирования. Примером может служить статья, посвященная описанию возможностей системы Polyanalyst в качестве инструмента анализа лингвистического материала на примере текстов о специальной военной операции (СВО). Показано, что предлагаемый авторами метод обладает возможностями текстовой аналитики: рассматривается как один из результативных инструментов, который целесообразно использовать при выделении основных тем и «сущностей» сообщений, семантических связей между ними, а также определения тональности высказываний, что позволяет оценить степень воздействия дискурсивных моделей. Статья сопровождается иллюстративным материалом, который позволяет раскрыть специфику и возможности предлагаемого инструмента. Проведенный анализ позволяет отнести систему Polyanalyst к эффективным методам отслеживания воздействующего ресурса через вербальные модели высказываний и настроений пользователей медиaproстранства в разные временные отрезки [4].

Текстовый анализ данных в политологии. Трансформации современных социально-политических процессов, переживание общества в состоянии изменившейся социальной реальности происходят на фоне глобальной цифровизации. Понятие «мобилизация» в обучении может быть интерпретировано двояко: как научное понятие и как определение процессов, происходящих в современном российском обществе (включая военно-

мобилизационный аспект). Мобилизация сетевых акторов в условиях реактивных отношений актуализирует использование в исследованиях интеллектуального анализа – интеллектуального анализа данных. Одной из целей применения текстовых данных в политологии может быть изучение возможностей исследования социальных и политических мобилизаций в международных сетевых сообществах, а также определение участников и траекторий мобилизации с помощью инструментов, управления интеллектуальным анализом данных.

В исследованиях применяются возможности аналитической платформы PolyAnalyst. Так, в одной из статей в качестве исследовательской базы авторы представили наборы актуальных данных региональных международных сообществ Кузбасса и Тюмени (без учета автономных территорий). Они включают 97 уникальных международных сообществ Кузбасса и 41 уникальное сообщество Тюменской области и тематически связаны с экономическими и политическими проблемами. Исследователи фиксируют, что региональные сетевые сообщества в состоянии следовать принципу объединения уникальных мнений и оценок и одновременно лучше фиксировать отношение пользователей к глобальным сетевым идентичностям, появляющимся в режиме реального времени. Изучение социальной и политической мобилизации в сети региональной среды позволяет претендовать на построение сравнительной модели с неочевидными внешними связями участников, внешними ресурсами, уникальной жизненной ситуацией и постоянными ситуациями, семантикой дискурса [6].

Еще одним направлением является измерение цифровой мобилизации региональных социальных и политических взаимодействий в сложившейся реальности. Помимо очевидных исследовательских инноваций, перехода к интеллектуальному цифровому исследовательскому алгоритму, поиска нетривиальных решений в изучении мобилизации ученые получают в свое распоряжение действенный инструмент n-кода измерений. Возможности цифровой свободы формируют новые широкие алгоритмы доступности сетевых инструментов и сервисов, превращают рядовых пользователей в акторов цифровой мобилизации.

В интересах визуализации результатов исходные данные преобразованы в программе PolyAnalyst. Получены тематические группы и проведена силовая укладка ребер и узлов акторно-сетевой модели сетевой пользова-

тельской коммуникации сообществ Кемеровской области – Кузбасса и Тюменской области. На основе сформированного датасета рассмотрены общие и различающиеся темы обсуждения. Выделены группы тем, объединяемые острой и насущной проблематикой обсуждаемых вопросов (инфраструктура, насилие, массовые мероприятия, правопорядок, специальная военная операция и пр.). В иллюстративном примере предложены приемы пользовательского поиска и обработки социальных запросов чат-бота искусственным интеллектом (рис. 3).

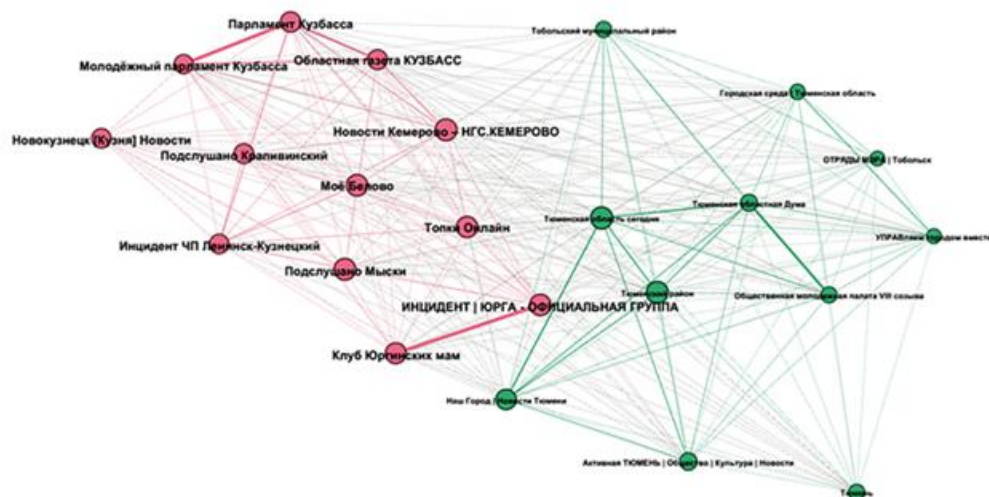


Рис. 3. Граф связей сообществ-лидеров сетевой коммуникации Кемеровской и Тюменской областей по общим подписчикам (декабрь 2022 года.) [2].

Результаты использования цифровых инструментов демонстрируют региональные особенности выбора доверительных источников информации, сетевых медиа, местных ресурсов и отработки связанных тем в пространстве Интернета. Пользовательское участие в сетевых обсуждениях порождает новые практики коммуникации – возможность получать, обсуждать и формировать актуальную тематическую повестку в режиме реального времени, подчас опережая возможности представителей власти и официальные СМИ [2].

Текстовый анализ данных в экономике. Широкое применение текстовых данных возможно и в экономике. Показательной является статья по анализу рынка сбыта продукции северного оленеводства России на основе анализа массивов больших данных. В арктическом регионе продукция оленеводства является неотъемлемым элементом культуры северных коренных народов, который внес значимый вклад в повышение адаптации человека к экстремальным климатическим условиям. Целью исследования стало изучение предложений о реализации продукции оленеводства в сети

Интернет для определения регионов Крайнего Севера РФ, наиболее активно использующих возможности информационных коммуникаций для продвижения своей продукции на внутреннем региональном и межрегиональном рынках, и для анализа ассортиментного ряда продукции арктического традиционного питания. Методологической основой исследования стали теоретические подходы пространственной экономики А. Г. Гранберга и теории свободной торговли П. Р. Кругмана. Применены методы анализа массивов больших данных на основе дата-майнинга в социальных сетях. Материалом исследования послужили сообщения в сети Интернет, тематически связанные с реализацией продукции оленеводства в регионах Крайнего Севера в 2021-2022 годах. Новизна исследования заключается в применении методов сбора и анализа массивов больших данных, что потребовало адаптации решений платформы PolyAnalyst к задачам конкретной научной области и совершенствования методологии анализа. В результате анализа предложений о реализации продукции оленеводства в сети Интернет были выявлены наиболее активные арктические регионы: Мурманская область, Ямало-Ненецкий автономный округ, Республика Коми, Чукотский автономный округ. «Мода» на Арктику способствовала формированию широкого ассортимента готовых блюд и способов переработки оленины, а также стала катализатором развития арктического гастрономического туризма, который создает важные предпосылки для экономического роста сельских арктических территорий. Организация продаж через сеть Интернет расширяет возможности экономики арктического этнического предпринимательства за счет преодоления относительной замкнутости арктических экономических систем и частичного перевода товарооборота в интернет-пространство. Соответственно, перспективами исследования может стать использование анализа больших данных рынка оленеводства для разработки мероприятий по управлению продажами продукции оленеводства [1].

Текстовый анализ данных в экологии. Весьма интересным представляется опыт применения методов текстового анализа данных в экологии. Так, в одной из статей представлены результаты эмпирического исследования сообщений в социальной сети «ВКонтакте» о качестве атмосферного воздуха в г. Череповце. Методология включает в себя отбор релевантных сообществ в социальной сети «ВКонтакте», относящихся к Череповцу;

выгрузку и автоматическую классификацию сообщений; анализ данных методом тематического моделирования и контент-анализ. Было выделено 48 сообщений, затрагивающих проблему качества атмосферного воздуха в этом городе. Чтобы проанализировать выбранные тексты, использовалась платформа для анализа данных PolyAnalyst (рис. 4).

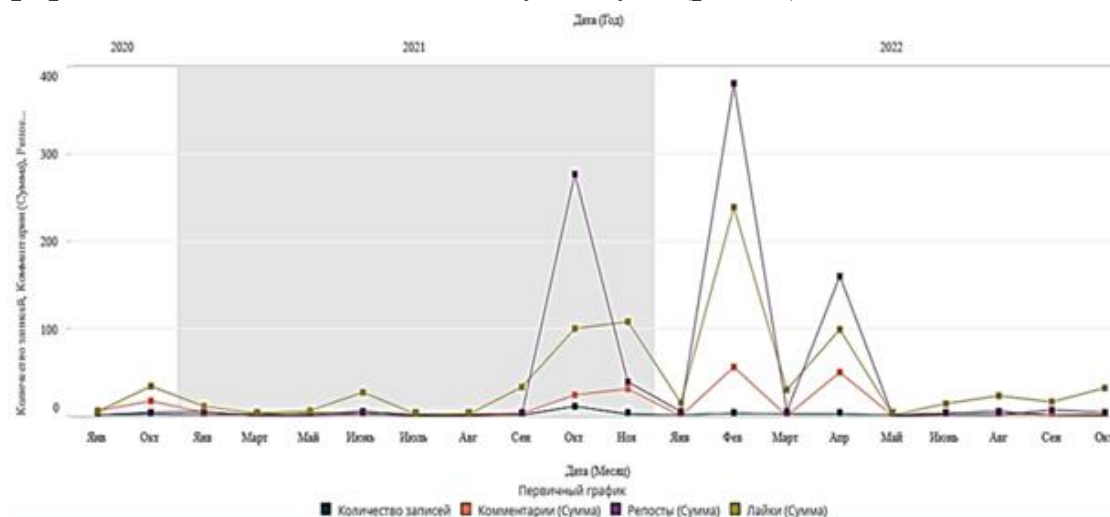


Рис. 4. Динамика цифровых следов [6].

Исследование охватывает период с 01.01.2020 по 31.10.2022. Представлено 4 наиболее популярных сообщения, в которых обсуждается проблематика качества атмосферного воздуха. Наиболее часто встречающиеся словосочетания в выделенном датасете: загрязняющее вещество, атмосферный воздух, вредное вещество, выброс аммиака, жидкие комплексные удобрения, концентрация аммиака. Выделены организации, которые встречаются в данном наборе сообщений. Результаты дают представление об оценках пользователями социальной сети «ВКонтакте» качества атмосферного воздуха в Череповце. Социальные сети могут служить хорошим инструментом для оперативного мониторинга интереса к экологическим проблемам, так как они прочно занимают первое место в формировании экологической повестки, опережая телевидение и другие источники информации [6].

Список использованных источников

1. Богданова Е. Н., Иванова М. В., Симашко Т. В. Анализ рынка продукции северного оленеводства России на основе анализа массивов больших данных // Север и рынок: формирование экономического порядка. 2023. Т. 26, № 2(80). С. 55-73. DOI 10.37614/2220-802X.2.2023.80.004. EDN: XNQYYS

2. Головацкий Е. В. Региональная специфика цифровой мобилизации социального и политического взаимодействия сетевых сообществ // Виртуальная коммуникация и социальные сети. 2023. Т. 2/ № 4(8). С. 246-254. DOI 10.21603/2782-4799-2023-2-4-246-254. EDN: VVYJAF

3. Романова Е. В. Проблемы российского студенчества в условиях карантина (на примере вузов Приморского края) / Е. В. Романова, С. А. Петрова // Международный научно-исследовательский журнал. 2020. № 12-2(102). С. 183-188. DOI 10.23670/IRJ.2020.102.12.069. EDN: UFOTSK

4. Стародубцева Е. А., Пикалов И. Ю., Кривко И. П. Система Polyanalyst как инструмент лингвистического анализа (на материале сообщений о специальной военной операции) // Теория языка и межкультурная коммуникация. 2023. № 1(48). С. 353-371. EDN: JGOERG

5. Калаврий Т. Ю., Романова Е. В., Григорьева М. В., Федорова Е. С. Текстовый анализ цифровых следов в социальных сетях для оценки мнения населения (на примере системы школьного образования в Республике Саха (Якутия)) // Современные технологии управления. – 2023. – № 4(104). EDN: LROMNV

6. Головацкий Е. В., Кранзеева Е. А., Донова И. В. [и др.] Цифровые перспективы социальной и политической мобилизации региональных сетевых сообществ Кузбасса и Тюмени с использованием data-mining // Власть. 2023. Т. 31, № 4. С. 30-40. DOI 10.31171/vlast.v31i4.9687. EDN: GCCLHW

7. Щекотин Е. В., Дунаева Д. О., Басина П. А., Вахрамеев П. С. Цифровые следы в экологии: опыт эмпирического исследования // Виртуальная коммуникация и социальные сети. – 2023. – Т. 2, № 4(8). – С. 255-263. – DOI 10.21603/2782-4799-2023-2-4-255-263. EDN: RXJPKL

8. Щекотин Е. В. Цифровые следы как новый источник данных о качестве жизни и благополучии: обзор современных тенденций // Вестник Томского государственного университета. 2021. № 467. С. 170-181. DOI 10.17223/15617793/467/21. EDN: JPNJIY

Практическая работа

Проанализируйте реализованные кейсы, которые представлены в галерее дашбордов (<https://gallery.pa6.megaputer.ru>).

Самостоятельная работа

Изучите основные публикации по применению методов текстового анализа данных в различных науках. Составьте конспект.

Текстовый анализ данных в экологии: Щекотин Е. В., Дунаева Д. О., Басина П. А., Вахрамеев П. С. Цифровые следы в экологии: опыт эмпирического исследования // Виртуальная коммуникация и социальные сети. – 2023. – Т. 2, № 4(8). – С. 255-263. – DOI 10.21603/2782-4799-2023-2-4-255-263. EDN RXJPKL.

Текстовый анализ данных в экономике: Богданова Е. Н. Иванова М. В., Симашко Т. В. Анализ рынка продукции северного оленеводства России на основе анализа массивов больших данных // Север и рынок: формирование экономического порядка. – 2023. – Т. 26, № 2(80). – С. 55-73. – DOI 10.37614/2220-802X.2.2023.80.004. EDN XNQYY5.

Текстовый анализ данных в политологии: Головацкий Е. В., Кранзева Е. А., Донова И. В. [и др.] Цифровые перспективы социальной и политической мобилизации региональных сетевых сообществ Кузбасса и Тюмени с использованием data-mining // Власть. – 2023. – Т. 31, № 4. – С. 30-40. – DOI 10.31171/vlast.v31i4.9687. EDN GCCLHW.

Текстовый анализ данных в лингвистике: Стародубцева Е. А., Пикалов И. Ю., Кривко И. П. Система Polyanalyst как инструмент лингвистического анализа (на материале сообщений о специальной военной операции) // Теория языка и межкультурная коммуникация. – 2023. – № 1(48). – С. 353-371. EDN JGOERG.

Текстовый анализ данных в образовании: Романова Е. В., Петрова С. А. Проблемы российского студенчества в условиях карантина (на примере вузов Приморского края) // Международный научно-исследовательский журнал. – 2020. – № 12-2(102). – С. 183-188. – DOI 10.23670/IRJ.2020.102.12.069. EDN UFOTSK.

Текстовый анализ данных в социологии: Щекотин Е. В. Цифровые следы как новый источник данных о качестве жизни и благополучии: обзор современных тенденций // Вестник Томского государственного университета. – 2021. – № 467. – С. 170-181. – DOI 10.17223/15617793/467/21. EDN JPNJIY.

Вопросы для самопроверки

1. Дайте характеристику основных направлений применения методов текстового анализа данных в гуманитарных науках.
2. Дайте характеристику основных направлений применения методов текстового анализа данных в общественных науках.
3. Дайте характеристику основных направлений применения методов текстового анализа данных в науках о Земле.

ТЕМА 2. Методика текстового анализа в общественной географии

Лекция 4. Направления текстового анализа в общественной географии

Объектом изучения общественной географии выступают пространственные формы организации общества, а предметом – «территориальные общественные системы» (ТОС). Грядущая эпоха цифровой экономики диктует «смену вех» в исследованиях общественной географии. Ставятся задачи по изучению информационного поля, его территориальной структуры, иерархии, функций, особенности взаимодействия с другими функциональными структурами ТОС. Обращается внимание на важность исследований нематериальных активов стран и регионов: имиджа, репутации, статуса, образа, брендов. Ставятся задачи по изучению механизма их формирования и оценки вклада в экономику, геополитику и общественное развитие [7]. При решении этих задач исследователи сталкиваются с проблемой обработки больших данных, обрабатывать которые до недавнего времени было крайне сложной проблемой. Еще сложнее было обрабатывать текстовые данные.

Применение текстовой аналитики является новым направлением в общественной географии, слабо представленным как в научных исследованиях, так и в образовательном процессе. Основные исследования проводятся за рубежом, а в России данный метод применяется в основном социологами, которые не затрагивают общественно-географические темы. В то же время анализ текстовых данных с помощью современных инструментов и методов позволяет изучать ТОС с новых точек зрения. Это и вносит определенную научную новизну, и способствует выявлению новых связей и закономерностей пространственной организации жизни людей, в том числе в цифровом пространстве.

Изучение миграций населения. Недостатком традиционных источников информации, характеризующих миграции населения (данные текущего учета и переписей населения) является отсутствие данных о внутри- и межмуниципальных миграциях. Большинство исследований, посвященных миграционным потокам, учитывают лишь региональный или национальный уровень. В то же время использование больших данных, которые генерируются социальными сетями, сотовыми операторами, банковскими транзакциями, интернет-подключениями и пр., позволяют получать данные

не об общем числе прибывших или выбывших на определенной территории, а рассматривать одновременно «виртуальные» миграционные контингенты и миграционные потоки. Также важным преимуществом «больших данных» является независимость от государства как «заказчика» информации.

Недостатки использования данных интернет-сообществ связаны с неодинаковой степенью вовлеченности представителей различных социально-демографических групп. Конечно, и у цифровых данных есть определенные недостатки. Так, некоторые большие данные, например данные интернет-сообществ, охватывает не всю генеральную совокупность социально-демографических групп, а лишь наиболее активную часть интернет-пользователей. Однако значительный охват населения и широкая распространенность социальных сетей дают возможность считать миграционные данные, полученные на их основе, репрезентативными при качественно разработанной методологии.

Приведем некоторые примеры использования больших данных для изучения миграционных процессов. В статье А. В. Смирнова цифровые двойники населения рассматриваются как инструмент социально-демографических исследований и управления территориями. Систематизирован опыт создания цифровых двойников населения и интерактивных веб-сайтов, посвященных демографической проблематике. Предложена методика создания «цифрового двойника» населения Арктики, основанная на трех методологических принципах: учете иерархии территорий, пространственном представлении данных, сочетании демографической статистики с новыми цифровыми источниками данных. Разработан «Цифровой двойник населения Арктики» – интерактивный веб-сайт, содержащий детальные данные об арктическом населении, включая данные муниципального и поселенческого уровней (рис. 5).

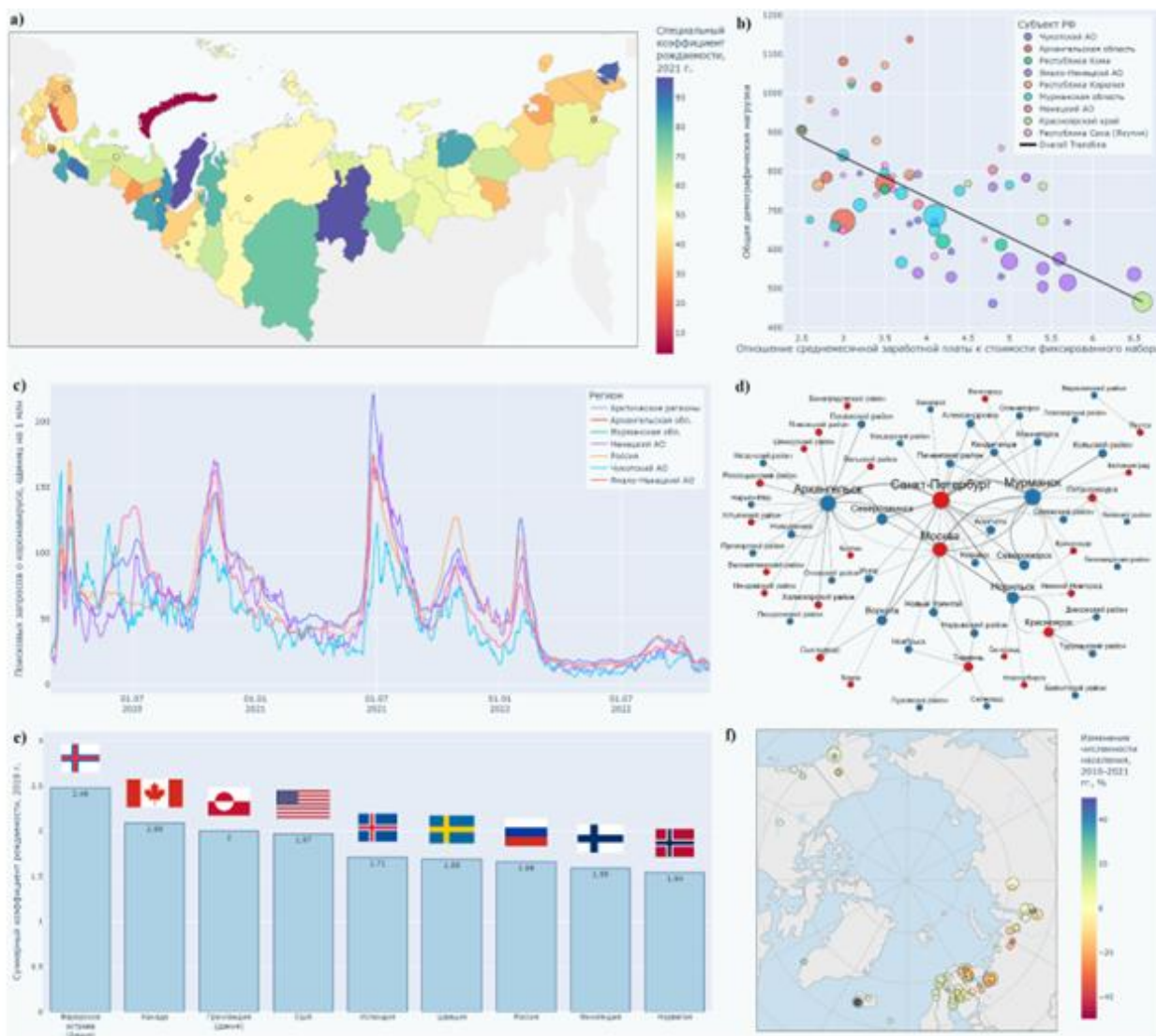


Рис. 5. Некоторые элементы дашборда цифрового двойника населения Арктики [3].

Сайт содержит демографическую статистику, итоги переписей и данные цифровых платформ. «Цифровой двойник населения Арктики» охватывает такую проблематику, как численность, динамика и состав населения, расселение, естественное и миграционное движение, труд и занятость, транспортные перемещения, наука и образование, влияние пандемии на демографические процессы. Для исследователей реализованы инструменты ранжирования, многомерного анализа, кластеризации и прогнозирования значений показателей. С точки зрения государственного и муниципального управления главный интерес представляют демографические профили регионов и территорий, отражающие подробную информацию о демографической ситуации. С помощью «Цифрового двойника населения Арктики» сделаны выводы о пространственных закономерностях демографического развития российской Арктики [3].

Оригинальная методика изучения миграционных процессов на межмуниципальном уровне предложена учеными из Кирова и Томска. В их статье предложен авторский подход к изучению миграции выпускников вузов, расположенных в регионах с высоким оттоком населения. На основе анализа цифровых следов пользователей социальной сети «ВКонтакте», являющихся выпускниками наиболее крупного вуза в регионе-доноре, был составлен рейтинг мест послевузовской миграции (рис. 6).



Рис. 6. Облако слов, характеризующих направление миграционных потоков [6].

Расчет коэффициентов оседлости, внутрирегионального возврата и столичности позволил осуществить типологию регионов-доноров [6].

Изучение качества жизни населения. В широком смысле качество жизни рассматривается как интегральное понятие, всесторонне характеризующее степень комфортности среды жизни и деятельности человека, уровень благосостояния, социального и духовного развития личности. Среди прочих составляющих качества жизни выделяются показатели, характеризующие востребованность объектов культуры, искусства и рекреации, а также экологическое состояние среды обитания человека, инфраструктурное обеспечение населения [2]. В традиционных методиках используются в основном интегральные оценки на основе количественных показателей, иногда дополняемые социологическим опросом [4; 5].

В свою очередь, в ряде исследований рассматривается возможность применения новых источников данных – цифровых следов пользователей в

сети Интернет, – для изучения качества жизни населения сельских территорий Российской Федерации. Так, в статье Е. В. Щекотина показано, что исследование различных аспектов качества жизни сельских территорий на основе анализа цифровых следов пользователей обладает как определенными преимуществами, так и методологическими ограничениями. Автором предложена методика анализа цифровых следов пользователей в российской социальной сети «ВКонтакте» для выявления проблем, с которыми сталкиваются жители сельских территорий и которые являются важными неблагоприятными факторами качества жизни. С помощью программного обеспечения PolyAnalyst в работе выявлены проблемы сельских территорий для десяти регионов РФ. К их числу относятся проблемы жилищно-коммунального хозяйства, благоустройства сельских поселений и транспортной инфраструктуры; экологические проблемы; проблемы с доступностью и неудовлетворительным качеством телекоммуникационной и банковской инфраструктуры; проблемы с системами здравоохранения и образования; проблемы безопасности; проблемы экономического характера; проблемы бездомных и безнадзорных животных (собак); потребление алкоголя и неэффективность деятельности органов власти сельских поселений [8].

Изучение туристско-рекреационного пространства. Еще одним перспективным направлением применения методов текстового анализа является изучение туристско-рекреационного пространства. Все чаще туристы оставляют отзывы о путешествии, рекомендации, выкладывают фотографии, создают тематические сообщества. И зачастую принятие решения о поездке основывается на представлениях о том или ином месте, ментальном образе территории. Выявить эти представления крайне важно как для туроператоров, так и для туристских администраций. Но сделать это, опираясь на традиционные методы социологические опросы или ментальное картографирование, достаточно трудоемко. Поэтому для оценки ментального туристско-рекреационного пространства предлагается использовать метод текстового анализа. Именно в отзывах туристов содержится фиксация эмоционального состояния, высказываются впечатления от взаимодействия с реальным компонентом туристско-рекреационного пространства. В статье Е. В. Конышева представлен аналитический обзор научных подходов к изучению ментального туристско-рекреационного пространства, при-

менения методов его оценки, анализа цифровых следов туристов, опыт применения текстового анализа к изучению отзывов туристов. Информационная база исследования формировалась из отзывов туристов и рекреантов с платформы Tripadvisor. Обработка базы отзывов и текстовый анализ осуществлялись с помощью программы PolyAnalyst. База данных состояла из 6300 отзывов. Осуществлялся расчет индекса тональности, извлечение ключевых слов и определялась связь терминов для определения ключевых атрибутов ментального туристско-рекреационного пространства. Проводился покомпонентный анализ индекса тональности и анализ его изменения с 2013 по 2021 год. Дополнительно был проведен географический анализ с выявлением территориальных особенностей восприятия туристско-рекреационного пространства городов Кировской области. Сделаны выводы о негативном восприятии транспортной инфраструктуры, природных объектов и рекреационных зон, благоустройства территории населенных пунктов Кировской области. Отмечено повышение доли негативных отзывов со временем. Результаты исследования могут быть использованы для устранения проблемных зон взаимодействия населения с компонентами реального туристско-рекреационного пространства и корректировки региональной туристской политики [1].

Список использованных источников

1. Конышев Е. В. Методика изучения ментального туристско-рекреационного пространства по отзывам туристов (на примере Кировской области) // Вестник Московского университета. Серия 5: География. 2022. № 5. С. 16-28. EDN: DWQVXS
2. Меркушев С. А. Качество жизни населения городских поселений Пермской области (территориальный анализ). Автореферат дис. на соискание уч. степ. канд. геогр. наук. Пермь, 1997. EDN: ZJGJRX
3. Смирнов А. В. «Цифровой двойник» населения Арктики в демографических исследованиях и управлении развитием территорий // Арктика и Север. 2023. № 53. С. 260-272. DOI: 10.37482/issn2221-2698.2023.53.260. EDN: IFVJFO
4. Тикунов В. С., Белоусов С. К. Интегральная оценка качества жизни населения городов и регионов России // Вестник Московского университета. Серия 5: География. 2022. № 2. С. 48-60. EDN: TAZCKZ

5. Трофимов А. М., Мальганова И. Г. Оценка качества жизни населения как форма изучения социально-географического пространства // Географический вестник. 2005. № 1-2. С. 36-43. EDN: NCSGEV

6. Митягина Е. В., Коньшев Е. В., Чернышев К. А., Сайфулин Э. Р. Цифровые следы выпускников вузов при исследовании миграции из регионов-доноров // Вестник Томского государственного университета. 2021. № 467. С. 144-155. DOI 10.17223/15617793/467/18. EDN: GCQDCH

7. Шарыгин М. Д., Столбов В. А. Перспективы применения цифровых методов в исследованиях территориальных общественных систем // Цифровая география : Материалы Всероссийской научно-практической конференции с международным участием. В 2-х томах. Пермь, 16–18 сентября 2020 года / Науч. ред. А. И. Зырянов, Т. В. Субботина, С. В. Копытов. Том II. Пермь: Пермский государственный национальный исследовательский университет, 2020. – С. 123-126. EDN: BGYVEV EV

8. Щекотин Е. В., Дунаева Д. О., Александрова Ю. К. Анализ качества жизни сельских территории России на основе цифровых следов пользователей // АПК: экономика, управление. 2022. № 11. С. 89-96. DOI 10.33305/2211-89. EDN: YAXMUK

Практическая работа

1. Изучите материал лекции.
2. Выберите направление исследования, в рамках которого может применяться текстовый анализ данных.
3. Дайте описание концепции исследования.

Самостоятельная работа

Проанализируйте статьи:

- Смирнов А. В. «Цифровой двойник» населения Арктики в демографических исследованиях и управлении развитием территорий // Арктика и Север. 2023. № 53. С. 260–272. DOI: 10.37482/issn2221-2698.2023.53.260 EDN: IFVJFO

- Митягина Е. В., Коньшев Е. В., Чернышев К. А., Сайфулин Э. Р. Цифровые следы выпускников вузов при исследовании миграции из регионов-доноров // Вестник Томского государственного университета. – 2021. – № 467. – С. 144-155. – DOI 10.17223/15617793/467/18. EDN GCQDCH.

- Щекотин Е. В., Дунаева Д. О., Александрова Ю. К. Анализ качества жизни сельских территории России на основе цифровых следов пользователей // АПК: экономика, управление. – 2022. – № 11. – С. 89-96. – DOI 10.33305/2211-89. EDN YAXMUK.

- Конышев Е. В. Методика изучения ментального туристско-рекреационного пространства по отзывам туристов (на примере Кировской области) // Вестник Московского университета. Серия 5: География. –2022. № 5. С. 16-28. EDN: DWQVXS

Дайте краткую характеристику проблемы, которая решается в данных исследованиях с помощью текстового анализа данных.

Вопросы для самопроверки

1. Что является объектом и предметом общественной географии?
2. Назовите преимущества и недостатки использования методов текстового анализа данных в общественной географии.

Лекция 5. Обзор реализованных кейсов. Оценка восприятия туристско-рекреационного пространства

В современном информационном обществе наблюдается взрывной рост объема и многообразия информации, необходимость обработки которой вызвало развитие передовых инструментов и технологий. В настоящее время под Big Data принято понимать не столько огромные объемы данных, сколько комбинацию подходов, инструментов и методов их обработки. В качестве источников данных могут выступать социальные сети, интернет вещей (IoT), промышленные и бытовые приборы, банковские транзакции, медицинские данные, собственные данные компаний, спутниковые снимки, поисковые запросы, файлы cookies, агрегаторы отзывов. Анализ данных осуществляется с помощью таких методов, как Data Mining, краудсорсинг, машинное обучение и нейронные сети, обработка естественного языка, анализ социальных сетей (Social Network Analysis), распознавание образов, статистический анализ, моделирование и предиктивная аналитика, визуализация аналитических данных. Среди научных направлений, в которых востребованы инструменты анализа больших данных, выделяется общественная география как наука, изучающая территориальную организацию общества, ее законы и закономерности.

Рекреационная география является частью общественной географии, имея при этом свой объект и предмет исследования. В качестве объекта исследования выступает туристско-рекреационное пространство, под которым понимается реальное пространство туристских и рекреационных объектов в сочетании с пространством физических связей и отношений между ними. Оно формируется в русле стратегии развития географического пространства, сохраняя внутреннюю целостность и функциональную структурированность. В отличие от смежных послойных образований общественно-географического пространства (экономического, социального, культурного и др.) оно формируется под влиянием не столько трудовой, бытовой и иной деятельности людей, сколько потребностей поддержания здоровья, умственного совершенствования, удовлетворения любознательности и желания общения с окружающим миром [5]. Оценка восприятия туристско-рекреационного пространства становится возможным через впечатления туристов, которые являются результатом реакции туриста на взаимодействие с отдельными компонентами туристской сферы [1]. Нередко впечатления остаются зафиксированными в едином информационном поле в виде цифровых следов.

Значение анализа цифровых следов для изучения поведения туристов и рекреантов возрастает с увеличением количества пользователей сети Интернет и все большим вовлечением их в виртуальную жизнь. В сфере туризма и рекреации изучение социальных сетей, определение их роли в принятии решений путешественниками, а также в туристических операциях и управлении является одной из «мега-тенденций» [2].

В качестве источника данных может рассматриваться платформа TripAdvisor, на которой после взаимодействия с компонентами реального туристско-рекреационного пространства люди оставляют текстовые сообщения – отзывы. Текстовые сообщения, как правило, содержат оценку результата взаимодействия, выраженную как в баллах, так и в виде эмоционального сообщения, что позволяет использовать инструменты текстового анализа.

Текстовый анализ данных для изучения сферы туризма стал применяться относительно недавно. В основном, применяя подходы машинного обучения для анализа отзывов, размещенных на платформе TripAdvisor, специалисты осуществляют сегментацию потребителей, учатся управлять при-

нятием решений о выборе потребителем дестинации и отелей, а также решать иные маркетинговые задачи [3]. В России изучение туристско-рекреационного пространства методами текстового анализа осуществляется учеными Пермского государственного национального исследовательского университета [4].

Рассмотрим кейс по изучению ментального туристско-рекреационного пространства методами анализа текстовых данных более подробно. Методика исследования разрабатывалась исходя из сущности и структуры туристско-рекреационного пространства, принятого в рекреационной географии, с учетом общего алгоритма текстового анализа.

Начальным этапом исследования стало формирование базы данных отзывов туристов об отдельных компонентах туристско-рекреационного пространства и классификация текстового контента отобранных в нем объектов. На данном этапе осуществляется формирование (парсинг информации с платформы TripAdvisor и иных источников) и первоначальная обработка информационной базы исследования. Было получено 7999 отзывов за период с 2014 по 2022 год. База данных представляет собой неструктурированные данные в виде отзывов туристов и рекреантов, скачанных с общедоступных агрегаторов и тематических сообществ. База данных состоит из следующих разделов:

- объект оценки включает следующие атрибуты: «название организации», «адрес организации», «ссылка на источник информации», «раздел на платформе TripAdvisor», «категория раздела на платформе TripAdvisor», «тип организации»;

- предмет оценки включает следующие атрибуты: «количество отзывов об организации», «заголовок», «отзыв (полный текст)», «тег», «оценка (в баллах)», «дата посещения организации», «дата размещения отзыва на платформе TripAdvisor», «количество лайков за отзыв»;

- субъект оценки включает следующие атрибуты: «имя пользователя», «ссылка на профиль на платформе TripAdvisor», «возраст пользователя», «пол пользователя», «количество подписок», «количество подписчиков», «количество публикаций», «количество благодарностей», «критик уровня», «количество посещенных городов», «место проживания», «год регистрации».

Структура базы данных позволяет получить развернутую информацию об объекте оценки (организация), субъекте оценки (пользователь платформы – турист или рекреант), а также провести текстовый анализ содержания отзыва (географические, структурные, социально-демографические, организационные и иные параметры).

Первоначальная обработка включает в себя разметку базы данных по регионам проживания респондентов (например, замена названия населенного пункта в столбике на название региона – субъекта РФ или название страны). Далее осуществляется подготовка базы исследования для покомпонентного (выделение из общей базы исследования массивов данных об отдельных компонентах туристско-рекреационного пространства – средствах размещения, организации общественного питания, транспорте, объекте показа и достопримечательностях, туристских маршрутах) и территориального анализа (выделение из общей базы исследования массивов данных о туристских центрах региона, а также баз данных респондентов из других регионов).

Далее осуществляется обработка первоначальной базы для дальнейшего анализа с использованием встроенных инструментов программного продукта PolyAnalyst: модификация колонок, индексирование, проверка орфографии, выявление уникальных записей и текстов (рис. 7).

Ограничения исследования:

- 1) Незначительное количество отзывов до 2014 года.
- 2) Не все предприятия туристско-рекреационной сферы зарегистрированы на платформе Tripadvisor.

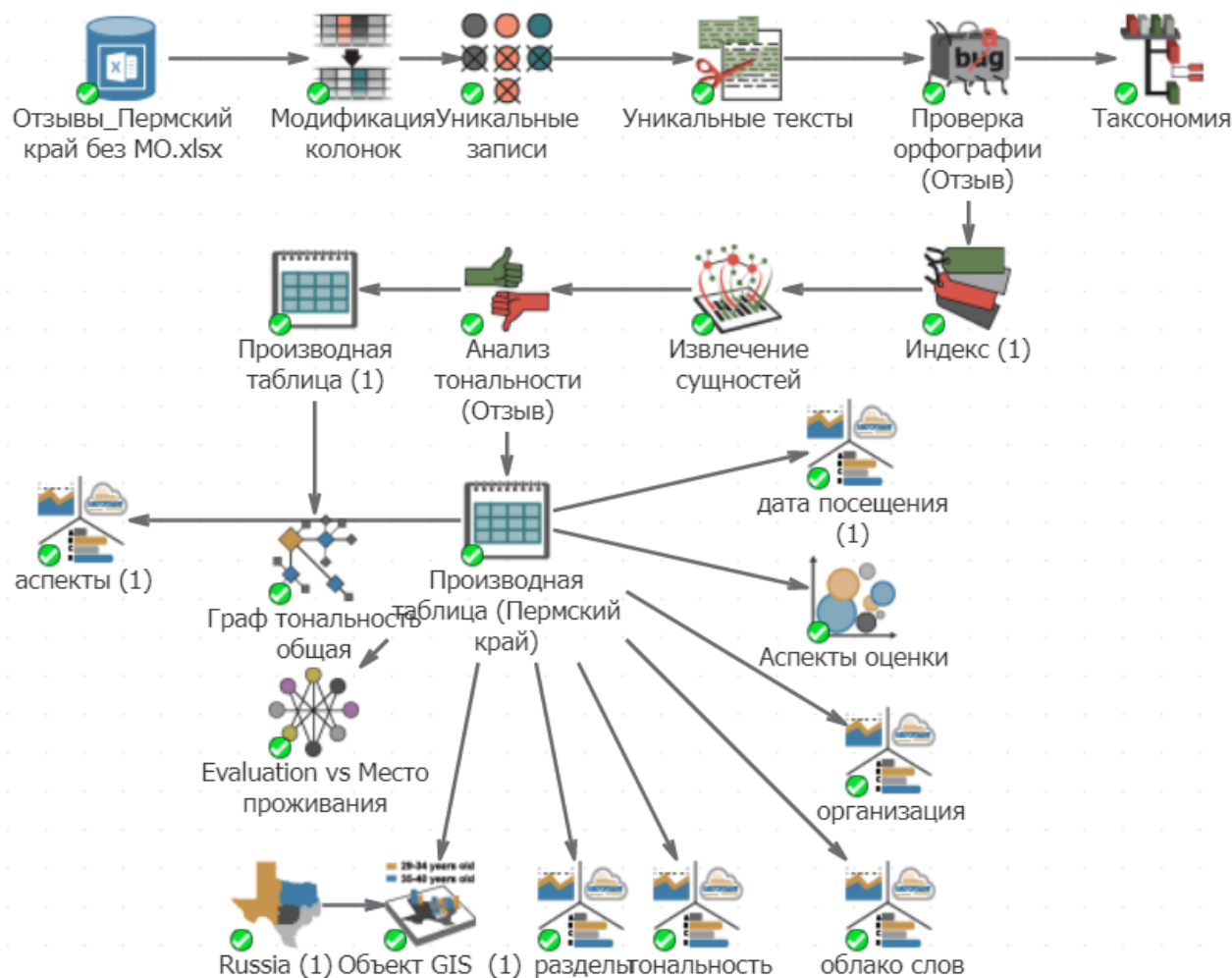


Рис. 7. Общая схема текстового анализа отзывов туристов на платформе PolyAnalyst (составлено автором).

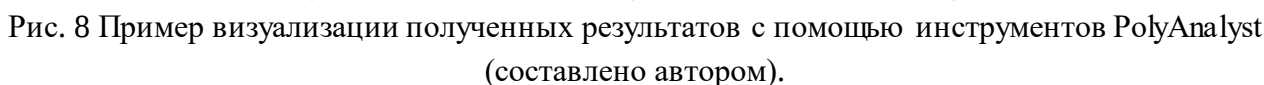
3) Как правило, цифровой след фиксируется в отношении тех компонентов туристско-рекреационного пространства, с которыми взаимодействует турист и рекреант и которые вызывают определенные эмоции. При формировании базы данных могут возникнуть сложности получения отзывов о тех компонентах реального туристско-рекреационного пространства, с которыми не осуществляется взаимодействие во время туристской поездки или отдыха (например, организационно-управленческий компонент). Не много отзывов и о рекреационных зонах, природных достопримечательностях.

4) Географический анализ отзывов потребителей затруднен из-за частичной анонимности пользователей.

Выявленные ограничения сужают временной диапазон проведения исследования, вносят определенные корректировки в интерпретации результатов по отдельным компонентам туристско-рекреационного пространства, однако, на наш взгляд, в целом не оказывают существенного влияния на достижение поставленной цели.

Обработка базы данных на платформе PolyAnalyst осуществляется через комбинации встроенных узлов. Так, узел «модификация колонок» позволяет изменить формат колонки отзывов со строкового на текстовый, что позволяет в последующем работать с более сложными узлами. Узел «индексирование» предназначен для разделения текстов на абзацы и предложения, осуществления разметки всех слов по частям речи. Данная операция требуется для более быстрой работы последующих узлов. Обязательной процедурой является исправление орфографических ошибок. Для данной базы был выбран порог достоверности в 70%. Результатом применения узлов «уникальные тексты» и «уникальные записи» является таблица, из которой удалены строки с дублирующим или похожим содержанием отзыва. После проведения необходимых процедур база исследования становится более подготовленной для обработки данных. Для быстрого понимания сути содержания отзывов предназначен инструмент «извлечение ключевых слов». Используя различные настройки, возможно составить собственный словарь для работы. Также инструмент включает в себя вкладку «облако ключевых слов», применение которой позволяет визуализировать полученные результаты в виде облака слов. Кроме того, визуализировать связи между ключевыми словами позволяет инструмент «связь терминов». Можно сгенерировать нужную связь терминов и настроить поиск связи в пределах предложения, абзаца или всего текста, а также установить нужное расстояние между словами и силу связи. В результате формируется граф связи, где сила связи рассчитывается как логарифм значения вероятности отклонения между двумя терминами. Чем больше сила связи, тем более значимы отношения. Инструмент «извлечение сущностей» направлен на решение задачи поиска именованных сущностей в тексте с помощью алгоритмов поиска последовательности слов и работы со словарями. В результате отображается список типов сущностей, среди которых существенный интерес представляют географические сущности. Инструмент «анализ тональности» позволяет оценить отношение автора отзыва к определенному объекту или ситуации. Оценка – это эмоционально-окрашенное суждение

Дополнительно можно настроить автоматическое обновление базы данных, ее обработку и представление результатов с помощью инструментов визуализации (рис. 8).



42

paclient/publication-view.html?reportUUID=52e2efb8-fb41-4b29-b93b-b2582d4e0de6&locale=rus#sheetId=1 и <https://pa.data.tsu.ru/polyanalyst/static/paclient/publication-view.html?reportUUID=af14628d-bc98-4d73-83bf-7d068c7749ab&locale=rus#sheetId=1>).

Таким образом, текстовый анализ данных является новым методом изучения географических явлений и процессов и демонстрирует хорошие результаты при применении комплексного подхода. Применение данного метода позволяет проводить покомпонентный и географический анализ восприятия туристско-рекреационного пространства с учетом его функциональной и пространственной структуры. Предложенный алгоритм текстового анализа позволяет получить результаты, которые могут использоваться как для совершенствования деятельности отдельных организаций, так и для улучшения свойств реального туристско-рекреационного пространства. Дальнейшие перспективы изучения отзывов туристов методами текстового анализа связаны как с совершенствованием инструментов исследования, так и алгоритмов машинного обучения с углублением анализа за счет выявления территориальных особенностей поведения туристов, изучения мнений разных возрастных групп, выявления особенности сезонного колебания индекса сезонности, раскрытия механизмов повышения туристской привлекательности регионов в цифровом пространстве.

Список использованных источников

1. Kowalczyk A. The phenomenology of tourism space. *Tourism*, 2014, vol. 24, no. 1, p. 9-15.
2. Leung D., Law R., Hoof van H., Buhalis D. Social Media. *Tourism and Hospitality: A Literature Review, Journal of Travel & Tourism Marketing*, 2013, vol. 30:1-2, p. 3-22, DOI: 10.1080/10548408.2013.750919.
3. Nilashi M., Ibrahim O., Yadegaridehkordi E., Samad S., Akbari E., Alizadeh A. Travelers decision making using online review in social network sites: a case on Tripadvisor. *Journal of Computational Science*, 2018, vol. 28, p. 168-179, DOI: 10.1016/j.jocs.2018.09.006. EDN: VIWEVY
4. Конышев Е. В. Методика изучения ментального туристско-рекреационного пространства по отзывам туристов (на примере Кировской области) // Вестник Московского университета. Серия 5: География. 2022. № 5. С. 16-28. EDN: DWQVXS

5. Шарыгин М. Д., Конышев Е. В. Туристско-рекреационное пространство: географические особенности формирования, территориализации и функционирования. / Пространственная организация общества: теория, методология, практика: сб. мат. Межд. науч.-практ. конф. / под ред. Т. В. Субботиной, Л. Б. Чупиной. Пермь, 2018. С. 115-119. EDN: ZCMRBZ

Практическая работа

1. Проанализируйте основные источники получения текстовых данных о компонентах туристско-рекреационного пространства. Какие атрибуты данных из них можно извлечь? Какие компоненты туристско-рекреационного пространства можно оценить с их помощью? В чем проявляются ограничения исследования?

2. Поработайте с дашбордом (<https://pa.data.tsu.ru:5043/polyanalyst/static/paclient/publication-view.html?reportUUID=52e2efb8-fb41-4b29-b93b-b2582d4e0de6&locale=rus&guest=1>).

3. Создайте базу данных отзывов туристов. Подготовьте ее для текстового анализа.

Самостоятельная работа

Изучение раздаточного материала для подготовки к тестированию.

Для более углубленного изучения темы рекомендуем ознакомиться со статьями:

1. Александрова А. Ю. Изменение туристского геопространства в эпоху всеобщей мобильности // Вестник Московского университета. Серия 5: География. 2020. № 2. С. 3-12. EDN: QDEQQJ

2. Nilashi M., Ibrahim O., Yadegaridehkordi E., Samad S., Akbari E., Alizadeh A. Travelers decision making using online review in social network sites: a case on Tripadvisor, Journal of Computational Science, 2018, vol. 28., p. 168–179, DOI: 10.1016/j.jocs.2018.09.006. EDN: VIWEVY. URL: <https://www.sciencedirect.com/science/article/pii/S1877750318303363>

Вопросы для самопроверки

1. Дайте определение понятия «туристско-рекреационное пространство». Какие подходы к его выделению вы можете назвать?
2. Дайте характеристику структуры туристско-рекреационного пространства.
3. Назовите основные источники данных, характеризующих развитие компонентов туристско-рекреационного пространства. В чем их преимущества и недостатки?
4. Какие проблемы развития туристско-рекреационного пространства можно решить с помощью методов текстового анализа данных?

Лекция 6. Анализ социального самочувствия населения в муниципальных образованиях

Изучение социального самочувствия осуществляется в контексте оценки качества жизни населения. «Качество жизни населения» представляет собой сложную синтетическую категорию, аккумулирующую все важнейшие для человека условия существования и развития; она выражает степень удовлетворения материальных и духовно-культурных потребностей людей (качество питания и одежды, обеспеченность жильем и его комфортность, качество системы здравоохранения и образования, доступность сферы обслуживания и разнообразие досуга, безопасность окружающей среды и возможность путешествовать, условия самореализации и творческого развития и др.) [1; 8].

Качество жизни населения – это сложная комбинация объективных жизненных условий и субъективной удовлетворенности своей жизнью отдельных индивидов, социальных групп или территориальных общностей. Поэтому современная методика оценки качества жизни населения должна базироваться на объективном и субъективном подходах. Эти подходы существенно взаимообогащают друг друга, поскольку объективная оценка выводится на основе статистических показателей, отражающих социальное развитие общества, а субъективная составляется на основе изучения степени удовлетворенности жизнью в ходе проведения социологических опросов населения.

При этом *социальное самочувствие* рассматривается как индикатор качества жизни и состояния территориальной общности людей и подробно рассмотрено в работе представителя пермской социологической школы В. А. Бурко [5].

Социальное самочувствие – эмоциональная оценка людьми своего общественного положения, уровня удовлетворения социально-экономических и духовных потребностей [3].

Дефиниция «социальное самочувствие», прочно войдя в научный оборот, имеет ряд синонимичных понятий: «настроение», «социальное положение», «жизненная ситуация» и др. Поскольку социальное самочувствие становится широко употребляемой и разносторонне изучаемой категорией, то, исходя из многообразия трактовок, предлагаются и различные методы его оценки [6].

Подходы к оценке социального самочувствия. Среди ученых-социологов, которые занимаются изучением социального самочувствия, ведущим методом исследования является опрос населения. При этом некоторые различия в методике проявляются в выборе показателей оценки и интерпретации результатов. Так, по мнению Е. Ю. Батуровой, большинство российских исследовательских компаний (ВЦИОМ, РОМИР, ФОМ и др.), проводящих социологические исследования социального самочувствия населения, применяют упрощенные операциональные определения, которые классифицируются по трем направлениям:

- оценка материального положения семьи на момент исследования;
- оценка уровня жизни семьи через один-два года;
- уровень адаптированности к происходящим в стране переменам [4].

Географы считают, что особую роль в формировании социального самочувствия играют природные факторы, определяющие образ жизни населения, особенности природопользования, традиции и обычаи, особенности поведения людей, формирующие менталитет. Не случайно одним из новых направлений развития географической науки стала поведенческая география, которая исследует территориальные особенности поведения людей, традиции, свойственные ТОЛ [9].

Оценка социального самочувствия в Пермском крае. Изучением социального самочувствия с применением метода опроса населения занимались ученые-географы на протяжении длительного периода. Так, в одной из публикаций раскрываются теоретические и методологические вопросы

изучения самочувствия населения как одного из показателей состояния общности людей. В качестве научного фундамента исследования выступает поведенческая география, которая представляет собой сравнительно новую и динамично развивающуюся ветвь общественной географии. Подчеркиваются связи социальной географии с бихевиористской, культурной, гуманитарной географией, особенности подходов и методов, показывающие предметное пересечение и взаимообогащение с социологией, культурологией, психологией и другими соответствующими науками. Усиление социальных приоритетов в развитии социальной географии требует поиска источников роста человеческого капитала, его социальных компонентов, которые находятся в фокусе внимания поведенческой географии. Методологической базой при изучении интеллектуального самочувствия населения в поведенческой географии принимается территориальная общность людей, представляющая собой концептуальную модель социума. Обоснована необходимость применения новых методов географической диагностики качества жизни с помощью ряда индикаторов, среди которых предпочтение отдано социальному самочувствию. В статье раскрывается понятие «социальное самочувствие населения» как совокупность параметров, характеризующих состояние территориальной общности людей. В работе предложены методы его диагностики и диптихи, показана ведущая роль социологических методов в изучении данного вопроса. Представлены результаты исследований, проведенных в Пермском крае в режиме диптихов с начала 1990-х до 2021 года. Подтверждена роль диагностики и оценка самочувственности членов общества как важного фактора сохранения человеческого капитала, развития экономики страны и региона [3].

Оценка социального самочувствия населения с применением методов текстового анализа. Для оценки социального самочувствия применялась адаптированная методика анализа социального самочувствия по В. Н. Аргуновой [2]. Но в отличие от методики В. Н. Аргуновой, где основным методом исследования был опрос, в качестве источника получения данных были выбраны социальные сети, а именно социальная сеть «ВКонтакте». Использование в качестве источника информации социальных сетей позволяет повысить охват населения, снизить временные затраты на сбор данных, уйти от «запрограммированности» ответов, сравнить результаты, извлеченные из официальных и неофициальных сообществ.

В постах сообществ содержится информация:

- о наиболее актуальных проблемах муниципального образования,
- их эмоциональная оценка,
- можно выявить индексы (социальной активности населения, миграционных настроений, информированности населения, социальной активности населения, популярности социальных инициатив, социальной напряженности, отношения к территории).

База данных для исследования формировалась на основании выгрузки постов официальных и неофициальных сообществ муниципальных образований Пермского края социальной сети «ВКонтакте» (табл. 1).

Были выгружены посты представителей данных сообществ за период с 1.01.2022 по 24.11.2022. Всего в базе данных 220 130 строк. К официальным сообществам относятся те, которые созданы и модерятся представителями муниципальных властей. Неофициальные сообщества, как правило, созданы по инициативе наиболее активных представителей населения местных сообществ. Можно предположить, что социальное самочувствие, определенное по результатам анализа официальных и неофициальных сообществ, будут различаться. Структура базы данных позволяет получить развернутую информацию об объекте оценки (сфера жизнедеятельности), а также выявить эмоциональное отношение жителей к той или иной проблеме.

Таблица 1

**Источники данных для оценки социального самочувствия
(официальные и неофициальные сообщества муниципальных образований)**

МО Пермского края	Официальная группа	Ссылка	Количество подписчиков	Неофициальная группа	Ссылка	Количество подписчиков
Александровский муниципальный округ Пермского края	Александровск Официально			[ПА] Подслушано Александровск		
				Александровск		
Бардымский муниципальный округ Пермского края	Администрация Бардымского муниципального округа			Подслушано Барда Бардымский район		

Обработка первоначальной базы для дальнейшего анализа с использованием встроенных инструментов программного продукта PolyAnalyst осуществлялась с помощью таких инструментов, как модификация колонок, индексирование, проверка орфографии, выявление уникальных записей и текстов (рис. 9).

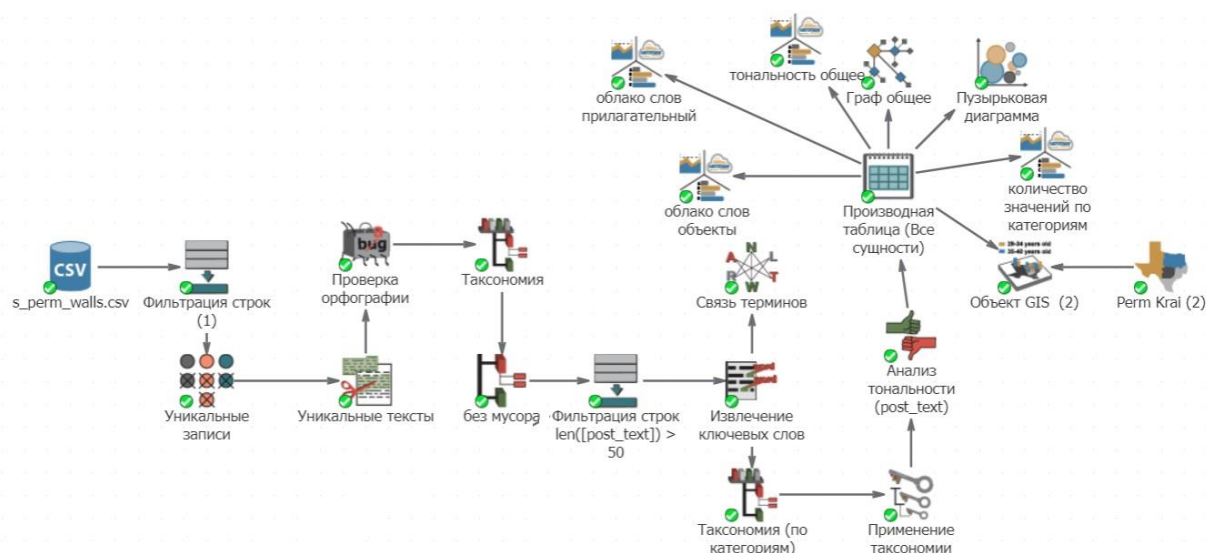


Рис. 9. Общая схема анализа данных в PolyAnalyst (составлено автором).

Узел **«модификация колонок»** позволяет изменить формат колонки отзывов со строкового на текстовый, что является обязательным условием для дальнейшей работы с более сложными узлами. С помощью узла **«фильтрация строк»** из общей базы для дальнейшего исследования были выделены посты из неофициальных сообществ. Количество строк в таблице, сформированной из постов только неофициальных сообществ, составило 152 063 единицы. Узел **«уникальные записи»** удалял дубликаты постов, в результате их количество сократилось до 121 554 единицы. Результатом применения узла **«уникальные тексты»** является таблица, из которой удалены строки с похожим содержанием отзыва. После применения узла было удалено 6623 поста, в которых похожесть текстов была более 85%, в результате общее количество строк в таблице составило 114 930 единиц (рис. 10).

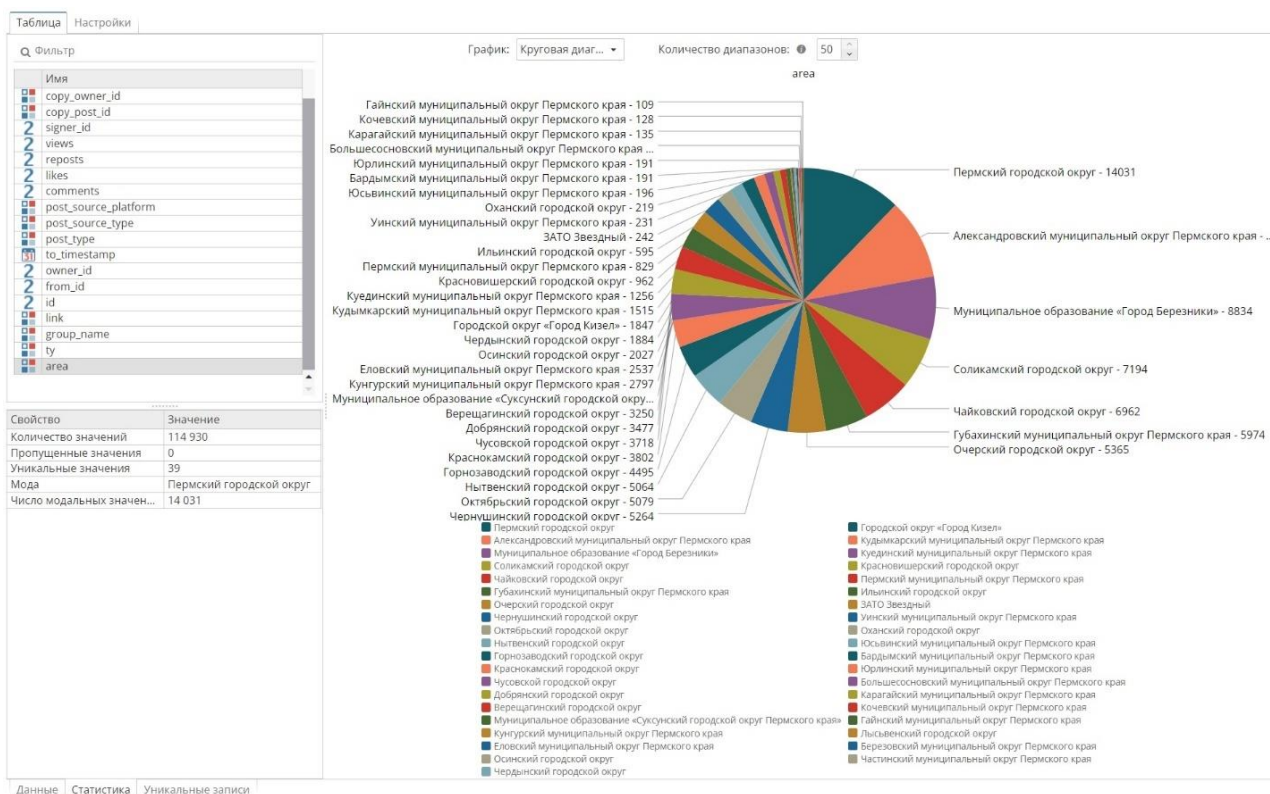


Рис. 10. Распределение постов по муниципальным образованиям Пермского края после применения узлов «уникальные записи» и «уникальные тексты» (составлено автором).

Обязательной процедурой подготовки базы исследования является исправление **орфографических ошибок**. Для данной базы был выбран порог достоверности в 70%. Всего было выявлено и исправлено 12 863 ошибки. После проведения необходимых процедур база исследования становится более подготовленной для обработки данных.

Для удаления из базы исследования «мусорных» текстов применялся узел «таксономия». «Мусорными» считались сообщения, тема или содержание которых не соотносились с темой исследования. Например, к ним относятся рекламные сообщения, участие в конкурсах, объявления о поездках, сообщения о пропажах животных и людей (рис. 11).

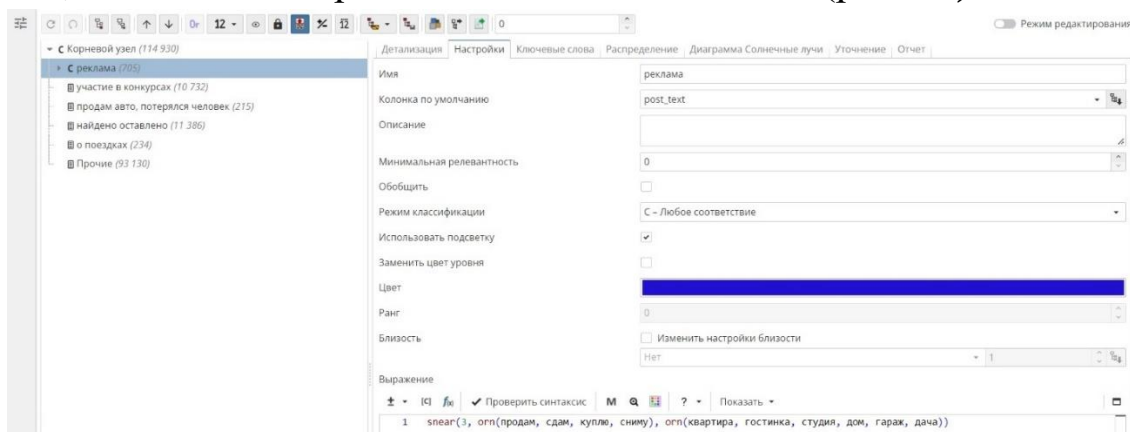


Рис. 11. Результаты очистки базы данных от «мусорных» сообщений (составлено автором).

В дальнейшем исследовании использовалась база, очищенная от «мусора» с количеством строк 93 130 единиц. Чтобы применять инструменты текстового анализа с помощью узла «фильтрация строк», были удалены короткие сообщения, размером меньше 50 слов, так как такие сообщения также не несли смыслов. В результате было удалено еще 15 988 строк. Таким образом, итоговый размер базы, который далее использовался для текстового анализа, составил 77 142 единицы, то есть 50,7% от общего числа сообщений, выгруженных из неофициальных сообществ муниципальных образований.

Информация о данных по постам в официальных сообществах муниципальных образований Пермского края:

- Количество строк в таблице с сообщениями из официальных сообществ муниципальных образований (после применения узла «фильтрация строк» – 65 008 ед.

- После узла «уникальные записи» – 57 689 ед.
- После узла «уникальные тексты» – 45 625 ед.
- Проверка орфографии исправила 3594 ошибки
- Результаты очистки от «мусора» (узел таксономия)
- Объем базы «без мусора» – 29542 ед.
- Объем базы после удаления сообщений длиной менее 50 слов – 28 280 ед.

Для группировки сообщений по сферам жизнедеятельности населения так же применялся узел «**таксономия**». В настройках узла «таксономия» в формулы вводились ключевые слова (определялись по узлу «**извлечение ключевых слов**»), представленные в табл. 2.

Таблица 2

**Ключевые слова для группировки сообщений
по категориям анализа (применение узла «таксономия»)**

Сферы жизнедеятельности	Параметры эмоциональной оценки	Ключевые слова
Экономическая сфера жизнедеятельности	удовлетворенность экономической ситуацией в стране, крае, муниципальном образовании	экономика, бизнес, заработать, предприятие, деньги, доход, стоимость, бабло, работа, рабочее место, ценность, промышленность, сельское хозяйство, отрасль, рентабельность, безработица, дотации, падение рубля, рубль, евро, доллар, имущество, налог, пенсия, уборка, благополучие, погашение долгов, вахта, вахтовик

Сферы жизнедеятельности	Параметры эмоциональной оценки	Ключевые слова
Политическая сфера жизнедеятельности	удовлетворенность политической ситуацией в стране, крае, муниципальном образовании	политика, политик, партия, власть, чиновник, чинуша, выборы, агитация, депутат, губернатор, мэр, глава района, администрация, глава города, сельское поселение, городской округ, избирательный округ, потерявший работу, мобилизация, война, Путин,
Социальная сфера жизнедеятельности	удовлетворенность различными параметрами функционирования социальной сферы муниципального образования	миграция, пособие, ЖКХ, жилищно-коммунальное хозяйство, физкультура, физическая культура, спорт, праздник, событийное мероприятие, здравоохранение, больница, врач, коронавирус, COVID, отдых, туризм, путешествие, образование, школа, учеба, кафе, рестораны, питание, голод, доктор, полиция, очередная волна, новая волна, водоснабжение, коммунальные службы, горе, алкоголь, водка, вино, убийство, грабеж, преступление, церковь, день подачи, полный рабочий день, соотечественник, первая помощь, музыка, несчастный случай, авария, колледж, инвалид
Духовная сфера и культура	отношение к культурной жизни	культура, искусство, православие, религия, верование, молитва, храм, книга, дворец культуры, кинотеатр, театр, театр юного зрителя, спектакль, опера, цирк,
Семейно-брачная сфера жизнедеятельности	удовлетворенность внутрисемейными отношениями	бабушка, родственник, подросток, подруга, муж, жена, развод, свадьба, похороны, дед, бабушка, тетя, дядя, сын, дочь, ребенок, дочка, малыш, брак
Среда, инфраструктура муниципального образования	испытываемые чувства по отношению к городу, желание в нем жить и растить детей	дом, дорога, транспорт, квартира, дорожка, асфальт, городской автобус, кладбище, освещение, поезд, забор, жилье, тротуар, крыша, больница, поликлиника, состояние дороги, главная дорога, детский садик, пожарная машина, билет, такси, отопление, время отключения, электроэнергия, рейс, городской центр, водопроводная сеть, мусор, мусорная реформа
Природные условия и экология	испытываемые чувства и отношение к реке	ветер, река, Кама, горы, лес, поле, световой день, мокрый снег, осадки, дождь, снег, пруд, озеро, ручей, родник, грязь, буря, воздух, закат, пляж, мороз, волна, пожар, понижение температуры

По наличию в сообщениях тех или иных ключевых строк все сообщения были распределены на группы (рис. 12).

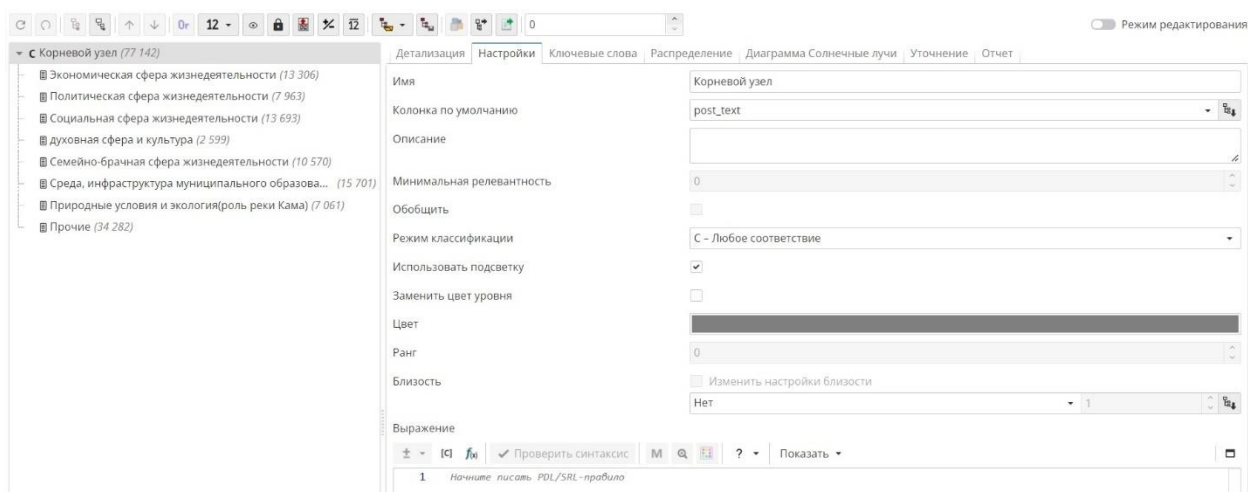


Рис. 12. Результаты группировки сообщений по сферам жизнедеятельности населения (составлено автором).

Инструмент «**анализ тональности**» позволяет оценить отношение автора отзыва к определенному объекту или ситуации. Оценка – это эмоционально-окрашенное суждение субъекта в отношении объекта (сферы жизнедеятельности населения). В настройках узла можно выбрать область исследования: общая, воздушный транспорт, финансы, гостеприимство, медицина и технологии. Для оценки социального самочувствия применялась область «общая». Встроенные алгоритмы анализа тональности определяют для каждого объекта оценки количество негативных и позитивных отзывов. Для этого анализируются слова, чаще всего прилагательные, которые соотносятся с объектом оценки и имеют свою степень отрицательности (от -1 до -5) и положительности (от +1 до +5). Соотношение количества положительных и отрицательных отзывов – это индекс тональности. Функционал узла предусматривает возможность построения графа для визуализации полученных результатов об отношении субъекта к объекту. В построенном графе серым пунсоном обозначаются аспекты (области) исследования, зеленым цветом отображаются положительные тональности, красным – отрицательные тональности.

Для визуализации полученных результатов и построения интерактивного дашборда использовались узлы для построения графиков и диаграмм, создания **облака ключевых слов**, графов связи и пр.

Список использованных источников

1. Айвазян С.А. Анализ синтетических категорий качества жизни населения субъектов Российской Федерации: их измерение, динамика, основные тенденции // Уровень жизни населения регионов России. 2002. № 11. С. 5-40. EDN: GRWEYJ
2. Аргунова В. Н., Кодина И. Н. Методика изучения социального самочувствия населения малого города // Вестник Нижегородского университета им. Н. И. Лобачевского. Серия: Социальные науки. 2013. № 1(29). С. 7-11. EDN: QLIVCB
3. Балина Т. А., Столбов В. А. Социальное самочувствие населения в контексте поведенческой географии // Географический вестник. 2023. № 1(64). С. 72-83. DOI 10.17072/2079-7877-2023-1-72-83. URL: <https://elibrary.ru/item.asp?id=50744061>. EDN: JJCULO
4. Батунова Е. Ю. Критический анализ подходов к измерению социального самочувствия: дис. магистра социологии Манчестерского университета. М., 2005. С. 73.
5. Бурко В. А. Социальное самочувствие в условиях трансформации российского общества (региональный аспект). Дис. на соискание уч. степ. канд. социол. наук, Пермь, 2000. 233 с. EDN: QDEVYV
6. Касьянов С., Горельшева М. Социальное самочувствие населения: Закономерности и механизмы // Стандарты и качество в психологии. 2020. № 2. С. 80-82.
7. Курышова Л. Н., Федосеева М. В. Факторы социального самочувствия населения региона // Огарев-online. 2020. № 9. EDN: OJPRLV
8. Мстиславский П. С. Вопросы теории и методологии анализа качества жизни // Уровень жизни населения регионов России. 2002. № 2. С. 5-17.
9. Столбов В. А., Шарыгин М. Д. Поведенческая география: учеб. пособие. Пермь, 2009. С. 8-16. EDN: QKIXTX

Практическая работа

1. Проанализируйте дашборд по оценке социального самочувствия в муниципальных образованиях Пермского края (<https://pa.data.tsu.ru/polyanalyst/static/paclient/publication-view.html?reportUUID=e353014f-c91f-43a1-883d-548894f4ae2b&locale=rus&guest=1#sheetId=1>).

Какие можно выделить пространственные особенности проявления аспектов социального самочувствия на муниципальном уровне?

2. Проведите типологию муниципальных образований по уровню социального самочувствия. Матрица по двум параметрам: 1) активность населения (количество постов в соцсетях) и 2) уровень социального оптимизма (тональность).

Типы (например):

1. Высокий уровень и социальный оптимизм.
2. Высокий уровень и социальная неопределенность.
3. Высокий уровень и социальный пессимизм.
4. Низкий уровень и социальный оптимизм.
5. Низкий уровень и социальная неопределенность.
6. Низкий уровень и социальный пессимизм.

Самостоятельная работа

Составьте словарь ключевых слов для различных компонентов оценки социального самочувствия.

Вопросы для самопроверки

1. Дайте определение понятия «социальное самочувствие».
2. Каковы особенности географического и социологического подходов к изучению социального самочувствия?
3. В чем состоят ограничения традиционных подходов к оценке социального самочувствия?
4. Какие источники данных применяются для оценки социального самочувствия с помощью методов текстовой аналитики?

Лекция 7. Анализ географического образа

Целью изучения географических образов является выявление наиболее ярких, запоминающихся черт, знаков, символов определенной местности, района и/или страны [2]. Традиционные методы изучения географических образов направлены на фиксацию представлений в сознании человека через применение ментальных карт, социологических опросов, анализа художественных текстов, географических карт и картоидов, произведений ис-

куства, фото- и видеоизображений. В последние годы, с развитием информационных технологий, предпринимаются попытки при изучении географического образа применять методы анализа больших данных.

Широкие перспективы применение методов анализа больших данных имеет и в сфере туризма. Все более совершенствуется методология и инструментарий исследования, расширяется география научного поиска и охват компонентов туристской индустрии [9]. При этом туристы, как активные пользователи сети Интернет и всевозможных онлайн-сервисов, добровольно участвуют в формировании больших данных, через оформление карт лояльности магазинов, отелей и других сервисных учреждений, оставляют отзывы и комментарии в социальных сетях, загружают фотографии и видео, сделанные в местах отдыха. Изначально основное внимание уделялось анализу количественных данных, например, с помощью сервиса Google Analytics на основе данных о посещаемости веб-сайтов составлялся прогноз изменения туристских потоков [8]. В то же время ряд исследователей подчеркивает, что использование лишь количественных данных снижает точность прогнозов и необходимо анализировать и иные типы данных [10].

С совершенствованием инструментария исследования возрастает ценность текстовой и визуальной информации, в которой содержится эмоциональная оценка восприятия компонентов туристско-рекреационного пространства, а также формируемый образ дестинации [7; 4]. Изучение виртуального освоения туристско-рекреационного пространства является важной задачей общественной географии в эпоху информационного общества. Источниками данных служат в основном либо специализированные сервисы для загрузки и хранения фотографий [5], либо социальные сети [6].

«ВКонтакте» является самой популярной социальной сетью в России с общим месячным охватом 100 млн пользователей по всему миру. При этом возможности социальной сети активно используют как обычные пользователи, так и бизнес-структуры и туристские администрации. Пользователи проводят в сети значительный объем свободного времени, подписываются на интересующие их сообщества, просматривают контент, оставляют цифровые следы в виде лайков, отзывов или комментариев. При этом цифровой след рассматривается не только как фиксирование действий автора, но и как источник формирования образа предстоящего отдыха для потенциальных туристов. Перед путешествием туристы вступают в сообщества,

посвященные конкретной дестинации, и на основании их анализа составляют первоначальное впечатление о территории. Кроме того, анализ цифровых следов представляет интерес и с точки зрения формирования туристской политики в регионе. И если пользователи в основном обращают внимание на зрительный образ, то для туристских администраций более ценной информацией являются прилагающиеся к фото текстовые данные (название, описание, комментарии), их пространственные характеристики и эмоциональные оценки.

Методы и методология. Источником для формирования базы данных послужило сообщество «Фотоальбом "Кировская область"» в социальной сети «ВКонтакте», где пользователи выкладывают фотографии, сделанные на территории Кировской области. В настоящее время в группе состоят около 20 000 подписчиков. База данных, полученная в ходе выгрузки постов из данного сообщества, охватывает период с 2019 по 2021 год состоит из 20 717 строк и включает следующие атрибуты:

1. Сообщения (посты) со стен сообществ социальной сети «ВКонтакте».
2. Профили подписчиков, указание, в какой группе они состоят, – результат («Файл с профилями пользователей»).
3. Количественные показатели реакций пользователей (лайки, репосты, комментарии).
4. Информация о дате и времени появления сообщения.
5. Портрет респондентов (город родной, город, в котором они живут сейчас, возраст, семейный статус, образование среднее, образование высшее, карьера, количество подписчиков).

Обработка данных осуществлялась с помощью аналитической платформы PolyAnalyst. Общая схема анализа и последовательность применения узлов представлена на рис. 13.

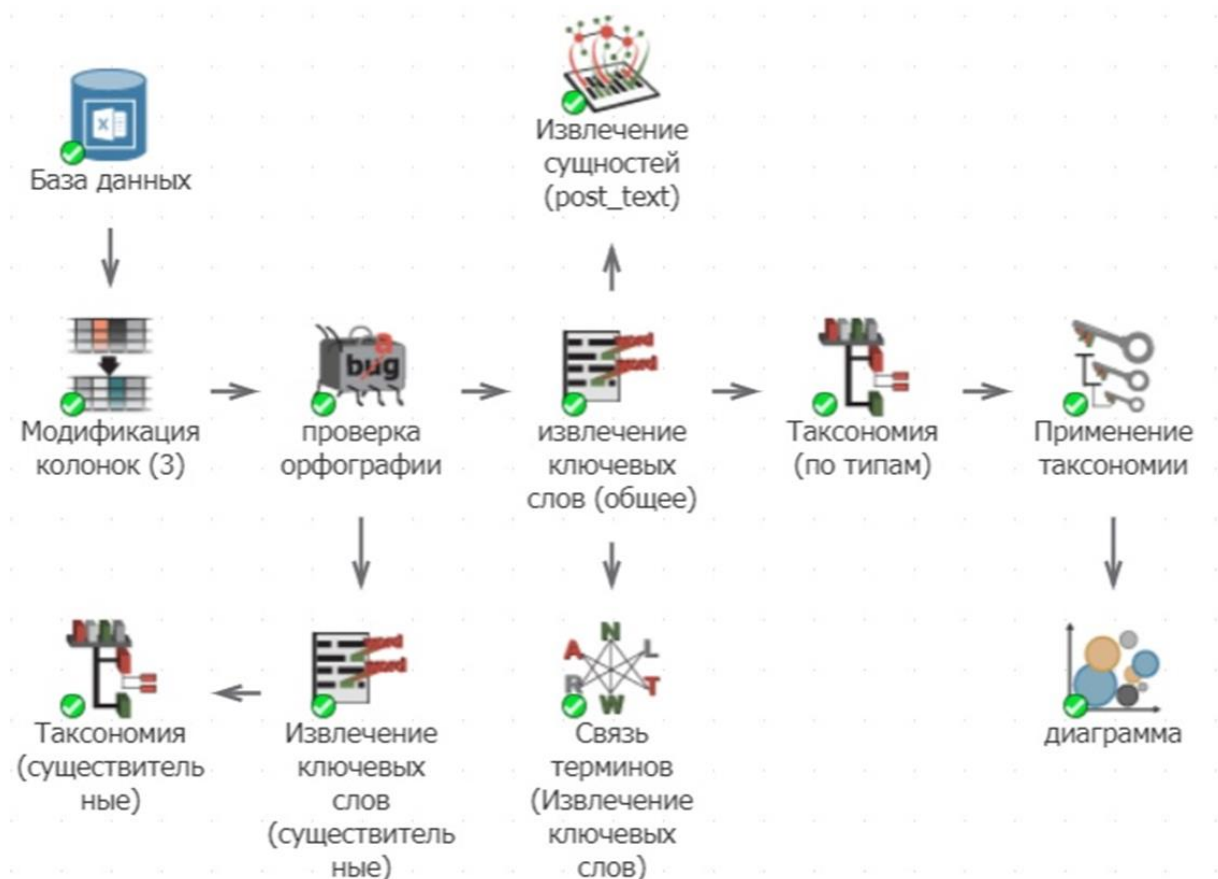


Рис. 13. Схема применения узлов для обработки данных в PolyAnalyst (составлено автором).

Узел «модификация колонок» позволяет изменить формат колонки отзывов со строкового на текстовый, что позволяет в последующем работать с более сложными узлами. Обязательной процедурой является исправление орфографических ошибок. Для данной базы был выбран порог достоверности в 70%. Для быстрого понимания сути содержания отзывов предназначен инструмент «извлечение ключевых слов». Используя различные настройки, можно составить собственный словарь для работы. Также инструмент включает в себя вкладку «облако ключевых слов», применение которой позволяет визуализировать полученные результаты. Кроме того, визуализировать связи между ключевыми словами позволяет инструмент «связь терминов». Можно сгенерировать нужную связь терминов и настроить поиск связи в пределах предложения, абзаца или всего текста, а также установить нужное расстояние между словами и силу связи. В результате формируется граф связи, где сила связи рассчитывается как логарифм значения вероятности отклонения между двумя терминами. Чем больше сила связи, тем более значимы отношения. Инструмент «извлечение сущностей» направлен на решение задачи поиска именованных сущностей в тексте с

помощь алгоритмов поиска последовательности слов и работы со словарями. В результате отображается список типов сущностей, среди которых немалый интерес представляют географические сущности. Узел «извлечение сущностей» позволяет выделить из текстов данные о геоадминистративных объектах, о природных и городских объектах.

Таблица 3

Ключевые слова для группировки текстов (фрагмент)

Типы текстов	Ключевые слова
Природный	ветер, река, река Вятка, река Юг, река Пушма, река Луза, Соколя гора, ...
Социальный	миграция, пособие, жкх, жилищно-коммунальное хозяйство, физкультура, физическая культура, спорт, праздник, событийное мероприятие, здравоохранение, ...
Культурный	культура, искусство, православие, религия, верование, молитва, храм, собор, церковь, книга, ...
Панорамный	деревня, поселок, город, дома, многоэтажка, панорама, избушка, урочище, ...
Бытовой	рыбалка, охота, поймать, сенокос, мед, пасека, банька, колодец, огород, стирка, ...
Индустриальный	экономика, бизнес, заработать, предприятие, деньги, доход, стоимость, бабло, работа, рабочее место, промышленность, сельское хозяйство, ...

В узле «таксономия» осуществляется настройка категорий для группировки текстов по заданным параметрам (ключевым словам). Все тексты были поделены на шесть основных типов (табл. 3). Для визуализации полученных результатов применялись узлы «облако слов», «диаграмма» и «связь терминов».

В пространстве России Кировская область (в повседневной жизни часто применяется наименование Вятский край) занимает место на северо-востоке Русской равнины. Географический образ Кировской области складывается прежде всего через представления о вятской природе. Она на первый взгляд проста и невыразительна. Отсутствие горных систем, удаленность от морей и океанов, незначительный перепад высот создают образ тянувшейся на многие километры увалисто-холмистой равнины, поросшей лесом, занятой возделываемыми и заброшенными сельхозугодьями. Но это только на первый взгляд! Сочетание протянувшихся в широтном и меридиональном направлении морфоструктур определяет региональные особенности рельефа. Для путешественников будут интересны формы рельефа, сформированные внешними факторами: скальные стенки и останцы, кар-

стовые воронки и пещеры, моренные гряды и даже дюны и пуги (столь экзотические для европейской части России). Порой сочетания ландшафтов создают уникальные пейзажи. Не случайно многие из них нашли отражение в картинах И. И. Шишкина, А. М. и В. М. Васнецовых. А «вятский крестьянин» Ф. И. Шаляпин говорил: «Поразительно, каких людей рожают на сухом песке еловые леса Вятки! Выходят из Вятских лесов... массивные духом, крепкие телом богатыри» [3].

Ориентация на природные объекты в географическом образе прослеживается и при анализе базы данных, состоящей из описаний размещенных пользователями фотографий и подписей к ним (рис. 14). Чаще всего фотографы стараются запечатлеть не просто природный объект, а некоторое пограничное или переходное состояние, связанный с ним процесс или явление, и затем через название кратко выразить основную идею снимка. Облако ключевых слов, составленных после применения узла «извлечение ключевых слов», показывает преимущественное использование в описании фотографий слов и фраз «туман», «осенний», «зимний», «луч света», «отражение».

равитель декабрьский лесная дорога неясный шарство верный тополь
ай октябрьский молодой голубой маленький жаркий широкий троп
восток рыбака соловья не любимый рыжий
дальний лёгкий последний вечерний
листва желтый спокойный старый эффектный
морозный белый тёплый снятой денек
бывший майский прекрасный лесной прошлый
каменный новый родной окрест
родецкий снежный красный светлый
чёрный весенний зимний холодный живо
асный Вятский туман зелёный пожелтый
нежный туман золотой летний счастливый природный
елый луна ночной большой солнечный русский синий
тёмный свет большой солнечный русский синий
рыбалка дорогой утренний сельский собор одинокий густ
костюм готовый розовый ранний отражение безупречный западный
необходимый радуга божеественный высокий бесконечный пожелтый

Облако ключевых слов по материалам
2021 года (все слова)



Осенний солнечный денек



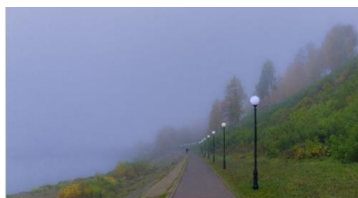
Вятка осенняя



Осеннее утро,
туман над
Ежовкой

дуют редактор забавы майский жук лучик дворец дворник разлив про
мла островок пешеходный мост правитель иней ле
два листок богатырь прожужжание сумерки ягода гладь мостик звон
лунда гусь проселочная дорога глубинка стро
земли малина лесная дорога птичка банди
оводы рыбака туман дымка
жолот огонёк крутой поворот собор гриб
акета тропа денёк исход око
зорок краса листва туча луна мед
ход снимка радуга рябина колол
ангел стихия охота отражение лебедь око
пару заря окраина луч света царство обаят
зное сель тополь рыбалка белый свет фонарь дере
ва ветерок восход снимок тропинка часовня зверь
вышивание язык переправа железнодорожный мост опушка калына
прек шиповник паломник подвесной мост зарисовка тройка
да бйство

Облако ключевых слов по
материалам 2021 года
(существительные)



Время туманов



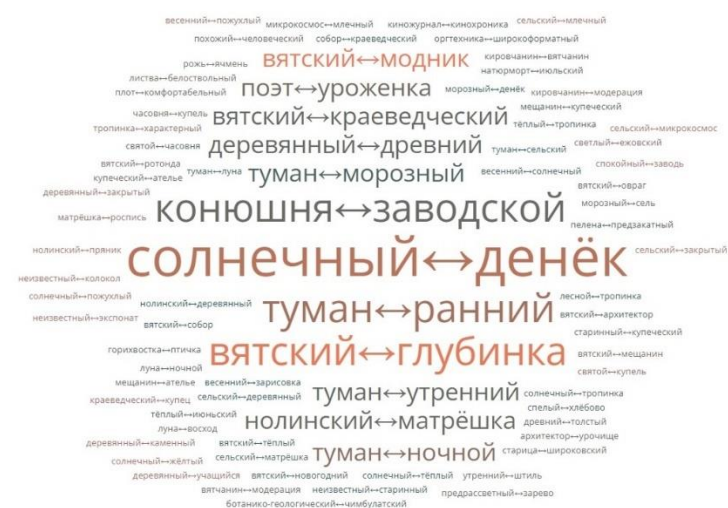
В паутине тумана



В тумане над рекой Б.
Холуница

Рис. 14. Облако ключевых слов по материалам 2021 года
(все слова и только существительные) (составлено автором).

Именно эти ключевые слова становятся в сознании человека триггером, запускающим ассоциативный ряд и проявляющим географический образ в виде набора представлений.



Погожий солнечный денёк на Вятке на исходе октября

В солнечный денёк

Рис. 15. Связь терминов по материалам 2019 года (составлено автором).

Узел «извлечение сущностей» помогает выявить географические объекты, где были сделаны фотографии. Встроенные алгоритмы PolyAnalyst различают объекты окружающей природной среды (реки, озера, моря, океаны, острова, горы, пустыни, леса и т. д.), а также объекты, которые относятся к местоположению и точкам в пределах города, места хозяйственной деятельности, например кафе, магазины, аэропорты и т. д. Также извлекаются местоположения, например, улицы, площади, трассы и участки. Области промышленности, такие как Services (услуги), Religion (религия), Energetics (энергетика), Transport (транспорт) и Culture (культура), также относятся к сущностям Facilities (объекты) там, где это возможно.

пруды, а также некоторые формы рельефа (Соколя гора, Большая гора) (табл. 4). Именно вокруг этих объектов и формируется так называемый типично русский ландшафт [1].

Таблица 4

Природные географические сущности (топ-10)

2021 год		2020 год		2019 год	
Природные объекты	ча- стота	Природные объекты	ча- стота	Природные объекты	ча- стота
река Вятка	23,7	Белохолуницкий пруд	19,1	река Вятка	22,8
река Юг	17,5	река Вятка	16,5	Белохолуницкий пруд	12,3
река Пушма	16,5	река Немда	13,9	озеро Холуново	10,5
Соколя гора	10,3	река Молома	9,6	река Быстрица	8,8
река Луза	9,3	река Песчанка	8,7	река Ивкинка	8,8
река Чепца	5,2	река Быстрица	7,8	река Омутная	8,8
Ежовское озеро	5,2	Омутнинский пруд	7,0	Ежовское озеро	8,8
Белохолуницкий пруд	4,1	Орловское озеро	7,0	река Молома	8,8
река Быстрица	4,1	Медведский бор	5,2	Большая гора	5,3
Черное озеро	4,1	озеро Чваниха	5,2	река Воя	5,3

В городской среде пристальное внимание фотографов приковано к объектам культового назначения, а также к городским паркам. Эти объекты традиционно являются точками притяжения для туристов и доминантами туристских маршрутов (табл. 5).

Таблица 5

Городские географические сущности (топ-10)

2021 год		2020 год		2019 год	
Экскурсионные объекты	ча- стота	Экскурсионные объекты	ча- стота	Экскурсионные объекты	ча- стота
Заречный парк	20,2	Трифонов монастырь	24,4	Заречный парк	25,3
Трифонов монастырь	19,0	Заречный парк	22,1	Спасский собор	16,1
Спасский собор	15,5	Музей истории и краеведения	8,1	Трифонов монастырь	12,6
Набережная Грина	13,1	Спасский собор	8,1	Набережная Грина	11,5
Старый мост	7,1	Набережная Грина	7,0	Церковь Архангела Михаила	10,3
Церковь Архангела Михаила	6,0	Свято-Троицкий собор	7,0	улица Ленина	5,7
Гагаринский парк	4,8	Ильинский храм	5,8	дом Детского творчества	4,6
Ильинский храм	4,8	Троицкий собор	5,8	улица Коммуны	4,6
Церковь Вознесения Господня	4,8	Церковь Архангела Михаила	5,8	Церковь Дмитрия Солунского	4,6
Спасская церковь	4,8	Церковь Николая Чудотворца	5,8	Церковь Николая Чудотворца	4,6

Следует отметить, что значимых изменений по популярности географических сущностей в тематике фотографий не выявлено. Однако изменения происходят на уровне типов географических образов (рис. 16).

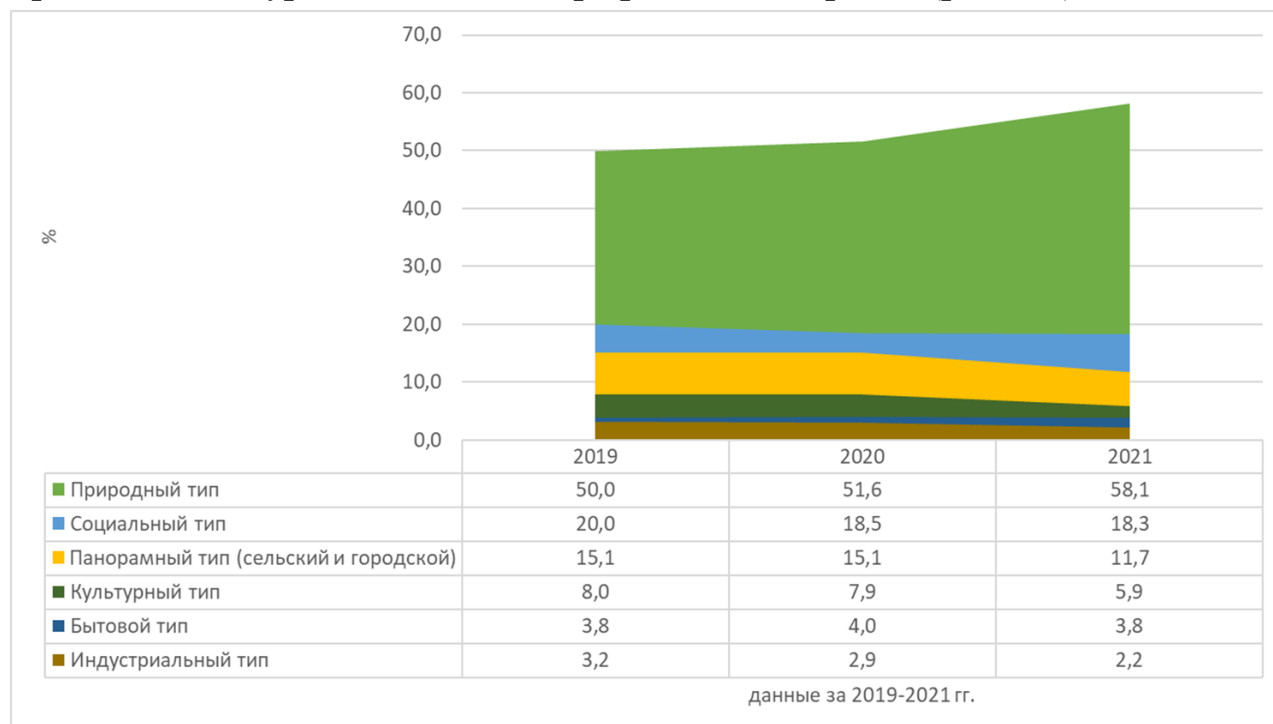


Рис. 16. Изменение географического образа Кировской области с 2019 по 2021 год (составлено автором).

Происходит усиление влияния природного типа в структуре географического образа Кировской области, прежде всего за счет сокращения доли панорамного и культурного типов.

Кировская область обладает стандартным туристско-рекреационным потенциалом, характерным для большинства регионов Европейской части страны. Сложившаяся еще в прошлом веке специализация на санаторно-оздоровительном туризме, определяет место в туристско-рекреационном пространстве страны и в настоящее время. В целях наращивания туристского потока и более эффективной реализации туристско-рекреационного потенциала, в регионе предпринимаются попытки развития отдельных направлений культурного туризма, спортивно-оздоровительного туризма, а также событийного туризма. Постепенно складывается образ Кировской области как одного из центров детского туризма. На работу с детской аудиторией ориентированы тематические парки, музеи, средства размещения и организации питания. Одним из механизмов реализации туристской политики в регионе является кластерный. Кроме того, область заключила соглашение с корпорацией «Туризм. РФ» на разработку мастер-плана.

Немаловажное значение в реализации туристской региональной политики является учет особенностей сформированного географического образа территории. Нужно понимать, что при выборе туристской дестинации потребитель руководствуется не только достоверной информацией, но и находится под воздействием собственных представлений. Проводя значительную часть времени в социальных сетях, потенциальный турист при выборе направления ориентируется на отзывы и комментарии других туристов, а также, просматривая фото- и видеоконтент, может изменить восприятие дестинации.

В социальных сетях хранятся и постоянно генерируются значительные массивы текстовых данных, обработка которых стала доступна сравнительно недавно благодаря появлению новых программных продуктов и технологий. Применение методов текстового анализа является новым направлением обработки данных, которые могут применяться для изучения географического образа. Размещая фотографии туристских объектов, пользователи дают им краткое, но содержательное название. В самом названии могут встречаться названия географических объектов, характеристика природных объектов и явлений, отсылка к культурным объектам или событиям. Эта информация является основой применения методов текстового анализа, для изучения географического образа территории. Кроме того, собранная база данных позволяет определить социально-демографический портрет участников сообщества, определить по количеству лайков и репостов наиболее и наименее популярные фотографии.

Структура географического образа Кировской области, выявленная с помощью методов текстового анализа данных включает в себя шесть основных типов: природный, социальный, панорамный, культурный, бытовой и индустриальный. Природный тип в составе географического образа вызывает наиболее яркие эмоции, и именно на его основе могут выстраиваться элементы туристского бренда Кировской области. Прикладное применение результатов исследования возможно не только в маркетинге, но и в территориальном планировании сферы туризма.

Список использованных источников

1. Веденин Ю. А. Информационные основы изучения и формирования культурного ландшафта как объекта наследия // Известия Российской академии наук. Серия географическая. 2003. № 3. С. 7-13. EDN: ONXGDP

2. Замятин Д. Н. Гуманитарная география: пространство и язык географических образов. СПб., 2003.

3. Конышев Е. В., Чернышев К. А. Кировская область / Современная Россия: географическое описание нашего Отечества. Европейская Россия и Урал. Том Книга 2. Регионы Европейской России и Урала. Москва: Русское географическое общество, 2021. С. 259-265. EDN: XZJJGU

4. Конышев Е. В. Методика изучения ментального туристско-рекреационного пространства по отзывам туристов (на примере Кировской области) // Вестник Московского университета. Серия 5: География. 2022. № 5. С. 16-28. EDN: DWQVXS

5. Соколова А. А. Виртуальное освоение и виртуальные образы региона (по данным Google Earth и Panoramio) // Известия Русского географического общества. 2010. Т. 142. № 6. С. 31-40. EDN: OJFOYD

6. Тикунов В. С., Белозеров В. С., Антипов С. О., Супрунчук И. П. Социальные медиа как инструмент анализа посещаемости туристических объектов (на примере Ставропольского края) // Вестник Московского университета. Серия 5: География. 2018. № 3. С. 89-95. EDN: YXHUSX

7. Chang Y. -, Ku C. -, & Chen C. -. (2020). Using deep learning and visual analytics to explore hotel reviews and responses. *Tourism Management*, 80 DOI:10.1016/j.tourman.2020.104129

8. Gunter U., & Önder I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, 61, p. 199-212. DOI: 10.1016/j.annals.2016.10.007

9. Li J., Xu, L., Tang L., Wang S., & Li L. (2018). Big data in tourism research: A literature review. *Tourism Management*, 68, p. 301-323. DOI: 10.1016/j.tourman.2018.03.009. EDN: VHBAMD

10. Weaver A. (2021). Tourism, big data, and a crisis of analysis. *Annals of Tourism Research*, 88 DOI:10.1016/j.annals.2021.103158. EDN: QWWXUJ

Практическая работа

1. Дайте характеристику исследований географических образов с помощью различных подходов и методов:

- Соколова А. А. Виртуальное освоение и виртуальные образы региона (по данным Google Earth и Panoramio) // Известия Русского географического общества. 2010. Т. 142. № 6. С. 31-40. EDN: OJFOYD

- Тикунов В. С., Белозеров В. С., Антипов С. О., Супрунчук И. П. Социальные медиа как инструмент анализа посещаемости туристических объектов (на примере Ставропольского края) // Вестник Московского университета. Серия 5: География. 2018. № 3. С. 89-95. EDN: YXHUSX

- Замятин Д. Н. Гуманитарная география: пространство и язык географических образов. СПб., 2003.

2. Выберите субъект РФ. Составьте список источников данных для оценки фотообразов территории.

3. Составьте таблицу ключевых слов для группировки текстов по типам фотообраза.

Самостоятельная работа

Изучение научных публикаций по применению методов текстового анализа в общественной географии.

Вопросы для самопроверки

1. Что является целью изучения географических отзывов?

2. Дайте характеристику основных методов анализа географических образов.

3. Какие функциональные узлы применяются для изучения географического образа методами текстового анализа?

4. Какие отличительные особенности географического образа Кировской области вы можете назвать?

ТЕМА 3. Сбор данных из открытых источников и разметка текста

Лекция 8. Принципы составления технического задания. Процедура сбора данных. Разметка данных

Техническое задание – это документ, в котором формулируется цель и задачи исследования, раскрывается план действий, указываются основные особенности сбора данных и предполагаемые источники их получения. Грамотно составленное техническое задание позволяет заказчику и исполнителю лучше понимать запросы и особенности работы друг друга. В нем заказчик излагает свои требования и описывает концепцию исследования, а исполнитель (чаще всего IT-специалист или аналитик данных) получает своеобразную инструкцию (рис. 17).



Рис. 17. Алгоритм составления технического задания (составлено автором).

Принципы составления технического задания:

1. Прежде чем составлять техническое задание, пользователь должен понять, что именно он хочет получить. Следует определить цель задания, ключевые особенности желаемого результата, нарисовать (написать, создать таблицу) для себя желаемый результат работы.

2. Избавиться в техническом задании от расплывчатых описаний, лишних слов, слов-паразитов. Проверить пунктуацию – зачастую ошибки в ней искажают смысл задания. Задание для проекта – это документ, и лексика в нем должна быть соответствующая. Однако не следует стараться написать все на техническом языке, если вы не владеете терминологией на должном уровне.

3. Обсудить написанное задание с непосредственным исполнителем, попытаться разрешить все вопросы, внимательно прислушиваясь к мнению собеседника. Не стоит забывать, что вы свою сферу деятельности знаете лучше и только вы можете точно объяснить, какой инструмент вам необходим для эффективной работы.

4. Проверить результат выполнения задания, протестировав его работу за некоторое время до срока сдачи проекта. Этот шаг необходим, чтобы исправить ошибки, если они будут выявлены при предварительном тестировании.

Пример составления технического задания в приложении А.

Разметка текста. Разметка данных (Data labeling), аннотирование данных (data annotation) – это процесс добавления тегов в сырые данные, чтобы показать модели машинного обучения целевые атрибуты (ответы), которые она должна предсказывать. Более подробно об этом можно прочитать в статье «Разметка данных в машинном обучении: процесс, разновидности и рекомендации» (<https://habr.com/ru/articles/678524/>).

Метка (label) или тег (tag) – это описательный элемент, сообщающий модели, чем является отдельный элемент данных, чтобы она могла учиться на примере. Допустим, модели нужно предсказывать музыкальные жанры. В таком случае обучающий массив данных будет состоять из множества композиций с метками жанров pop, jazz, rock и т. п.

Таким образом, размеченные данные выделяют признаки (характеристики) данных, чтобы помочь модели анализировать информацию и идентифицировать паттерны в исторических данных для выполнения точных предсказаний на новых, относительно близких входящих данных. Процесс разметки данных – один из важнейших этапов в подготовке данных для процессов машинного обучения с учителем.

Сложности в разметке данных:

1. Высокие затраты времени и труда.
2. Необходимость экспертных знаний.
3. Риск несогласованности.
4. Риск ошибок.

Подходы к разметке данных:

1. Разметка данных внутри компании.
2. Краудсорсинг.
3. Аутсорсинг.

4. Автоматическая разметка данных.

5. Создание синтетических данных.

Алгоритм разметки данных представлен в табл. 6.

Таблица 6

Этапы и содержание работ по разметке данных

№ п/п	Этап	Содержание работ
	Сбор данных	Начальным этапом любого проекта ML является сбор нужного количества сырых данных (изображений, аудиофайлов, видео, текстов и т. п.). Источники данных у разных компаний могут различаться. Некоторые организации сами годами накапливали информацию, другие используют публично доступные массивы данных. Как бы то ни было, такие данные часто бывают несогласованными, поврежденными или попросту неподходящими для конкретного случая. Поэтому перед созданием меток их необходимо подвергнуть очистке и предварительной обработке. Как правило, для обеспечения более точных результатов модели необходимо предоставить большой объем разнообразных данных.
	Аннотирование данных	Это основная часть всего процесса. Специалисты изучают данные и добавляют к ним метки. Таким образом они добавляют осмысленный контекст, который модель может использовать в качестве эталонных данных (ground truth) – целевых переменных, которые должна предсказывать модель. Например, это могут быть теги на изображении, описывающие запечатленные на нем объекты.
	Контроль качества	Данные должны быть качественными, надежными, точными и согласованными. Качество массивов данных в обучении модели ML определяется тем, насколько точно метки добавлены к конкретному обучающему примеру данных. Для обеспечения точности меток и их оптимизации должны быть реализованы постоянные проверки качества (QA).
	Обучение и тестирование модели	Логическим продолжением предыдущих этапов является использование для обучения модели размеченных данных, содержащих правильные ответы. В этом процессе обычно применяется тестирование модели на неразмеченном массиве данных для проверки того, обеспечивает ли она ожидаемые прогнозы или оценки. В зависимости от конкретного случая используются метрика уверенности или уровни точности. Допустим, модель считается успешно обученной, если ее точность составляет 96 и более процентов, то есть она делает 960 правильных прогнозов для 1000 примеров.

Разметка текста для задач обработки естественного языка

Обработка естественного языка (Natural language processing, NLP) –

это подмножество искусственного интеллекта, позволяющее машинам интерпретировать человеческий язык.

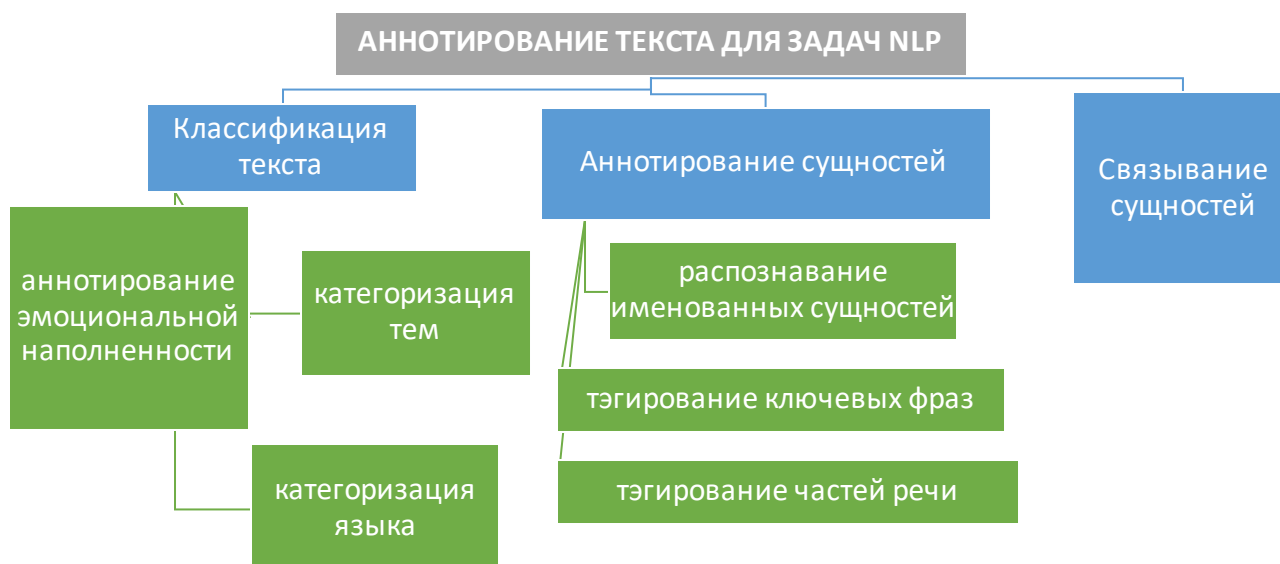


Рис. 18. Подходы к аннотированию текста (составлено автором на основании [2]).

В NLP мощь лингвистики, статистики и машинного обучения используется для исследования структуры и правил языка, а также для создания интеллектуальных систем, способных понять смысл текста и речи. Алгоритм может анализировать различные лингвистические аспекты: семантику, синтаксис, прагматику и морфологию, а затем применять эти знания для выполнения необходимых задач. Выделяют несколько способов аннотирования текста для задач NLP (рис. 18).

Разметка текста может применяться для тегирования сообщений как спам и не-спам при распознавании спама; объяснения значения текста и связи слов для машинного перевода; классификации документов; проектирования бесед для чат-ботов; извлечения ощущений и мнений потребителей из обзоров для анализа эмоциональной наполненности (приложение Б) [1].

Список использованных источников

1. Подольский А. И., Идобаева О. А., Щербакова В. В. и др. Подростки и молодежь во взаимодействии с интернет-контентом : Междисциплинарное исследование с использованием больших данных М.: Изд-во «Перо», 2021. 210 с.

2. Разметка данных в машинном обучении: процесс, разновидности и рекомендации (<https://habr.com/ru/articles/678524/>)

Практические и семинарские занятия

Ознакомьтесь с «Инструкцией по использованию API «ВКонтакте» для загрузки постов онлайн-групп».

Выполните следующие действия:

Составление технического задания. Составление таблицы по источникам данных. Сбор данных. Аннотирование данных. Контроль качества. Обучение и тестирование модели. Разметка текста для задач обработки естественного языка (классификация текста, аннотирование эмоциональной наполненности, категоризация тем, категоризация языка). Аннотирование сущностей. Извлечение и тегирование ключевых фраз или слов. Тегирование частей речи.

Инструкция по использованию API «ВКонтакте» для загрузки постов онлайн-групп

1. Все действия будем проводить в «песочнице» VK API, для этого необходимо перейти на страницу метода wall.get, который позволяет загружать содержимое страницы пользователя или группы/паблика (<https://dev.vk.com/method/wall.get>).

2. На странице метода wall.get расположено краткое описание основных полей для формирования запросов, а также внизу страницы имеется окно «Пример запроса».

3. Для работы с VK API необходим access_token, который позволит серверу «ВКонтакте» определить, какой пользователь с какого IP-адреса собирается получить данные через API.

Для получения access_token скопируйте и введите в новом окне строку ниже
(https://oauth.vk.com/authorize?client_id=5440699&display=page&redirect_uri=vk.com/callback&response_type=token&v=5.131).

Сеть «ВКонтакте» попросит разрешения предоставить общую информацию приложению "bigdata4vk", это пустое тестовое приложение необходимо для получения токена.

4. Если все сделано правильно, то в поисковой строке появится ответ с указанием полученного токена (например, https://api.vk.com/blank.html#access_token=vk1.a.wAyL7Q1XpXv8y31OZEpfSdXwi3LRpJC-Nd2XF-FEg6qmPROYvkuSjo5Z3frbexOkzlp-t_zQrLZXmBQkC7eLNpjQ4fsujNhGDia1E9wlyqX-vb1YMaMTA13U4v-

48Bh0pFEDx8nUwg-

xlTDSJHgyIC99PTNsUFgFCiYtJzBepoAjASfLuqKYho75QGbf2q3l&expires_in=86400&user_id=67538041),

где между "access_token=" и "&expires_in" и находится наш токен.

Для удобства его лучше скопировать отдельно в специальное место на следующей строке: access_token=

5. Для загрузки постов нам необходим идентификатор сообщества, его можно получить, кликнув на пост на странице сообщества. В адресной строке появится примерно такая строка (https://vk.com/cybersport.university?w=wall-208693767_329), где между "wall" и нижним подчеркиванием расположен идентификатор сообщества. В нашем случае это сообщество Н.А.У.К.А (Национальная Ассоциация Университетского Киберспорта, <https://vk.com/cybersport.university>) с идентификатором -208693767.

6. Вставляем токен и идентификатор в окно "Пример запроса", указываем count (количество постов для загрузки) равным 10 и запускаем.

Самостоятельная работа

Самостоятельное завершение лабораторного занятия. Составьте техническое задание по выбранному вами направлению.

Вопросы для самопроверки

1. Назовите основные принципы составления технического задания на проведение общественно-географических исследований с использованием методов текстового анализа данных.
2. Опишите алгоритм разметки данных.
3. Дайте характеристику основных подходов к аннотированию текста.

ТЕМА 4. Обработка данных low-code и no-code инструментами

Лекция 9. Основные функциональные возможности Orange

Как считают эксперты, к 2026 году более 40% компаний будут использовать в основе своих сервисов low-code/no-code (LCNC, зерокод), сократив при этом усилия по их развертыванию на 33%, более 65% разрабатываемых приложений будет построено с помощью инструментов low-code/no-code. Преимуществами инструментов low-code/no-code является то, что они сводят к минимуму усилия по написанию программного кода и упрощают настройку.

No-code – создание сайтов, приложений и сервисов без кода путем моделирования в графическом интерфейсе или сборка продукта с помощью конструктора.

Low-code – метод разработки цифровых решений, в котором есть возможность как использовать готовые блоки, так и вручную написать код любой сложности и на любом языке программирования. У low-code-платформ есть важная особенность – визуализация алгоритма. Все, что в обычном программировании скрыто за строчками кода, здесь нарисовано в виде схемы. Это дает возможность аналитикам, владельцам продукта, программистам, проектировщикам общаться на одном языке, глядя на схему процесса [2].

Одним из доступных low-code инструментов, который применяется для анализа структурированных и неструктурированных данных является аналитическая платформа Orange. **Orange** – это инструмент для визуализации данных, машинного обучения и интеллектуального анализа данных, который является визуальным фронтом программирования для исследовательского анализа данных и визуализации интерактивных данных. Видео о возможностях программы (<https://picktech.ru/product/orange/>) доступно в сети Интернет. Визуальное программирование осуществляется через интерфейс, в котором рабочие процессы создаются путем объединения стандартных или пользовательских наборов виджетов, в то время как опытные пользователи могут использовать Orange в качестве библиотеки Python для манипулирования данными и изменения виджетов.

Orange является пакетом программного обеспечения с открытым исходным кодом, распространяемым по лицензии GPL, и доступен для использования на GitHub. Версии до 3.0 включают в себя основные компоненты на C++ с обертками в Python. Начиная с версии 3.0, Orange использует общие Python-библиотеки с открытым исходным кодом для научных вычислений, такие как Numpy, SciPy и scikit-learn, в то время как его графический пользовательский интерфейс работает благодаря кросс-платформенному Qt framework.

По умолчанию включено машинное обучение, предварительная обработка и алгоритмы визуализации данных в шести наборах виджетов. Дополнительные функциональные возможности доступны в виде дополнений (рис. 19).

Orange поддерживается на MacOS, Windows и Linux, а также может быть установлен из репозитория пакетов Python. По состоянию на 2016 год, стабильная версия 3.3 работает с Python 3, в то время как предыдущая версия 2.7, которая работает с Python 2.7, по-прежнему доступна.

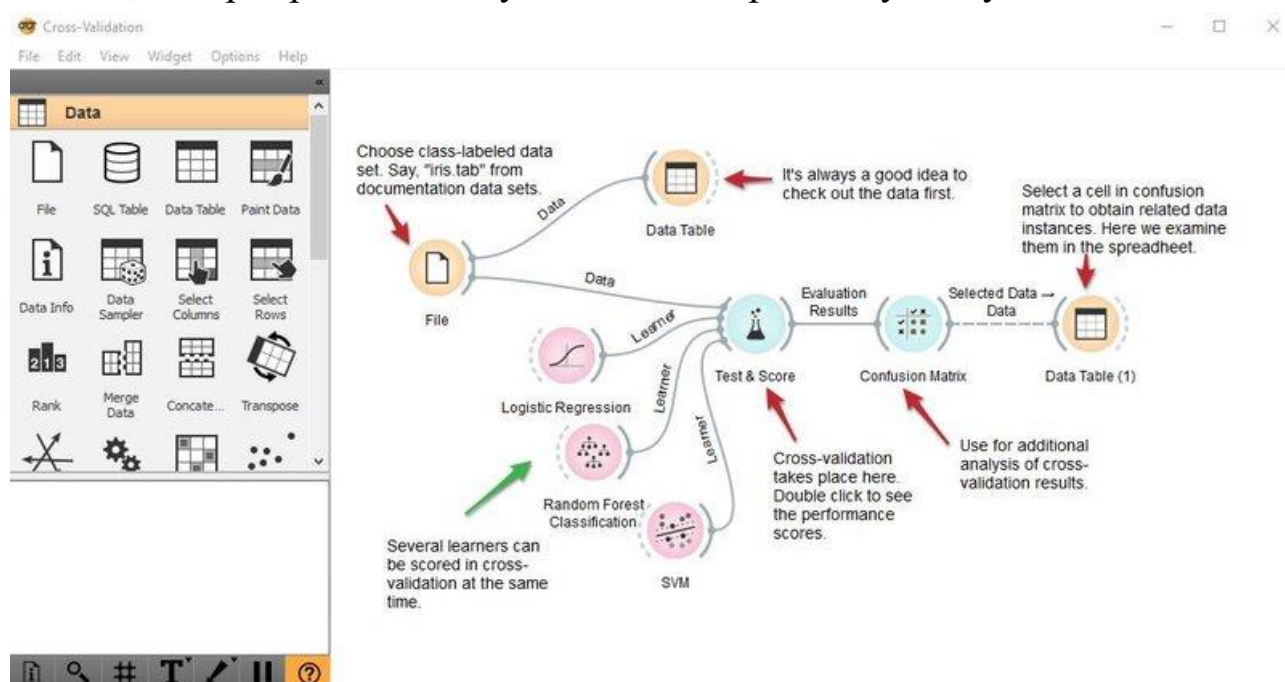


Рис. 19. Интерфейс Orange [1].

Программа предоставляет платформу для выбора эксперимента, систем рекомендаций и прогнозного моделирования. Она используется в биомедицине, биоинформатике, геномном исследовании и обучении. В научной деятельности программа используется в качестве платформы для тестирования новых алгоритмов машинного обучения и внедрения новых методов в области генетики и биоинформатики [1].

Список использованных источников

1. Официальный сайт программного продукта. Orange - Визуализация данных и машинное обучение. URL: <https://xn--90abhbolvbbf9aje4m.xn--p1ai/orange-vizualizaciya-dannyh-i-mashinnoe-obuchenie/>

2. Верховцева Мария. Тренд на low-code/no-code: как разработка без кода влияет на рынок, и почему она не заменит опытных программистов. Дата публикации 16 января 2023 года. URL: <https://habr.com/ru/companies/netologyru/articles/710728/>

Лабораторная работа

Установите программу Orange, следуя инструкции:

1. Перейдите на сайт <https://orangedatamining.com/>
 2. Перейдите по кнопке «DownloadOrange».
 3. Выберите вашу платформу и скачайте приложение.
 4. Откройте скачанный файл и установите программу, согласно предлагаемым рекомендациям (дефолтным настройкам).
 5. После того, как программа будет установлена, она должна открыться в автоматическом режиме. Если программа не открылась самостоятельно, сделайте это сами.
 6. В открывшемся окне программы закройте приветственное окно “Welcome to Orange”.
 7. Перейдите во вкладку “Options” и выберите “Add-ons”.
 8. В открывшемся окне выберите “Text” и нажмите “ОК” – таким образом у вас загрузится необходимый пакет для обработки естественного языка.
 9. После установки пакета, программа перезапустится, и раздел анализа текстов отобразится на панели с другими инструментами.
- Также вы можете загрузить и другие пакеты, которые будут полезны в исследовательской деятельности.
- Загрузите базу данных. Создайте алгоритм исследования. Произведите подключение и настройку функциональных узлов.

Самостоятельная работа:

Посмотрите дополнительные видео по функциональным возможностям Orange (https://www.youtube.com/playlist?list=PLUfHxBkkFMSfvgd_rAzt8iM9foo9sbcFx).

Пример кейсов, решаемых с помощью Orange (<https://www.youtube.com/watch?v=mVbb5P1bg9Q&t=1196s>).

Кейс по определению сорта вина (<https://elibrary.ru/item.asp?id=39546969>).

Кейс в образовании (<https://elibrary.ru/item.asp?id=50487301>).

Обзор возможностей Orange для решения классификационных задач:

- * логистическая регрессия;
- * наивный байесовский метод;
- * деревья решений;
- * нейронные сети;
- * оценка и сравнение моделей (https://www.youtube.com/watch?v=S1EitRI4w_4).

Обзор возможностей Orange, видеокурс (https://www.youtube.com/playlist?list=PLUfHxBkkFMSfvgd_rAzt8iM9foo9sbcFx).

Вопросы для самопроверки

1. В чем преимущества инструментов low-code/no-code в сравнении с традиционными программами обработки данных?
2. Опишите функциональные возможности Orange.
3. Какие задачи в области общественной географии можно решать с помощью Orange?

Лекция 10. Основные функциональные возможности Gephi

Gephi – самый известный инструмент для визуализации графов и сетевого анализа. *Сетевой анализ* (анализ социальных сетей, network analysis) – это метод исследования, который лежит в основе Gephi.

Изначально сетевой анализ – метод социологии для изучения сообществ, основанный на математической теории графов. Основа сетевого анализа, что логично, – сети, а точнее – представление данных в виде сети. Сеть (тот же граф) – это какое-то количество объектов и связь между ними, следовательно, она состоит из двух важнейших элементов:

- узлы (nodes) – это сами объекты или элементы, которые мы исследуем;
- ребра (edges) – это формализованные связи между нашими элементами.

В целом сетевой анализ представляет собой анализ данных в виде графа. Его преимущество как метода исследования в том, что могут проявиться скрытые и неочевидные связи между изучаемыми элементами. Кроме того, сетевой анализ универсален и имеет потенциал в общественной географии. Одно из главных достоинств метода – сочетание наглядности и репрезентативности. За визуализацией данных стоят цифры, которые можно представить отдельно в виде таблички и изучить их [2].

Подготовка данных

Программа скачивается с официального сайта Gephi (на этом же сайте есть страничка, которая поможет решить возможные проблемы при установке). Готовые наборы данных можно скачать с хранилища, которое создали разработчики Gephi (<https://github.com/gephi/gephi/wiki/Datasets>), а также воспользоваться коллекцией Стэнфордского университета: (<https://snap.stanford.edu/data/>).

Работа с графами требует файлов определенного вида и формата. Самый простой формат представления сетей в понятном компьютеру виде – это таблицы. Самый простой формат таблицы, который можно записывать и читать хоть в блокноте (но и в excel тоже можно) – это формат вида CSV (comma separated values). В CSV наша информация записана просто в строчку, но каждый элемент отделен запятой.

Графовая аналитика потребует большей предварительной работы над исходными данными, а именно: для загрузки в Gephi необходимо подготовить два файла:

1. Файл вершин (вершины графа являются объектами), в котором обязательно должно быть поле ID. Этого файла достаточно, чтобы построить неориентированный граф.

2. Файл ребер (ребра являются связями), в котором должны быть поля источника (Source – исходной, отправной точки) и назначения (Target – цели, точки, куда направлен поток информации на графе).

Загрузка исходных данных происходит в несколько этапов. На каждом этапе Gephi проверяет корректность данных. Загруженные в Gephi данные можно схожим с MS Excel образом фильтровать, а также можно фильтровать количество узлов и ребер графа с помощью регулярных выражений и логических операторов. Кроме того, для узлов/ребер графа и надписей можно задавать размер и цвет как через указание значения переменной при загрузке данных, так и генерироваться в настройках Gephi. Узлам графа можно даже придать объемный вид (рис. 20).

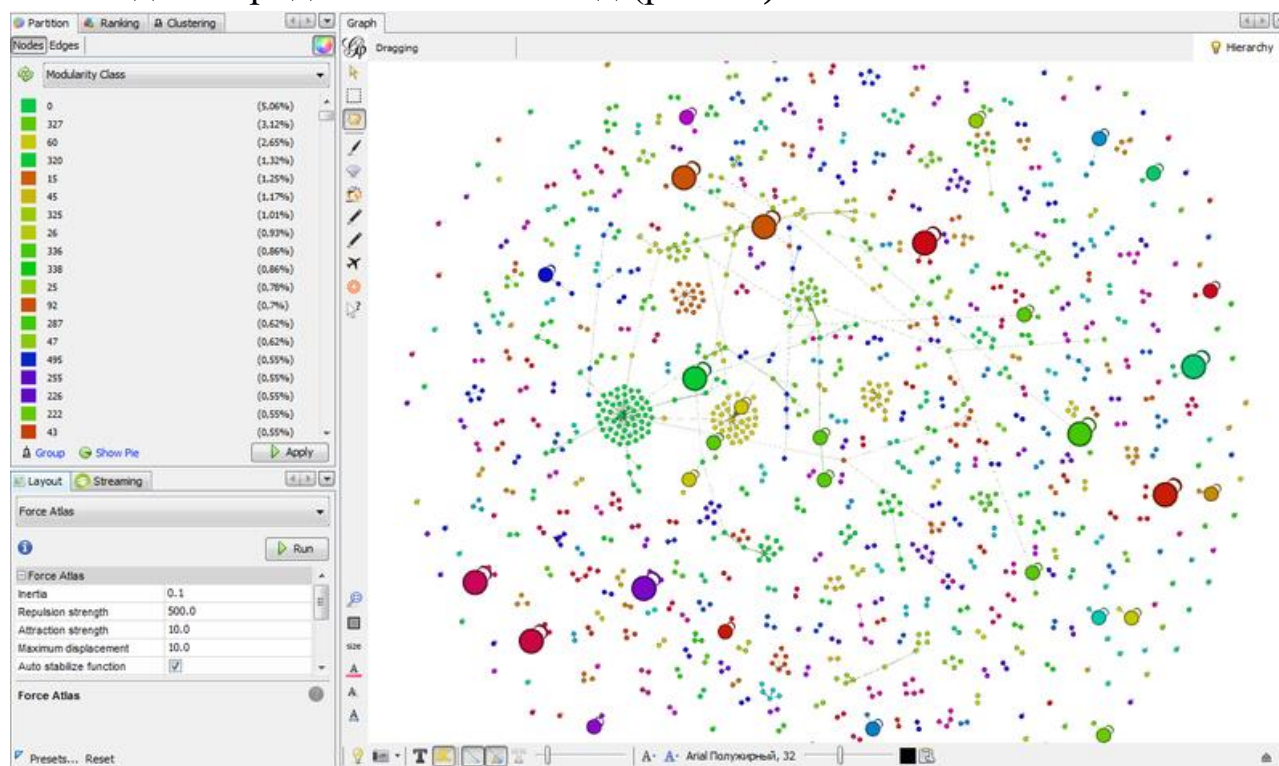


Рис. 20. Визуализация графа социальной сети:
анализ событий блогосферы перед декабрем 2011 года [1].

Алгоритмы кластеризации и ранжирования

Кластеризация относится к такому процессу классификации узлов или ребер графа, что узлы или ребра с аналогичными атрибутами и свойствами объединяются в один сегмент. Подобное разбиение графа и является кластеризацией.

Ранжирование узлов или ребер в графе – еще одна операция, которая широко выполняется во время анализа графов. Ранжирование может производиться либо по узлам, либо по ребрам, в зависимости от того, что хочет найти пользователь при исследовании графа.

Метрики и фильтры

Gerhi предоставляет пользователю в распоряжение несколько готовых способов изучения статистических свойств графов. Эти свойства включают в себя как оценку графа в целом, так и свойства отдельных узлов и ребер.

Исследование динамических и многослойных графов

Статические графы – это графы, параметры которых в узлах и ребрах фиксированы, не изменяются во времени. Система, моделируемая подобными графами, не меняет ни своих сущностей, ни отношений между ними с течением времени. Кроме того, все графы, которые мы рассмотрели в примерах выше, были однослойными или одноуровневыми, то есть отношения между объектами не иерархичны.

В Gerhi встроены инструменты анализа графов, моделирующих динамические системы.

Кроме того, Gerhi позволяет визуализировать динамику графа путем добавления проигрывателя, показывающего динамику изменения отношений между объектами графа во времени.

В Gerhi встроены инструменты обработки графа с целью ранжировать его на уровни, чтобы промоделировать иерархические зависимости в реальной системе с помощью многоуровневого графа.

Таким образом Gerhi – это открытый инструмент, постоянно дорабатываемый и обогащаемый новыми функциями решения современных задач прикладного моделирования и анализа данных с помощью результатов и последних достижений современной науки о сетях (Network Science).

Список использованных источников

1. Визуализация графа социальной сети: анализ событий блогосферы перед декабрем 2011 года. Дата публикации: 28 декабря 2012 года. URL: <https://habr.com/ru/articles/164307/>

2. Мингазов Михаил. Как плести виртуальные сети с помощью Gephi. Дата публикации: 31 января 2022 года URL: <https://sysblok.ru/knowhow/kak-plesti-virtualnye-seti-s-pomoshhju-gephi/>

Лабораторная работа

Загрузка и обработка данных в Orange. Анализ социальных сетей с помощью Gephi.

1. Обоснуйте применение сетевого анализа для изучения общественно-географических проблем.

2. Подготовьте данные (наборы данных, которые собрали разработчики Gephi (<https://github.com/gephi/gephi/wiki/Datasets>), Коллекция Стэнфордского университета (<https://snap.stanford.edu/data/>).

3. Установите приложение. Инструкция по установке Gephi:

Шаг 1. Загрузка файлов.

Для Windows: загрузите приложения Java и Gephi 0.9.2 на свой компьютер.

Для Mac: загрузите приложения Java и Gephi 0.9.2 на свой компьютер.

Шаг 2. Установка Java.

Запускаем установочный файл java, нажимаем кнопку «установить» и ждем окончания установки. Если у вас появится предупреждение, что уже имеется ранее установленные версии Java, рекомендуем удалить старые, оставив только эту.

Шаг 3. Установка Gephi.

Запускаем установочный файл gephi, принимаем лицензию, выбираем путь для установки (или оставляем как есть), в окне выбора ассоциаций для файлов обязательно устанавливаем все флажки.

После окончания установки у нас есть возможность запускать программу.

4. Выполните настройку узлов. Проведите визуализацию сетевых данных в Gephi.

Самостоятельная работа

Проанализируйте кейсы:

Кейс «Визуализация сетевых данных в Gephi» (<https://www.youtube.com/watch?v=ofPQevNXrms>).

Как уложить граф на географическую карту в Gephi (<https://www.youtube.com/watch?v=25Y8NBI4hGw>).

Кейс анализа социальной сети: отображение соединений и сетей с использованием аналитики социальных сетей (<https://fastercapital.com/ru/content/Анализ-социальной-сети--отображение-соединений-и-сетей-с-использованием-аналитики-социальных-сетей.html>).

Изучите прочие low-code и no-code инструменты.

Вопросы для самопроверки

1. Дайте характеристику Gephi и его функциональных возможностей.
2. Опишите направления использования сетевого анализа в общественно-географических исследованиях.
3. Опишите алгоритм подготовки данных для сетевого анализа.

ТЕМА 5. Текстовая аналитика на платформе PolyAnalyst

Лекция 11. Функциональные узлы платформы PolyAnalyst для работы с текстовыми данными. Обзор возможностей платформы PolyAnalyst

Аналитическая платформа для анализа и обработки данных PolyAnalyst (<https://www.megaputer.com/polyanalyst/>) была разработана российской компанией «Мегапьютер Интеллидженс» (Megaputer Intelligence). Разработка началась в 1994 году. PolyAnalyst предоставляет набор удобных инструментов для текстовой аналитики. Программа позволяет импортировать большие объемы данных (в данном случае текстов с возможностью сопровождения их метаданными) и обрабатывать их следующим образом: с одной стороны, обеспечивая отлаженную поисковую систему по запросу пользователя (возможны запросы самых разных типов), с другой стороны, извлекая и визуализируя содержимое совокупности текстов (вновь — по самым разным параметрам), объективируя его вне зависимости от селективных запросов пользователя. PolyAnalyst не требует владения специальными навыками программирования и доступен для освоения пользователями, не имеющими IT-подготовки (полезным при этом оказывается хорошее знание программы MS Excel). PolyAnalyst имеет графический интерфейс. Основным инструментом платформы являются «узлы» (визуально представленные в виде иконок), между узлами пользователь устанавливает связи, выстраивая, таким образом, аналитическое решение своей задачи (рис. 21).

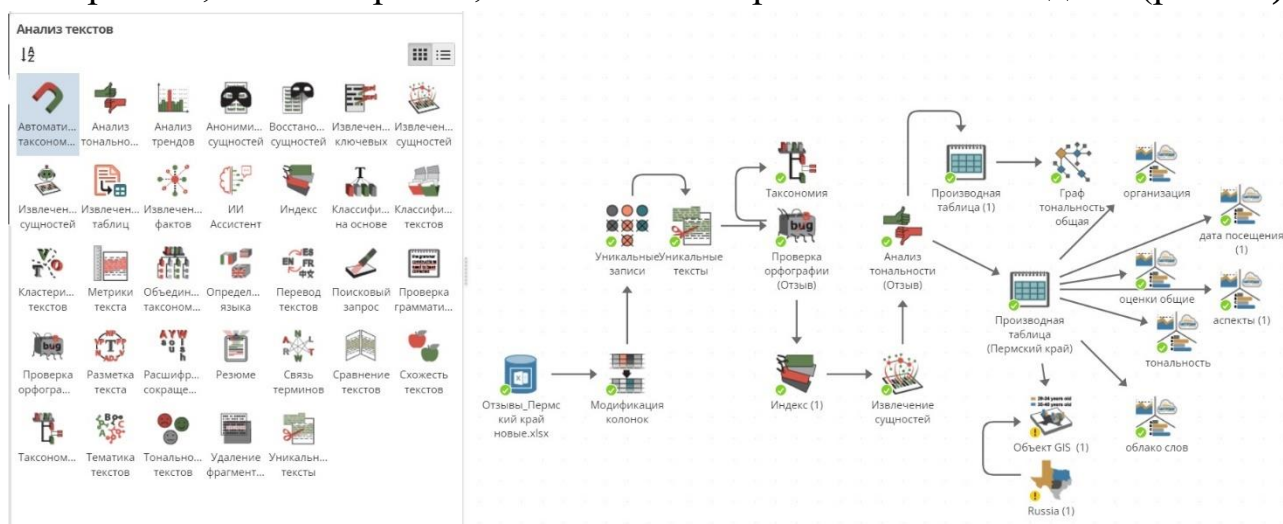


Рис. 21. Узлы для текстового анализа и пример аналитического решения в PolyAnalyst (составлено автором).

Хотя PolyAnalyst обладает инструментами сбора данных (парсинг интернет-страниц, загрузка данных из социальных сетей и др.), однако в

первую очередь эта платформа предназначена для обработки и анализа данных, а не для их сбора. Оптимальнее загружать в систему предварительно собранные данные. Помимо этого, платформа предусматривает импортирование скриптов (кода), написанных на Python, R, и встраивание их в аналитическое решение.

Работа в программном приложении начинается, как правило, с импортирования документа. Платформа позволяет загружать документ с большим объемом текстов. В реальной практике исследователя перечень текстов может представлять собой, например, выборку всех сообщений из социальной сети по определенной теме; полное собрание сочинений одного автора; большую подборку текстов одного жанра; полную подборку индексируемых в заданной международной базе научных статей по теме; полный архив журнала/сайта; выборку газет за определенный период и т. д. С помощью PolyAnalyst пользователь сможет увидеть системные для данной коллекции текстов характеристики.

Спектр инструментов обработки текстового контента, предоставляемых платформой PolyAnalyst, достаточно широк.

Некоторые узлы («производная таблица» и др.) носят технический характер и нужны только для обеспечения функционирования других узлов. Другая условно выделяемая группа узлов объединяет команды, также в определенной мере вспомогательные, адаптирующие текстовые данные для удобства их дальнейшего анализа, но уже дающие содержательные результаты, которые могут быть интерпретированы. Так, платформа позволяет отфильтровать колонки или строки в загруженном документе по заданным параметрам; уменьшить объем текстов, создав случайную выборку; найти/удалить одинаковые или похожие тексты (выявив их процент) и др. Так, узел «исправление орфографии» необходим в первую очередь для обеспечения дальнейшей идентификации лексической единицы программой и корректного ее отображения в результатах других узлов, однако может быть востребован, например, для изучения грамотности в интернете, давая интересный статистический материал. К данной же группе примыкают узлы визуализации: извлеченная из загруженного документа информация может быть наглядно представлена в виде самого разного типа диаграмм, графов, схем.

Остановимся выборочно на операциях (третья группа), которые могут представлять наибольший интерес для исследователя и обеспечивают материал для научного описания текстовых данных.

Узел «поисковый запрос» позволяет находить в большом массиве текстов заданные слова, словосочетания, типы грамматических конструкций. Результатом узла являются отфильтрованные по данному критерию тексты, с которыми далее можно работать, сгенерировав соответствующее подмножество. Возможно задавать как точный запрос, так и примерный; последний касается как формального, так и семантического признаков. Пользователь определяет важность/неважность при запросе словоформы, регистра, особенностей графики, учета синонимов. При поиске сочетаний слов можно учитывать границы контекста, степень удаленности данных слов друг от друга, важность/неважность порядка слов. Возможно также сочетание лексического и грамматического критериев: например, задать поисковый запрос, при котором перед конкретным словом или его синонимом будет стоять любое слово заданной части речи. Существует возможность поиска одушевленных и неодушевленных объектов. Узел удобен при сборе экспериментального материала для исследований в области семантики, концептологии, изучения дискурса и многих других.

Программное приложение позволяет также автоматически извлекать из загруженного документа *ключевые слова*. По результатам соответствующего узла можно быстро и легко понять суть текстовых данных. Функция ценна тем, что позволяет при отборе материала избежать субъективизма пользователя (в отличие от «поискового запроса»), опираясь на заложенные в алгоритм параметры поддержки, частоты и значимости слова. Частотные, но не значимые слова (например, местоимения) в результатах отображаться не будут. Извлеченные ключевые слова текста могут быть ранжированы по убыванию по каждому из названных признаков, что позволяет анализировать отношения между ними в рамках исследуемой коллекции текстов. Пользователь может настраивать объединение однокоренных слов под одним ключевым словом; допустимость словосочетаний в качестве «ключевых слов»; общее количество ключевых слов, включенных в отчет; минимальное количество словоупотреблений для признания слова ключевым. Возможно также удаление из результатов не актуальных для искомой задачи ключевых слов с помощью создания «стоп-листа».

Любопытные результаты может давать функция обнаружения и визуализации связей между ключевыми словами (узел «связь терминов»). Пользователю необходимо задать границы контекста, в пределах которого узел будет определять слова как связанные (например, два предложения), и минимум записей, в которых данные два слова встретились вместе. На графе будет визуализирована «сила связи» ключевых слов как относительная и измеряемая характеристика. Интерактивный характер графа позволяет подробно анализировать связи любого слова, представляющего особый интерес для исследователя. Результаты узла, помимо графа, будут представлены в таблице слов с полной статистикой; область детализации содержит исходные тексты, иллюстрирующие каждую «связь».

В рамках узла «извлечение ключевых слов» возможно извлечение как слов, так и их сочетаний, включая предикативные. Пользователь настраивает максимальное количество слов, допустимые части речи учитываемых элементов. Результатом узла станет список наиболее частотных и значимых в данной текстовой коллекции фразовых фрагментов. Узел может быть полезен для автоматического извлечения разного типа клише, а также разработки методологии извлечения устойчивых нарративов.

Из загруженной в систему коллекции текстов могут быть извлечены не только слова, но и «сущности» – смысловые объекты одного класса, упоминаемые в тексте.

Узел «анализ тональности» используется для выявления отношения автора текста к упоминаемым в тексте объектам. По умолчанию заданы позитивная и негативная тональности. Результаты представляются в текстовой форме отчета и визуализируются в виде графа. Если для бизнес-компаний данная функция актуальна в связи с задачами маркетинга, мониторинга мнений потребителей, учета жалоб, имиджмейкинга, то в контексте научных задач узел оказывается, в частности, удобным для исследования поляризации мнений в обществе по отношению к социально значимым явлениям (например, дистанционному образованию, пенсионной реформе, проблеме изменения климата) на материале больших данных. Возможно также многоаспектное изучение персональных предпочтений, ценностей и оценок автора импортированного в систему текста.

Ряд функций, предоставляемых программной платформой PolyAnalyst, направлен на тематическую классификацию текстов. Так, узел «кластеризация» автоматизирует объединение похожих текстов в классы

(кластеры) – без вмешательства пользователя. В результатах отображаются обнаруженные кластеры, статистическая информация по ним и область детализации с возможностью просмотра текстов каждого кластера. Узел «классификация» разбивает тексты по тематическим категориям, задаваемым пользователем: для его обеспечения необходима обучающая выборка.

PolyAnalyst обладает функцией создания *интерактивных вебо отчетов*. Веб-отчет представляет собой интернет-страницу, на которой можно разместить результаты своего аналитического решения. Любой пользователь может перейти на него по ссылке и изучить. Отчет полностью интерактивный, с возможностью детализации данных (drill down). Если «читателя» заинтересует какой-либо термин (слово, фраза) на графе или диаграмме, кликнув на него, он может перейти в текст, где тот употребляется, и более детально изучить.

Список использованных источников

1. Петров Е. Ю., Саркисова А. Ю. Ресурс аналитической платформы Polyanalyst в социогуманитарных научных исследованиях // Открытые данные–2021. Материалы форума. Севастополь, 30 сентября – 02 октября 2021 года / Под редю А. Ю. Саркисовой. Томск: Национальный исследовательский Томский государственный университет, 2021. С. 94-104. EDN: MSPLDG

Практические и лабораторные занятия

1. Установите аналитическую платформу Polyanalyst.
2. Произведите настройку и загрузите данные.
3. Примените следующие инструменты: модификация колонок, фильтрация строк, поисковые запросы, индексирование, проверка орфографии, выявление уникальных записей и текстов, извлечение ключевых слов, таксономия, кластеризация, создание облака ключевых слов, связь терминов, извлечение сущностей, анализ тональности. Интерпретация результатов.

Самостоятельная работа

1. Самостоятельное завершение лабораторного занятия.

2. Обзор возможностей платформы PolyAnalyst: просмотр видео
Практический анализ данных с помощью PolyAnalyst
(<https://www.megaputer.ru/project-gallery/>). Обзор возможностей
PolyAnalyst. Реализованные кейсы. Вопросы и ответы
(<https://www.megaputer.ru/project-gallery/>).

Вопросы для самопроверки

1. Дайте характеристику возможностей аналитической платформы PolyAnalyst.

2. Какие функциональные узлы применяются для текстового анализа данных?

3. Какие функциональные узлы применяются для анализа количественных данных?

ТЕМА 6. Инструменты создания пользовательских отчетов

Лекция 12. Инструменты визуализации PolyAnalyst и Visiology. Шаблоны визуализаций. Построение и настройка дашборда

В сфере Business Intelligence (BI-платформ) ценится не только интерактивность и многослойное представление данных, но и возможность исследовать данные: переключиться с одной метрики на другую с помощью кнопок, изучить данные в разных разрезах с помощью Drill Down, проанализировать дополнительную информацию во всплывающем окне, выбрав категорию на одном графике и отфильтровав по ней все остальные графики на дашборде. Дашборд представляется уже в виде готового инструмента для анализа показателей компании и принятия data-driven решений на их основе.

По разным оценкам, российский рынок BI за 2021 год составил от 30 до 40 млрд рублей. Основная доля рынка пришлась на решения с использованием зарубежных BI-платформ, таких как Tableau, Power BI и Qlik. Однако в 2022 году отрасль претерпела значительные изменения и перед многими компаниями остро встал вопрос замещения функционала лидеров рынка BI-платформ в части создания отчетов и проведения визуальной аналитики. С исследованием рынка российских (и не только) BI-платформ, можно познакомиться в сети Интернет ((<https://habr.com/ru/companies/axenix/articles/695960/>)).

В нашем учебном пособии будут рассмотрены функциональные возможности визуализации результатов аналитики больших данных и построения интерактивных дашбордов с помощью двух платформ.

Visiology (ссылка для скачивания и установки: <https://ru.visiology.su/>) — это платформа, построенная на микросервисной архитектуре. Она разворачивается в виде набора Docker-контейнеров. Frontend архитектура реализована с помощью JavaScript (с подключением любых соответствующих библиотек), JQuery, Angular, Highcharts, D3.js и собственного документированного JavaScript API.

BI-платформа может использоваться для Self-Service разработки несложных дашбордов. Для создания дашбордов среднего+ уровня сложности, BI-специалисту необходимо владеть JavaScript для разработки и кастомизации визуализаций и Python для обработки данных.



Рис. 22. Пример дашборда в Visiology [2].

В **Visiology** присутствуют все базовые визуализации (рис. 22). Изображение выглядит достаточно современно. При необходимости использования более редких видов визуализаций можно купить готовые виджеты у партнеров Visiology или разработать самостоятельно на JS. Особенностью Visiology является то, что большую часть кастомизации и форматирования визуализаций и внешнего вида дашборда необходимо выполнять с помощью JS и CSS. Для загрузки и трансформации данных, их интеграции есть собственный in-memory OLAP-движок ViQube для хранения данных и создания модели данных. Встроенных инструментов ETL в Visiology нет. Данные лучше подготовить заранее на уровне СУБД или других ETL-инструментах. Также можно использовать в качестве источника данных Python и C# скрипты. Встроенный движок позволяет агрегировать данные (sum, count и т. п.), а также выполнять несложные расчеты с логическими и математическими операторами. Для более сложных вычислений необходимо использовать Python. В Visiology доступна гибкая настройка кросс-фильтрации на уровне связанных визуализаций, а не на уровне датасета. Доступен функционал Drill Down, однако для удобного представления иерархичных данных необходима доработка в JS. Настройка условного форматирования выполняется с помощью JS-кода, без кода можно задать цвета по категориям. Отсутствуют возможности по настройке всплывающих окон. В целом платформа имеет интуитивно понятный интерфейс для

создания базового дашборда, однако большая часть настроек и кастомизации ведется с помощью JS, а обработку данных необходимо проводить на Python, либо выносить на сторону СУБД. Это увеличивает порог входа для ВІ-специалистов. Для освоения функционала платформы Visiology предлагает программы обучения и сертификации по работе со своей платформой.

PolyAnalyst позволяет добавлять пользовательские процедуры для загрузки, преобразования и выгрузки данных на языках R и Python. Также предоставляется возможность обрабатывать и анализировать данные, применяя алгоритмы машинного обучения.

Инструменты визуализации в платформе PolyAnalyst:

1. Узел «OLAP-таблица».
2. Узел «агрегированный график».
3. Узел «граф».
4. Узел «диаграмма рассеяния».
5. Узел «диаграмма связей».
6. Узел «иерархическая диаграмма».
7. Узел «карта».
8. Узел «коробчатая диаграмма».
9. Узел «линейный график».
10. Узел «матрица диаграмм рассеяния».
11. Узел «многомерная матрица».
12. Узел «облако тегов».
13. Узел «объект GIS».
14. Узел «параллельные координаты».
15. Узел «пузырьковая диаграмма».

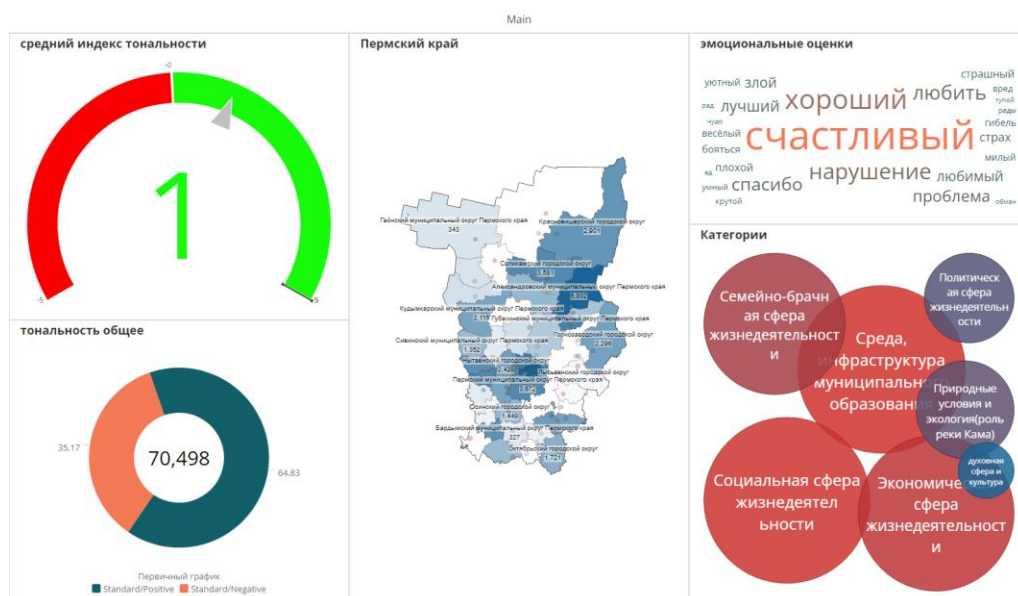


Рис. 23. Пример дашборда в PolyAnalyst (составлено автором).

Для PolyAnalyst характерно большое количество встроенных визуализаций, в том числе редко встречающихся. Визуализации и готовые дашборды в целом выглядят современно. Имеются широкие возможности по кастомизации и форматированию визуализаций по принципу no-code. Доступна гибкая настройка текстовых карточек, как в Tableau. Есть свой in-темогу движок для работы с данными. Функционал работы с данными доступен путем создания ETL-пайплайнов с возможностью последующего вывода показателей по заданным срезам в дашборд (рис. 23). Для создания каждой расчетной метрики необходимо создавать отдельную ветвь ETL-пайплайна. Есть продвинутый функционал, такой как работа с контекстными функциями, также есть собственный язык расчетов со справочником. Доступна гибкая настройка кросс-фильтрации на уровне связанных визуализаций, а не на уровне датасета. Есть функционал Drill Down и возможность переключения визуализаций на дашборде с помощью создания «вкладок». Настройки условного форматирования доступны на небольшом количестве редко используемых диаграмм (диаграмма рассеяния, карта, облако тегов и т. п.). Отсутствует возможность настройки всплывающих окон. В целом интерфейс платформы в целом интуитивно понятный, но выглядит немного устаревшим. Подход no-code снижает порог входа и сложность переобучения ВІ-специалистов. Все расчеты и преобразования данных проводятся в графическом интерфейсе ETL-пайплайна, что может увеличивать время на разработку отчета за счет необходимости строить ETL-пайплайна для каждого среза/функции.

Более подробно об инструментах PolyAnalyst для визуализации данных официальном сайте компании Мегапьютер (<https://www.megaputer.ru/project-gallery/>).

Список использованных источников

1. Руководство по эксплуатации на официальном сайте производителя. Знакомство с VISIOLOGY 3.X (<https://ru.visiology.su/lms/znakomstvo-s-visiology-3>).

2. Остапенко Анастасия. Исследование рынка российских (и не только) ВІ-платформ. Дата публикации: 27 октября 2022 года. (<https://habr.com/ru/companies/axenix/articles/695960/>).

Практические и лабораторные занятия

Установите одну из программ для визуализации и работы с дашбордами. Протестируйте возможности платформ. Выполните следующие действия: приемы визуализации данных, процедура выбора и настройки цветовой палитры, типа визуализации, области детализации, включения графических объектов, построение и настройка дашборда, публикация отчетов.

Самостоятельная работа

Самостоятельное завершение практического занятия.

Изучите функциональные возможности и основные преимущества инструментов визуализации:

- Построение и настройка дашборда (<https://habr.com/ru/companies/axenix/articles/695960/>).
- Подробно о Visiology (<https://ru.visiology.su/>).
- Подробно об инструментах PolyAnalyst для визуализации данных (<https://www.megaputer.ru/project-gallery/>).

Вопросы для самопроверки

1. Какие тенденции на рынке ВІ в России формируются в последнее время?
2. Раскройте основные преимущества и недостатки Visiology.
3. Раскройте основные преимущества и недостатки PolyAnalyst.

3. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Освоение дисциплины требует систематического изучения всех тем в той последовательности, в какой они указаны в рабочей программе.

Основными видами учебной работы являются аудиторные занятия. Их цель – расширить базовые знания обучающихся по осваиваемой дисциплине и создать систему теоретических ориентиров для последующего более глубокого освоения программного материала в ходе самостоятельной работы.

Обучающемуся важно помнить, что контактная работа с преподавателем эффективно помогает ему овладеть программным материалом благодаря расстановке необходимых акцентов и удержанию внимания интонационными модуляциями голоса, а также подключением аудио-визуального механизма восприятия информации.

Самостоятельная работа преследует следующие цели:

- закрепление и совершенствование теоретических знаний, полученных на лекционных занятиях;
- формирование навыков подготовки текстовой составляющей информации учебного и научного назначения для размещения в различных информационных системах;
- совершенствование навыков поиска научных публикаций и образовательных ресурсов, размещенных в сети Интернет;
- самоконтроль освоения программного материала.

Обучающемуся необходимо помнить, что результаты самостоятельной работы контролируются преподавателем во время проведения мероприятий текущего контроля и учитываются при промежуточной аттестации.

Обучающимся с ОВЗ и инвалидностью предоставляется возможность выбора форм проведения мероприятий текущего контроля, альтернативных форм, предусмотренным рабочей программой дисциплины. Предусматривается возможность увеличения в пределах одного академического часа времени, отводимого на выполнение контрольных мероприятий.

Процедура оценивания результатов обучения инвалидов и лиц с ОВЗ по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации.

При проведении текущего контроля применяются оценочные средства, обеспечивающие передачу информации от обучающегося к преподавателю, с учетом психофизиологических особенностей здоровья обучающихся.

4. ВОПРОСЫ ДЛЯ ПОДГОТОВКИ К ИТОГОВОМУ КОНТРОЛЮ

1. Общее понятие о больших данных. Источники данных.
2. Правовые и этические аспекты использования данных.
3. Принципы текстовой аналитики и разработка концепции исследования.
4. Обзор направлений применения методов текстового анализа в общественных науках.
5. Направления текстового анализа в общественной географии.
6. Применение текстовой аналитики в политологии.
7. Применение текстовой аналитики в социологии.
8. Применение текстовой аналитики в иных науках.
9. Выявление проблемы и разработка концепции исследования.
10. Обзор реализованных кейсов в общественной географии.
11. Оценка восприятия туристско-рекреационного пространства.
12. Анализ социального самочувствия населения в муниципальных образованиях.
13. Принципы составления технического задания. Составление технического задания.
14. Составление таблицы по источникам данных. Сбор данных. Аннотирование данных. Контроль качества.
15. Процедура сбора данных. Разметка данных.
16. Основные функциональные возможности Orange.
17. Основные функциональные возможности Gephi.
18. Анализ социальных сетей с помощью Gephi.
19. Применение low-code и no-code инструментов в общественной географии.
20. Обзор возможностей платформы PolyAnalyst.
21. Функциональные узлы платформы PolyAnalyst для работы с текстовыми данными.
22. Применение инструментов платформы PolyAnalyst.
23. Инструменты визуализации PolyAnalyst. Шаблоны визуализаций.
24. Инструменты визуализации Visiology. Шаблоны визуализаций.
25. Построение и настройка дашборда.

5. СПИСОК РЕКОМЕНДУЕМЫХ ИСТОЧНИКОВ

Рекомендуемая литература (обязательная)

1. Блануца В. И. Общественная география: цифровые приоритеты XXI века М.: ИНФРА-М, 2022. 252 с. EDN: GWTFQA
2. Миркин Б. Г. Введение в анализ данных: учебник и практикум М.: «Юрайт», 2019. 174 с.
3. Титов А. Н. Python. Обработка данных: учебно-методическое пособие Казань: Изд-во КНИТУ, 2022. 104 с.

Рекомендуемая литература (дополнительная)

1. Горбанев В. А. Общественная география зарубежного мира и России: учебник для студентов вузов, обучающихся по специальностям «Экономика», «Социально-экономическая география» и «Природопользование» М: ЮНИТИ-ДАНА, 2018. 567 с.
2. Протодяконов А. В., Пылов П. А., Садовников В. Е. Алгоритмы Data Science и их практическая реализация на Python: учебное пособие – М., Вологда: Инфра-Инженерия, 2022. 392 с. EDN: BITUQD
3. Чубукова И. А. Data Mining: учебное пособие. М., Саратов: Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020. 469 с.

ПРИЛОЖЕНИЯ

Приложение А

Пример составления технического задания

Тема: Текстовая аналитика в общественной географии.

Цель: Выявить навыки текстовой аналитики, которые необходимо включить в тематический план учебной дисциплины «Текстовая аналитика в общественной географии».

Задачи:

1. Составить актуальную базу вакансий в сфере территориального планирования и управления, где требуются навыки работы с текстовой аналитикой.
2. Проанализировать ключевые параметры рынка труда: распределение вакансий по навыкам, компаниям, условиям труда, зарплатным предложениям.
3. Определить список ключевых навыков, запрашиваемых работодателями в вакансиях.
4. Написать рекомендации для учебного плана онлайн-курса «Текстовая аналитика в общественной географии».

Данные:

Вакансии с должностями, указанными в целевых вакансиях.

ТЗ на выгрузку данных:

Критерий	Характеристики
Регион	Россия
Период	01.01.2022 – 30.04.2023
Поля для выгрузки	дата публикации вакансии ID статус вакансии (открыта/закрыта) название (должность) описание требования (если это отдельное поле) заработная плата (нижний и верхний уровень) график работы опыт работы компания-работодатель регион ключевые навыки
Целевые вакансии	руководитель отдела статистики; начальник отдела планировки; начальник отдела транспортного развития территории; начальник отдела концептуальных решений по благоустройству территории; начальник отдела стратегического развития; начальник отдела пространственного планирования (развития); начальник градостроительного отдела; начальник отдела разработки градо-экологического каркаса

Результат выгрузки:

Таблица в формате xls, в которой для каждой вакансии указаны соответствующие поля.

Форма предоставления результатов:

Аналитический отчет в формате кейса, в котором представлены:

1. Рекомендации о необходимых компетенциях для работы в данных организациях.

2. Рекомендации по структуре тематического плана учебной дисциплины «Текстовая аналитика в общественной географии».

3. Заключение о востребованности навыков работы с большими данными для выпускников магистратуры (направление подготовки – География, профиль магистратуры «Территориальное планирование и управление»).

Практическая значимость: С учетом результатов исследования будет скорректирован тематический план онлайн-курса «Текстовая аналитика в общественной географии».

Пример выполнения разметки:

Цель разметки – поиск в социальной сети «ВКонтакте» постов и комментариев студентов, которые отражают их мнение, настроение, эмоциональную оценку по следующим тематическим категориям:

- Учебная вовлеченность.
- Взаимодействие студентов с ППС.
- Социально-бытовые условия/проблемы.
- Финансовые условия/проблемы.
- Стресс и психологическое благополучие.

Категория «Мусор»

Мусор (нерелевантное сообщение) – это сообщение, которое не относится к исследуемым тематическим категориям и в котором содержится текст следующего содержания:

- **Вопросы** (например, *«Добрый вечер, кто подскажет телефон деканата?»*).
- **Информационные/официальные сообщения (анонсы, релизы, афиши)** (например, *«Открыли юбилейный сезон. Команда ВГУВТ успешно выступила в первом этапе всероссийских студенческих соревнований по академической гребле в Санкт–Петербурге»*);
- **Опросы** (например, *«Уважаемые студенты второго курса, пройдите, пожалуйста, опрос по ссылке, очень важно для ВКР!!!»*).
- **Частные объявления** (продам, сдам, куплю, нашел/потерял, знакомства, помощь в выполнении учебных заданий, сдаче определенных предметов, получении контактов преподавателей, размер стипендии, сроки экзаменов и т. д., например, *«Найден пропуск во 2 корпусе на Плехановской, отнес на вахту»*).
- **Общение между студентами** (кино, любовь, отношения, мошенничество, собаки, кошки, одежда, парни, девушки, вечеринки, заведения, игры, межличностные (семейные) отношения, политические взгляды, формы досуга, поездки домой и др.).
- **Реклама, поздравления, акции, розыгрыши** (например, *«Наше все. С днем рождения, Саша!»*).
- **Ироничные сообщения, мемы, шутки.**
- **«Пустое» повествование** (относящееся к тематическим категориям, но текст которого не является ценным и полезным в рамках исследования, содержание которого неоднозначно) (например, *«Впереди выходные и опустевшие корпуса политеха»*).

Примеры нерелевантных сообщений:

- ✓ Всем привет, как зовут высокого парня с рыжими волосами? Видела тебя в 5 корпусе, отзовись
- ✓ Кто с факультета журналистики проходил практику – отзовитесь, пожалуйста, имеется пара вопросов.
- ✓ Кто-нибудь продает гладильную доску?
- ✓ Заседание Студсовета СПбГУ от 14.05.2021
- ✓ Кто самый добрый препод в универе?
- ✓ Лингвисты из ИГХТУ в числе экспертов II Международного конкурса научно-исследовательских работ «Моя интеллект-траектория»! Работа в составе жюри дала преподавателям кафедры иностранных языков и лингвистики возможность познакомиться с новыми форматами проведения научных студенческих конкурсов, завязать новые научные контакты, оценить возможности участия в подобных конкурсах среди студентов-гуманитариев представителей нашего университета. Как написать заявление на социальную стипендию? Платят ли стипендию летом?
- ✓ Есть ли военная кафедра??
- ✓ Всем привет кто хорошо разбирается в экономике и сможет выполнить несколько задач. Всем привет! Пожалуйста, помогите студенту-социологу собрать выборку. Если вы готовы пройти один опрос, то поставьте лайк на запись, я вам напишу. Заранее спасибо)
- ✓ У кого есть лабы по документоведению? (Виноградова)
- ✓ Препод: *исключает меня с онлайн-пары* я: вновь появляюсь, потому что пригласил одногрупп (мем)
- ✓ Когда перешел на дистанционное обучение, но увидел количество заданий: (мем/подпись под картинкой)

Таблица 1

Маркировка текста по категориям

Тематические категории	Примеры сюжетов	Примеры маркеров
Особенность взаимодействия преподавателей / научных руководителей и студентов/аспирантов	отношение к преподавателям и качеству преподавания в университете	препод, консультация, сдача, график приема, пропуск, экзамен...
Учебная вовлеченность студентов и аспирантов, НИР	системные проблемы и организация образовательного процесса (сессия, домашние задания, лекции, самостоятельная работа студентов, экзамены, дедлайны, документооборот, техническое обеспечение учебного процесса, личные кабинеты, канал интернета и т. д.) – как в условиях перехода на дистант, так и вне его	расписание, сессия, занятия, пара, восстановиться, очка, заочка, дистант, семестр, трафик, сайт, аудитория, корпус, обратился, техподдержка, конференция, публикация, конкурс, НИР, научный...

Тематические категории	Примеры сюжетов	Примеры маркеров
Стресс и психологическое благополучие в условиях дистанционного и очного обучения	самочувствие и психологическое состояние студента, в том числе особенности коммуникации с остальными студентами физиологическое состояние	горько, тоска, одиноко, тревога, боюсь, плохо, ненавижу, убить, страшно, унижительно, стыдно, виноват, уверен, плакать, депрессия, устал, надеюсь, не хочу, они, вы, отстой, зачем, требую, умоляю, воодушевлен, доволен, прекрасно, счастье, ура, класс, круто, жду, восхищен, гордость, супер, спасибо, лучшие, восторг, волшебю, красота...

Таблица 2

Маркировка текста по категориям (продолжение)

Тематические категории	Примеры сюжетов	Примеры маркеров
Социально-бытовые условия/проблемы	проблемы студентов в общежитии проживание в кампусе, его доступность; организация питания и медицинского обслуживания	общага, кампус, столовка, заселение, выселение, проживание, пребывание, отключение, обслуживание, шум, админ, комендант, медики, медкабинет, справка...
Финансовые условия/проблемы	стипендии и соцвыплаты, матпомощь возможность совмещения работы и учебы оплата за обучение	стипендия, оплата, материалка, социалка, стоимость обучения..

Составлено на основе монографии: Подольский А. И., Идобаева О. А, Щербакова В. В. и др. Подростки и молодежь во взаимодействии с интернет-контентом. Междисциплинарное исследование с использованием больших данных – М.: Изд-во «Перо», 2021. 210 с.

Учебное издание

Автор-составитель:
Евгений Валерьевич **Коньшев**

Текстовая аналитика в общественной географии

Учебно-методическое пособие

Редактор *А. А. Арустамова*
Корректор *Л. И. Семицветова*
Компьютерная верстка: *Е. В. Коньшев*

Объем данных 3,63 Мб
Подписано к использованию 20.05.2025

Размещено в открытом доступе
на сайте www.psu.ru
в разделе НАУКА / Электронные публикации
и в электронной мультимедийной библиотеке ELiS

Управление издательской деятельности
Пермского государственного
национального исследовательского университета
614068, г. Пермь, ул. Букирева, 15