ОБРАБОТКА НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ БИБЛИОТЕКИ SPACY

Москалев М.Г.[1], Газизов Т.Т.[2]
[1]Томский государственный университет систем управления и радиоэлектроники, аспирант, e-mail: moskalev@tspu.edu.ru
[2]Томский политехнический университет, зам. проректора по цифровизации, e-mail: gtt@tpu.ru

Аннотация

В статье рассматривается проблема обработки и анализа неструктурированных данных в условиях быстрого роста их объема. Обсуждается важность обработки неструктурированных данных для различных сфер деятельности. Рассматривается возможность применения методов извлечения ценных знаний из неструктурированных текстовых данных на примере использования технологий на языке Python для выделения ключевых слов из выпускных квалификационных работ.

Ключевые слова: неструктурированные данные, обработка данных, данные, информация, Python.

Введение

С распространением использования информационных технологий, наблюдается рост объема данных, доступных для обработки. В 2020 году количество информации в сети интернет оценивалось в объем 51 Зеттабайта (10^{21} байт), по состоянию на 2022 год объем данных в сети интернет равен примерно 90 Зеттабайтам, а уже к 2025 году прогнозируют рост объема данных до 175 Зеттабайт [1]. Часть этой информации является созданной людьми, на естественных или формальных языках, однако, с ростом развития интернет вещей появилось множество машиногенерируемой информации, которая создается различными устройствами и датчиками, а также содержит телеметрическую информацию для работы данных устройств [2]. При этом, более 80 % объёма информации является неструктурированными или слабоструктурированными данными [3].

Обработка неструктурированных данных позволяет извлекать ценные знания, выявлять тенденции, принимать обоснованные решения и повышать конкурентоспособность. Технологии обработки естественного языка, компьютерного зрения, обработки речи и другие методы позволяют анализировать и классифицировать неструктурированные данные, а также обеспечить эффективный поиск, который необходим для пользователей, ищущих информацию в интернете или внутри организаций.

Исследования по поиску и обработке неструктурированных данных также имеет огромное значение в контексте современной медицины. С постоянным увеличением объема медицинских данных, таких как клинические записи, изображения, результаты обследований и отчеты, возникает необходимость в разработке эффективных методов и инструментов для их анализа и обработки [4]. Исследования в этой области могут способствовать разработке инновационных алгоритмов машинного обучения и обработки естественного языка, которые позволят автоматизировать процессы анализа медицинских данных. Это может привести к улучшению диагностики заболеваний, предсказанию риска возникновения различных патологий и разработке персонализированных методов лечения. Целью работы является использование технологии на языке Python для выделения ключевых слов из выпускных квалификационных работ с целью обеспечения более эффективного поиска по полнотекстовым. Это также позволит быстрее понять, о чем рассказывается в работе и какие темы в ней затрагиваются.

Неструктурированные данные

Существуют различные понятия структурированности. Саймон Г. и Ньюэлл А. дают следующее понятие слабоструктурированных (или частично структурированных) задач: задачи, которые содержат как качественные, так и количественные элементы, причем качественные, малоизвестные и неопределенные стороны проблем имеют тенденцию доминировать [5]. Под слабоструктурированными данными в данной статье понимаются любые структуры данных между строгой структурированностью и её полным отсутствием.

Неструктурированные данные (или неструктурированная информация) — информация, которая либо не имеет заранее определенной структуры данных, либо не организована в установленном порядке. К неструктурированным относятся данные, произвольные по форме, включающие тексты и

графику, мультимедиа (видео, речь, аудио). Сложность анализа такого типа данных обусловлена невозможностью или трудностью использования традиционных программ для анализа структурированных данных. Для анализа такого типа данных используются особые методы, которые позволяют интерпретировать неструктурированную информацию, например, интеллектуальный анализ данных, обработка естественного языка и интеллектуальный анализ текста [6].

Структурированные текстовые данные можно рассмотреть на примере паспорта гражданина Российской федерации. Он содержит информацию о гражданине, такую как имя, дата рождения, место жительства, серию и номер паспорта, дату выдачи, выдавший орган, пол. Все эти данные представлены в виде структурированного текста. Эти данные в паспорте организованы по определенным структурированным правилам и формату, что делает их легко интерпретируемыми и доступными для обработки компьютерными системами. Благодаря структурированности информации в паспорте, возможно автоматизированное считывание и обработка данных, что облегчает множество административных процедур, включая регистрацию, идентификацию личности, оформление документов и другие операции, требующие использования данных из паспорта.

В 2020 году количество информации в сети интернет оценивалось в объем 51 Зеттабайта (10^{21} байт) (рис. 1) [7], недостаточная обработка и анализ неструктурированных данных может привести к следующим последствиям:

1. потеря информации. Неструктурированные данные могут содержать ценную информацию, которая может быть утеряна при неправильной обработке и анализе;
2. риск принятия неправильных решений. Недостаточная обработка и анализ неструктурированных данных может привести к принятию неправильных решений, что может негативно сказаться на бизнесе или научных исследованиях;
3. ухудшение качества жизни. Неструктурированные данные могут содержать информацию о здоровье, образовании, культуре и других сферах жизни людей. Если эти данные не обрабатываются и не анализируются должным образом, это может привести к ухудшению качества жизни населения.

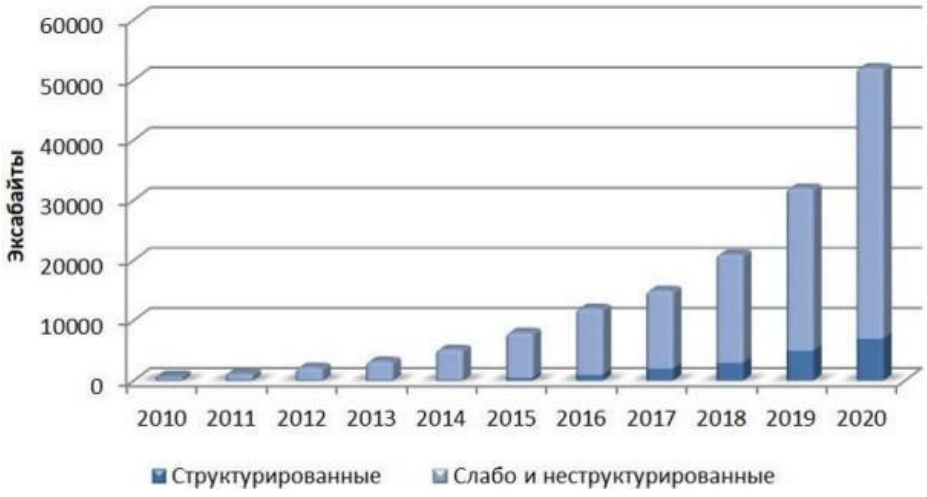

Рис. 1. Рост количества данных различной структуры

В государственной программе «Цифровая экономика Российской Федерации» от 28.07.17 № 1632-р, утвержденной по распоряжению Председателя Правительства РФ, определяются цели и задачи по созданию необходимых условий для развития в России цифровой экономики в рамках «Стратегии развития информационного общества в Российской Федерации на 2017 – 2030 годы» [8]. Одним из важных аспектов утвержденной программы является задача оптимизации систем обработки и обмена информацией. Такая задача ставится в связи с широким распространением систем электронного документооборота, и, в частности, ростом количества используемых текстовых документов, содержащих неструктурированные данные.

Существующие исследования в области обработки неструктурированных текстовых данных используют несколько основных подходов:

1. Машинное обучение. Машинное обучение играет важную роль в задачах обработки неструктурированных текстовых данных, позволяя создавать модели, которые могут автоматически

извлекать и анализировать информацию из текста. Методы машинного обучения, такие как классификация, кластеризация и регрессия, могут быть применены для решения различных задач обработки неструктурированных текстовых данных. Основной задачей применения данного подхода является исключение необходимости программировать и создавать шаблоны для обработки документов. Алгоритмы настраиваются (или обучаются) на основе подготовленных примеров обработки документов. Важным преимуществом данного подхода является возможность регулярного дополнительного обучения алгоритмов новыми примерами [9].

2. Статистические методы. Этот подход к обработке неструктурированных текстовых данных основан на статистических моделях и теории вероятностей для анализа данных, выявления связей и проверки гипотез. Они могут использоваться для описания и анализа данных, проверки статистических гипотез, а также для построения прогностических моделей. Статистические методы обработки неструктурированных данных обычно требуют хорошей предварительной обработки данных, выбора подходящих статистических моделей и анализа результатов для получения значимой информации из неструктурированных данных [10].

3. Глубокое обучение и нейронные сети: Исследования в области обработки неструктурированных текстовых данных все чаще включают в себя применение глубокого обучения и нейронных сетей. Эти методы позволяют создавать модели, способные автоматически извлекать признаки из текста на разных уровнях абстракции, что позволяет решать сложные задачи, такие как машинный перевод, анализ тональности, генерация текста и другие [11].

Одним из основных методов обработки неструктурированных текстовых данных является токенизация, которая представляет собой процесс разделения текста на отдельные слова, фразы или токены. Токенизация позволяет преобразовать текст в структурированный формат для дальнейшего анализа. После токенизации может быть проведена лемматизация, которая приводит слова к их базовой форме. Это позволяет уменьшить вариативность форм слова и упростить дальнейший анализ. Другим важным методом является определение частей речи. Анализ частей речи в тексте позволяет определить роль каждого слова в предложении. Например, это может помочь выделить ключевые слова или фразы, а также понять синтаксическую структуру текста. Определение частей речи является важным этапом для многих задач обработки неструктурированных текстовых данных на естественном языке, таких как извлечение информации, анализ синтаксических зависимостей и генерация текста. Методы анализа тональности позволяют определить эмоциональную окраску текста, выявить положительные, отрицательные или нейтральные отзывы, комментарии или новости.

При выборе между машинным обучением, нейронными сетями и статистическими методами следует учитывать несколько факторов. Во-первых, машинное обучение обладает большей гибкостью и способностью к адаптации к различным типам данных и задачам, что делает его универсальным инструментом для решения широкого спектра проблем. В отличие от статистических методов, которые могут быть ограничены в своей способности обработки сложных зависимостей, и нейронных сетей, которые могут требовать большого количества данных и вычислительных ресурсов, машинное обучение предоставляет баланс между гибкостью, производительностью и эффективностью. Кроме того, машинное обучение позволяет автоматически извлекать признаки из данных, что может быть критически важно в случаях, когда природа зависимостей в данных неизвестна заранее или когда данные имеют высокую размерность. Таким образом, использование машинного обучения оправдано в случаях, когда требуется универсальный и гибкий подход к анализу данных, способный работать с разнообразными типами информации и решать сложные задачи прогнозирования, классификации и кластеризации. Таким образом, для поставленной задачи выделения ключевых слов из выпускных квалификационных работ, в данной статье будут рассматриваться библиотеки машинного обучения.

Использование технологии SpaCy для выделения ключевых слов

Обработка неструктурированных текстовых данных будет рассматриваться на примере использования библиотек SpaCy – написанной на языке Cython [12] и NLTK – полностью разработанной на Python [13]. Эти библиотеки являются наиболее обширными и популярными для обработки неструктурированных текстовых данных на русском языке. В контексте ВКР, SpaCy и NLTK могут быть использованы для выделения ключевых слов, которые играют важную роль в определении тематики и структуры работы, что позволит быстрее понять, о чем рассказывается в работе и какие темы в ней затрагиваются.

Для работы с русскоязычными данными при использовании библиотеки SpaCy, обычно выбирают модель «ru_core_news_sm», которая представляет собой небольшую модель для русского

языка. Эта модель включает в себя различные компоненты, такие как токенизатор, тэггер, синтаксический анализатор и другие, что делает ее подходящей для обработки и анализа текстов на русском языке. Модель «ru_core_news_sm» была выбрана для работы с русскоязычными данными, потому что она предоставляет хорошее сочетание скорости и качества обработки текста на русском языке. Она достаточно легкая, чтобы обеспечить быструю обработку текста, при этом обладает достаточной точностью для выделения ключевых слов, определения частей речи и других задач обработки текста. Таким образом, выбор модели для работы с русскоязычными данными обусловлен балансом между скоростью обработки и качеством результатов, что делает ее подходящим выбором для многих задач обработки текста на русском языке.

В отличии от библиотеки SpaCy, NLTK не предоставляет предварительно обученных моделей и сконцентрирована на предоставлении инструментов для обработки русскоязычных текстов. Для лемматизации и определения частей речи в русском языке, NLTK предоставляет ряд инструментов и ресурсов, таких как WordNetLemmatizer для лемматизации и модуль «averaged_perceptron_tagger» для определения частей речи. Они могут быть использованы для анализа текста на русском языке и проведения различных задач обработки текста. Кроме того, NLTK также предоставляет доступ к корпусам текстов на русском языке, таким как «ru_stopwords», который содержит стоп-слова на русском языке, что может быть полезно для фильтрации стоп-слов в тексте.

В зависимости от поставленных задач могут так же использоваться дополнительные параметры обработки, такие как выделение именованных сущностей или анализ синтаксических зависимостей. Принцип работы данных библиотек и выбранных наборов параметров схож, поэтому в статье будет описана обработка текстовых данных на примере использования библиотеки SpaCy.

Выделение ключевых слов с помощью SpaCy состоит из нескольких этапов. Во-первых, необходимо загрузить модель языка, которая соответствует теме работы. В случае использования библиотеки SpaCy для обработки выпускных квалификационных работ используется библиотека русского языка. После загрузки модели языка необходимо подготовить документ для дальнейшей обработки. Для работы программы необходимо перевести документ в текстовый файл «.txt». Так как чаще всего выпускные квалификационные работы хранятся в формате «.pdf», то для перевода в нужный формат будет использоваться библиотека PDFMiner [14]. PDFMiner – представляет собой простую в использовании библиотеку Python с открытым исходным кодом, предназначенную для работы с файлами PDF и позволяет конвертировать их в различные текстовые форматы. Указывается путь до папки с файлами ВКР и передается в библиотеку PDFMiner для обработки. Следующим этапом обработки является передача полученных текстовых данных в библиотеку SpaCy. Она разбивает весь текст на отдельные слова и фразы, что позволяет выделить ключевые слова характеризующие ВКР. Далее происходит выделение леммы для преобразования слов в начальную форму и объединения возможных склонений. На следующем шаге можно использовать функцию для определения частей речи каждого слова. Это может помочь определить, какие слова являются ключевыми и наиболее важными для темы работы. Например, глаголы, существительные и прилагательные часто являются наиболее значимыми словами в тексте. Во время выполнения кода программы можно отслеживать статус выполнения и приблизительное оставшееся время. Ключевые слова по каждой ВКР записываются в виде массива в файл «.txt» с названием идентичным названию файла работы. Это позволит в дальнейшем использовать полученную информацию без необходимости повторного запуска программы. Пример работы программы показан на рис. 2.

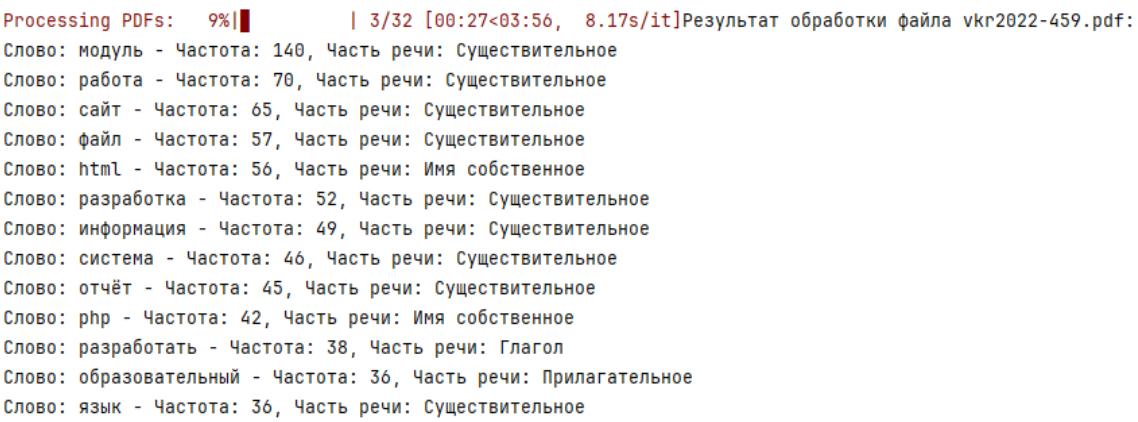

Рис. 2. Пример работы программы по определению ключевых слов в ВКР

Для проведения эксперимента использовались документы 32 выпускных квалификационных работ в pdf формате. Данные были переведены в одну строку текста с удалением переносов, знаков пунктуации и пустых строк. Результаты апробации библиотек представлены в таблице 1. Полученные данные включают: количество успешно определенных слов (токенов), скорость определения токенов и скорость определения частей речи для каждой библиотеки. Данное сравнение позволяет определить эффективность библиотек в обработке текстовых данных и определении частей речи на большом объеме документов, что имеет важное значение для выбора наиболее подходящей библиотеки для конкретных задач обработки текста.

Таблица 1

Результаты выделения ключевых слов в ВКР

Библиотека	**Количество определенных слов (всего)**	**Скорость определения слов (всего)**	**Скорость определения частей речи (всего)**
SpaCy	303272 слова	256,05 секунд	0.09224 секунд
NLTK	295300 слов	84,35 секунд	12.08061 секунд

Заключение

В данной статье рассмотрены проблемы роста объема неструктурированных данных, обосновывается актуальность работы с неструктурированными текстовыми данными. Дается определения слабоструктурированных и неструктурированных данных. Приводятся риски, связанные с недостаточной обработкой неструктурированных данных. Рассматривается пример использования библиотек SpaCy и NLTK на языке Python для выделения ключевых слов из выпускных квалификационных работ. Целью этого подхода является улучшение понимания тематики и структуры работы без необходимости прочтения полного текста. Также он направлен на повышение эффективности при поиске данных в выпускных квалификационных работах. Результаты сравнения двух библиотек показывают, что для задачи выделения ключевых слов лучше подходит библиотека NLTK, так как она осуществляет процесс токенизации в три раза быстрее чем библиотека SpaCy. Однако с задачей определения частей речи библиотека SpaCy справляется гораздо быстрее за счет наличия предобученых моделей русского языка. Таким образом, можно сделать вывод что для более эффективной обработки неструктурированных данных на русском языке можно использовать библиотеку SpaCy для определения частей речи, в то время как для задачи токенизации библиотека NLTK может оказаться более предпочтительной из-за своей скорости. Однако, выбор между этими библиотеками также зависит от конкретных потребностей и характеристик задачи, а также от других факторов, таких как поддержка языков, наличие необходимых моделей и удобство использования. Данные, полученные в ходе работы программ, могут быть использованы для последующего анализа. Дальнейшие исследования по этой теме планируют внедрение классификаторов для проведения оценки выделенных ключевых слов.

Список использованных источников

1. IDC. How IDC's Industry CloudPath & SaaSPath Surveys Can Inform Your Cloud & SaaS Strategy. — URL: https://blogs.idc.com/2019/09/04/how-idcs-industry-cloudpath-saaspath-surveys-can-inform-your-cloud-saas-strategy/ (дата обращения: 10.02.2024).

2. IDC. Future of industry ecosystems shared data and insights. — URL: https://blogs.idc.com/2021/01/06/future-of-industry-ecosystems-shared-data-and-insights (дата обращения: 10.02.2024).

3. Forbes. The big unstructured data problem. — URL: https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/?sh=4bd0fcf2493a (дата обращения: 12.02.2024).

4. Li B., Li J. Jiang Y., Lan X. Experience and Reflection from China’s Xiangya Medical Big Data Project // Journal of Biomedical Informatics. – 2009 – Vol. 93.

5. Astera. Unstructured Data Challenges for 2023 and their Solutions. – URL: https://www.astera.com/type/blog/unstructured-data-challenges/ (дата обращения: 13.02.2024)

6. Орлова, К.И. Неструктурированные данные и технологии их обработки / К.И. Орлова, А.А. Орлюк // Актуальные направления фундаментальных и прикладных исследований: материалы XI международной научно-практической конференции, North Charleston, 27–28 февраля 2017 года / НИЦ «Академический». Том 1. – North Charleston: CreateSpace. – 2017. – С. 149-152. – EDN YAAIXJ.

7. Azad P., Navimipour N., Rahmani A., Sharifi A. The role of structured and unstructured data managing mechanisms in the Internet of things. / // Cluster Computing. – 2020. – No. 23. – P. 1-14

8. Об утверждении программы «Цифровая экономика Российской Федерации». – URL: http://government.ru/docs/28653/ (дата обращения: 13.02.2024).

9. Лисенков И.А., Кузнецов В.А., Леонова Н.М. Обработка неструктурированной текстовой информации с помощью алгоритмов машинного обучения. Вестник НИЯУ МИФИ. – 2020. – Т. 9(4). – С. 376–384. https://doi.org/10.56304/S2304487X20040057

10. Клячкин В.Н. Статистические методы анализа данных / В.Н. Клячкин, Ю.Е. Кувайскова, В.А. Алексеева. – Москва : Издательство "Финансы и статистика", 2016. – 240 с. – ISBN 978-5-279-03583-0.

11. Войтко Е.Д. Применение нейросетевых технологий для обработки неструктурированной текстовой информации / Е.Д. Войтко, Н.П. Курочка // Информационная безопасность : сборник статей конференции, Анапа, 17 апреля 2019 года. – Анапа: Федеральное государственное автономное учреждение "Военный инновационный технополис "ЭРА", 2019. – С. 23–35. – EDN RJPJLF.

12. SpaCy. URL: https://spacy.io/ (дата обращения: 14.02.2024).

13. NLTK. URL: https://nltk.org, свободный (дата обращения: 14.02.2024).

14. PDFMiner. URL: https://github.com/pdfminer/pdfminer.six (дата обращения: 14.02.2024).