Интеллектуальные системы извлечения текста и структур из документов

ТКАЧУК АНДРЕЙ ИВАНОВИЧ

Гродненский государственный университет им.Янки Купалы,
факультет математики и информатики,
кафедра системного программирования и компьютерной безопасности
студент 3 курса специальности «Компьютерная безопасность»

Научный руководитель Косарева Екатерина Владимировна, к. ф.-м. н., доцент

В работе исследуются методы и реализации, помогающие при анализе естественных языков и неструктурированных данных в целом. Используются имеющиеся знания для определения тональности текста. Изучались методы анализа неструктурированных данных, а также методы анализа семантики естественного языка, связанные с созданием векторно-семантических моделей. Областью применения является анализ неструктурированных данных, обнаружение в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Ключевые слова: МОДЕЛИ РАСПРЕДЕЛЕННОГО ПРЕДСТАВЛЕНИЯ, АНАЛИЗ ЕСТЕСТВЕННОГО ЯЗЫКА, ТОНАЛЬНОСТЬ ТЕКСТА, АНАЛИЗ ТЕКСТА, АНАЛИЗ ДОКУМЕНТОВ.

Тема анализа неструктурированных данных является актуальной уже долгое время. Теперь же, во времена больших данных этот вопрос является еще более востребованным. Постоянный рост хранимых данных, как их объема, так и разнообразия самих данных, которые не только хранятся, но и так или иначе обрабатываются, довольно сильно усложняют задачу управления корпоративными данными. В этих обстоятельствах умение качественно анализировать корпоративную информацию и оперативно реагировать на любые несоответствия её хранения политикам и требованиям бизнеса является ключевым показателем зрелости информационной стратегии организации [1].

Неструктурированные данные — это информация во многих различных формах, которая не соответствует традиционным моделям данных и, таким образом, обычно не подходит для основной реляционной базы данных. Благодаря появлению альтернативных платформ для хранения и управления такими данными, они все более распространены в ИТ-системах и используются организациями в различных приложениях для бизнес-аналитики и аналитики.

Одним из наиболее распространенных типов неструктурированных данных является текст. Неструктурированный текст генерируется и собирается в самых разных формах, включая документы Word, сообщения электронной почты, презентации PowerPoint, ответы на опросы, стенограммы взаимодействий центра обработки вызовов и сообщения из блогов и сайтов социальных сетей.

Другие типы неструктурированных данных включают изображения, аудио и видео файлы. Машинные данные — это еще одна категория, которая быстро растет во многих организациях. Например, файлы логов с веб-сайтов, серверов, сетей и приложений, особенно мобильных, дают множество данных об активности и производительности. Кроме того, компании все чаще собирают и анализируют данные с датчиков производственного оборудования и других устройств, подключенных к Интернету (IoT).

В силу того, что большинство информации, которую мы можем получить, является неструктурированными данными, плюсы анализа этих данных нельзя переоценить. Анализ

приносит множество действенных идей о том, как можно улучшить производительность компаний или каких-либо услуг.

Если мы говорим, что структурированные данные “большие”, то неструктурированные данные, безусловно, “огромные”. Компьютеры хороши в обработке структурированных данных, потому что структурированные данные – это преимущественно число в своей основе — сколько раз страница была посещена, сколько времени кто-то был на вашем сайте, какие продукты они купили, откуда они пришли. Однако какую интерпретацию компьютер может получить из необработанного текста?

Многие компании используют несколько разных подходов к этому вопросу, один из них — применение грубой силы и ручного труда для чтения необработанного текста, интерпретации эмоций и тональности, а затем преобразования их в структурированные данные и визуализация. К сожалению, грубая сила тратит много времени и неэффективна в целом. На этом фоне родилась концепция анализа неструктурированных данных. Этот анализ можно представляет собой набор процессов преобразования неструктурированных данных, в данные пригодные для анализа. Итак, вопрос в том, почему мы должны преобразовывать данные? Почему мы не можем просто передать данные на компьютер? Ответ заключается в том, что компьютер — это машина, которая не имеет возможности четко знать, как обрабатывать текст или слова. Тогда как мы должны преобразовать данные, чтобы они могли обрабатываться компьютером как машиной? Чтобы получить понятные информацию и знания из текста, необходимо предпринять несколько шагов: предварительная обработка, извлечение признаков, обработка посредством алгоритмов машинного обучения и процесс аналитики. И в конце процесса мы можем получить информацию, которая позволяет нам делать, например, анализ тональности для обзора продукта или жалоб клиентов, мониторинг социальных сетей, фильтр спама в электронной почте или получение главной информации из текста.

Работа с неструктурированными данными

Так как неструктурированные данные обычно хранятся в форме электронных документов, программы для анализа содержания или управления документами предпочитают классифицировать скорее целые документы, чем производить манипуляции внутри документов. Таким образом, программы для обработки такого типа данных обычно представляют средства для создания коллекций документов с неструктурированной информацией. Однако сегодня существуют также решения, работающие с атомарными элементами меньшими, чем целый документ.

Интеллектуальный анализ данных

Интеллектуальный анализ данных — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин введён Григорием Пятецким-Шапиро в 1989 году [2].

Data mining – это в первую очередь различные методы моделирования, классификации и прогнозирования, реализуемые с помощью с помощью нейронных сетей, деревьев решений, эволюционного программирования, ассоциативной памяти, нечёткой логики. Также часто можно столкнуться с упоминаниями статистических методов, таких как корреляционный, дескриптивный, регрессионный, факторный, дисперсионный, компонентный, дискриминантный анализы, анализ временных рядов, анализ выживаемости, анализ связей.

Обработка естественного текста

Обработка естественного языка (Natural Language Processing, NLP) — общее направление искусственного интеллекта и математической лингвистики. Оно изучает

проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека [3].

Проще говоря, NLP служит мостом между компьютером и естественным языком человека (речи и текста).

Интеллектуальный анализ текста

Интеллектуальный анализ текстов (ИАТ, англ. text mining) — направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка. Название «интеллектуальный анализ текстов» перекликается с понятием «интеллектуальный анализ данных» (ИАД, англ. data mining), что выражает схожесть их целей, подходов к переработке информации и сфер применения; разница проявляется лишь в конечных методах, а также в том, что ИАД имеет дело с хранилищами и базами данных, а не электронными библиотеками и корпусами текстов [4].

Text mining — это метод извлечения полезной информации из неструктурированных данных посредством идентификации и исследования больших объемов текста. Проще говоря, интеллектуальный анализ текста — это метод извлечения структурированной информации из неструктурированного данных (в частности, текста).

Реализация анализа неструктурированных данных. Библиотека word2vec

Word2vec — это хорошо известная концепция, используемая для генерации векторного представления из слов. Модели машинного обучения принимают векторы (массивы чисел) в качестве входных данных. При работе с текстом первым делом мы должны придумать стратегию преобразования строк в числа (или «векторизации» текста) перед подачей его в модель.

Библиотека doc2vec

Целью doc2vec является создание числового представления документа независимо от его длины. Но в отличие от слов документы не имеют логической структуры, такой как при анализе слов, поэтому необходимо найти другой метод.

Концепция предусматривает модель, схожую с модель word2vec за тем исключением, что был добавлен еще один вектор — идентификатор параграфа. То есть вместо того, чтобы использовать только слова для предсказания следующего слова, был также добавлен еще один вектор, который является уникальным для документа.

Библиотека fast.ai

Библиотека fast.ai ориентирована на использование предварительно обученных языковых моделей и их точную настройку, с целью обработки посредством NLP не только английского языка. Выполняется это в три этапа:

• Предварительная обработка данных в минимальном объеме.

• Создание языковую модель с предварительно подготовленными весами, которые вы можете точно настроить в своем наборе данных.

• Создание других моделей, такие как классификаторы, поверх языковой модели.

Библиотека DeepPavlov

DeepPavlov — это библиотека с открытым исходным кодом для разработки чат-ботов и виртуальных помощников. Она имеет всеобъемлющие и гибкие инструменты, которые позволяют разработчикам и исследователям NLP создавать готовые к разговору навыки общения и сложные многофункциональные разговорные помощники.

Также библиотека предусматривает работу с текстом на русском языке: она способна работать с предварительно обученными моделями BERT:

• RuBERT для русского языка;

• Slavic BERT для болгарского, чешского, польского и русского языков;

• Conversational BERT для разговорного русского.

Заключение

В рамках работы мы ознакомились с понятием неструктурированных данных, а также поняли, почему так важно иметь возможность их анализировать. Определили основные способы анализа данных и что нужно для извлечения информации из неструктурированных данных. Также были показаны примеры реализации библиотек для анализа текста естественного языка и основные понятия, необходимые для понимания их принципов работы. Были упомянуты цели и задачи, реализуемые посредством анализа естественного языка (NLP), и то, как они помогают в основных сферах жизни. Одна из сфер применения NLP — анализ тональности текста — была выбрана в качестве главной задачи для практической части. Для данной задачи были приведены примеры библиотек для обработки естественного языка на русском языке, описан общий принцип их работы и представлен набор данных для реализации поставленной задачи.

Литература

1. Анализ неструктурированных данных и оптимизация их хранения [Электронный ресурс] / https://habr.com/ru/company/hpe/blog/265499/. Режим доступа: свободный. (дата обращения: 09.11.2019)

2. Data mining [Электронный ресурс] / https://ru.wikipedia.org/wiki/Data_mining. Режим доступа: свободный. (дата обращения: 11.11.2019)

3. Обработка естественного языка [Электронный ресурс] / https://ru.wikipedia.org/wiki/Обработка_естественного_языка Режим доступа: свободный. (дата обращения: 11.11.2019)

4. Интеллектуальный анализ текста [Электронный ресурс] / https://ru.wikipedia.org/wiki/Интеллектуальный_анализ_текста - Режим доступа: свободный. (дата обращения: 09.11.2019)