ПРОЕКТИРОВАНИЕ КОНВЕЙЕРА АВТОМАТИЗИРОВАННОГО ИЗВЛЕЧЕНИЯ НАРРАТИВОВ ИЗ ТЕКСТОВЫХ ДАННЫХ НА ПРИМЕРЕ ОТКРЫТЫХ ИСТОЧНИКОВ TELEGRAM-КАНАЛОВ

Пройдаков М.К.
Томский политехнический университет, ИШИТР, студент гр. 8ИМ2М, e-mail: mkp3@tpu.ru

Аннотация

В работе рассматривается задача автоматизированного извлечения нарративов из текстовых данных для последующего проблемно-ориентированного анализа. Описано формализованное представление нарратива и архитектура конвейера обработки данных для его извлечения, включая используемые на различных этапах обработки AI-методы и технологии.

Ключевые слова: автоматизированное извлечение нарратива, конвейер данных, NLP.

Введение

В русском языке слово «нарратив» является заимствованным и происходит от латинского narrare - рассказывать, являясь родственным латинскому gnarus - знать. Таким образом, уже в этимологии этого понятия зафиксирована идея перевода имеющегося "знания" в "рассказывание". Рассказывая, человек не только прослеживает последовательность событий, но и интерпретирует, познает окружающую действительность и самого себя [1].

Нарратив является одним из наиболее универсальных видов речевого взаимодействия, играет важную роль в нашей жизни, формируя представления о мире, о себе и о других людях, а также занимает важное место в процессах осмысления и передачи человеческого опыта.

Нарратив предстает как окно в индивидуальный человеческий опыт, выраженный через повествование, историю. В условиях межкультурной коммуникации нарратив становится носителем национально-культурной специфики, отождествляя говорящего с определенной лингвокультурной общностью [2].

Автоматизированное извлечение нарративов позволит обрабатывать гораздо большее количество данных, нежели ручной способ, что важно для таких областей как маркетинг и реклама (отзывы потребителей, комментарии в социальных сетях и других источниках), социальные и гуманитарные науки (анализ литературных произведений, исторических документов, интервью и т. д.), медицина и здравоохранение (психотерапия, форумы пациентов и др.), образование и обучение (учебные материалы, студенческие работы).

Целью проекта являлось проектирование конвейера данных, который предоставил бы возможность автоматизированного извлечения нарратива.

Основная часть

Существует огромное количество дефиниций нарратива. В данной статье будет использоваться следующее определение: нарратив – монолог-повествование, рассказ о каком-либо произошедшем ранее событии [3].

Для извлечения нарратива из текста необходимо определить составляющие его части. У. Лабов выделял шесть элементов структуры нарратива, каждый из которых несет определенную функциональную нагрузку:

1. Краткое изложение для слушателя существа происходящих событий, о которых пойдет речь.
2. Ориентация слушателя относительно времени, места, ситуации, участниках событий.
3. Передача последовательности событий (что случилось).
4. Оценка значимости и смысла действий, выражение отношения рассказчика к событиям и действиям.
5. Разрешение ситуации (что случилось в конце концов).
6. Кода, возвращающая к настоящему времени [4].

Исходя из вышеописанного, были определены основные составляющие аспекты нарратива, необходимые для извлечения из текста:

1) тема;
2) действующие лица;
3) временная линия;
4) события;
5) мнение автора.

Для решения подзадач извлечения нарративов необходимо использовать обработку естественного языка. Обработка естественного языка (Natural Language Processing, NLP) – это область искусственного интеллекта, которая занимается взаимодействием между компьютерами и естественными языками (языками, которыми говорят люди).

Для решения задачи извлечения нарратива был спроектирован конвейер данных (Data Pipeline) – набор элементов обработки данных, соединенных последовательно, где выход одного элемента является входом следующего [5]. Попадая в модуль (который является набором методов обработки естественного языка), данные подвергаются изменению, переходя от модуля к модулю, пока не завершатся все необходимые преобразования, после чего готовые данные доставляются конечному пользователю. Схема конвейера данных изображена ниже на рис. 1.

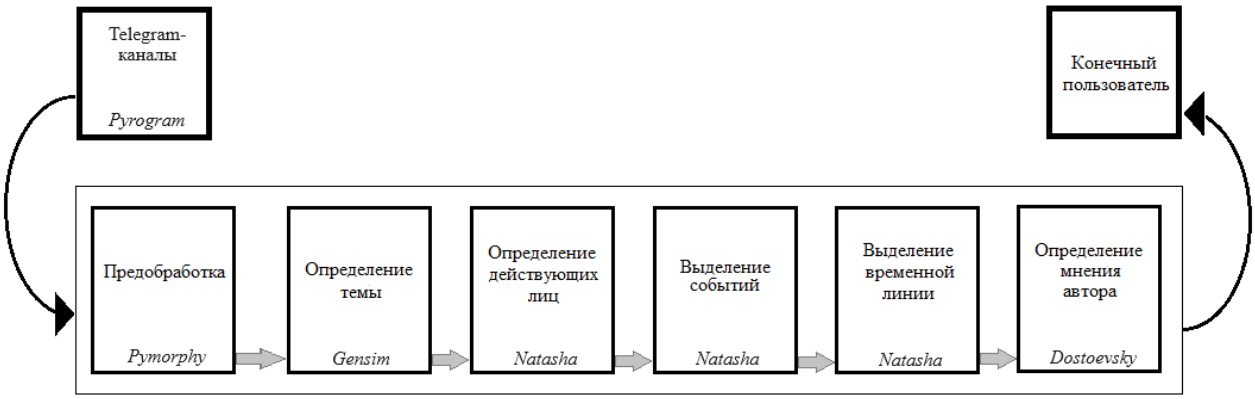

Рис. 1. Схема конвейера данных

В общем случае конвейеры данных состоят из трех основных компонентов: источника, одного или нескольких этапов обработки и пункта назначения.

В спроектированном нами конвейере источником данных являются Telegram-каналы. Для получения данных из каналов используется библиотека Pyrogram.

Сырые данные попадают в модуль предобработки, где подвергаются первой трансформации. Поскольку алгоритмы машинного обучения не могут работать с текстом на естественном языке, для работы с ним нужна предварительная подготовка текстов. Предобработка текста переводит текст на естественном языке в формат удобный для дальнейшей работы. Предобработка состоит из различных этапов, которые могут отличаться в зависимости от задачи и реализации. Первым этапом подготовки текста является его нормализация, которая заключается в приведении текстов к одному регистру, удалении знаков пунктуации, чисел и пробельных символов. Это необходимо для стандартизации методов обработки текста. Затем следует этап токенизации, который включает разбиение длинных строк на более короткие, обычно по словам. После токенизации происходит удаление стоп-слов, то есть слов, не несущих смысловой нагрузки, таких как союзы, предлоги и связки [6]. Далее идет процесс лемматизации, который приводит слова к их словарной форме (лемме). Для предварительной обработки текста использовалась библиотека Pymorphy.

Следующим модулем является модуль определения темы. Для определения темы была использована библиотека Gensim. Используя латентное размещение Дирихле (Latent Dirichlet Allocation, LDA), вероятностную модель, которая предполагает, что каждый документ в коллекции представляет собой смесь различных тем, а каждая тема представляет собой распределение слов, LDA ищет такие темы и их распределение в документах, чтобы объяснить наблюдаемые данные.

Модуль определения действующих лиц использует распознавание именованных сущностей (Named Entity Recognition, NER) – процесс обнаружения и классификации именованных сущностей в тексте. В данном модуле задача NER заключается в том, чтобы определить находится ли слово или группа слов в тексте в предопределенных категориях (персоналии, организации, локации и т. д.). Распознавание именованных сущностей реализовано средствами библиотеки Natasha.

Выделение событий из текста происходит с помощью определения части речи слов в тексте с помощью морфологического анализа, и извлечения отношений – извлечения информации (Information Extraction) о событиях путем выявления отношений между сущностями в тексте. Например, извлечение отношений «субъект-глагол-объект». Для этой задачи также была использована библиотека Natasha.

Модуль выделения временной линии использует упомянутое выше распознавание именованных сущностей, однако определяет находится ли слово в категориях дата, время и т.д.

Определение мнения автора является сложной задачей, устоявшихся решений для которой в настоящий момент не существует. В рамках нашего конвейера данных была использована библиотека Dostoevsky и её анализ тональности (Sentimental analysis) – определение тональности текста, то есть его эмоциональной окраски (позитивной, негативной или нейтральной).

В конце концов полностью преобразованные данные предоставляются конечному пользователю в виде JSON-файла. Формат JSON удобен для хранения получившихся данных, а также при необходимости дальнейшей работы пользователя с данными, например, для анализа получившихся результатов.

Заключение

На данном этапе реализации проекта результатами являются формализованное представление нарратива и спроектированная архитектура конвейера обработки данных. Следующим этапом является разработка конвейера.

Список использованных источников

1. Репьевская М.В. Подходы к изучению нарратива // Cyberleninka: сайт. – 2012. – URL: https://cyberleninka.ru/article/n/podhody-k-izucheniyu-narrativa.

2. Обдалова О.А., Левашкина З.Н. Понятие «наратив» как феномен культуры и объект дискурсивной деятельности // Cyberleninka: сайт. – 2019. – URL: https://cyberleninka.ru/article/n/ponyatie-narrativ-kak-fenomen-kultury-i-obekt-diskursivnoy-deyatelnosti.

3. Лапнина Е.В. Полный словарь лингвистических терминов. – Ростов-на-Дону : Феникс, 2010. – С. 225.

4. Евстигнеева Н.В., Обремко О.А. Модели анализ нарратива // Cyberleninka: сайт. – 2007. – URL: https://cyberleninka.ru/article/n/modeli-analiza-narrativa.

5. Data Engineering: ETL, ELT, Data Pipeline, Data Warehouse, Data Lakes, Data Marts // Иван Шамаев: сайт. – 2021. – URL: https://ivan-shamaev.ru/data-engineering-etl-pipeline-data-warehouse-datalake.

6. Акжолов Р.К., Верига А.В. Предобработка текста для решения задач NLP // Cyberleninka: сайт. – 2020. – URL: https://cyberleninka.ru/article/n/predobrabotka-teksta-dlya-resheniya-zadach-nlp.