Интеллектуальный анализ текстовых данных для решения задачи категоризации информации

Д. М. Лосева
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
d.m.loseva@gmail.com

***Аннотация*. В докладе рассматривается применение алгоритмов Text Mining таких как семантический анализ текста и поиск ключевых слов для решения задачи категоризации данных, поступающих в информационную систему в виде коротких сообщений в текстовом формате. Приведен пример применения подобных алгоритмов в информационной системе обработки сообщений пользователей.**

Ключевые слова: Text Mining; ключевые слова; семантический анализ текста; нейронные сети

I. Введение

Анализ текстовых данных – актуальная задача, так как люди по всему миру ежесекундно генерируют информацию, и это не только видео и картинки. Текст – это деловая переписка, договора и сообщения в соцсетях. Именно поэтому для того, чтобы быть способным максимально эффективно использовать имеющиеся аналитические ресурсы, нужно уметь анализировать текстовые данные.

II. Извлечение понятий (поиск ключевых слов вручную)

Самый очевидный способ поиска ключевых слов в тексте – ручной метод, когда человек самостоятельно анализирует текст и выписывает слова и словосочетания, определяющие смысловую нагрузку предложенных для изучения текстовых данных. В зависимости от тематики текста может потребоваться участие специалистов разных сфер.

Выделяют несколько видов ключевых слов для извлечения:

1) Ключевые слова с длинным хвостом – конкретные и длинные ключевые слова, ориентированные на свою нишу (то есть соответствующие сфере текста). Чем длиннее и точнее будут условия поиска, тем легче будет ранжировать этот термин. Ключевые слова, которые являются более конкретными и, в большинстве случаев, более длинными, обычно называют ключевыми словами с длинным хвостом. Поисковые фразы с длинным хвостом очень актуальны для конкретных ниш и обычно более точно отражают намерения поисковика по сравнению с общими фразами.

2) Поисковое намерение – цель поиска выяснить, чего на самом деле хочет поисковик. Без четкого понимания цели поиска посетителей даже самая хорошо финансируемая кампания почти наверняка потерпит неудачу. Однако, используя поисковое намерение для маркетинга, основанного на ключевых словах, рекламодатели могут не только увеличить посещаемость своих сайтов, но и привлечь потенциальных клиентов, увеличивая продажи.

Почему так важно извлечение ключевых слов? С помощью извлечения ключевых слов вы можете найти самые важные слова и фразы в огромных наборах данных за считанные секунды. Эти слова и фразы могут дать ценную информацию о темах, о которых говорят ваши клиенты.

Учитывая, что более 80 % данных, которые мы генерируем каждый день, являются неструктурированными, то есть они не организованы заранее определенным образом, что чрезвычайно затрудняет анализ и обработку, компаниям необходимо автоматическое извлечение ключевых слов, чтобы помочь им обрабатывать и анализировать данные о клиентах более эффективно и оптимальным способом.

Извлечение информации выявляет соответствующие фрагменты данных при поиске в различных документах. Он также ориентирован на извлечение структурированной информации из произвольного текста и сохранение этих сущностей, атрибутов и информации о взаимосвязях в базе данных. [2]

Общие подзадачи извлечения информации включают:

- Выбор функций или выбор атрибутов – это процесс выбора важных функций (измерений), которые в наибольшей степени способствуют выходу модели прогнозной аналитики.
- Извлечение признаков – это процесс выбора подмножества признаков для повышения точности задачи классификации. Это особенно важно для уменьшения размерности.
- Распознавание именованных сущностей (NER), также известное как идентификация сущностей или извлечение сущностей, направлено на поиск и категоризацию определенных сущностей в тексте, таких как имена или местоположения. Например,

NER определяет «Санкт-Петербург» как место, а «Марию» – как женское имя. [1]

III. Поиск ключевых слов

Метод Text Mining, по сути, использует результаты тематического индексирования для поиска документов, отвечающих указанным требованиям, в частности, содержащих указанные пользователем ключевые слова. [2]

В зависимости от базы данных эти данные могут быть организованы как:

- Структурированные данные – данные, стандартизированые (то есть структурированные) в табличный формат, что упрощает их хранение и обработку для анализа и алгоритмов машинного обучения.
- Неструктурированные данные – данные, у которых нет предопределенного формата. Это может быть текст из социальных сетей, из обзоров продуктов, извлеченный из мультимедийных файлов, такие как видео и аудио файлы, и так далее.
- Полу-структурированные данные – смесь структурированных и неструктурированных форматов данных. Хотя у него есть некоторая организация, у него недостаточно структуры для удовлетворения требований реляционной базы данных. Примеры полуструктурированных данных включают файлы XML, JSON и HTML.

Ключевое слово в Text Mining определяется как набор слов, отражающих и представляющих содержимое текста. Существует множество лингвистических и математических методов, позволяющих находить ключевые слова; наиболее распространенный из них – анализ частоты появления слов в тексте.

Определять ключевые слова в автоматическом режиме позволит алгоритм подсчета встречаемости слов в том или ином сегменте соответствия текста ответу. Каждое слово подсчитывается для отдельно взятой группы текстов, соотнесенных по смыслу, и вычленяются лидеры по количеству использований в отдельно взятом варианте текста, за исключением предлогов и определенных системой и разработчиками стоп-слов.

Стоп-слово подразумевает под собой слово, которое не будет использовать при подсчете в конкретном наборе текста. [1]

IV. Тематическое индексирование

Под термином индексирование первоначально понималось присвоение текстовым данным тематических индексов, отражающих некие атрибуты их классификации (по принципу библиотечных каталогов). С развитием направление приобрело смысл процесса своеобразного «перевода» описаний текстовых данных с естественного языка на формализованный, когда эти описания представляют собой перечни ключевых слов и словосочетаний, отражающие их тематическое содержание. Такая форма получила название поискового образа описаний; при этом поисковый образ запроса представляет собой логическую конструкцию, где слова и словосочетания соединены при помощи логических и синтаксических операторов.

Одной из важных частей индексирования является обработка естественного языка. Обработка естественного языка, которая возникла из компьютерной лингвистики, использует методы из различных дисциплин, таких как информатика, искусственный интеллект, лингвистика и наука о данных, чтобы позволить компьютерам понимать человеческий язык как в письменной, так и в устной формах. Анализируя структуру предложения и грамматику, подзадачи НЛП позволяют компьютерам «читать». Общие подзадачи включают:

- Обобщение: этот метод обеспечивает синопсис длинных фрагментов текста для создания краткого, связного резюме основных положений документа.
- Тегирование части речи (или тематическое индексирование): этот метод присваивает тег каждому токену в документе на основе его части речи, то есть обозначает существительные, глаголы, прилагательные и так далее. Этот шаг обеспечивает семантический анализ неструктурированного текста. Под токеном при этом понимается замещение объекта неким обозначением и/или идентификатором.
- Классификация текста: эта задача, также известная как классификация текста, отвечает за анализ текстовых документов и их классификацию на основе заранее определенных тем или категорий. Эта подзадача особенно полезна при классификации синонимов и сокращений.
- Анализ настроений: эта задача обнаруживает положительные или отрицательные настроения из внутренних или внешних источников данных, позволяя отслеживать изменения в отношении клиентов с течением времени. Обычно он используется для предоставления информации о восприятии брендов, продуктов и услуг. Эти идеи могут подтолкнуть компании к налаживанию связи с клиентами и улучшить процессы и взаимодействие с пользователем. [3]

V. Семантический анализ текста

Семантический анализ – это один из пунктов в последовательном алгоритме автоматического понимания текстов, основывающийся в выделении семантических отношений, формировании семантического представления текстов. Один из возможных вариантов представления – структура, состоящая из так называемых текстовых фактов. [3]

В общем случае семантическое представление является графом, семантической сетью, отражающим бинарные отношения между двумя узлами – смысловыми единицами текста. Глубина семантического анализа может быть

разной, а в реальных системах чаще всего строится только лишь синтаксико-семантическое представление текста или отдельных предложений.

Показатели семантического анализа можно разделить на следующие категории:

1. Частота встречаемости ключевых слов – частота употребления в тексте слов, составляющих семантическое ядро.
2. Стоп-слова – количество в тексте слов, не несущих смысловой нагрузки (предлоги, союзы, местоимения, наиболее часто употребительные в интернете существительные, глаголы и др.). Стоп-слова при индексации не учитываются нашей системой, о чем следует помнить, вычисляя процент плотности в тексте ключевых слов. Стоп-слова относятся к так называемой воде.
3. «Вода» – совокупность в тексте незначимых слов и выражений. [4]

VI. Пример примененя

Сфера, в которой автоматизация процесса анализа текста действительно незаменима, – это работа технической поддержки. Ежедневно операторы получают огромное количество информации, причем это не только сообщения от пользователей, но и различные уведомления о проводимых работах, уведомления систем. При этом информация не всегда поступает в чистом текстовом формате, но и в виде аудио-сообщений (звонков), например. Выбрав оптимальный метод для предварительной обработки информации, можно все возникающие задачи для анализа текста передавать на обработку универсальному алгоритму. В задачах технической поддержки может хватить даже простого метода поиска ключевых слов, при более чуткой настройке используемой для анализа системы.

При этом можно утверждать, что внедрение подобных решений повышает эффективность работы технической поддержки в разы. [5]

VII. Заключение

Рассмотренные методы проведения интеллектуального анализа текстовых данных для решения задачи категоризации информации являются многообещающими способами для оптимизации и автоматизации многих процессов, в том числе и рассмотренного примера с технической поддержки. Учитывая тенденцию к повсеместному использованию Big Data, без аналитики текста нельзя будет обойтись.

Список литературы

[1] Subramanian D. Text Mining in Python: Steps and Examples, 2019 // Towards AI Co. [Электронный ресурс]. URL: https://towardsai.net/p/data-mining/text-mining-in-python-steps-and-examples/ (дата обращения: 10.06.2021).

[2] Кондратюк А.И. Разработка системы поддержки принятия решений для информационной системы управления обращениями в ООО «ГАЗПРОМ ПЕРЕРАБОТКА»: магистерская диссертация. СПб. СПбГЭТУ «ЛЭТИ», СПб, 2021.

[3] Филиппова Е. Основные технологии text mining, 2014 // datareview.info [Электронный ресурс]. URL: http://datareview.info/article/osnovnyie-tehnologii-text-mining// (дата обращения: 10.06.2021).

[4] Janakiev N. Practical Text Classification With Python and Keras // Real Python [Электронный ресурс]. URL: https://realpython.com/python-keras-text-classification/ (дата обращения: 10.06.2021).

[5] Loseva Daria M., Kondratiuk Anton I.. Neural Network Development for IT-users Requests Processing // 2020 IEEE Conference of Russian Young Researches in Electrical and Electronic Engineering (ElConRus), January 27 - January 30, 2020, St.Petersburg, 2020.