УДК 004.912

ОСНОВНЫЕ ПРИНЦИПЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Бабаев А.М.,
студент
Институт сферы обслуживания и предпринимательства (филиал) ДГТУ
Шахты, Россия

Аннотация. Статья посвящена рассмотрению основных этапов анализа текстовых данных. Перечислены задачи, решаемые в сфере обработки естественного языка, а также главные проблемы, возникающие при анализе текста. Рассмотрены методы, применяемые для очистки текста и его векторизации. Приведен алгоритм, позволяющий представить текст в виде структурированных данных, пригодных для использования в моделях машинного обучения.

Ключевые слова: обработка естественного языка, векторизация текста, лемматизация, стемминг, машинное обучение.

BASIC PRINCIPLES OF NATURAL LANGUAGE PROCESSING

Babaev A.M.,
student
Institute of Service and Entrepreneurship (branch) of DSTU
Shakhty, Russia

Abstract. The article is devoted to the consideration of the main stages of the analysis of text data. The tasks that are solved in the field of natural language processing, as well as the main problems that arise in the analysis of the text are listed. The methods used for cleaning text and its vectorization are considered. An

algorithm is presented that allows you to present text in the form of structured data suitable for use in a machine learning model.
Keywords: natural language processing, text vectorization, lemmatization, stemming, machine learning.

Все, что человек выражает в письменном виде, несет в себе огромное количество информации. Тема, порядок слов, длина предложений – всё это представляет данные, которые можно интерпретировать, и из которых можно извлечь некоторую ценность. С развитием Интернета и средств сканирования документов люди стали генерировать огромное количество текстовой информации, доступной для анализа с помощью вычислительных устройств.

Текст отзыва на сайте интернет-магазина, новостная статья, запись пользователя социальной сети и все прочие разновидности доступного для исследования текста являются неструктурированными данными. Они не вписываются в традиционную для баз данных структуру строк и столбцов и не поддаются анализу с помощью традиционных методов изучения структурированных данных. Однако достижения в области машинного обучения позволяют структурировать текст и затем интерпретировать его. Этим занимается область обработки естественного языка (Natural Language Processing, далее NLP).

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 [1]. Конкретными примерами задач, которые решаются NLP, являются:

- машинный перевод;

– классификация текстов по категориям (сортировка новостей по рубрикам);

– сентиментный анализ (классификация отзывов на товар на отрицательные, положительные и нейтральные);

– извлечение именованных сущностей (поиск словосочетаний, несущих определенный смысл, например, названия организаций или имена людей);

– вопросно-ответные и диалоговые системы (Siri, Алиса), при создании которых решают все приведенные выше задачи [2].

При решении приведенных выше задач помимо уже указанной неструктурированности данных возникает также ряд других проблем:

– полисемия – наличие у слова нескольких значений, взаимосвязанных по смыслу (остановка – процесс или здание);

– омонимия – наличие у слова не связанных по смыслу значений (ключ, замок);

– анафора – разные варианты интерпретации одного выражения в зависимости от другого выражения, ранее встречавшегося в тексте;

– синонимия;

– порядок слов в предложении.

Модели машинного обучения работают со структурированными данными, поэтому набор текстовых данных сначала необходимо привести к некоторой структуре. В настоящее время исследователи в области NLP сформировали алгоритм анализа текста, который является актуальным для большинства задач.

1. На первом этапе необходимо разделить текст на более мелкие составляющие. В зависимости от решаемой задачи он может быть разбит на отдельные предложения – это сегментация (например, машинный перевод), или на отдельные слова – это токенизация (например, сентиментный анализ).

Необходимо отметить, что в список токенов помимо слов также включаются знаки препинания.

2. После разбиения текста на составные части необходимо выполнить очистку списка токенов от шума. Здесь могут быть применены следующие операции:

– удаление нерелевантных символов (например, знаков препинания);

– удаление нерелевантных сочетаний символов (адреса сайтов, номера телефонов и т.д.);

– перевод всех слов к нижнему регистру, иначе в дальнейшем модель машинного обучения будет рассматривать слова в разных регистрах как совершенно разные;

– лемматизация текста – процесс приведения слова к его словарной форме путем изменения суффиксов и окончаний слова;

– стемминг текста – более грубый способ приведения слов к одной форме, который заключается в отсечении всех частей слова от его корня.

Лемматизация позволяет сохранить контекст предложения и, следовательно, несет больше информации. Однако он более вычислительно сложен, чем стемминг. Кроме того, иногда применение лемматизации не повышает качество решения задачи в сравнении со стеммингом. В таких случаях исследователи выбирают более грубый, но быстрый способ преобразования формы слова [3].

При чистке списка токенов широко применяются регулярные выражения – это последовательности символов, которые облегчают и ускоряют поиск необходимых шаблонов текста.

3. Далее необходимо выполнить удаление стоп-слов – слов, которые в условиях задачи не несут смысловой нагрузки, например, предлоги, частицы или союзы. Также в него могут входить другие слова, которые встречаются в большинстве текстов выборки, так как, например, при решении задачи

классификации они не смогут стать отличительным признаком класса. Таким образом, итоговый вид списка стоп-слов определяется поставленной задачей.

4. Теперь следует выбрать подходящее представление данных. Так как модели машинного обучения принимают на вход матрицы, состоящие из чисел, то исследователю необходимо выполнить векторизацию – кодирование каждого текста с помощью числовых значений.

Метод Bag of Words («Мешок слов») рассматривает каждое слово как отдельный категориальный признак. Сначала выполняется построение списка всех уникальных слов для всех текстов набора данных. Далее каждый отдельный текст кодируется следующим образом: если слово встречается в данном тексте, то значение элемента вектора признаков, соответствующее данному слову, будет увеличено на единицу. Таким образом, будет получен одномерная матрица с числом вхождений каждого слова в данный текст (рисунок 1).

	любить	писать	код	не	запомнил
Я люблю писать код. ►	1	1	1	0	0
Я не запомнил код. ►	0	0	1	1	1

Рис. 1 – Векторизация текста методом Bag of Words

Проблема описанного выше метода векторизации текста в том, что слова с наибольшей частотностью будут иметь самую высокую оценку. В итоге при обучении модели более важные для классификатора слова могут быть пропущены. Этого недостатка лишен метод term frequency — inverse document frequency (далее TF-IDF) [4]. В нём оценка слова увеличивается с ростом частоты появления слова в документе, но при этом также учитывается

количество документов, содержащих это слово. Например, формула вычисления оценки w слова x в документе y приведена ниже.

$$w_{x,y} = tf_{x,y} * \log\frac{N}{df_x}$$

где $tf_{x,y}$ – частота x в y (отношение числа вхождений x к общему числу слов документа);

df_x – число документов, содержащих x;

N – число документов.

5. Описанные выше шаги позволяют представить текст в структурированном виде. Выполненные ранее этапы универсальны для большинства задач NLP, но могут дополняться в зависимости от специфики проблемы. Далее начинается работа с моделями машинного обучения, и теперь исследователю необходимо руководствоваться рекомендациями по решению анализа текста в определенной сфере. Хотя в настоящее время существуют архитектуры искусственных нейронных сетей, которые частично можно назвать универсальным решением [5].

В данной статье были рассмотрены основные принципы решения задач в области обработки естественного языка. Были описаны универсальные операции, которые применяется для структурирования текстовой информации практически в любой задаче в области NLP. Однако приведенный выше алгоритм не гарантирует самое высокое качество работы модели машинного обучения, так как он не может охватить особенности всех решаемых в NLP проблем.

Библиографический список

1. Jain K. Deep Learning for Natural Language Processing: Creating Neural Networks with Python / K. Jain, P. Goyal, S. Pandey. – Apress, 2018. – 296 pp.

2. NLP. Основы. Техники. Саморазвитие. Часть 1. Коллективный блог «Хабр», Блог компании ABBYY. URL: https://habr.com/ru/company/abbyy/blog/437008/ (дата обращения 15.12.2019).

3. Бонцанини М. Анализ социальных медиа на Python / пер. с анг. А.В. Логунова. – М: ДМК Пресс, 2018. – 288 с.

4. Feature extraction. Scikit-learn. URL: https://scikit-learn.org/stable/modules/feature_extraction.html (дата обращения: 16.12.2019).

5. Rao D. Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning / D. Rao, McMahan B. – O'Reilly Media, 2019. – 256 pp.

Оригинальность 93%