Сафаров Лазиз Сайимович

Старший реподаватель кафедры «Алгоритмы и технологии программирования»

Каршинский государственный университет

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИИ TEXT MINING ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТА

Аннотация: в статье рассматривается использование технологии text mining при автоматической обработке текста.

Ключевые слова: анализ текста, классификация документов, модель векторного пространства, терм-документная матрица, TF-IDF, нейронная сеть.

Safarov Laziz Sayimovich

Senior Lecturer at the Department of Algorithms and Programming Technologies

Karshi State University

USING TEXT MINING TECHNOLOGY IN AUTOMATIC TEXT PROCESSING

Abstract: the article discusses the use of text mining technology in automatic text processing.

Keywords: text analysis, document classification, vector space model, term-document matrix, TF-IDF, neural network.

В настоящее время во многих сферах человеческой деятельности применяются интеллектуальные информационные технологии, призванные повысить эффективность решения различного рода задач. Одной из таких технологий является Text Mining – интеллектуальный анализ текстовой информации – алгоритмический процесс обнаружения не известных ранее знаний из текста, а также выявления основных понятий и взаимосвязей между ними. Как правило, процесс извлечения новых знаний из текстовой информации является нетривиальным и достаточно трудоемким.

Одной из задач интеллектуального анализа данных является задача классификации. В частности, классификация текстовых документов – задача информационного поиска, которая заключается в определении одной из нескольких категорий для документа на основании его содержания. Процесс классификации текстов может быть осуществлен тремя разными способами: вручную, автоматически на основе заданного экспертом набора правил и автоматически на основе методов машинного обучения. В последнем случае набор правил принятия решений по классификации текстовых документов формируется автоматически на основе обучения классификатора на данных из обучающей выборки.

Данные для обучения представляют собой набор эталонных образов из всех классов текстовых документов. При анализе текстов на основе методов машинного обучения, таких как нейронная сеть, в обучающей выборке необходимо вручную назначить класс для каждого обучающего примера (документа). Назначение класса является более простой задачей по сравнению с экспертным формированием набора правил классификации. При этом метка класса может формироваться во время функционирования системы. Так, например, в электронных почтовых клиентах имеется возможность указывать тип письма («спам» / «не спам»). Это позволяет сформировать обучающую выборку для дальнейшего построения нейросетевого классификатора (спам-фильтра) [???]. Следовательно, классификация документов на основе методов машинного обучения представляет собой пример обучения с учителем (человеком), определяющим набор классов и формирующим обучающую выборку.

Большинство методов автоматической классификации текстов основаны на концепции «похожести» различных документов одного класса. Такие документы содержат в себе похожие слова и их сочетания.

Анализ текстовых документов методами Text Mining выполняется в 5 шагов:

1) Поиск информации. На этом шаге происходит определение документов, подготавливаемых для дальнейшей обработки и анализа. При небольшом количестве исходных документов пользователи информационной системы могут сами выбрать нужные документы для анализа. Если документов достаточно много, то желательно использовать алгоритмы автоматического выбора документов.

2) Предобработка текстов. Происходит преобразование текста документа в форму, удобную для применения алгоритмов Text Mining. На выходе данного этапа формируется текст без лишних слов, не влияющих на результат анализа.

3) Извлечение требуемой информации. Данный этап предназначен для формирования набора основных понятий (терминов) обрабатываемого текста для их дальнейшего анализа.

4) Применение методов Text Mining. Это главный шаг анализа, на котором формируются новые знания и скрытые в тексте закономерности.

5) Анализ и интерпретация полученных результатов. Представление результатов анализа в форме, удобной для пользователя, например, на естественном языке или в графическом виде.

Рис. 1 - Этапы анализа текстовых документов методами Text Mining

Рассмотрим приемы, используемые на этапе предварительной обработки. На данном этапе одним из основных приемов является токенизация текста, т.е. разбиение текстового документа на отдельные

абзацы (токенизация по абзацам), отдельные предложения (токенизация по предложениям), отдельные слова (токенизация по словам). Результаты данного разбиения называются токенами.

После токенизации, как правило, следует фильтрация stop-слов, которые не содержат в себе никакого смысла, например, союзы, предлоги, артикли, междометия, частицы и т.п. Список stop-слов составляется заранее в зависимости от языка обрабатываемого текста. В данном приеме предварительной обработки stop-слова удаляются из текста.

Следующим шагом является стэмминг или лемматизация, где происходит нормализация слов. Все слова в текстовом документе приводятся к нормальной форме, в частности, в единственном числе, именительном падеже, без особенностей устной речи. Недостатком в данном приеме может являться нарушение семантики предложений, словосочетаний, поэтому необходимо так же, как и в токенизации, учитывать язык текста. Наиболее известным алгоритмом нормализации слов русского языка является Snowball, основная идея которого заключается в нахождении однокоренных слов и отсечения у них окончаний, суффиксов и т.п.

Таким образом, нейронная сеть распознала все внесенные ошибки в тестовых выборках. Полученные в настоящей работе результаты будут использованы для построения интеллектуальной системы анализа и классификации судебных документов, что позволит достичь следующих требований:

- исключить процесс ручного распределения дел по категориям судебных споров;

- снизить количество ошибок определения категорий споров;

- улучшить контроль рассмотрения дел применительно к конкретным категориям споров, исходя из установленных для них сроков;

- осуществить поддержку принятия решений путем информирования о существенных обстоятельствах, подлежащих установлению для конкретной категории спора.

Использованные источники:

1. Якубов С. Х., Бозорова И. Ж. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОПТИМИЗАЦИИ ФОРМЫ ТРЕХШАРНИРНЫХ АРОК ПРИ СЛОЖНЫХ УСЛОВИЯХ ЗАГРУЖЕНИИ //The Scientific Heritage. – 2022. – №. 82-1. – С. 71-73.

2. Jumanazarovna B. I. Electronic Educational Resources as a Component and Conditions of Interaction of the Subjects of the Educational Process //International Journal of Innovative Analyses and Emerging Technology. – 2022. – Т. 2. – №. 3. – С. 39-43.

3. Jumanazarovna B. I. Electronic Educational Resources as a Component and Conditions of Interaction of the Subjects of the Educational Process //International Journal of Innovative Analyses and Emerging Technology. – 2022. – Т. 2. – №. 3. – С. 39-43.

4. Raximov N., Primqulov O., Daminova B. Basic concepts and stages of research development on artificial intelligence //2021 International Conference on Information Science and Communications Technologies (ICISCT). – IEEE, 2021. – С. 1-4.

5. Даминова Б. Э., Якубов М. С. Проблемы защиты от внешних и внутренних информационных угроз //Труды Северо-Кавказского филиала Московского технического университета связи и информатики. – 2013. – №. 1. – С. 306-308

6. Якубов М. С., Даминова Б. Э. СОВЕРШЕНСТВОВАНИЕ СИСТЕМЫ ОБРАЗОВАНИЙ НА ОСНОВЕ ПРИМЕНЕНИЕ ЦИФРОВЫХ ТЕХНОЛОГИЙ //Eurasian Journal of Mathematical Theory and Computer Sciences. – 2022. – Т. 2. – №. 4. – С. 31-44.