УДК 004.62

РАЗРАБОТКА АЛГОРИТМА ДЛЯ РАСШИФРОВКИ И ОПРЕДЕЛЕНИЯ АББРЕВИАТУР В ТЕКСТЕ

Тирский Б.Д. (Университет ИТМО), **Рунушкина О.Р.** (Университет ИТМО), **Верко Р.А.**
Научный руководитель – к. п. н., доцент Авксентьева Е. Ю.
(Университет ИТМО)

Введение. На сегодняшний день почти каждая информационная система подразумевает работу с данными в виде текстов. Для более точной постановки задачи или анализа прочитанного требуется корректно расшифровывать данные, которые могут быть записаны в совсем разных видах. Примером различного написания являются аббревиатуры устоявшихся словосочетаний и сокращения слов. Расшифровка и извлечение аббревиатур позволяет улучшить производительность процесса работы с текстами, также более конкретно понимать содержимое предложений [1]. Но так как русский язык является одним из самых сложных языков перед нами возникает ряд нетривиальных задач и вопросов. Например, как извлечь аббревиатуры из текста, отличить аббревиатуры от обычных слов, которые начинается с большой буквы в начале каждого предложения, с помощью каких инструментов можно расшифровать прочитанное.

Основная часть. Для написания алгоритма был выбран язык разработки Python и среда программирования Jupyter Notebook. Были сформулированы и описаны конкретные задачи для написания алгоритма:

1) Подобрать и преобразовать к нужному виду словарь сокращений [2].
2) Сформировать корпус текстов с имеющимися аббревиатурами.
3) Определиться с критериями идентификации аббревиатур.
4) Написать алгоритм по извлечению аббревиатур из корпуса текстов.
5) Провести оценку по TF-IDF контекстной части текста.
6) Обучить модель по расшифровке аббревиатур.
7) Получить результат и провести оценку точности определения.

Для идентификации аббревиатур, было решено проводить анализ по регистру написанных слов, если не меньше половины букв в слове написаны в верхнем регистре, то алгоритм определял это слово как аббревиатуру. После идентификации предложения рассматривались как три части, контекст до аббревиатуры, аббревиатура и контекст после аббревиатуры. Контекст необходимо сохранить для оценки по TF-IDF, с помощью данной меры определяется важность используемого слова в конкретном предложении [1]. После оценки TF-IDF модель обучается и расшифровывает аббревиатуры.

Выводы. В ходе работы были исследованы ряд словарей с сокращениями русских слов, был написан алгоритм по парсингу подходящего нам словаря, был написан сам алгоритм по извлечению аббревиатур и обучению расшифровке аббревиатур.

Список использованных источников:

1. Шилов И. М. . Автоматическое выявление и расшифровка аббревиатур и сокращений в тексте [диссертация из интернета]. Санкт-Петербург: Санкт-Петербургский государственный университет (СПбГУ); 2016. Ссылка: https://nauchkor.ru/pubs/avtomaticheskoe-vyyavlenie-i-rasshifrovkaabbreviatur-i-sokrascheniy-v-tekste-587d36365f1be77c40d58984
2. Словарь сокращений. Ссылка: https://sokr.ru/

Тирский Б.Д. (автор)

Авксентьева Е.Ю. (научный руководитель)