УДК 004.912

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ И РАСШИФРОВКА АББРЕВИАТУР ИЗ ПРОФЕССИОНАЛЬНЫХ ТЕКСТОВ

Егоров М. П. (Университет ИТМО)
Научный руководитель – Функнер А.А.
(Университет ИТМО)

Аннотация: Данная работа посвящена извлечению и расшифровке аббревиатур из профессиональных текстов. Для тестирования используется несколько корпусов естественных медицинских текстов, в том числе отчеты организации БУ «Медицинский информационный-аналитический центр» и анамнезы болезни пациентов с ОКС (острым коронарным синдромом).

Введение. Проблема заключается в том, что аббревиатуры неоднозначно определены, существуют аббревиатуры с несколькими возможными расшифровками в зависимости от контекста предложения. Результат данной работы помог бы с определением контекста предоставленных текстов, тем самым улучшив точность предсказаний моделей машинного обучения. На данный момент нет готовых модулей для решения поставленной задачи.

Основная часть. Для решения данной задачи предлагается обучить модель XGBoost – библиотека с открытым кодом для построения моделей машинного обучения, использующие градиетный бустинг решающих деревьев. Предварительно данные нужно подготовить: очистить данные от пунктуаций, нормализовать слова, а также дополнительно извлечь и расшифровать представленные аббревиатуры.
Для обучения модели будут подаваться предложения с нерасшифрованными аббревиатурами с их расшифровками. После обучения модель будет выводить вектор аббревиатур с их вероятностями для конкретных предложений.
Для проверки гипотезы мы сравнили модели, предсказывающие возраст пациентов с ОКС, в зависимости от анамнеза, обученные с расшифрованными и нерасшифрованными анамнезами текстах.

Выводы. В результате данной работы была получена пред-обученная модель, которая будет предоставлять возможные расшифровки аббревиатур из профессиональных текстов. Она поможет улучшить качества предсказаний других моделей, работающих с корпусами естественного языка.

Егоров М. П. (автор) Подпись

Функнер А. А. (научный руководитель) Подпись