

Е. В. Буйницкая
(ГрГУ им. Я. Купалы, Гродно)

МЕТОДЫ И СРЕДСТВА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ПОИСКА ОНЛАЙН КУРСОВ

Обработка естественного языка (NLP) – это область искусственного интеллекта, которая позволяет машинам читать, понимать и извлекать смысл из человеческих языков [1]. Ценность этой технологии исходит из случаев её использования: она позволяет распознавать и предсказывать заболевания на основе электронных медицинских карт, может использоваться для отсеивания спама в электронных письмах и т.д.

Сейчас всё большую актуальность приобретают образовательные курсы с различных веб-платформ, поскольку они дают возможность дистанционного изучения материалов лучших университетов мира. Они могут быть классифицированы как слабоструктурированные данные. Для того, чтобы сделать выбор курса более точным и подходящим к потребностям студента можно использовать NLP.

В работе исследуются данные об образовательных курсах с платформ Coursera, edX и FutureLearn, собранные с помощью веб-

скрапера. На основе анализа этих данных осуществляется выдача рекомендаций курсов, наиболее релевантных запросам студентов.

Предварительная обработка данных включает в себя удаление стоп-слов, лемматизацию и токенизацию, осуществляется с помощью библиотек gensim и nltk на языке Python 3. Далее для представления текста в векторном виде и обучения используется модель word2vec и, основанная на ней модель doc2vec. Эти модели позволяют представлять курсы схожие по контенту векторами близкими в пространстве.

После проведения обработки данных и обучения модели можно оценить сходство образовательных курсов на основе косинусного сходства их векторных представлений и рекомендовать курсы на основе ключевых слов в запросах.

Литература

1 Gupta, S. Practical Enterprise Data Lake Insights / S. Gupta. – apress, 2018. – 327 p.