УДК 004.852

МОДИФИКАЦИЯ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ С ИСПОЛЬЗОВАНИЕМ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ТЕКСТА ОПИСАНИЙ РЕКОМЕНДУЕМЫХ ОБЪЕКТОВ

Горбунова И.А.
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., доцент Хлопотов М.В.
(Университет ИТМО, г. Санкт-Петербург)

Введение. Любой онлайн бизнес, в котором существует большое разнообразие продаваемых товаров, сталкивается с проблемой отбора товаров для отображения пользователям. Сейчас все больше компаний решают эту задачу с помощью персональных рекомендаций. Эксперименты показывают, что наличие в сервисе персональных рекомендаций на основе алгоритмов машинного обучения повышает средний чек, возвращаемость клиентов и другие измеримые бизнес метрики. Качество предсказаний рекомендательной системы зависит от правильного извлечения и представления признаков о пользователях и товарах. Отдельное внимание в подготовке данных для обучения модели уделяется получению признаков из текста на естественном языке – моделированию языка и обучению представлений. В данной работе рассматривается существующая рекомендательная система фильмов, которая в качестве признаков не учитывает описание контента.

Цель работы. Целью данной работы является исследование современных подходов векторного представления текстов на естественном языке о товарах и модификация существующей модели рекомендаций фильмов с помощью добавления признаков на основе их сюжетных аннотаций, полученных в ходе исследования представлениями.

Общие положения исследования. Гипотеза исследования состоит в предположении о том, что представления, полученные с помощью различных моделей, по-разному влияют на качество предсказания модели, которые эти представления принимает в качестве признака. Предполагается, что вектор, полученный с помощью с помощью нейронной сети, может улучшить качество рекомендаций.

Существующие методики. Самый простой способ численно представить текст на естественном языке это - «мешок слов». Каждый документ представляется вектором размерности |V|, где i-ая компонента вектора является счетчиком встречаемости i-ого токена в рассматриваемом документе, если какой-либо токен не встретился в рассматриваемом документе, то соответствующая компонента будет равна нулю. Два документа, отличающиеся лишь порядком токенов, будут иметь одинаковые векторы, так как порядок токенов игнорируется.

Не все элементы (токены) одинаково значимы: например, союзные слова, очевидно, не несут никакой полезной нагрузки. Поэтому при определении числа совпадающих элементов в двух векторах все измерения нужно предварительно взвешивать по их значимости. Данную задачу решает хорошо известное преобразование TF-IDF, которое назначает больший вес более редким элементам. TF вычисляется, как доля документов, в которых присутствует токен, а IDF как инверсия частоты, с которой некоторое слово встречается в документах коллекции. Вес токена в документе вычисляется как произведение TF*IDF.

В последние годы популярность завоевал подход Transfer Learning - он
позволяет адаптировать заранее обученную модель к конкретной задаче с использованием относительно небольшого объема данных. Ключевыми разработками, использующими этот

метод, можно назвать ULMFiT, ELMO, OpenAl Transformer, и Google BERT – все они появились в 2018 году и показали выдающиеся результаты в различных NLP задачах.

Промежуточный результат. Проведено сравнение получения векторных представлений текста на естественном языке.

Практический результат. Модифицирована рекомендательная система благодаря новым контентным признакам – векторным представлениям, полученным в ходе исследования.

Горбунова И.А. (автор)	Подпись
Хлопотов М.В. (научный руководитель)	Подпись