

АНАЛИЗ ИНСТРУМЕНТОВ ВИЗУАЛЬНОГО ПРОГРАММИРОВАНИЯ НА ОСНОВЕ ОТЗЫВОВ ИСПОЛЬЗУЯ ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТА

Ходненко И. В. Университет ИТМО,
Научный руководитель – доцент ИДУ, Иванов С. В..
Университет ИТМО

В данной работе предлагается анализ инструментов визуального программирования в сфере машинного обучения с помощью интеллектуального анализа отзывов. Отзывы извлекаются с сайта IT Central Station и содержат в себе основную идею, а также плюсы и минусы для каждого инструмента. Преобразованные в векторное пространство слова отзывов из групп с различными оценками кластеризуются и формируют наглядную оценку общих достоинств и недостатков существующих инструментов.

Введение. Визуальное программирование (ВП) в машинном обучении – это анализ данных или создание моделей с помощью манипуляции графа. С 2015 по 2020 года было создано шесть популярных программ, предоставляющих ВП подход. В литературе наиболее часто встречаются частные сравнения созданных решений друг с другом. Но анализ инструментов в целом, как, например сводная таблица, найти так и не удалось. В данной работе предлагается подход, использующий «интеллектуальный анализ текста» на отзывах об инструментах, результатом которого будут выявленные в кластеры как общие, так и индивидуальные плюсы и минусы большинства популярных решений.

Основная часть. IT Central Station – сервис, представляющий обзор инструментов а также содержащий в себе отзывы о них. Отзывы состоят из нескольких частей и наиболее интересными для данной работы является раздел «плюсы и минусы». В нем рецензенты коротко описали достоинства и недостатки каждого решения по пунктам. Остальные части отзывов могут использоваться в дальнейшем, для более детального анализа инструментов.

Идея метода – преобразовать тесты отзывов в векторное пространство для последующего анализа. Для преобразования использовалось два метода. Первый TF-IDF – для анализа важности слов в контексте отзыва. Этот подход достаточно прост и позволяет при оценке важности слов не выходить за рамки исследуемой работы. Вторым методом для преобразования является word2vec. Данный метод интересен тем, что он после преобразования может улавливать некоторые семантические свойства слов, что так же является интересным предметом исследования. Таким образом, после преобразования, на выходе имеются векторы вместо каждого слова, которые уже можно анализировать применяя стандартные подходы.

На этапе кластеризации ищутся две группы кластеров и для достоинств, и для недостатков инструментов. Первая группа – это выбросы. Какие-либо индивидуальные слова или фразы при обзоре, которые характеризуют индивидуальность инструмента. Вторая группа – кластеры с несколькими общими точками для разных отзывов. Здесь задача найти общие плюсы и минусы для инструментов, избегая при этом общие слова, которые и так встречаются в каждом отзыве как, например, местоимения.

Выводы. Таким образом, данный анализ дает и общую оценку инструментов по достоинствам и недостаткам, так и индивидуальную по каждому из них. Полученные оценки в дальнейшем помогут разработчикам при создании новых программ использующих визуальное программирование ориентироваться не только на частные обзоры в литературе между парами

инструментов, но и на общую картину в целом. Еще одним достоинством произведенного анализа является его автоматизированность. Большая часть анализа выполняется автоматически и только последняя часть требует внимание исследователя и интерпретации результатов, что позволит обновлять оценку инструментов при добавлении новых отзывов в систему IT Central Station.

Ходненко И. В. (автор)

Подпись

Иванов С. В. (научный руководитель)

Подпись