УДК 004.622

ОЦЕНКА РЕЛЕВАНТНОСТИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ ДЛЯ LLM И АНАЛИЗА

Мищенко М.Ю. (ИТМО), **Мустафин Д.Э.** (ИТМО), **Унтила А.А.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Федоров Д.А. (ИТМО)

Введение. В современном мире всё большую роль играют данные. Объёмы информации увеличиваются от года к году [1]. С развитием интернета и социальных медиа огромное количество данных становится доступным для анализа и обучения LLM. Однако, с ростом объёма информации, появляется необходимость в предварительной оценке её релевантности. Сложности вносит тот факт, что большее число данных неструктурированно, в них отсутствует схема и упорядоченность [2].

Основная часть. Необходимость в предварительной оценке релевантности данных заключается в нескольких факторах. Наиболее значительными среди них являются качество обучения модели, защита от смещения, а также эффективное использование вычислительных ресурсов. Качество обучающей выборки оказывает наибольшее влияние на способность модели понимать и генерировать текст, а также на её способность к выделению закономерностей в неструктурированных данных. Оценка релевантности данных даёт возможность избежать смещения в данных, что могло бы значительно исказить результаты работы модели. Это особенно важно в сфере LLM, где даже малое смещение может привести к грубейшим ошибкам в обучении модели. Помимо прочего, оценка релевантности и исключение лишних данных оказывает положительное влияние на скорость работы с моделью.

Первым шагом к увеличению релевантности входных данных для модели мы обозначили обработку данных (далее - документов). Мы применили алгоритм стемминга и лемматизацию [3] для приведения словоформ к леммам - их нормальным (словарным) формам.

Затем из документов были удалены слова из стоп-словаря с часто встречающимися, но не несущих смысловой нагрузки слов (артикли, местоимения, предлоги) [4]. Этот метод позволяет повысить точность анализа в общем случае.

Далее была применена TF-IDF [5] метрика, позволяющая оценить важность слова в контексте одного документа, относительно всего набора входных данных. Конкретней, вес некоторого слова прямо пропорционален его частоте в одном документе и обратно пропорционален его частоте во всех документах данных. Важным моментом было определить порог, начиная с которого слова помечались как малозначимые. С помощью полученной метрики была построена матрица признаков, определяющая релевантность каждого слова. Эта матрица была использована в роли входных данных для алгоритма DBSCAN. Он позволил выделить кластеры точек, которые были наиболее плотно сконцентрированы в пространстве признаков. Это позволило выбрать наиболее релевантные слова и отбросить малозначимые слова и аномалии.

В общем случае, каждый из этих методов по отдельности оказывает положительный эффект на качество модели LLM. Но комбинацию методов разработки необходимо подбирать в зависимости от задачи анализа.

Выводы. Проведен анализ возможных методов оценки релевантности неструктурированных данных для обучения LLM. Исследованы комбинации методов обработки данных для улучшения оценки релевантности данных для анализа.

Список использованных источников:

1. Назаренко Ю. Л. Обзор технологии" большие данные"(Big Data) и

программно-аппаратных средств, применяемых для их анализа и обработки //European science. – 2017. – №. 9 (31). – С. 25-30.

2. Захарова А.А., Подвесовский А.Г., Лагерев Д.Г. Визуальная аналитика и когнитивные методы для обработки и анализа гетерогенных данных мультисенсорных систем: проблемы и тенденции //Информационные и математические технологии в науке и управлении. – 2019. – №. 4 (16). – С. 60-74

3. КОРЮКИН А. В. ИССЛЕДОВАНИЕ ВЛИЯНИЯ СТЕММИНГА И ЛЕММАТИЗАЦИИ НА КАЧЕСТВО БИНАРНОЙ КЛАССИФИКАЦИИ ПО ТОНАЛЬНОСТИ КРАТКИХ ТЕКСТОВЫХ КОММЕНТАРИЕВ //Актуальные исследования. – 2021. – С. 10.

4. Гращенко Л. А. О модельном стоп-словаре //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2013. – №. 1. – С. 40-46.

5. Rajaraman A., Ullman J. D. Mining of massive datasets. – Cambridge University Press, 2011.