Дедупликация карточек товаров

Введение

Семантическая дедупликация — это задача, в которой объекты, обладающие схожим смыслом, должны быть объединены в одну категорию, и один из них может быть удалён как дубликат. Обычно эта задача решается путём представления объектов во векторном пространстве (эмбеддингах), где семантическая близость между объектами измеряется с помощью различных метрик. При превышении заданного порога близости объекты считаются дубликатами.

Задача дедупликации остаётся актуальной в различных областях, особенно при построении больших датасетов. Она позволяет сократить объём данных, повысить качество обучения моделей и избежать переобучения на повторяющихся примерах. В связи с этим разрабатываются всё более эффективные методы, которые отличаются подходами к построению семантических представлений и метриками сравнения таких представлений.

В задаче дедупликации товарных карточек на e-commerce платформах необходимо учитывать мультимодальные данные: текст (названия, описания, характеристики), изображения и, в некоторых случаях, видео. Объединение этих различных типов данных требует использования мультимодальных моделей или согласования векторов, полученных из разных источников. Поскольку на текущий момент видео встречаются не так часто, и являются более сложной задачей, сосредоточим наше внимание на тексте и изображениях.

Ключевая сложность задачи заключается в выборе подходящего пространства эмбеддингов, которое способно отразить смысл объектов. Для текстовых данных эмбеддинги могут являться, выходами из моделей-трансформеров: BERT [1] или Sentence-BERT [2]. Или же можно использовать более современные решения, такие как GTR [3, 4], и модели от OpenAI [5]. Также, существуют и простые практические решения задачи семантической дедупликации текста — SemanticDeduplicator [6]. Для изображений используют эмбеддинги, полученные с помощью моделей вроде ResNet [7], CLIP [8] или более современных решения.

Но получение векторов только этап, для решения задачи дедупликации используют такие методы, как SemDeDup [9] или FairDeDup [10] которые позволяют эффективно удалять семантические дубликаты в больших коллекциях изображений или текстов.

Обзор методов

Рассмотрим пару методов, которые можно применить для решения задачи дедупликации карточек товаров.

Sentence-BERT

Sentence-BERT (SBERT) — это модификация оригинального BERT, разработанная для получения плотных векторных представлений предложений, пригодных для семантического сравнения. Вместо стандартной архитектуры BERT, которая неэффективна

для сравнения пары предложений, SBERT использует сиамскую или триплетную архитектуру, что позволяет выполнять семантическое сравнение пар предложений с помощью простых метрик, таких как косинусное расстояние. Это особенно важно при массовом сравнении объектов, как в задаче дедупликации товарных карточек.

SBERT позволяет предварительно вычислить эмбеддинги всех объектов в коллекции и затем выполнять быстрый поиск схожих объектов по заранее построенному индексу. Это существенно повышает производительность и масштабируемость.

SemDeDup

SemDeDup — это подход к удалению дубликатов, основанный на семантическом сходстве, с упором на масштабируемость и применимость к большим наборам данных. Он был предложен для очистки веб-данных, таких как LAION и C4, перед использованием их в обучении моделей машинного обучения. SemDeDup использует мощные эмбеддинговые модели (например, CLIP для изображений, BERT для текста), применяет приближённый поиск ближайших соседей и адаптивные стратегии отбора дубликатов.

В отличие от традиционных методов, SemDeDup не ограничивается только точным совпадением или синтаксическим сходством, а удаляет семантически избыточные данные. Это особенно ценно при работе с шумными данными в e-commerce. SemDeDup также предоставляет возможность адаптивной настройки порогов и гибкой фильтрации, что делает его применимым в различных сценариях очистки данных.

FairDeDup

FairDeDup — это расширение идеи SemDeDup, разработанное для борьбы с проблемами справедливости (fairness) в задачах семантической дедупликации мультимодальных данных. Авторы поднимают вопрос, что традиционные методы дедупликации могут по-разному влиять на различные демографические или визуальные группы, создавая перекосы в итоговом датасете.

FairDeDup предлагает методику обнаружения и смягчения этих перекосов при помощи адаптивного отбора кандидатов на дедупликацию с учётом семантической информации и анализа справедливости.

Метод оценивает не только степень сходства объектов, но и потенциальную утрату разнообразия при удалении данных. Это делает его особенно полезным при работе с чувствительными данными в области e-commerce с большим числом пользовательских категорий.

Заключение

Задача семантической дедупликации карточек товаров в e-commerce требует комплексного подхода, учитывающего мультимодальность данных и эффективные методы сравнения. Ключевая сложность заключается в выборе оптимального эмбеддингового пространства, обеспечивающего точное семантическое сравнение текстовых и визуальных данных. Для решения этой задачи успешно применяются такие методы, как BERT, CLIP и GTR

для векторизации, а также метода по типу SemDeDup или FairDeDup для поиска дубликатов. Дальнейшее развитие методов дедупликации связано с совершенствованием мультимодальных моделей, автоматизацией подбора порогов сходства и возможностью включения в данные видео. Грамотное применение этих подходов позволяет не только повысить качество данных, но и оптимизировать работу поисковых и рекомендательных систем в e-commerce.

Литература

1. Devlin J., Chang M.—W., Lee K., Toutanova K. BERT: Pre—training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. — 2018. — URL: https://arxiv.org/abs/1810.04805

2. Reimers N., Gurevych I. Sentence—BERT: Sentence Embeddings using Siamese BERT—Networks // arXiv preprint arXiv:1908.10084. — 2019. — URL: https://arxiv.org/abs/1908.10084

3. Ni J., Qu C., Lu J., Dai Z., Hernandez Abrego G., Ma J., Zhao V., Luan Y., Hall K.B., Chang M.—W., Yang Y. Large Dual Encoders Are Generalizable Retrievers // arXiv preprint arXiv:2112.07899. — 2021. — URL: https://arxiv.org/abs/2112.07899

4. Google Research. T5X Retrieval [Электронный ресурс] // GitHub. — URL: https://github.com/google—research/t5x_retrieval

5. OpenAI. Embeddings [Электронный ресурс] // OpenAI Platform. — URL: https://platform.openai.com/docs/guides/embeddings

6. gkamradt. SemanticDeduplicator [Электронный ресурс] // GitHub. — URL: https://github.com/gkamradt/SemanticDeduplicator

7. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // arXiv preprint arXiv:1512.03385. — 2015. — URL: https://arxiv.org/abs/1512.03385

8. OpenAI. CLIP: Connecting text and images [Электронный ресурс] // OpenAI. — URL: https://openai.com/index/clip/

9. Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, Ari S. Morcos. SemDeDup: data-efficient learning at web-scale through semantic deduplication // arXiv preprint arXiv:2303.09540. — 2023. — URL: https://arxiv.org/abs/2303.09540

10. Slyman E., Lee S., Cohen S., Kafle K. FairDeDup: Detecting and Mitigating Vision-Language Fairness Disparities in Semantic Dataset Deduplication // arXiv preprint arXiv:2404.16123. — 2024. — URL: https://arxiv.org/abs/2404.16123