Описание метода построения пар для задачи сопоставления товаров.

Для предсказания сопоставления товаров необходимы размеченные данные вида product1, product2, match_type, где product1 - информация о первом товаре, product2 - информация о втором товаре, match_type - тип сопоставления (0 - сопоставления нет, 1 - сопоставление есть).

После этапа сбора данных с маркетплейсов данные хранятся как список, состоящий из информации о продуктов. Данный список необходимо привести в вид размеченных данных, который был указан выше.

Для реализации данной трансформации необходимо проделать 2 шага:

1) сгенерировать предположительные положительные и отрицательные пары

2) разметить полученные пары с помощью веб-интерфейса

Генерация положительных пар.

Для генерации положительных пар были проделаны следующие шаги:

1) перевод названия в векторное представление

2) уменьшение размерности с помощью алгоритма umap

3) кластеризация товаров с помощью алгоритма hdbscan (количество кластеров определялось автоматически)

4) выбор кластеров, содержащие от 6 до 150 элементов (в ином случае либо невозможно построить пары, либо в них будет много шума)

5) выбор n товаров для каждого кластера (n зависит от количества товаров в кластере: n = 2 при 30 товарах и меньше, n = 3 при 40 товарах и меньше, n = 4 при 50 товарах и меньше и т.д до 100)

6) генерация n товаров в кластере случайно так, чтобы элементы были не похожи друг на друга, чтобы составить как можно больше разнообразных пар

7) построение попарной близости по названию, цене, а также по 5 первым словам в описании для каждого товара со всеми товарами, кроме его самого, в том же кластере. Для измерения близости текстовых данных использовалась косинусная близость, для измерения близости цен использовалась формула: 1-(|pricaA - priceB|/max(priceA, priceB))

8) выбор пар:

- 2 пары с максимальной близостью по названию и цене
- 1 случайная пара из середины (средняя близость по названию и цене)
- 2 пары с самой низкой близостью по названию и цене
- 2 случайные пары из всего кластера

Также для разметки данных о смартфонах была использована специфика характеристик данной категории.

Для генерации дополнительных положительных пар использовался код производителя, по которому производилась группировка товаров. В случае, если в группе оказалось больше одного товара, то для каждого товара внутри группы строились пары со всеми товарам из этой же группы.

Также для генерации дополнительных положительных пар использовался такой набор признаков как: модель, емкость аккумулятора, диагональ экрана, объем встроенной памяти, объем оперативной памяти. Так как не у всех товаров указаны данные характеристики, то выбирались только те товары, у которых они имеются. Далее проводилась группировка по этим признакам. После чего для каждого товара внутри группы строились пары со всеми товарам из этой же группы.

Все эти три подхода были объеденины, в результате чего удалось получить ~3.3k пар из 7.6k товаров.

Генерация отрицательных пар

Генерация отрицательных пар производилась похожим образом с положительным. Первые 5 шагов повторяются, далее на 6ом шаге высчитывается близость со всеми товарам из предыдущего, а также из следующего кластера. Выбираются следующие пары:

- 1 пара с максимальной близостью по названию и цене
- 1 случайная пара с товаром из кластера

При данном подходе удалось получить 1.8k негативных пар.

Разметка полученных пары с помощью веб-интерфейса

Для отображения, а также разметки полученных пар использовался фреймворк flask и база данных postgres.

В базе данных имеется две главные таблицы: products и pairs.

Таблица products хранит всю информацию о товарах, полученную после парсинга маркетплейсов: id, brand, title, price, description, specifications, category, marketplace, url, stars, sku.

Таблица pairs хранит всю информацию о парах: pair_id, id_1, title_1, id_2, title_2, match_type.

На странице, для удобного пользования, отображаются 4 пары, для этого составляется запрос в таблицу pairs и берутся первые 4 товара с match_type = -1 (сигнал о том, что пара еще не была размечена). Далее по id товара отправляется запрос в таблицу products для получения всей информации о товарах. После чего товары, их характеристики, а также поле для указания типа сопоставления товаров отображается на странице.

После того как пользователь разметил все товары на странице нажимается кнопка “отправить результаты”, данные о типе сопоставления пар в таблице pairs обновляются, а в интерфейсе отображаются новые неразмеченные данные.