Разработка системы с web-интерфейсом для сопоставления характеристик товаров маркетплейса с их эталонными значениями

Московские зайцы

Данные

- Очистка от спецсимволов
- Эталон эталона - он сам
- Честное разделение на данные для обучения и отложенную тестовую выборку
- Название + характеристики для создания эмбеддинго


```
{'product_id': '0007302f2fe1d54d',
 'name': 'Классическая сплит-система ROYAL CLIMA PANDORA RC-PD28HN,
иса, комплект',
 'props': ['Класс  энергоэффективности\tA',
  'Мощность  кондиционера\t9 BTU',
  'Уровень  шума внутреннего блока\t21.5 дБ – 38 дБ',
  'Режим   работы\тохлаждение / обогрев',
  'Фильтр тонкой очистки\тесть',
  'Доп.   режимы турбо, экорежим, осушение, ночной, вентиляция'],
 'is_reference': False,
 'reference_id': 'f497219eb0077f84'}
```


Пример данных

Структура алгоритма

1

Получение эмбеддингов при помощи предобученных моделей

2

Обучение KNN на базе данных

3

Получение результатов для тестовых данных

KNN –
поиск ближайшего соседа

- Мера близости - косинусная
- Поиск и среди продуктов, и среди эталонов. Экспериментально доказано, что так получается лучшая метрика.
- Общий алгоритм для всех моделей эмбеддингов.
- После обучения не требует эмбеддингов базы данных, нужны только правильно проиндексированные ID эталонов.
- Скорость работы моментальная. Крайне простая имплементация.

Модели эмбеддингов

Название модели	Accuracy
fasttext (cc.ru.300)	0.471
fasttext (unsupervised)	0.389
DeepPavlov/rubert-base-cased-conversational	0.361
sberbank-ai/ruBert-base	0.426
bert-base-multilingual-uncased	0.428
all-MiniLM-L6-v2	0.602
labse (sentece-transformers)	**0.842**
tfidf	0.883

Неподходящий словарь

Лучший вариант

Сломается от любой буквы

Как улучшить эмбеддинги? AAE или ArcFace.

В схеме ArcFace убирается этап retrieve field vector и заменяется на модуль классификации и ArcFaceLoss.

ArcFace - один из наиболее эффективных подходов для задач поиска и идентификации.

Вместо обычного ArcFace использовалась модификация ElasticFace.

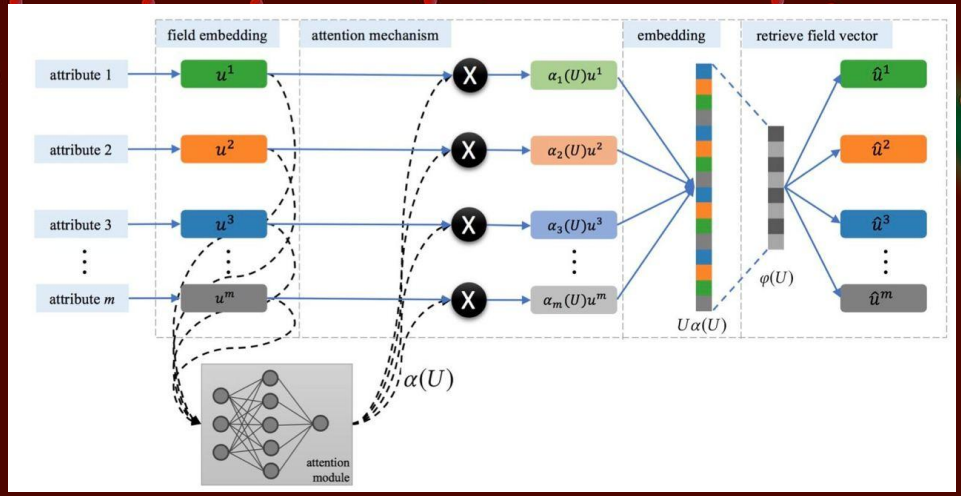


Схема Attention Auto-Encoder

Название модели	Accuracy
arcface (elastic) with labse	**0.963**

Прирост > 14%

Самое важное для данных - аугментации

name_embedding
prop_embedding_1
prop_embedding_2
...
prop_embedding_3
prop_embedding_4
...

name_embedding
prop_embedding_8
prop_embedding_3
...
prop_embedding_16
prop_embedding_1
...

Без аугментаций модель обучалась до Accuracy ~ **0.35**

С аугментациями Accuracy ~ **0.95**

Про API

API написано на Flask и полностью готово к использованию.
Запускается как с GPU, так и без,
но с потерей скорости.

Сервер	Скорость
CPU + 8GB RAM	14 секунд / 100 товаров
CPU + GPU + 16GB RAM (Colab)	2 секунды / 100 товаров

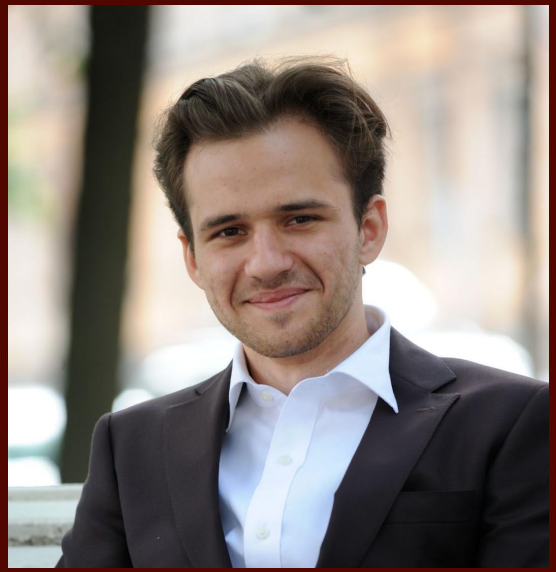
Возможности для дальнейшего улучшения алгоритма

- Протестировать большее количество моделей для получения эмбеддингов.
- Обучить AAE для чистоты эксперимента.
- Обогатить набор данных для обучения модели.
- Оптимизировать API.

Материалы:

- https://arxiv.org/pdf/1904.05985.pdf
- https://www.kaggle.com/code/lextoumbourou/happywhale-tpu-baseline-to-0-804-elasticface/notebook
- https://github.com/UKPLab/sentence-transformers

Команда



Степанов Даниил

ML Engineer & Data Scientist
РСХБ-Интех & НИТУ “МИСиС”

Беляева Анна

Data Scientist, ui/ux-дизайнер
РЭУ им. Г.В. Плеханова