Разработка системы рекомендаций по улучшению карточек товара на маркетплейсе

Подготовил: Скрыпник Дмитрий Юрьевич
Руководитель: к.т.н., Семендяев Родион Юрьевич

Санкт-Петербург, 2025 г.

Актуальность проблемы

- Маркетплейсы, например, **Wildberries используют систему оценки карточек товара**, которая влияет на позицию товара в выдаче.
- Продавцу демонстрируется лишь общий рейтинг карточки и не предоставляются рекомендации по ее улучшению.
- Позиция товара в выдаче напрямую влияет на количество просмотров и покупок, товары которые не видят - не покупают.

Цель и задачи

Целью работы является: разработать систему рекомендаций, которая на основании анализа товаров в выдаче по ключу, будет определять что конкретно можно улучшить в карточке товара для продвижения ее в поисковой выдаче по этому ключу

Задачи:

- Парсинг и очистка данных
- Исследование признаков влияющих на поисковую выдачу
- Разработка системы рекомендаций по улучшению карточек

Парсинг и очистка данных. Карточка товара

Какие признаки можно изменить в карточке

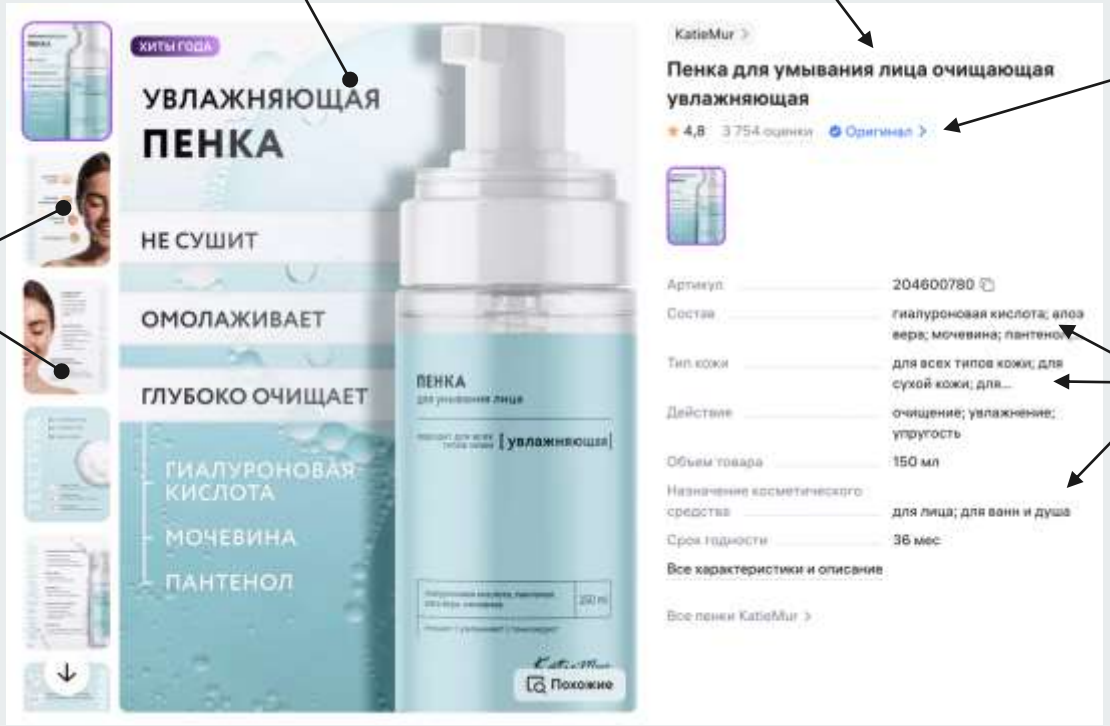
Титульное изображение

Наименование

Наличие сертификата

Количество фотографий и наличие либо отсутствие видео

Характеристики



Парсинг и очистка данных. Процесс получения и очистки данных

Получение карточек товара в формате JSON и титульных изображений посредством запросов к API сайта

Извлечение полей, которые могут представлять интерес для анализа, из JSON объектов при помощи моделей Pydantic

Преобразование моделей в словари, а словарей в датафрейм Pandas

Удаление рекламной выдачи, товаров из других категорий которые попали в выдачу, первичный анализ и удаление зависимых и мусорных признаков

Подготовленный датафрейм

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4052 entries, 0 to 4051
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype 
---  ------                          --------------  ----- 
 0   description                     4044 non-null   object
 1   title_image                     4052 non-null   object
 2   position                        4052 non-null   int16 
 3   Комплектация                    3448 non-null   object
 4   Тип_кожи                        3008 non-null   object
 5   Состав                          3377 non-null   object
 6   Действие                        3912 non-null   object
 7   Страна_производства             3697 non-null   object
 8   Срок_годности                   3563 non-null   object
 9   Назначение_косметического_средства  3799 non-null   object
 10  Объем_товара                    3630 non-null   object
 11  photo_count                     4052 non-null   int8  
 12  has_video                       4052 non-null   bool  
 13  is_autoplaying_video            4052 non-null   bool  
dtypes: bool(2), int16(1), int8(1), object(10)
memory usage: 336.5+ KB
```

Исследование признаков влияющих на поисковую выдачу. Детекция текста

Процесс выделения легко читаемых слов

Обнаружение и распознавание текста при помощи EasyOCR

['УВЛАЖНЯЮЩАЯ',
'ПЕНКА',
'НЕ',
'СУШИТ',
'ОМОЛАЖИВАЕТ',
'ГЛУБОКО',
'ОЧИЩАЕТ',
'ТИРЛУРОНОВАЙ',
'КИСЛОТА',
'МОЧЕВИНА',
'ПАШТЕНол']

['[ЫЯ',
'[я',
'УМЫВАНИЯ',
'[ИЯООШНЯГИДОМ',
'ГН_к,:',
'ПЧНШЕНИЕ',
'КаЛЛЖНЕННЕ',
'@БЪЕМ',
'ЬЫ',
'аЧаЯГЖЕНИЕ',
'ПОДАРОК',
'маашан',
'7е',
'в']

Выделение легко читаемых слов длиной более трех символов через сопоставление обнаруженного текста с существующими словами при помощи pymorphy2

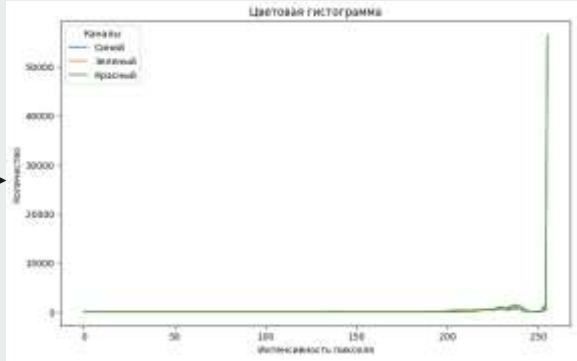
['УВЛАЖНЯЮЩАЯ',
'ПЕНКА',
'СУШИТ',
'ОМОЛАЖИВАЕТ',
'ГЛУБОКО',
'ОЧИЩАЕТ',
'КИСЛОТА',
'МОЧЕВИНА']

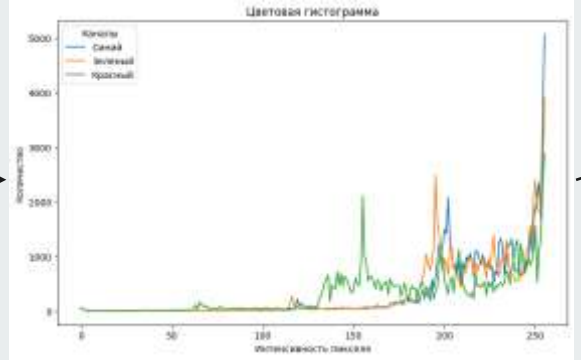
['УМЫВАНИЯ',
'ПОДАРОК']

Исследование признаков влияющих на поисковую выдачу. Выделение цвета

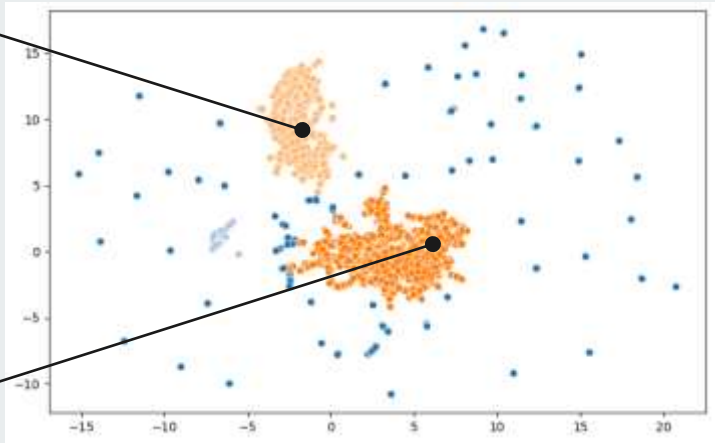
Процесс выделения цвета

Извлечение гистограмм интенсивности цветовых каналов при помощи OpenCV

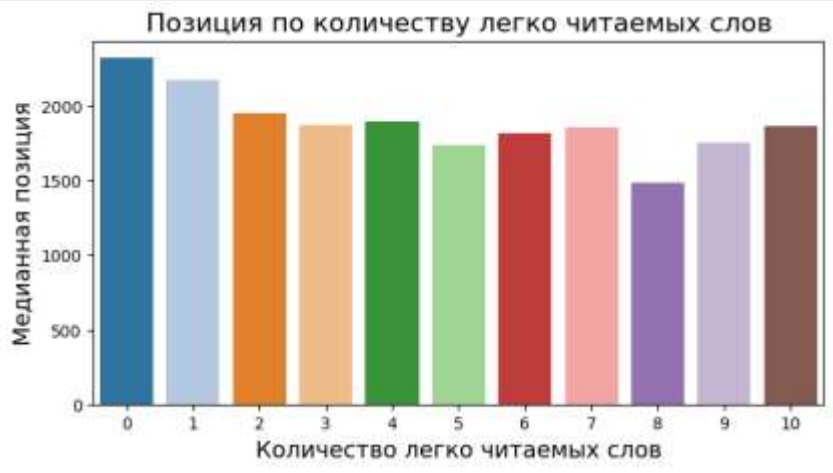




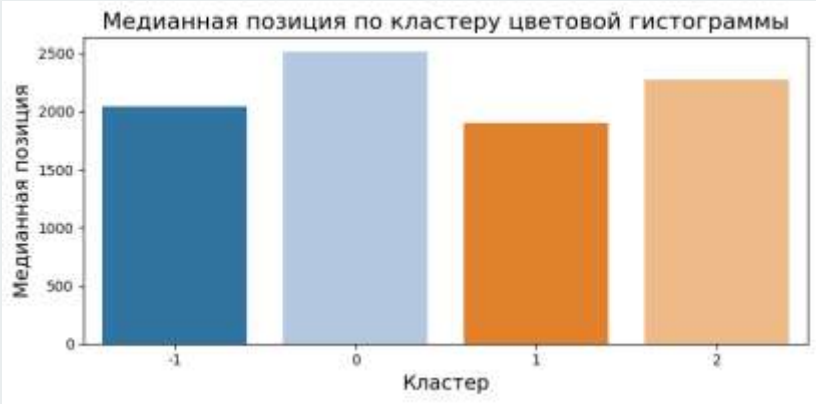
Уменьшение размерности гистограмм цветовых каналов при помощи UMAP и кластеризация при помощи HDBSCAN



Исследование признаков влияющих на поисковую выдачу. Влияние признаков изображения



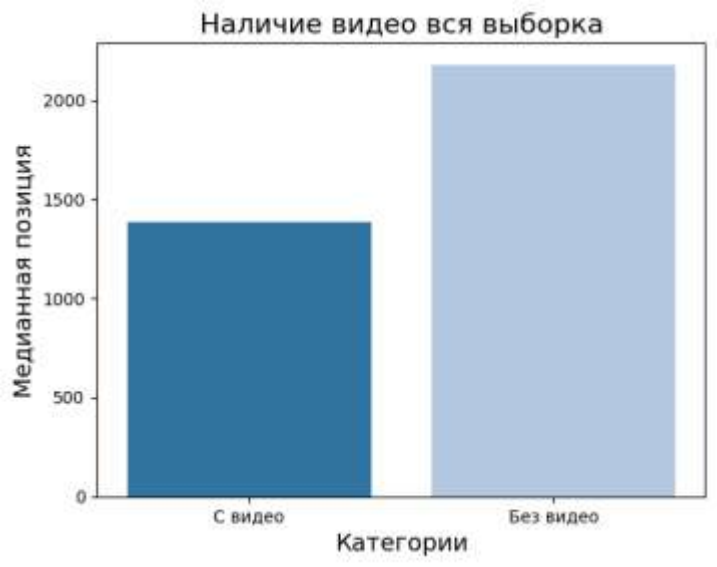
Присутствует слабая отрицательная корреляция между позицией и длиной легко читаемых слов на картинке
Коэффициент корреляции: -0.2
p-значение: 0

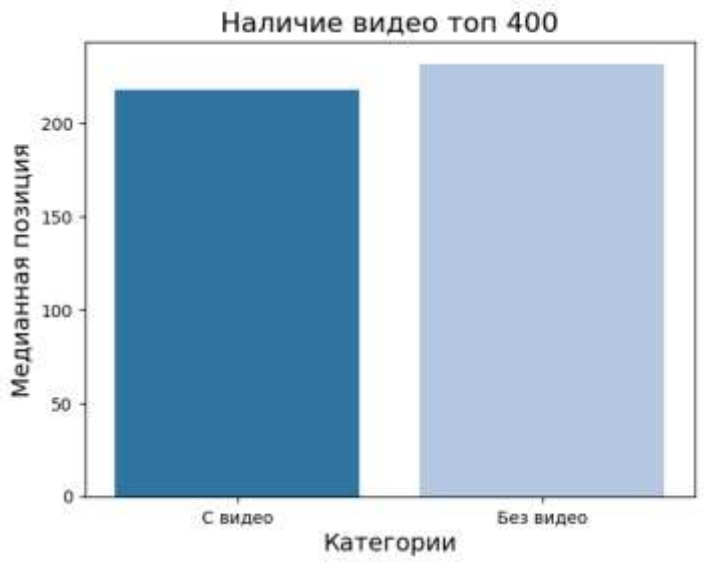


Результаты теста Крускала-Уоллиса (статистика: 55.6, **p-значение: 0**) показали наличие статистически значимых различий в выдаче между группами товаров отнесенных к разным кластерам

Исследование признаков влияющих на поисковую выдачу. Влияние наличия видео

Результаты теста **Манна-Уитни** (U-статистика: 772797.5, **p-значение: 0**) показали наличие статистически значимых различий в выдаче между группами товаров с видео и без него. Однако результаты теста (U-статистика: 23792.5, **p-значение: 0.41**) на топ 400 товаров показали отсутствие статистически значимой разницы. Что может свидетельствовать о том что продавцы из топа рейтинга более склонны загружать видео товаров, а само наличие видео не является важным критерием для этой категории товаров.

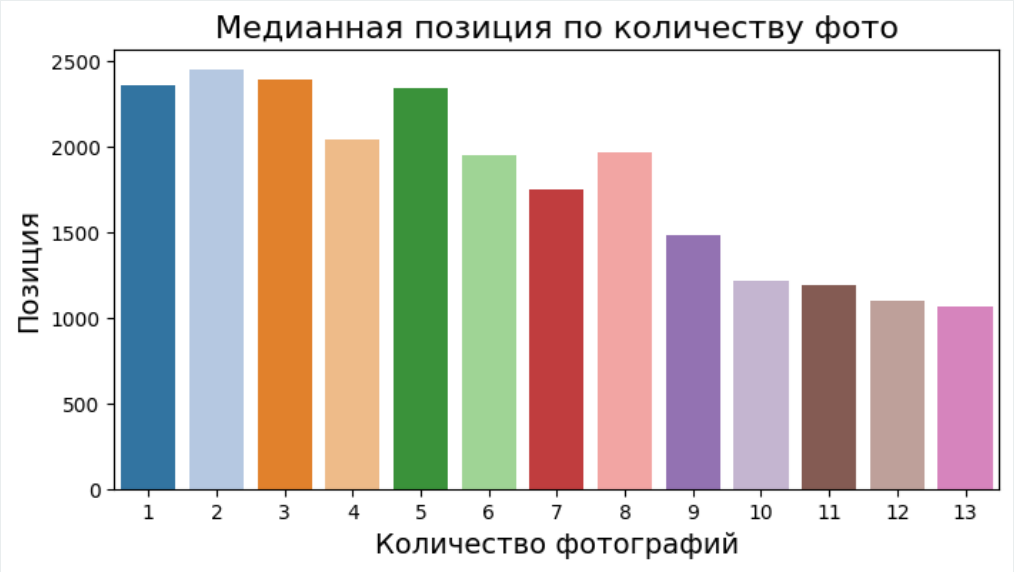




Исследование признаков влияющих на поисковую выдачу. Количество фотографий

Результаты **теста Крускала-Уоллиса** (статистика: 210, **p-значение: 0**) показали наличие статистически значимых различий в выдаче между группами товаров с разным количеством фотографий.

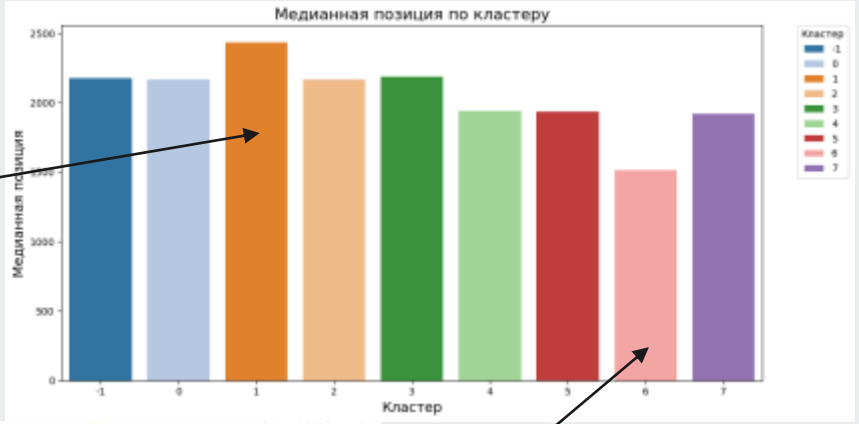
Позиция имеет слабую отрицательную корреляцию (статистика: - 0.2, p-значение: 0) с количеством фотографий

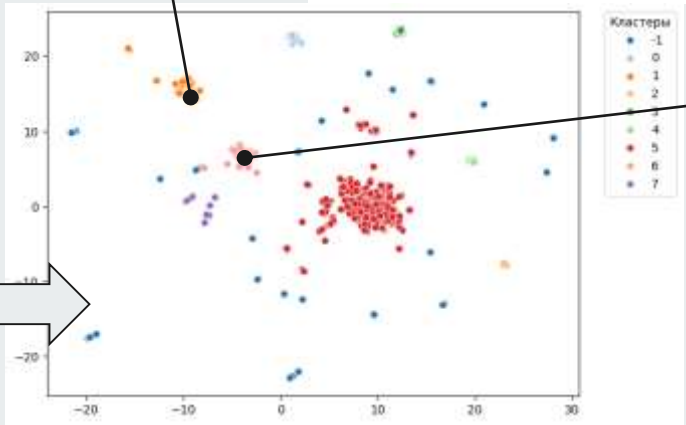


Исследование признаков влияющих на поисковую выдачу. Состав товара

Существует слабая отрицательная корреляция между диапазоном длины поля “состав” и позицией
Коэффициент корреляции: -0.15

“Состав в описании”



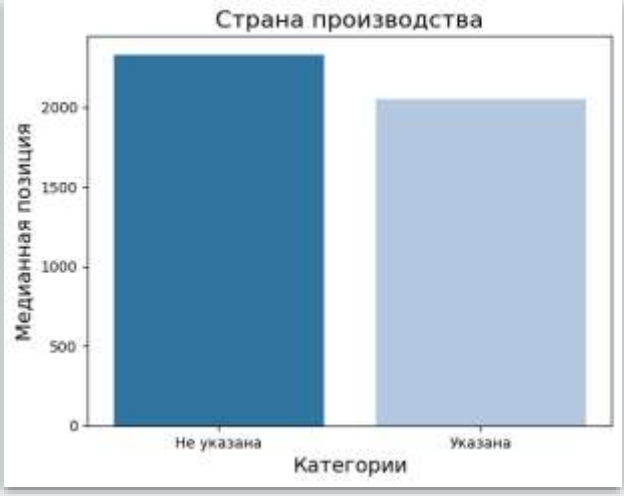


Описание состава в этом кластере состоит из латинских наименований ингредиентов

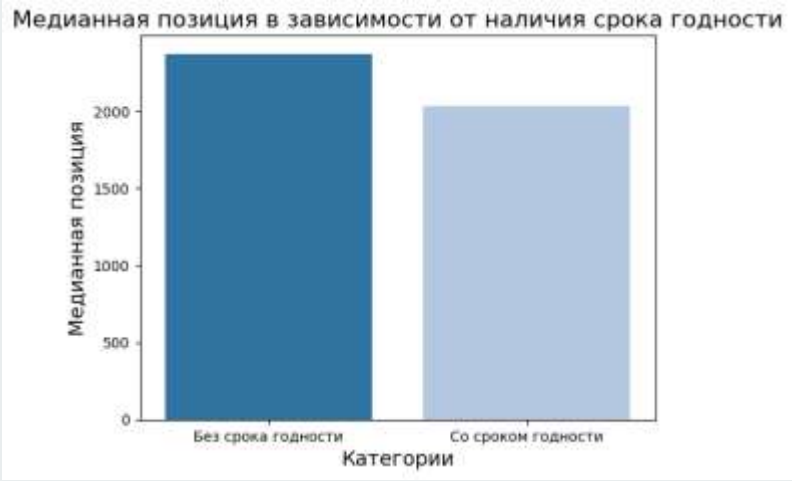
Извлечение признаков текста TfidfVectorizer библиотеки sklearn, уменьшение размерности UMAP с метрикой “cosine”, кластеризация HDBSCAN min_cluster_size = 50

Исследование признаков влияющих на поисковую выдачу. Срок годности, страна производства

Результаты теста **Манна-Уитни** (U-статистика: 746257, **p-значение: 0**) показали наличие статистически значимых различий в выдаче между группами товаров с заполненной характеристикой “страна производства” и без нее.

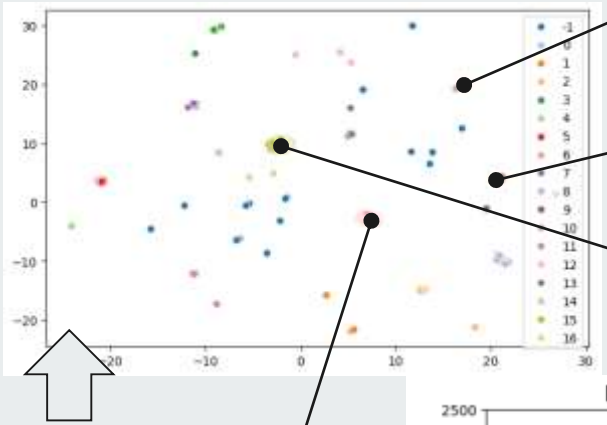


Результаты теста **Манна-Уитни** (U-статистика: 1016601, **p-значение: 0**) показали наличие статистически значимых различий в выдаче между группами товаров с заполненной характеристикой “срок годности” и без нее.



Исследование признаков влияющих на поисковую выдачу. Тип кожи

Результаты теста Манна-Уитни (U-статистика: 16350088, p-значение: 0.046) показали наличие статистически значимых различий в выдаче между группами товаров с заполненной характеристикой “тип кожи” и без нее



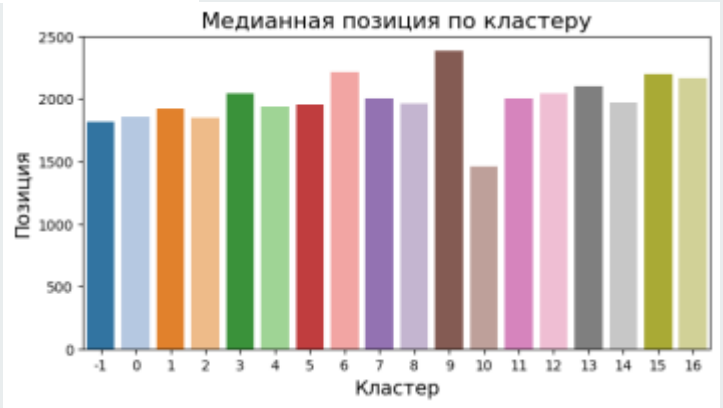
Кластер *10*
“для всех типов кожи; для сухой кожи; для чувствительной кожи”

Кластер *9*
“для всех типов кожи; для нормальной кожи; для комбинированной кожи”

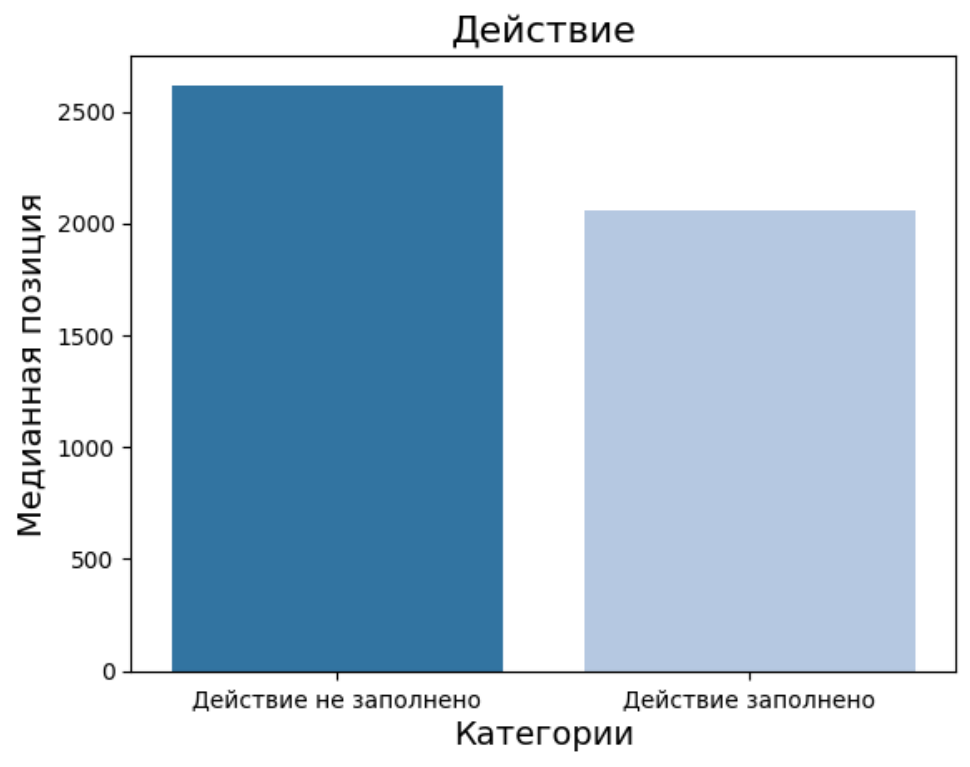
Кластер *15*
Характеристика отсутствует

Извлечение признаков текста TfidfVectorizer библиотеки sklearn, уменьшение размерности UMAP с метрикой “cosine”, кластеризация HDBSCAN min_cluster_size = 50

Кластер *6*
“для всех типов кожи”



Исследование признаков влияющих на поисковую выдачу. Действие



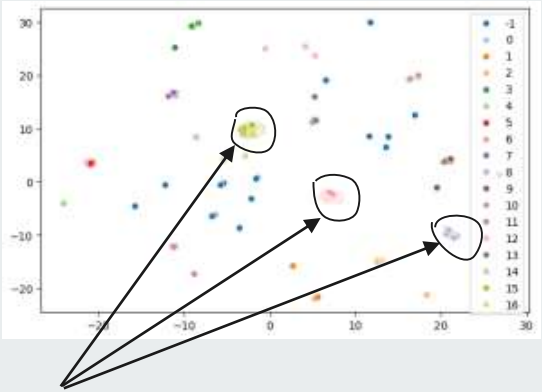
Результаты теста **Манна-Уитни** (U--статистика: 323160.5, **p-значение: 0**) показали наличие статистически значимых различий в выдаче между группами товаров с заполненной характеристикой “Действие” и без нее.

Результаты анализа поисковой выдачи

- Заполненность состава, типа кожи, действия, срока годности, объема товара и страны производства влияет на поисковую выдачу.
- Лучшее заполнение состава - перечисление латинских наименований компонентов, худшее - “Состав в описании”.
- Лучшее заполнение типа кожи - указание специфических особенностей кожи.
- Заполнение характеристики “Комплектация” не влияет на поисковую выдачу в данной категории товара.
- Оптимальное количество фотографий в карточке 10-13.
- На титульном изображении стоит размещать легко читаемый текст.
- Оптимальное количество легко читаемых слов на титульном изображении 5-9.
- Титульное изображение с фоном лучше чем без него
- Наличие видео не влияет на поисковую выдачу в данной категории товара.
- Несмотря на то, что статистический тест Манна-Уитни (U--статистика: 506966.5, p-значение: 0.14) не показал наличия статистически значимой разницы между товарами с заполненным и незаполненным значением “назначение косметического средства”, процент товаров с незаполненным значением 6% и поле лучше заполнить.

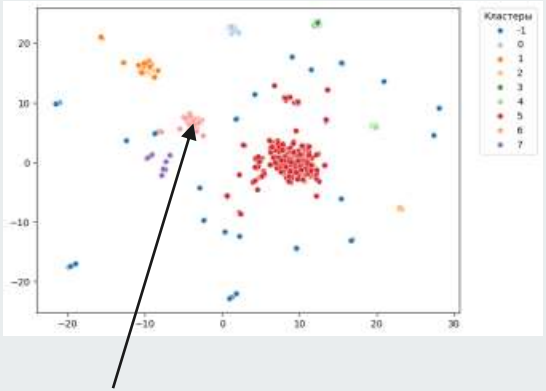
Разработка классификаторов

Тип кожи



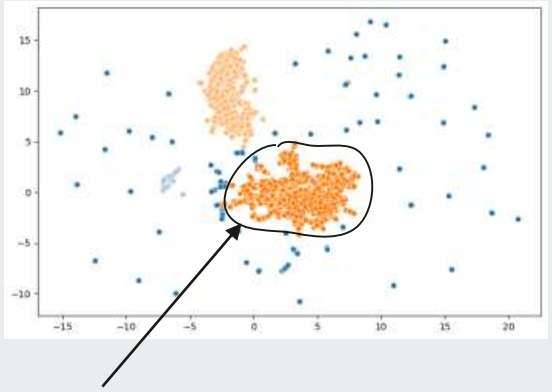
Кластеры без информации о специфике типа кожи, и статистически занимающие худшие позиции в поисковой выдаче получают метку 1, остальные метку 0.

Состав



Кластер с перечислением латинских наименований ингредиентов, статистически имеющий более высокие позиции в поисковой выдаче получает метку 1, остальные 0.

Цветовая гистограмма



Изображения с фоном, которые статистически занимают более высокие позиции получают метку 1, остальные 0

Для построения классификаторов использовался метод k-ближайших соседей, все классификаторы, при кроссвалидации, показали метрики близкие к 1, ввиду того что кластеры имеют четкие границы между группами.

Схема системы рекомендаций