ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Исследовательский проект на тему:

Сжатие словарей для нейросетевого анализа исходных кодов программ

Выполнил студент:

группы #БПМИ171, 4 курса Гусев Андрей Алексеевич

Принял руководитель ВКР:

Чиркова Надежда Александровна
Научный сотрудник
Факультет компьютерных наук НИУ ВШЭ

Москва 2024

Содержание

Аннотация

Ваша аннотация на русском языке.

Ключевые слова

Глубинное обучение, разреживание моделей, рекуррентные нейронные сети

1 Введение

В современном мире маркетплейсы стали неотъемлемой частью электронной коммерции, предоставляя платформы для продажи товаров и услуг различным производителям и ритейлерам. В 2023 году маркетплейсы продолжили быть главной движущей силой российской онлайн-торговли. Рост объема трат на маркетплейсах в 1,5 раза по сравнению с предыдущим годом свидетельствует о том, что интерес потребителей к онлайн-покупкам только укрепляется. На это влияют общерыночные факторы: продолжают развиваться альтернативные каналы поставок продукции ушедших брендов, улучшаются условия доставки, повышается удобство пользования платформами, расширяется сеть пунктов выдачи.

Согласно исследованиям Tinkoff Ecommerce, количество транзакций на маркетплейсах за год выросло на 63% (см. рисунок 1.1). Лидерами по росту количества покупок стали Мегамаркет (число транзакций выросло в 4,3 раза), Wildberries (в 2 раза) и Ozon (в 1,6 раза).

Рис. 1.1: Динамика покупок на маркетплейсах в регионах.

Появился тренд на рост популярности маркетплейсов в российских регионах. В 2023 году жители российских городов стали значительно активнее совершать покупки на маркетплейсах: выросло как количество транзакций на онлайн-площадках, так и их сумма (см. рисунок 1.2). По количеству совершенных транзакций особенно заметен рост в таких городах, как Омск (+91%), Красноярск (+88%), Новосибирск (+79%), Челябинск (+79%) и Волгоград (+75%). В Москве зафиксирован наименьший прирост числа транзакций (+41%). Увеличение интереса жителей регионов к маркетплейсам объясняется рядом причин: расширением географии присутствия площадок, развитием сетей ПВЗ и логистических сервисов,

улучшением условий доставки.

Рис. 1.2: Динамика покупок на маркетплейсах в регионах.

Вместе с тем выросло количество селлеров на 8%. Рынок становится более зрелым: место неопытных продавцов занимают более профессиональные. Они ведут бизнес более уверенно, укрепляют свои позиции на площадках и торгуют на нескольких платформах одновременно. Количество селлеров, ведущих торговлю на двух и более маркетплейсах, за год увеличилось на 17%. Самой привлекательной платформой для старта бизнеса является Wildberries: 63% продавцов в конце 2023 года выбирали ее в качестве первой площадки (см. рисунок 1.3).

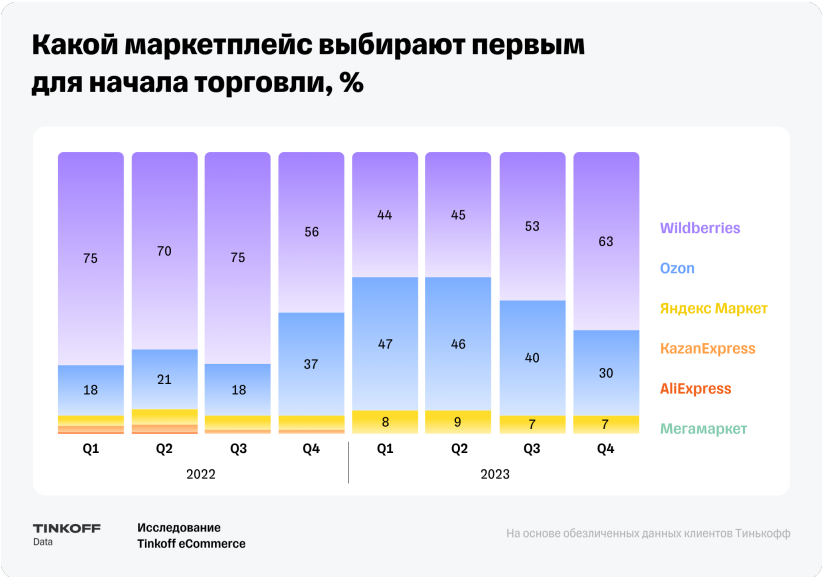

Рис. 1.3: Популярность маркетплейсов за 2023 год.

С увеличением конкуренции продавцы все чаще обращаются к системам, которые могут помочь им в продажах, а также автоматизировать процесс работы. Одним из ключевых

факторов успешной продажи становится эффективная SEO-оптимизация карточек товаров. Подбор наиболее подходящей категории и создание продаваемого описания, содержащего ключевые слова, позволяет улучшить видимость товаров в результатах поиска как на самом маркетплейсе, так и в поисковых системах, что напрямую влияет на увеличение продаж.

На данный момент существуют несколько сервисов (например, TurboTextPro, CopyMonkey, Gerwin), позволяющих сгенерировать описание товаров по характеристикам, ключевым словам или фото. Однако, качество сгенерированных описаний не всегда позволяют использовать их в системах автономного управления. В настоящее время российский рынок не предлагает специальных технологий и решений для подбора наиболее подходящей категории товара на маркетплейсе. Таким образом, задача создания качественного инструмента для эффективной SEO-оптимизации карточек товаров является актуальной.

Цель данной работы — разработать сервис, в основе которого будет реализован функционал, способный классифицировать товары на маркетплейсе на основе их фотографий и генерировать соответствующие к ним описания. Сервис реализуется для маркетплейса Wildberries, так как он представляют наибольшую популярность у продавцов и покупателей. В качестве дальнейшего развития проекта можно рассмотреть другие площадки.

Для достижения поставленной цели необходимо решить следующие задачи:

- подготовить набор данных, содержащий изображения товаров и соответствующие им категории и текстовые описания;
- провести исследование текущих архитектур нейронных сетей, используемых для классификации изображений и генерации текста, и на основе этого исследования выбрать наиболее подходящую архитектуру или их комбинацию для решения поставленной задачи;
- обучить выбранную нейронную сеть на подготовленных данных, оптимизировать и настроить параметры модели для повышения её производительности и качества результатов, а затем оценить эффективность реализованной архитектуры нейронной сети;
- интегрировать модель в программное обеспечение.

2 Данные

Данные для поставленной задачи собирались самостоятельно, поскольку в открытых источниках не имелось удовлетворяющего всем требованиям датасета. (тут можно описать, какие варианты есть). Как было сказано ранее, выбор маркетплейса пал на «Wildberries». Соответственно данные собирались согласно особенностям структуры данного маркетплейса. Для удобства введем некоторые термины, которыми будем оперировать далее:

- **Карточка товара** – это страница продукта на маркетплейсе, где размещена информация о товаре, фотографии, описание цены и кнопка «Купить»
- **Конечная категория** – это категория, на которых располагаются карточки товаров
- **Материнская категория** – это категория, которая содержит конечные и материнские категории и на которой не располагаются карточки товаров

Каталог «Wildberries» разделен на категории, в которых размещены карточки товаров одного типа. Категории выстроены по принципу дерева. Есть основные «широкие» категории, такие как «Женщинам», «Дом», «Продукты», которые объединяют внутри себя более мелкие подкатегории. Например, в разделе «Дом» имеются подкатегории «Ванная», «Кухня», «Спальня» и тд, которые в свою очередь могут подразделяться на еще более маленькие подкатегории.

Требуемые данные располагались на товарных карточках, в которые можно попасть только зная конечную категорию товара. Поэтому было принято решение разделить сбор данных на 2 этапа. На первом этапе был произведен сбор всех имеющихся на «Wildberries» конечных категорий. На втором – сбор необходимой информации с карточек товаров.

2.1 Этап 1

«Wildberries» предлагает 22 основные категории (см. рисунок 2.1), из которых одна является конечной категорией. Данные категории в дальнейшем будем называть категориями первой вложенности. Их подкатегории, соответственно, будут называться категориями второй вложенности. И так далее, спускаясь все ниже по дереву категорий. Экспериментальным путем была выявлена максимальная глубина вложенности – 5.

Для правильного формирования таргета для классификации при сохранении ссылки на конечную категория нужно было учитывать весь путь по дереву категорий, начиная с первой вложенности. Решением данной задачи стало создание таблицы, где отражалось какие

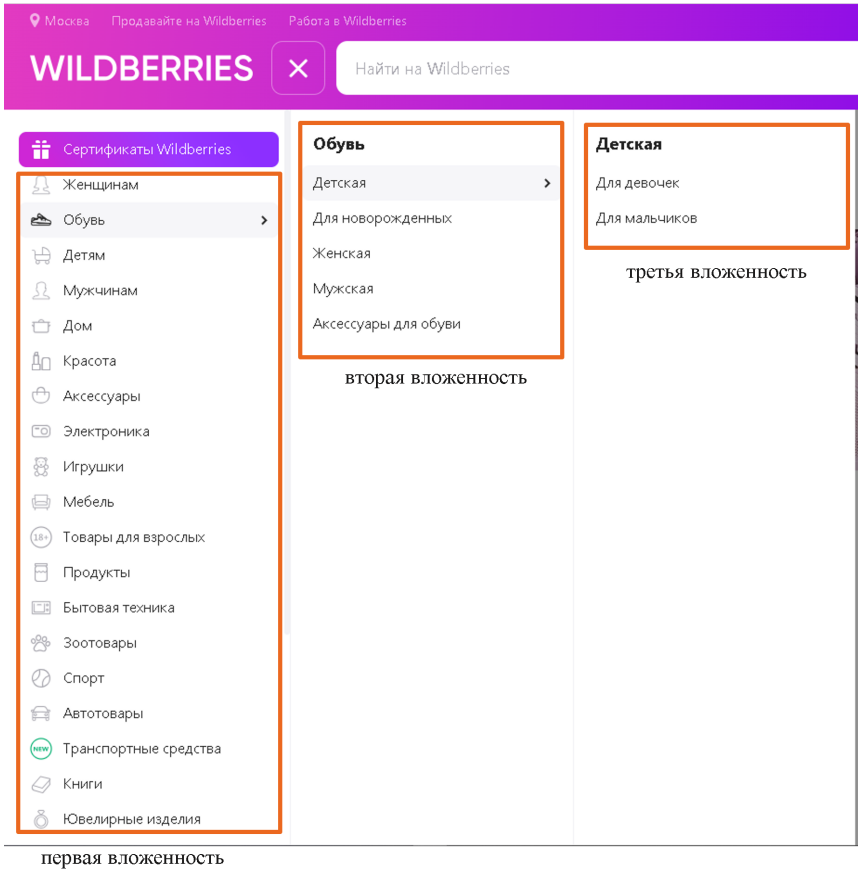

Рис. 2.1: Структура каталога «Wildberries» на примере категории «Обувь».

категории было предшествовавшими конкретной конечной категории (см. таблицу 2.1). При отсутствии более глубокой вложенности на месте данных категорий ставились «NaN». Таким образом, было собрано 1668 конечных категорий.

Таблица 2.1: Фрагмент таблицы, полученной после первого этапа сбора данных.

	$category_1$	$category_2$	$category_3$	$category_4$	$category_5$
437	Дом	Предметы интерьера	Фоторамки и фотоальбомы	Фотоальбомы	NaN
438	Дом	Предметы интерьера	Картины и постеры	Рамы для постеров	NaN
439	Дом	Предметы интерьера	Картины и постеры	Постеры	Детская тематика
440	Дом	Предметы интерьера	Картины и постеры	Картины	Арт и абстракция
441	Дом	Предметы интерьера	Картины и постеры	Постеры	Фэнтези

Составление данной таблицы производилось посредством парсинга данных с сайта «Wildberries» через Python с использованием библиотек selenium и BeautifulSoup. Блокировок со стороны маркетплейса замечено не было. Особенность и неудобством парсинга была динамическая подгрузка страниц, которая вынуждала выдерживать паузы в несколько секунд для удовлетворяющей прогрузке страницы. Данное обстоятельство привело с значительному увеличению времени парсинга данных.

При анализе собранной таблицы были выявлены некоторые особенности категори-

альной политики «Wildberries». Во-первых, категории у данного маркетплейса не фиксированы. Например, было отмечено, что часть категорий активно перемещается из раздела в раздел, какие-то категории могут пропадать, также могут появляться новые категории. Данные, собранные в текущем датасете, актуальны на конец января 2024 года. Однако для поддержания списка категорий в актуальном состоянии необходимы механизмы регулярного обновления данных. Во-вторых, на маркетплейсе имеются конечные категории, ссылающиеся на одни и те же url страницы. Подобные категории будут называться дублирующими. Подобные дубляжи могли иметь разное происхождение: особенности маркетинга и неудачное время парсинга, выпавшее на перемещение категорий. С точки зрения маркетинга подобные дублирования оправданы, поскольку потенциальный покупатели могут по своим соображениям относить одни и те же товары к разным категориям. Для примера, категория «Коврики» находилась в разделе «Автотовары_Коврики» и «Электроника_Автоэлектроника&и&навигация_Коврики». Для корректной работы модели была написана отдельная процедура удаления подобных дублирующих категорий. Выбор, какой из дубликатов оставлять, производился вручную. Всего было найдено 69 дублирующих ссылок, которые могли встречаться 2 и более раза. Таким образом, после удаления в таблице осталось 1580 категорий.

Далее можно было переходить ко 2му этапу.

2.2 Этап 2

Второй этап сбора данных заключался в прохождении по собранному ранее списку конечных категорий и сбора из каждой из них информации с карточек товаров. Было принято решение брать по 20 товаров из каждой конечной категории. Из каждой карточки товара сохранялось первое фотография от продавца, первая фотография из отзыва и описание товара (см. рисунок 2.2). Первая фотография от продавца бралась по причине ее обязательного присутствия в карточке товара, а также гарантированного качественного изображения товара на ней. Однако поскольку разрабатываемый сервис рассчитан на работу в большинстве случаев с фотографиями от пользователей, все дефекты, присущие любительским фотографиям могут иметь место быть. Поэтому для стабильности предсказаний классификационной модели, было решено подавать в нее также фотографии из отзывов, которые максимально близко будут похожи на фотографии, с которыми будет работать в дальнейшем сервис. Описания товара были нужны для задачи генерации текстовых описаний к изображениям. В итоге при полном наборе для каждой конечной категории имелось 40 фотографий (20 фотографий от

продавцов и 20 фотографий из отзывов) и 20 описаний.

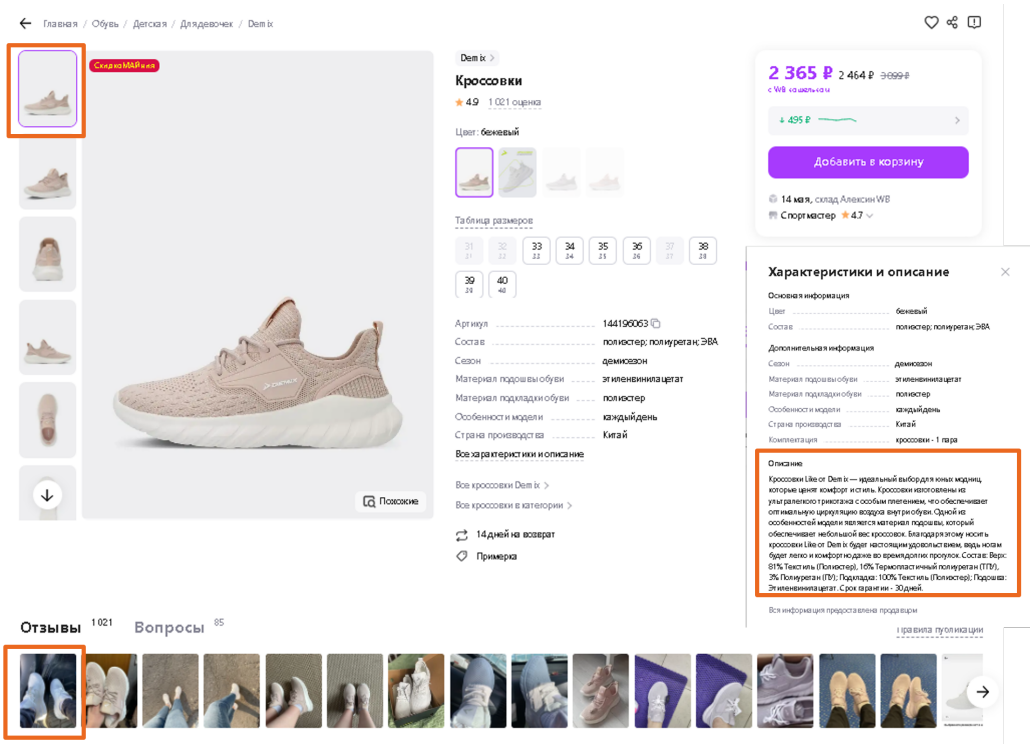

Рис. 2.2: Пример карточки товара на «Wildberries». Верхний левый прямоугольник – первая фотография от продавца. Нижний левый прямоугольник – первая фотография из отзыва. Правый прямоугольник – текстовое описания товара.

При выполнении данного этапа было несколько трудностей. Во-первых, стандартная на «Wildberries» динамическая прогрузка страниц, увеличивающая время парсинга более чем в 3 раза. Приходилось делать паузы при открытии страницы с карточкой товара, при открытии описания, лежащее в отдельной вкладке, и пролистывание страницы вниз для подгрузки информации об отзывах. Примерное время парсинга данных, затраченное на второй этап, равнялось 2м неделям. Во-вторых, имели товары с меткой «18+», для которых требовалось дополнительное нажатие кнопки, подтверждающее достижение указанного возраста. В-третьих, некоторые поля в карточке товара заключали в себе картинки, которые вынуждали в определенных случаях дополнительно пролистывать страницу вниз. В-четвертых, для нажатия кнопки с целью получения описание к товару, выдвигалось требование расположение кнопки в зоне видимости экрана. Это приводило к еще более тонкой настройке пролистывания страницы, подобранной под конкретный размер экрана компьютера.

Для сохранения данных из карточек товара была придумала специальная структура с целью дальнейшего удобства использовании в задаче классификации и генерации текста. Все товары, собранные из одной конечной категории, сохранялись в отдельную папку, содержащую следующие элементы:

- папку «card», куда складывались фотографии от продавцов
- папку «feedbacks», куда складывались фотографии из отзывов
- файл «descriptions.csv», где сохранялись описания к товарам

Название данной папки определялось посредство таблицы 1 и складывалось из всех материнских категорий, участвовавших в пути к конечной категории. Например, для конечной категории «Фотоальбомы» (см. таблицу 2.1) название папки было следующее: «Дом_Предметы а для категории «Фэнтези» - «Дом_Предметы&интерьера_Картины&и&постеры_Постеры_Фэнт Более подробно об использовании подобной структуры ранения данных будет описание в главе ___ в разделе ___.

Первичный анализ собранных данных выявил, что не у всех товаров имелись отзывы с фотографиями и описания. Описания имелись в 99.8% проценте случаев. В таблице 2.2 приведены некоторые статистические данные о собранных фотография от продавца и из отзыва. Можно заметить, что некоторые конечные категории были полностью без фотографий в отзывах. Однако, опираясь на перцентили, можно сделать вывод, что таких категорий было довольно мало. Касательно фотографий от продавцов можно сделать 2 вывода. Во-первых, есть категории, представленные менее чем 20ю товарами. Во-вторых, есть как минимум одна категория, в которой имеется только 1 товар. Подобные категории нас не устраивают, потому что далее будет производиться деление каждой категории на 2 части, и категории с одним товаром невозможно будет разделить.

Таблица 2.2: Описательная статистика по фотографиям от продавца (столбец «card») и фотографиям из отзыва (столбец «feedbacks»).

	card	feedbacks
count	1580	1580
mean	19.98	17.72
std	0.63	4.23
min	1	0
25%	20	18
50%	20	19
75%	20	20
max	20	20

Всего категорий, представленных менее 20 товарами, было выявлено 5 штук (см. таблицу 2.3). Из них представляли наибольший интерес _ и _, из-за чересчур малого количества товаров. Категорию с одним товаром было решено удалить. Таким образом, осталось 1579 конечных категорий, с которыми шла вся дальнейшая работа.

Таблица 2.3: Таблица с категориями, имеющими менее 20 товаров.

кол-во товаров	категория
18	Дом_Кухня_Кухонный&текстиль_Чехлы&для&ручек&холодильников
4	Дом_Освещение_Лифты&для&люстр
19	Мебель_Гардеробная&мебель_Ящики
1	Мебель_Офисная&мебель_Перегородки&офисные
19	Мебель_Офисная&мебель_Шкафы

При более детально рассмотрении собранных данных было замечено, что фотографии из отзывов довольно шумные (см. рисунок 2.3). Очень много одинаковых фотографий, фотографий, где не очень понятно, что изображено. Поэтому для дальнейшей работы использовались только фотографии товара от продавца.

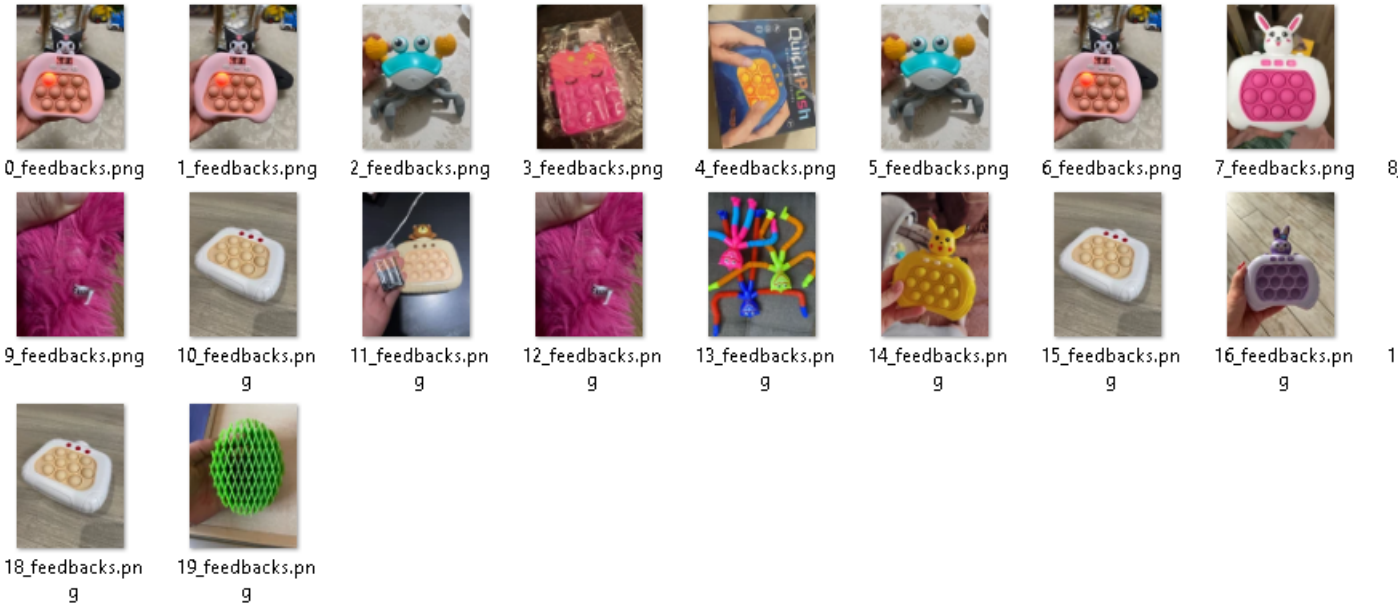

Рис. 2.3: Фотографии из отзывов в категории «Игрушки_Антистресс».

Далее интересно было рассмотреть количество собранных данных в разрезе категорий первой вложенности (см. рисунок 2.4). Из круговой диаграммы можно заметить, что категории довольно несбалансированный. Например, категория «Дом» вмещает в себя порядка 6000 пример. В то время как в категории «Ювелирные&изделия» только 320 примеров.

Подобную картину можно наблюдать в категориях всех вложенностей (см. приложение А).

После появление базового понимания данных надо было приступать к их детальному изучению. Как правило, парсинг большого количества данных, тем более с постоянно меняющихся маркетплейсов, не проходит идеально. В сохраненных данных могут быть оплошности, которые помешают грамотно решать задачу классификации и генерации текста. Приведем некоторые примеры, выявленных особенностей, требуемых принятие решений с нашей сто-

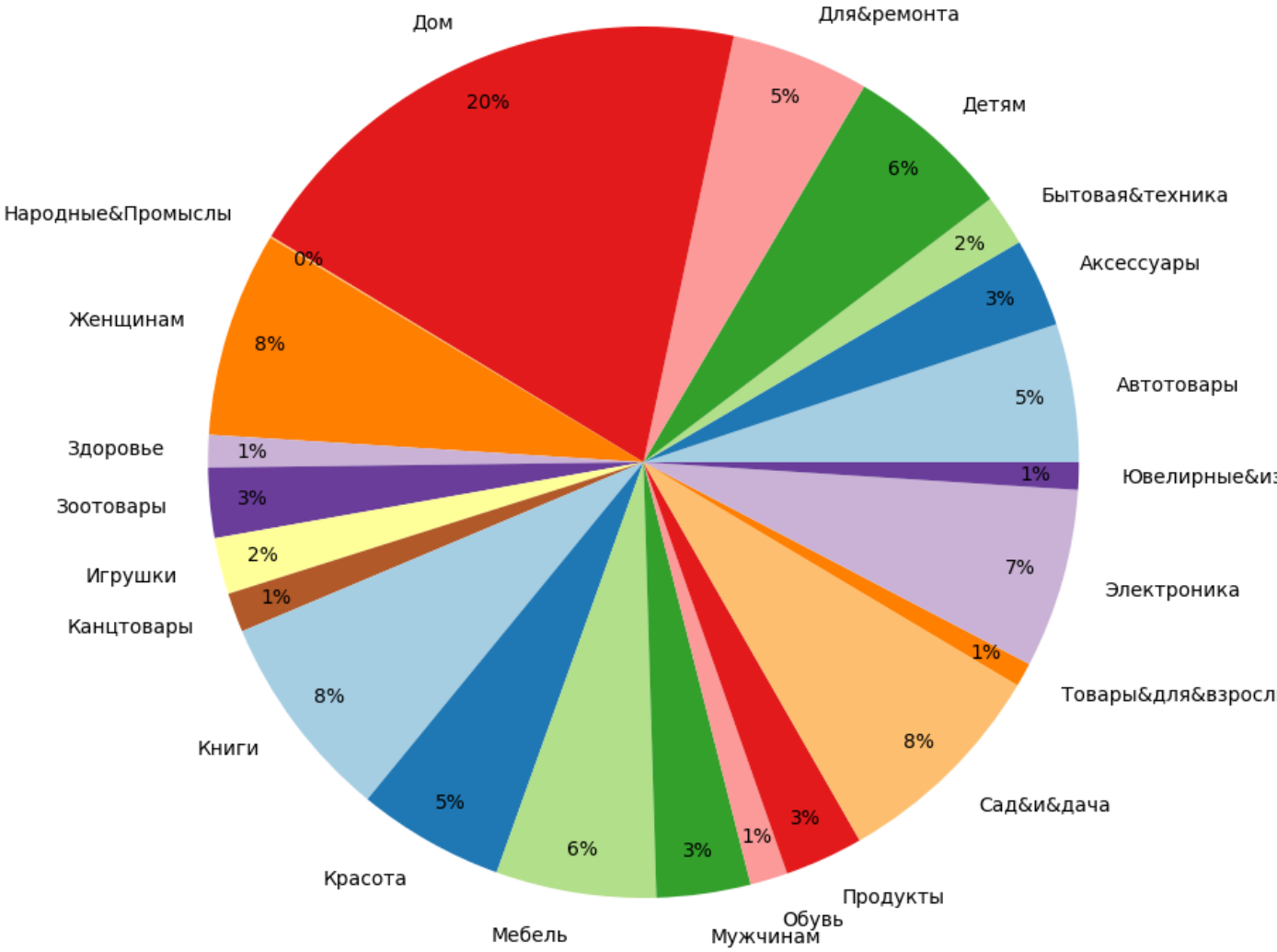

Рис. 2.4: Распределение данных по категориям первой вложенности.

роны:

- Встречаются категории, которые с точки зрения категорийного менеджмента, имеют место быть. Одна для классификационной модели машинного обучения такие категории будут не осиливаемыми. Например, в разделе «Женщинам» есть подкатегории «Женщинам_Белье» и «Женщинам_Большие&размеры_Белье». Если детально изучить фотографии, которые в этих категориях присутствуют, можно сделать вывод, что особых различий между ними нет. Единственное, что было подмечено, что на очень немногих фотографиях стоит надпись «4XL» или что-то подобное, указывающее, что у данного товара имеются большие размера. Более того, в обеих категориях присутствуют одинаковые товары.

- В данных попадались «мусорные» категории. Предположительно, подобное возникало из-за изменения url ссылок на категории со стороны «Wildberries». В подобных категориях находились товары, собранные случайным образом из всевозможных категорий с маркетплейса.

- Распределение товаров по категориям не очень четкая задача. В связи с этим встречались одинаковые товары, находящиеся в разных категориях. Например, одна и та же продукт мог находиться в категориях «Зоотовары_Груминг&и&уход», «Зоотовары_Для&кошек_Груминг&и&уход» и «Зоотовары_Для&собак_Груминг&и&уход».

- В части материнских категорий встречались разделы «Подарки» (например, материнские категории «Мужчинам_Подарки&мужчинам» и «Женщинам_Подарки&женщинам»), куда были собраны товары из совершенно разных категорий, таких как «Аксессуары», «Дом», «Продукты» и тд.

- Поскольку при парсинге из каждой категории брались первые 20 товаров, появляется неконтролируемый фактор того, какие товары стоят вначале. Как правило, пользователи смотрят только на первые товары в выдаче. Поэтому на маркетплейсах существуют множество механизмов и правил отбора товаров, которые будут показаны пользователю в начале. В нашем случае было замечено, что некоторые категории стали более шумными из-за сезонных товаров. Например, в категории «Зоотовары_Фермерство» были найдены пасхальные яйца.

- Бывали категории, которые по смыслу имели место быть как отдельные категории, однако в них были собраны не совсем подходящие товары. Например, в категории «Мужчинам_Религиозная_Православие» находились обычные рубашки и штаны, часть из которых присутствовала также в категории «Мужчинам_Рубашки» и «Мужчинам_Брюки».

Приведенные особенности сохраненных данных требовали ручной очистки датасета. Необходимо было применять следующие действия: полное удаление категории, удаление конкретной фотографии из отзыва, удаление товара полностью (фотографию от продавца, из отзыва и описание к нему) и произведение полного переноса товара (фотографию от продавца, из отзыва и описание к нему) из одной категории в другую. Для удобства и быстроты данной процедуры были написаны функции, позволяющие механизмами Python вносить изменения в собранные данные.

По окончании данной процедуры было подмечено, что категории требовали очистки в разной степени. Какие-то категории, как «Зоотовары», требовали практически полного

переформирования. Другие категории обходились легкой очисткой, например «Продукты». Некоторые категории совсем не требовала вмешательства. Таким образом, финальный датасет стал состоять из 1459 конечных категорий (см. таблицу 2.3).

Таблица 2.4: Описательная статистика по фотографиям от продавца после очистки собранных данных. *Примечание:* статистика по фотографиям из отзывов не приведена, поскольку, как упоминалось ранее, решено было в дальнейшем работать только с фотографиями от продавца.

	card
count	1459
mean	20.09
std	2.32
min	4
25%	20
50%	20
75%	20
max	54

A Приложение 1

Распределение данных по категориям второй вложенности.

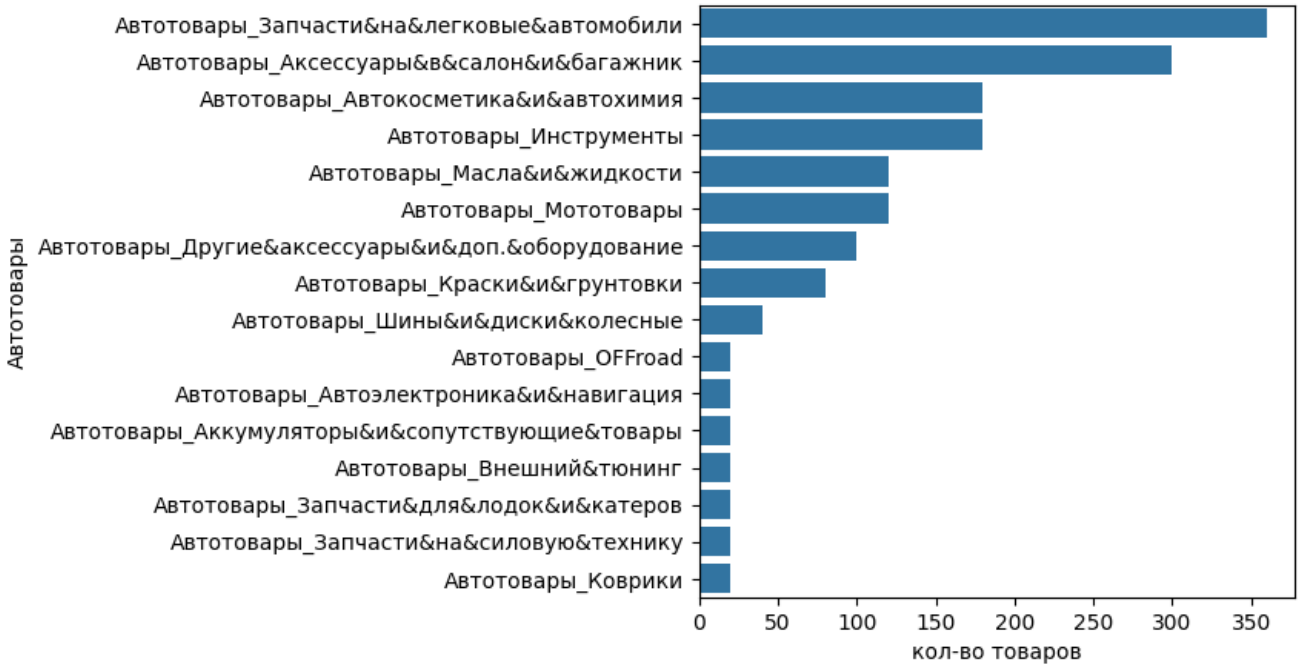

Рис. А.1: Распределение данных в категории «Автотовары».

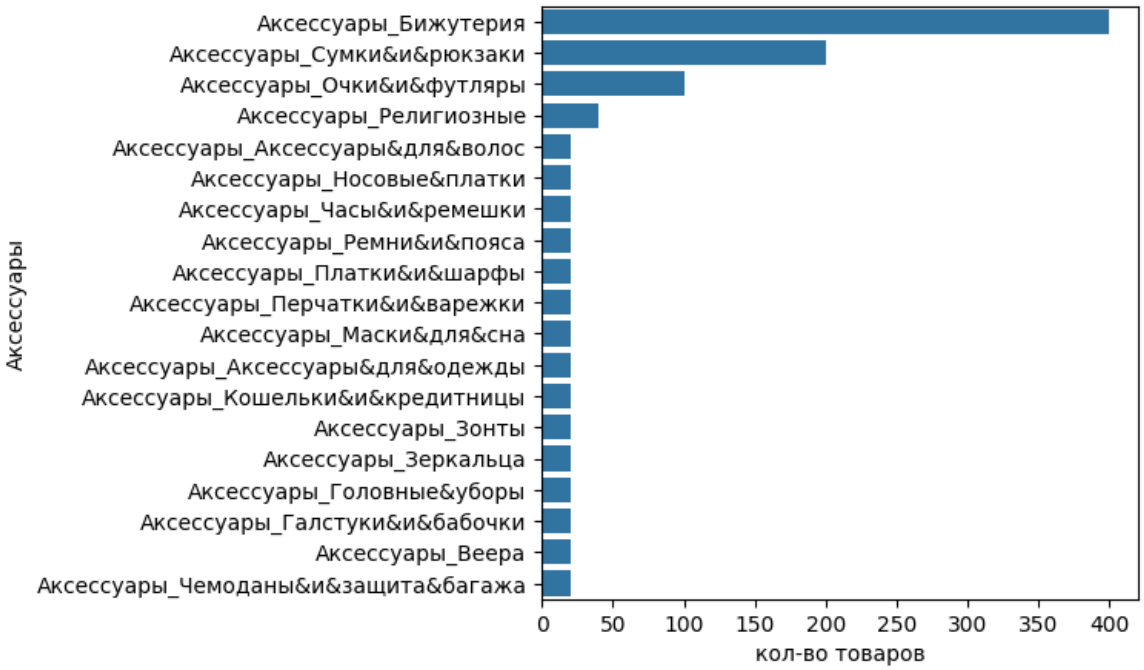

Рис. А.2: Распределение данных в категории «Аксессуары».

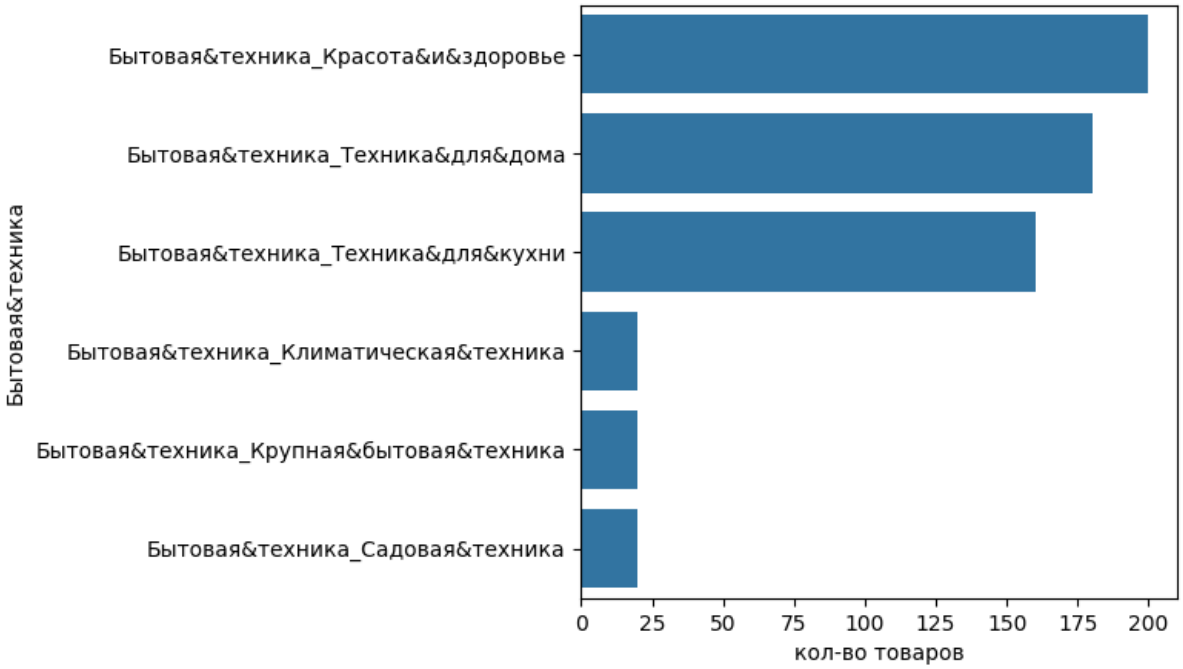

Рис. А.3: Распределение данных в категории «Бытовая&техника».

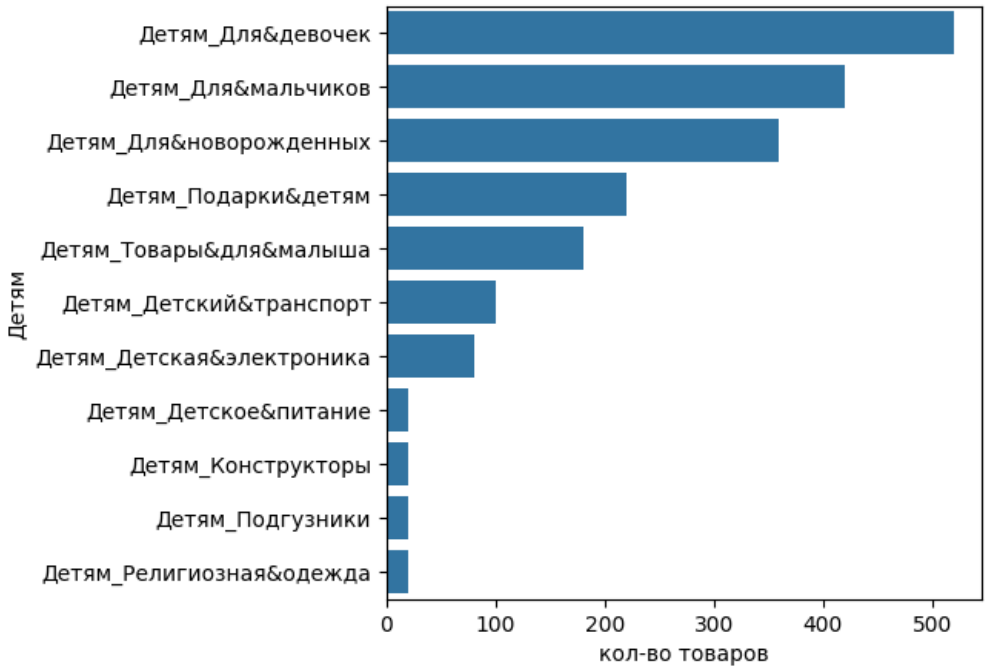

Рис. А.4: Распределение данных в категории «Детям».

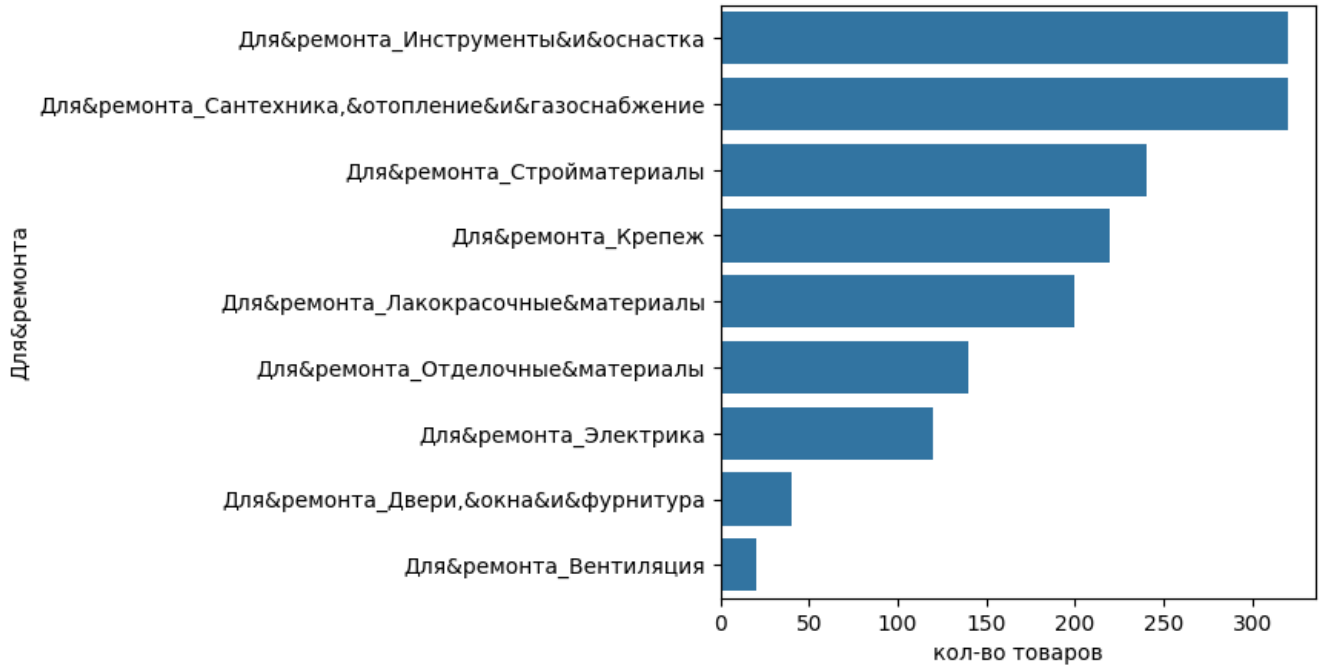

Рис. А.5: Распределение данных в категории «Для&ремонта».

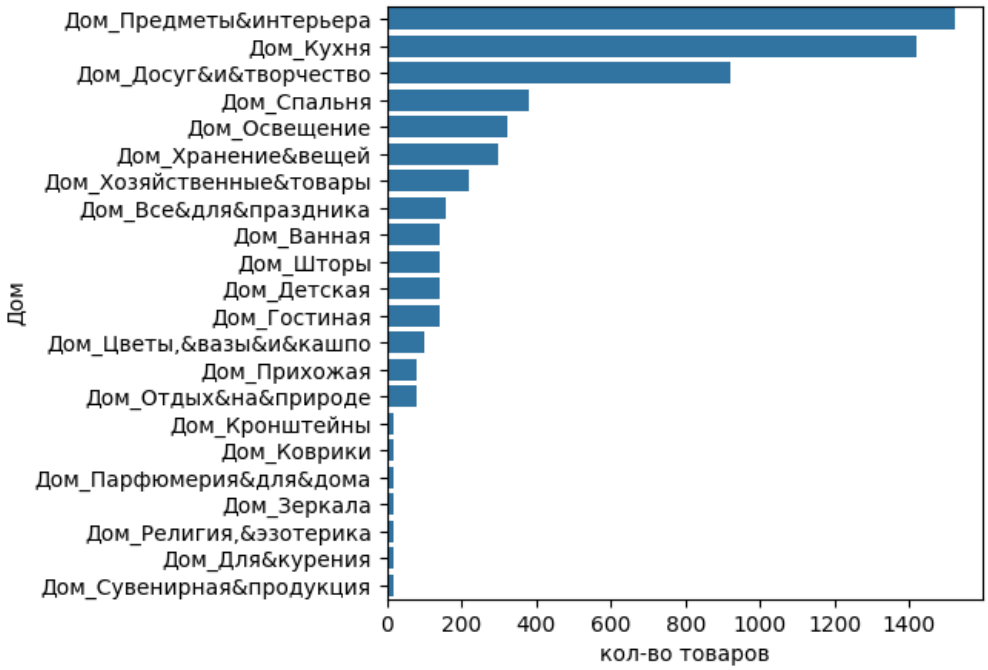

Рис. А.6: Распределение данных в категории «Дом».

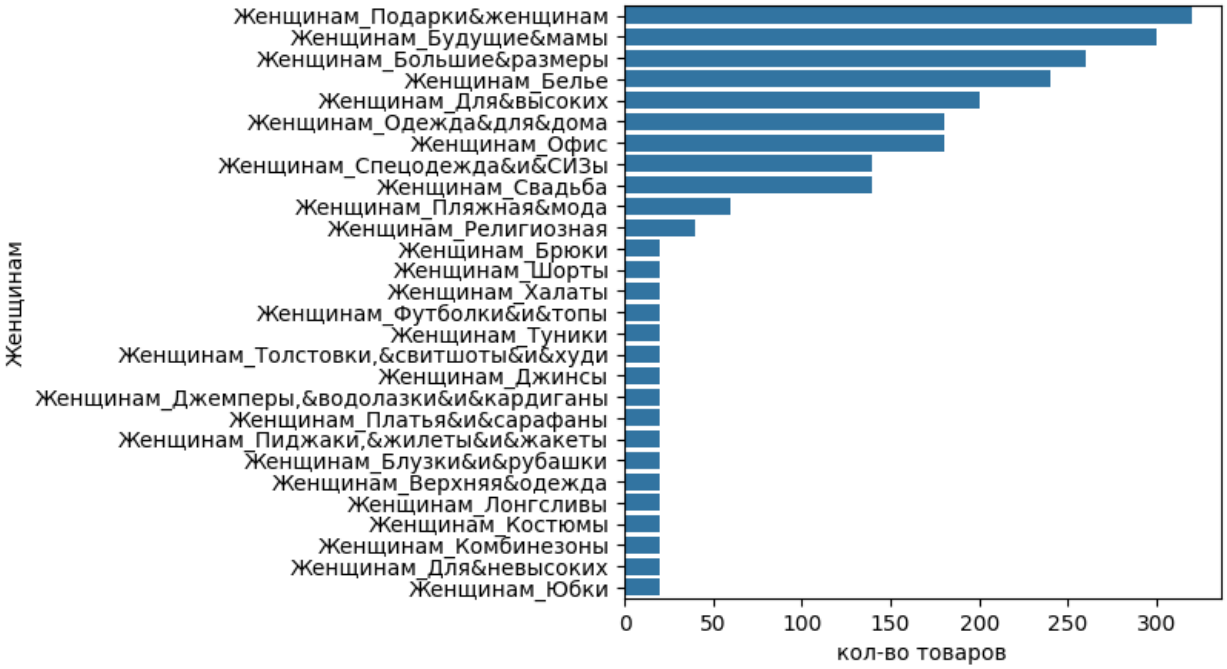

Рис. А.7: Распределение данных в категории «Женщинам».

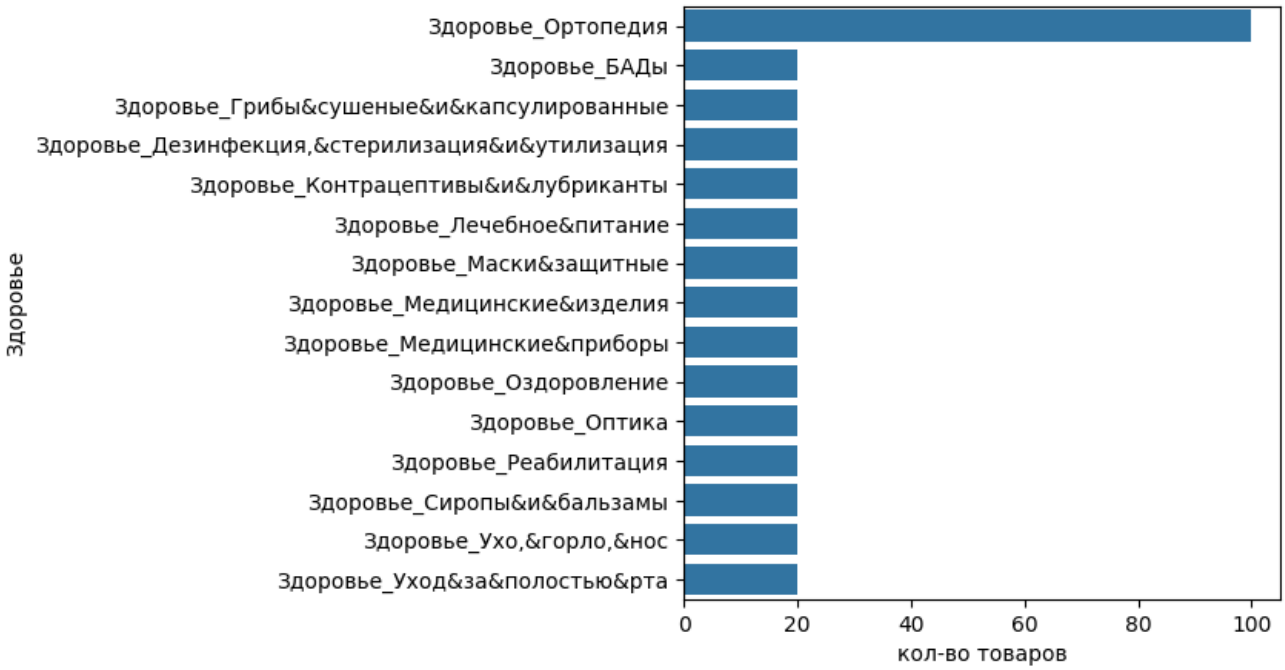

Рис. А.8: Распределение данных в категории «Здоровье».

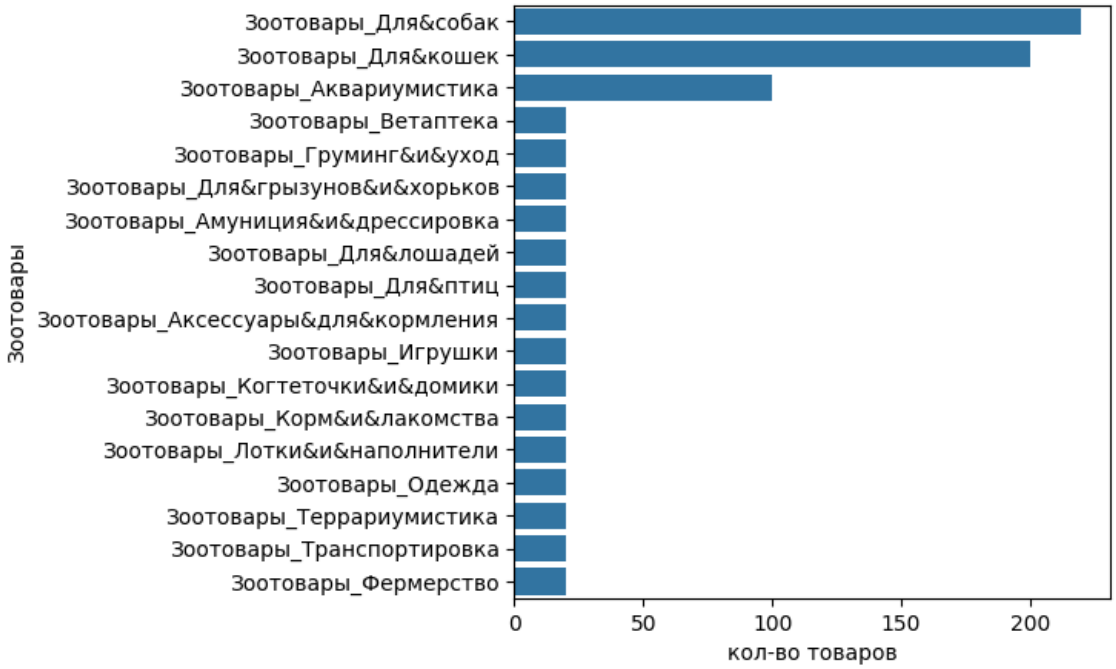

Рис. А.9: Распределение данных в категории «Зоотовары».

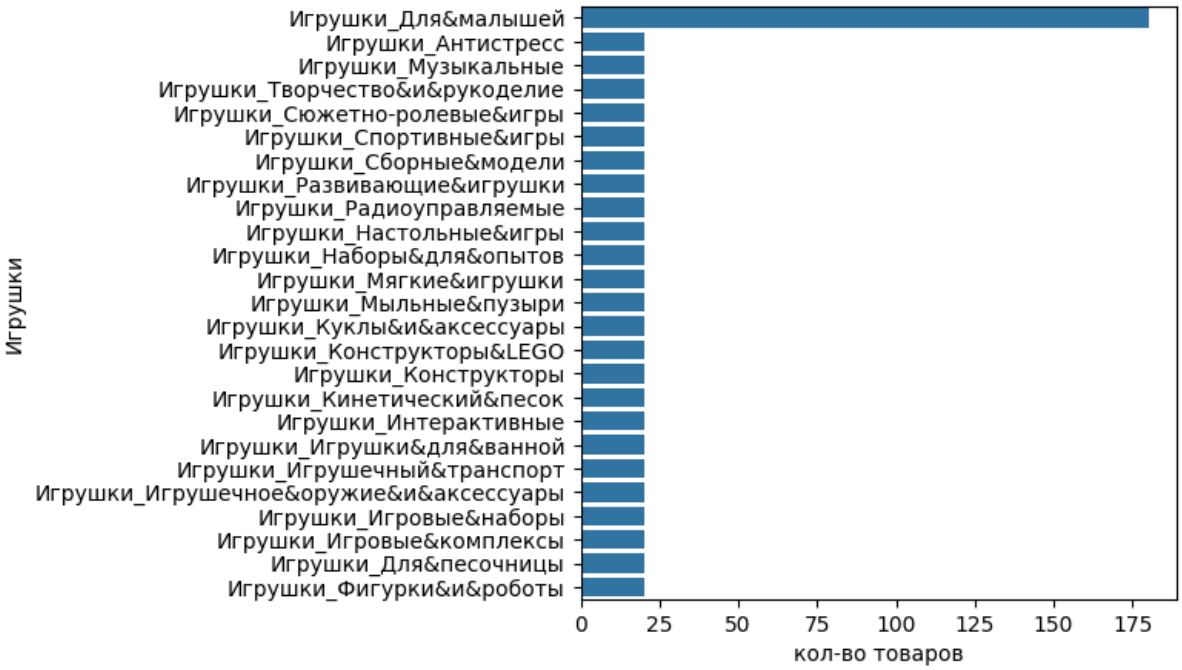

Рис. А.10: Распределение данных в категории «Игрушки».

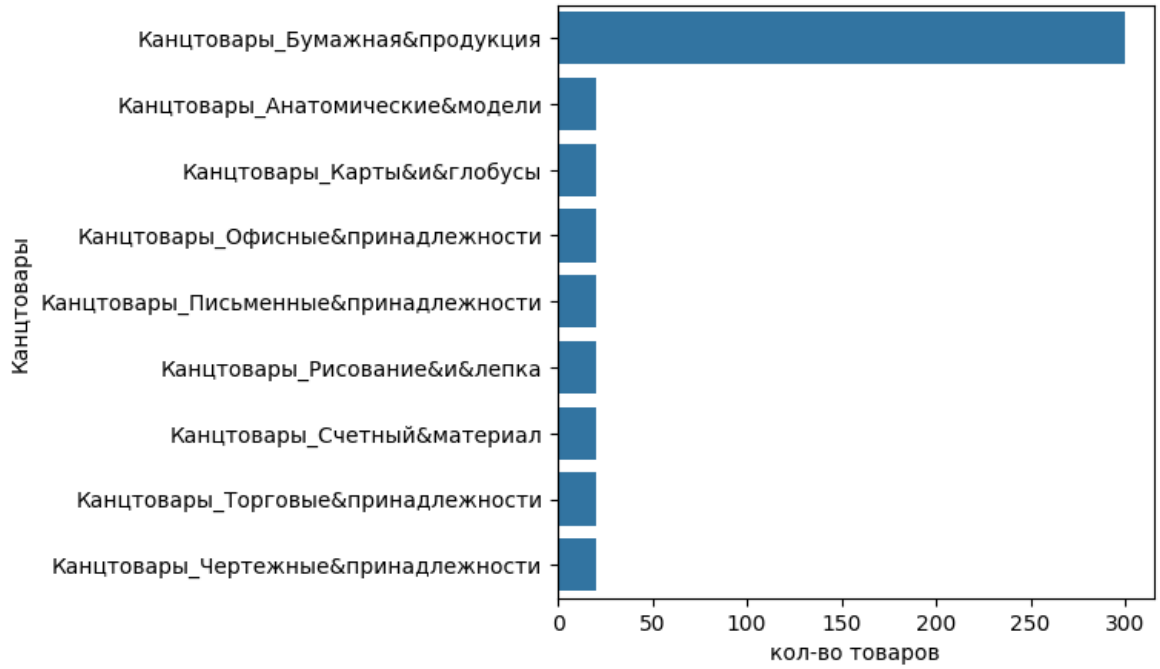

Рис. А.11: Распределение данных в категории «Канцтовары».

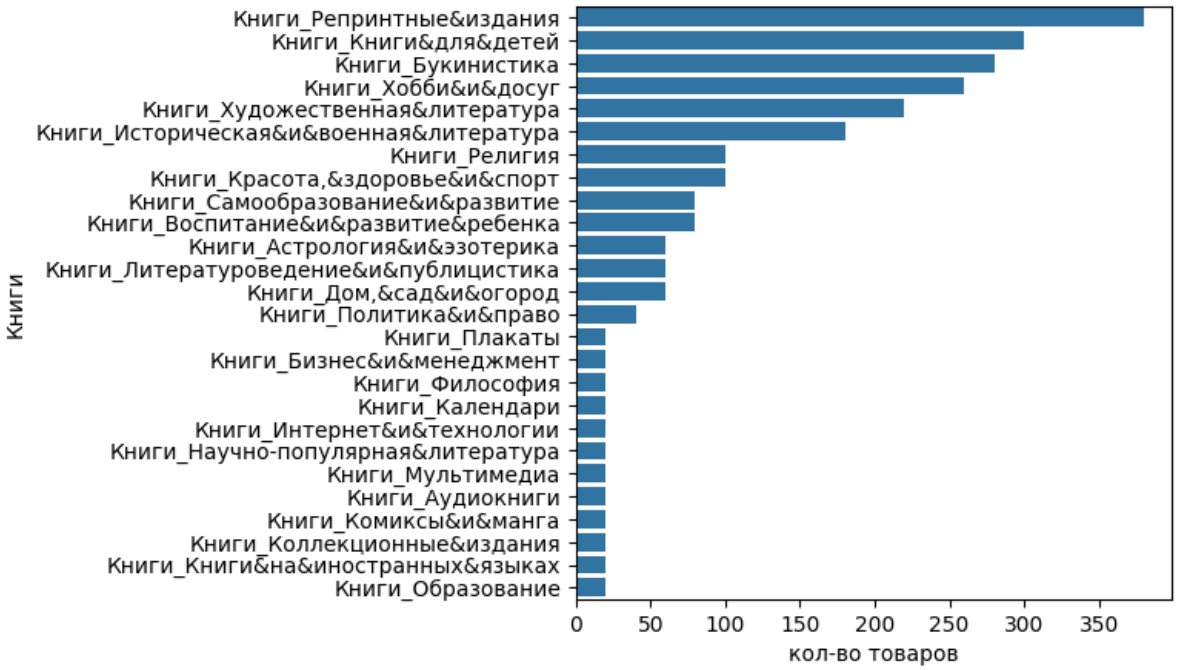

Рис. А.12: Распределение данных в категории «Книги».

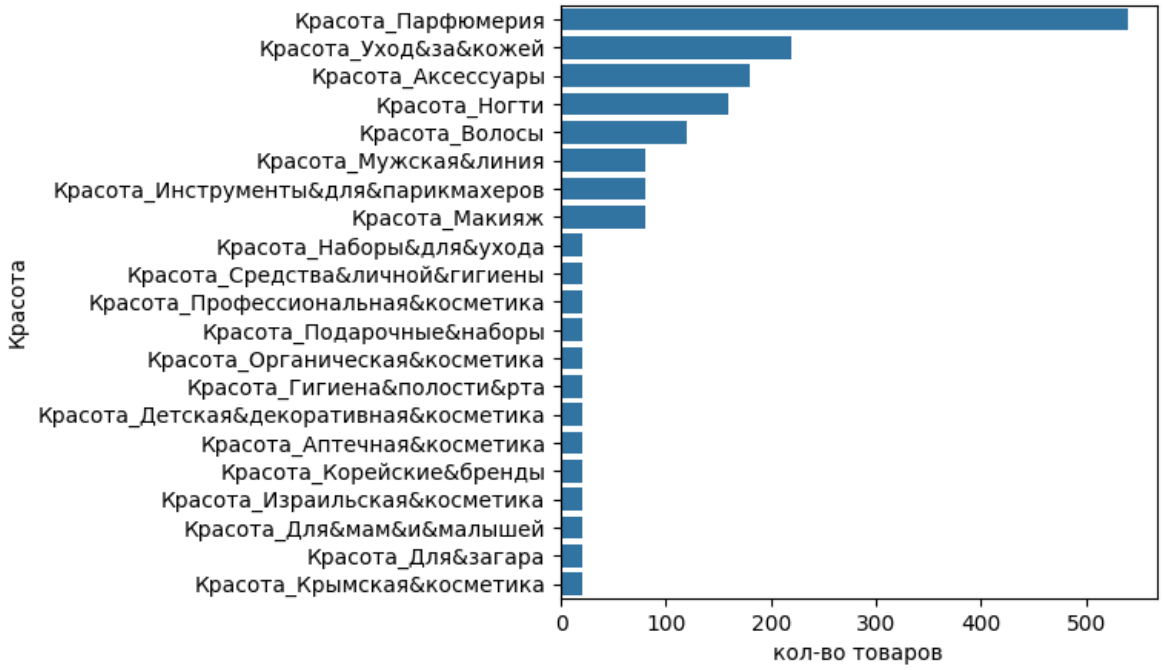

Рис. А.13: Распределение данных в категории «Красота».

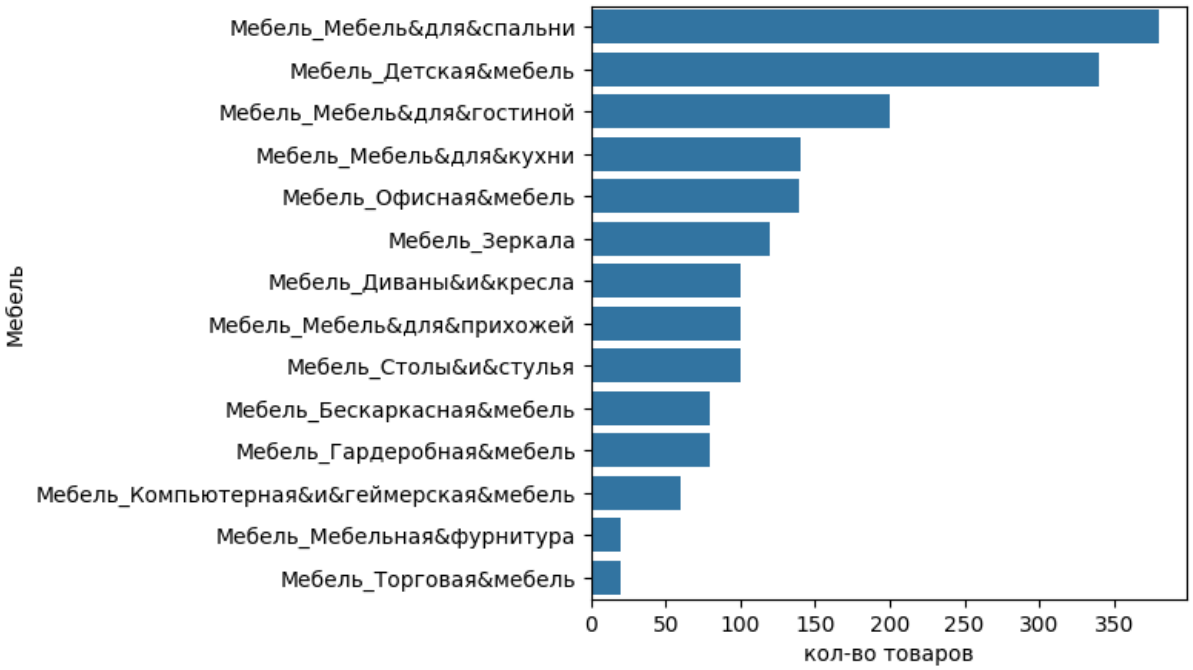

Рис. А.14: Распределение данных в категории «Мебель».

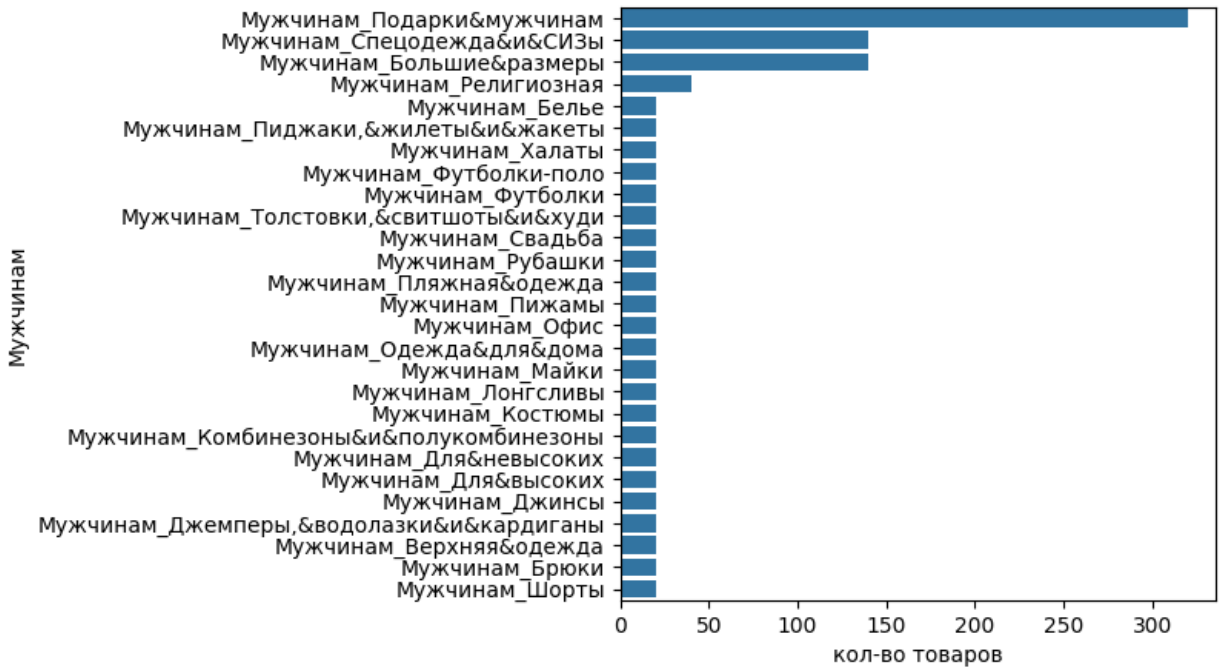

Рис. А.15: Распределение данных в категории «Мужчинам».

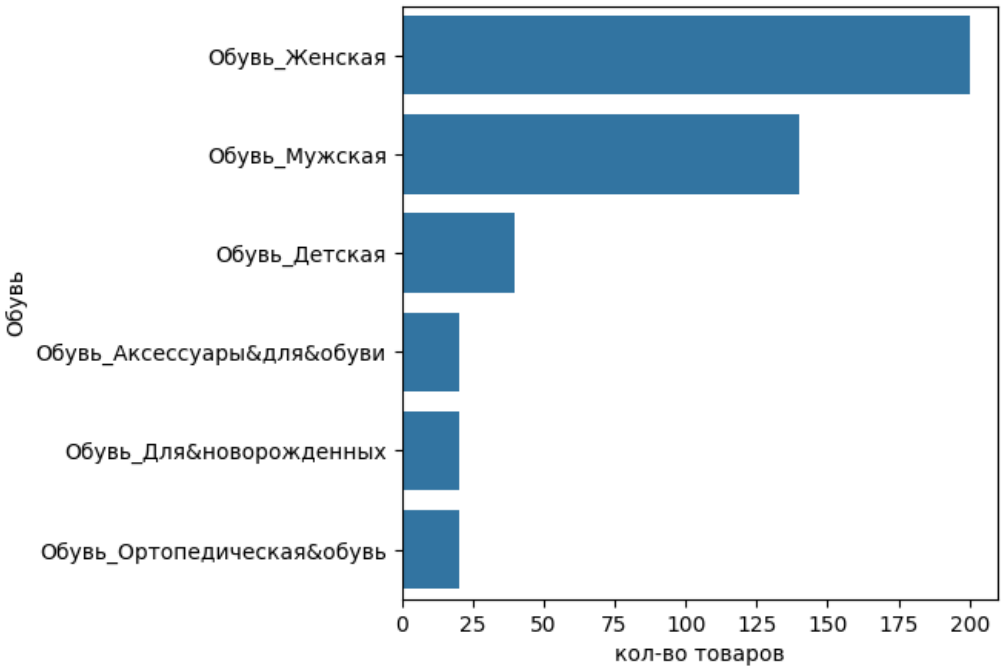

Рис. А.16: Распределение данных в категории «Обувь».

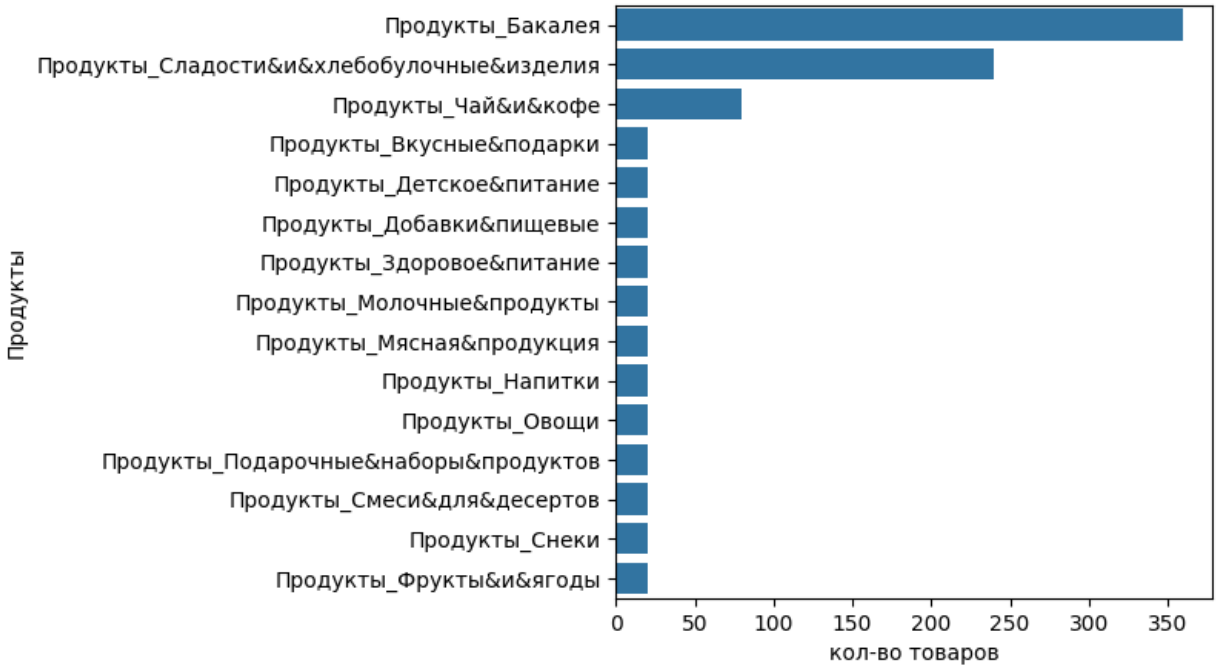

Рис. А.17: Распределение данных в категории «Продукты».

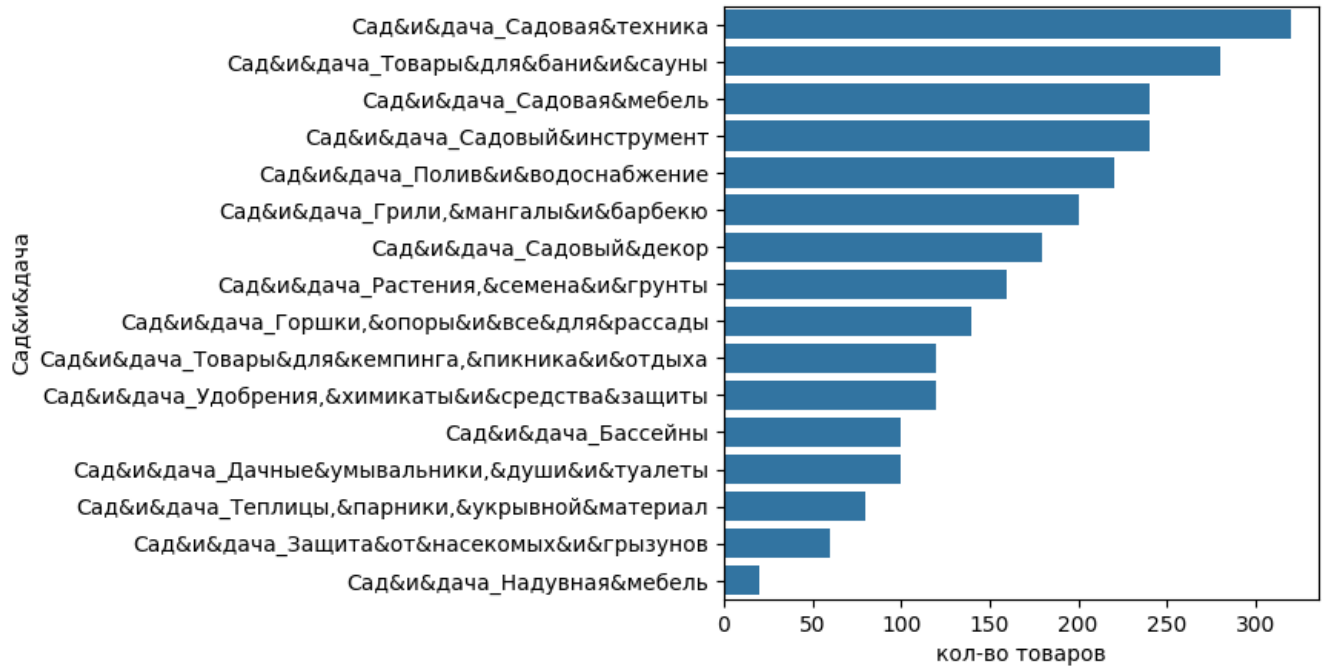

Рис. А.18: Распределение данных в категории «Сад&и&дача».

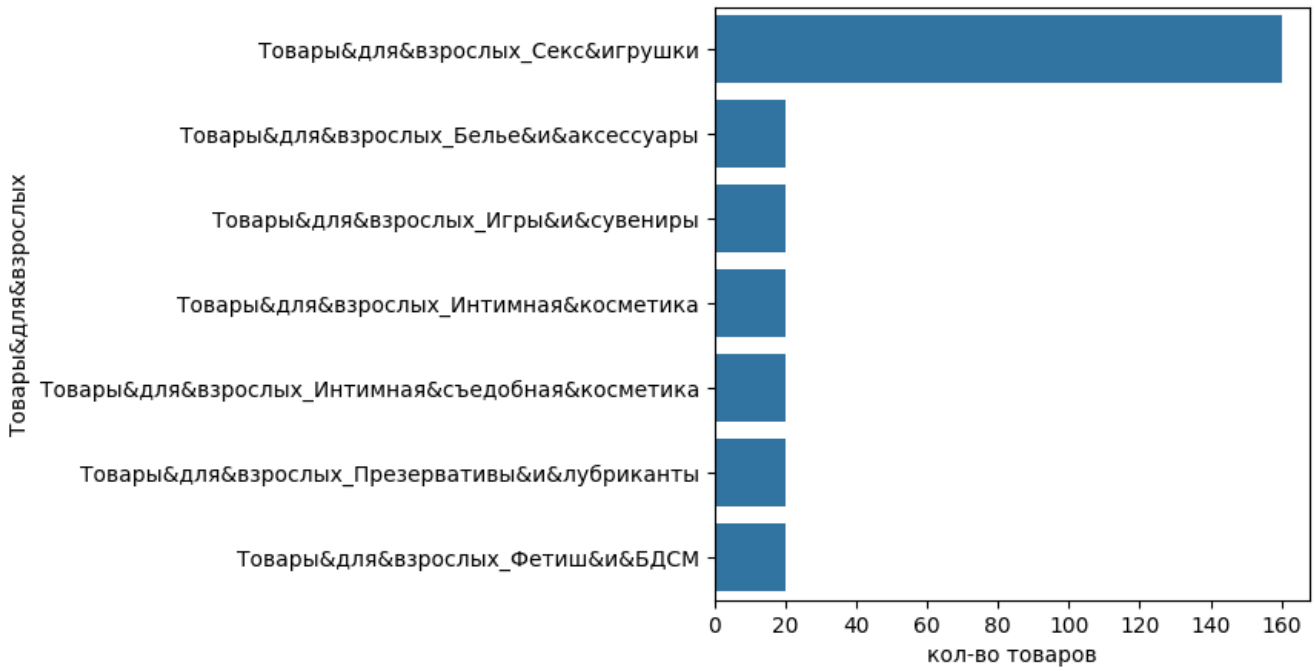

Рис. А.19: Распределение данных в категории «Товары&для&взрослых».

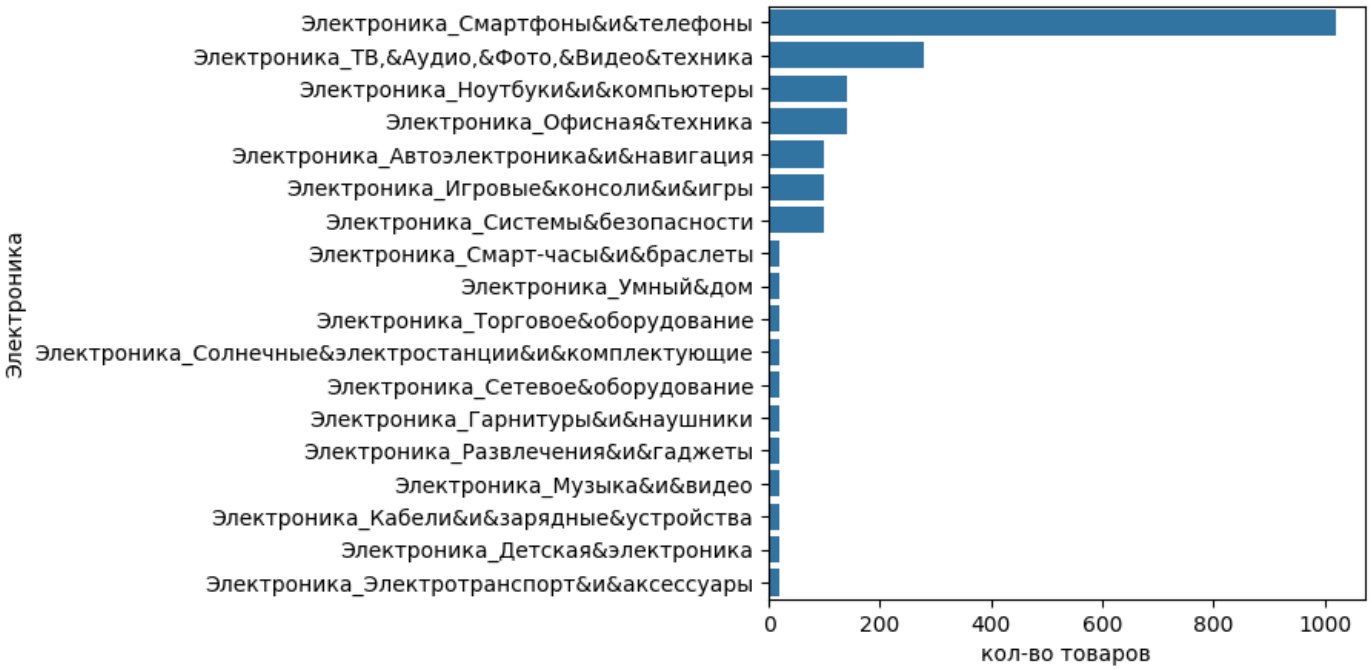

Рис. А.20: Распределение данных в категории «Электроника».

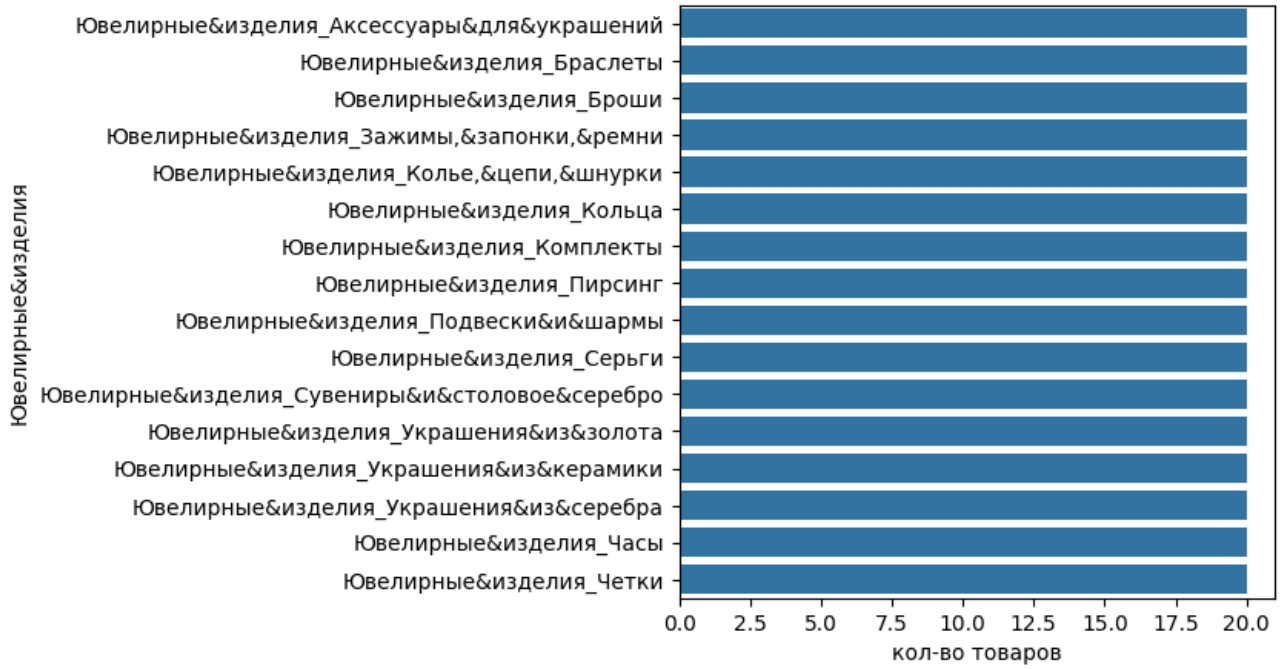

Рис. А.21: Распределение данных в категории «Ювелирные&изделия».