**Введение**

Компания Proanalytics занимается мониторингом цен на товары в интернет-магазинах, и предоставляет различные отчеты для магазинов конкурентов по собранным данным о товарах с различных сайтов. Для компании важно иметь собственный классификатор товаров, который на данный момент составляется вручную. Это представляет огромные затраты и большой объем рутинной работы. На данный момент существуют технологии, которые позволяют автоматизировать эту работу.

Данная выпускная квалификационная работа решает реальную проблему бизнеса. Цель этой работы — разработать систему, позволяющую гораздо быстрее классифицировать товары с минимальными затратами ресурсов.

Для достижения этой цели решено применить методы интеллектуального анализа, и были поставлены следующие задачи:

1) Изучить теоретические основы интеллектуального анализа данных, основные технологии и библиотеки для практического применения методов

2) Собрать и проанализировать данные

3) Выбрать метрику и способ оценки качества моделей

4) Построить и оценить различные модели

5) Внедрить модель в промышленное применение

В ходе разработке будут исследованы различные способы подготовки данных и методы машинного обучения.

Более того, данная работа может оказаться полезной не только для конкретной компании, поскольку в данной работе будет проведено исследование различных методов. Это может дать пользу последующим исследованиям в области обработки естественного языка.

Оглавление

1. **Бизнес-анализ**

2. Сбор данных

3. Подготовка и анализ данных

4. Моделирование

5. Оценка

6. Внедрение

7. Дальнейшее развитие

Оглавление

[1. Бизнес-анализ 3](#_Toc513028410)

[2. Сбор данных 5](#_Toc513028411)

[3. Анализ и подготовка собранных данных 8](#_Toc513028412)

[3.1. Поиск выбросов в наименовании товара 10](#_Toc513028413)

[3.2. Поиск выбросов в тексте категории товара 12](#_Toc513028414)

[3.3. Поиск выбросов в тексте характеристик 14](#_Toc513028415)

[3.4. Анализ каталога 18](#_Toc513028416)

[4. Построение моделей машинного обучения 21](#_Toc513028417)

[4.1. Постановка задачи машинного обучения 21](#_Toc513028418)

[4.3. Метрики качества классификации 22](#_Toc513028419)

# Бизнес-анализ

Компания ProAnalytics ежедневно собирает информацию о миллионах цен в день с различных сайтов — интернет-магазинов. В нее входит наименование цен, текст категории, изображение товара, характеристики, цена, описание, наличие. Для отдельных заказчиков может собираться дополнительно другая информация.

Клиентами компании являются вендоры либо интернет-магазины. Как правило они заказывают отчеты по ценам на товары конкурентов. Для многих отчетов требуется информация по товарам в разрезе определенных категорий. Проблема заключается в том, что у каждого интернет-магазина своя структура каталога товаров. Поскольку собирается информация с различных интернет-магазинов, собранные товары требуется классифицировать в единый каталог.

На данный момент эта классификация производится вручную работниками компании. Т.е. работники вручную привязывают по определенному списку некоторые товары к категориям, к которой эти товары относятся.

Целью является автоматизация этого процесса. Необходимо разработать программный продукт, который снизит затраты, и позволит более оперативно классифицировать товары, повышая довольство клиентов.

За период существования компании было привязано порядка миллиона товаров. Накопленные данные можно исследовать, и на их основе создать классификатор, который смог бы автоматически классифицировать товар.

Процесс ручной классификации выглядит следующим образом. Работник открывает специальную панель, на который выпадает список товаров, для которых необходимо определить категорию. Для них он выбирает нужную категорию. Для некоторых товаров автоматически предлагаются категории по ключевым словам, составленным вручную, и работник только выбирает «Да» или «Нет».

В процессе сбора данных было предложено дополнительно собирать нажатие кнопки «Нет». Т.е. при нажатии «Нет» в базе должна сохранятся запись, которая говорит о том, что конкретный товар не принадлежит конкретной категории. Эти данные позволят улучшать модель машинного обучения.

# Сбор данных

Первый этап разработки это просмотр и анализ имеющейся информации. А так же выбор нужной информации для использования в построении модели машинного обучения.

Вся информация по товарам хранится в единой базе данных. В качестве СУБД используется MySQL.

Был проведен опрос стейкхолдеров, для выяснения того, как собираются данные, в каких таблицах можно найти ценную информацию. Так же было необходимо выяснить насколько зашумленными могут быть данные. В итоге удалось выяснить следующее.

Всего накоплено 1770580 товаров привязанных к товару.

Данные, которые собираются с интернет-магазинов по товарам:

1. Наименование
2. Текст категории
3. Статус (В наличии / Нет в наличии)
4. Цена
5. Ссылка на изображение
6. Характеристики товаров
7. Описание
8. Артикул товара на сайте
9. Дата сбора
10. Ссылка на товар
11. Компания (интернет-магазин), с которой был собран товар
12. Бренд

У всех товаров собираются:

1. наименование
2. текст категории
3. статус
4. цена, ссылка на изображение
5. артикул товара на сайте
6. дата сбора
7. ссылка на товар
8. ссылка на изображение

И только для некоторых компаний собираются:

1. Характеристики товаров
2. Описание
3. Бренд

????

Здесь привести количество товаров с хар-ми

С описанием

С брендами

????

Данные о привязках верны, и без ошибок. Так утверждал стейкхолдер, но, как оказалось, этому утверждению нельзя доверять.

Для построения первой версии модели машинного обучения было решено использовать текстовую информацию, поскольку она наиболее характеризует категорию товара:

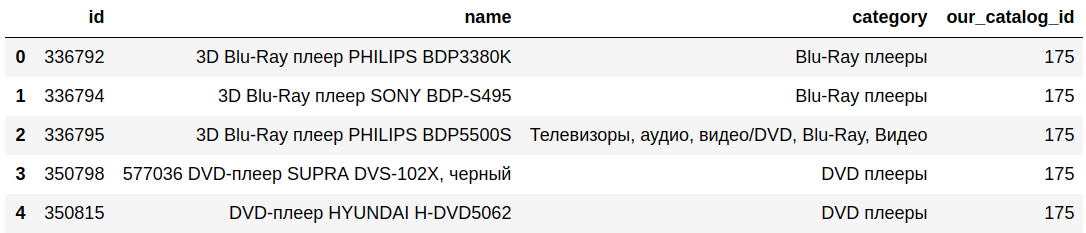
1. Наименование
2. Текст категории
3. Характеристики

Поскольку многие товары в обучающей выборе не имеют описания и бренда, эту информацию решено не включать в обучающую выборку.

Картинки так же можно использовать, но хорошее качество в классификации они способны показать лишь при использовании глубокого обучения, на многослойных нейронных сетях, которые способны генерировать высокоуровневые признаки из изображений. Но для их обучения требуется большое количество изображений и вычислительных ресурсов. Поэтому для начала было решено их не использовать.

# Анализ и подготовка собранных данных

Для начала посмотрим как выглядят наименования и категории товаров. Для этого выведем 5 товаров в виде таблицы.

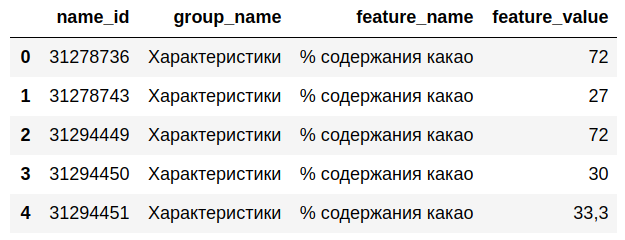


В таблице id — это идентификатор товара в базе, name — наименование, category — текст категории, our\_catalog\_id — идентификатор каталога.

Как мы видим наименования и текст категории содержат различный регистр символов, символы пунктуации, и прочие символы, которые необходимо заменить на пробел. Символы верхнего регистра необходимо перевести в нижний. Необходимо оставить только буквы, числа и пробелы. К тому же в тексте встречаются слова, состоящие только из чисел, такие слова тоже удалим.

Структура каталога имеет 4 уровня вложенности. Но товары относятся только к последнему 4-му уровню. Т.е. к первым трем уровням привязок товаров нет.

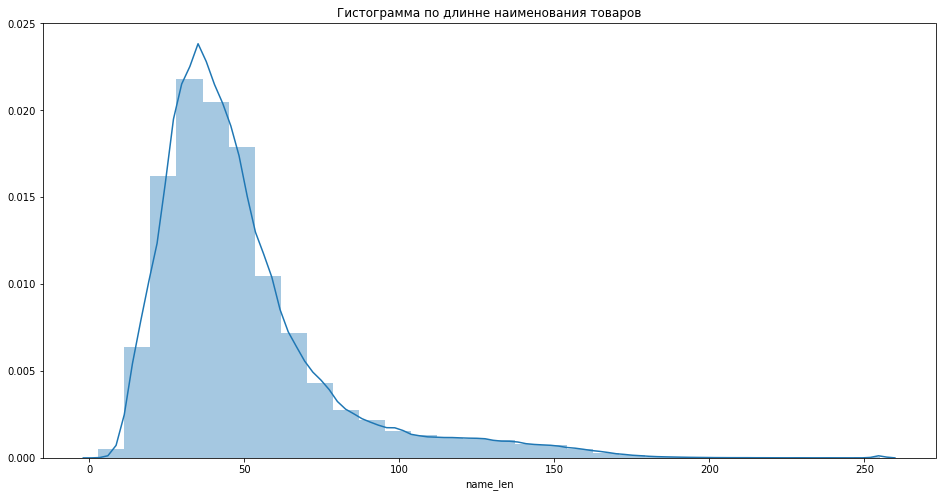
К тому же были собранны характеристики товаров. Эти характеристики иметь привязку один товар ко многим характеристикам. Т.е. несколько записей характеристик могут относиться к одному товару. Посмотрим записи собранных из базы характеристик.



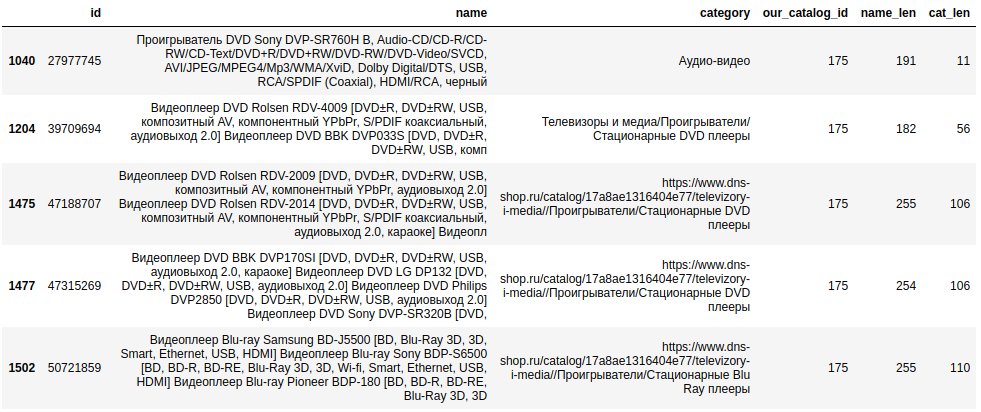
Здесь group\_name — это наименование группы характеристик. Характеристики могут быть сгруппированы. feature\_name — название характеристики, feature\_value — значение характеристики.

## 3.1. Поиск выбросов в наименовании товара

Для того чтобы найти выбросы, посмотрим наименования с слишком короткой длинной наименования и слишком длинной. Для этого для начала построим гистограмму, по которой можно увидеть распределение количества товаров из выборки по длине наименования.

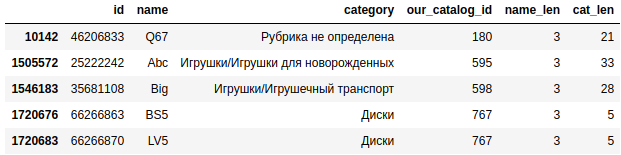


По гистограмме видно, что количество товаров, у которых длинна наименования меньше 180, резко уменьшается, практически до 0. Но в конце после 250 увеличивается. Это подозрительно. Посмотрим на товары у которых длинна наименования больше 180. Для примера возьмем 5 товаров.



Видно, что у этих товаров в наименовании присутствуют характеристики. К тому же у них присутствуют ошибки в тексте категории. Поскольку таких товаров немного, удалим их из выборки.

Далее посмотрим на слишком короткие наименования. Возьмем товары, с длинной наименования менее 4-х символов.

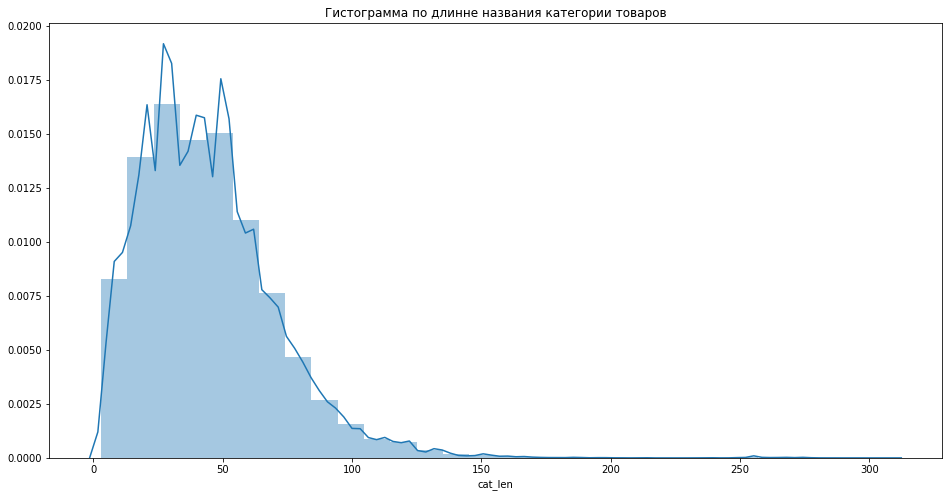


У этих товаров в наименовании только текст модели. Также удалим их, поскольку их не много, а текст модели никак не характеризует категорию.

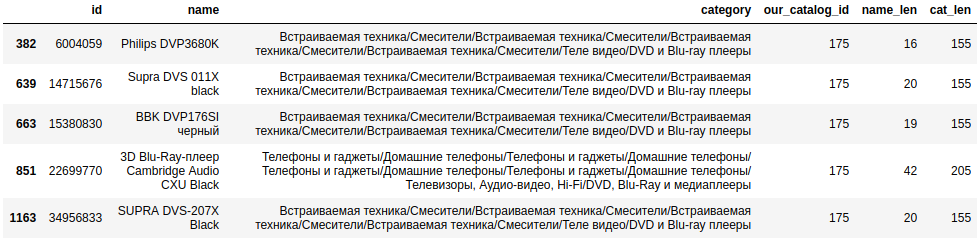
Так же мы видим, что у товаров может быть категория «Рубрика не определена». Для товаров где в тексте категории содержится «Рубрика не определена», будем оставлять пустой текст.

## 3.2. Поиск выбросов в тексте категории товара

Для поиска выбросов в тексте категории построим такую же гистограмму, как в пункте выше, только для текста категории.

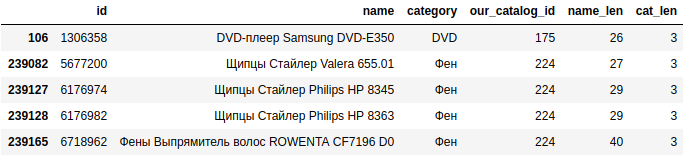


Видим что есть аномальные, по длине категории, товары, при длине больше 150 символов. Посмотрим на эти товары.



В этих товарах дубли категории. Удалим их, поскольку их также немного.

Посмотрим на слишком короткие категории, с длинной менее 4-х символов.

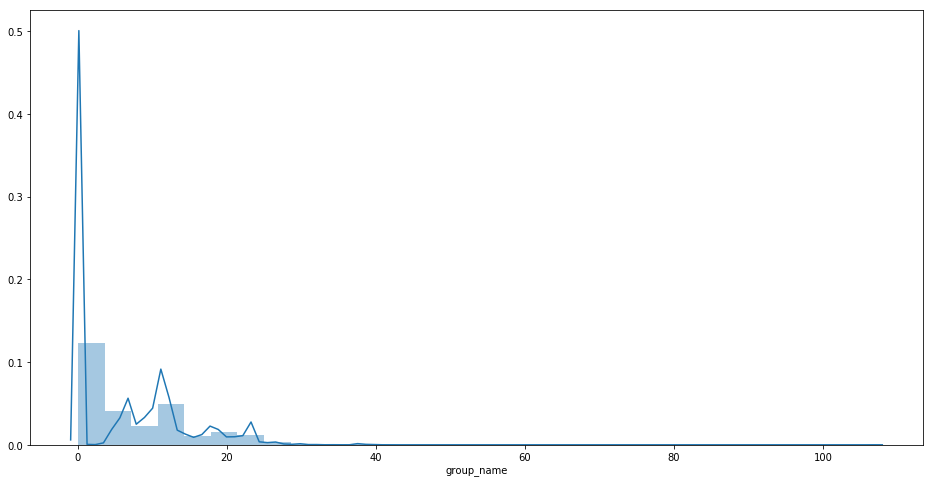


У этих товаров все нормально, поэтому их можно оставить.

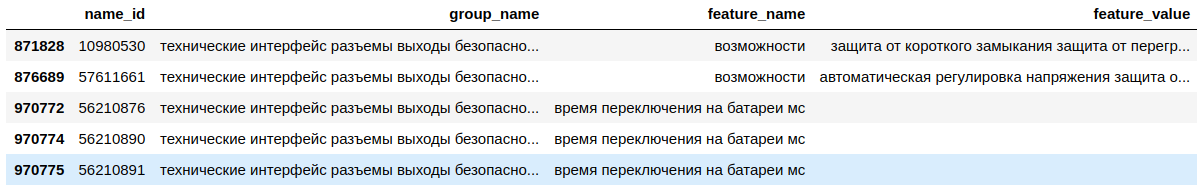
## 3.3. Поиск выбросов в тексте характеристик

Прежде чем искать выбросы, сразу удалим из характеристик различные спецсимволы, такие как «%», «-», «@», знаки пунктуации, и другие. Переведем все в нижний регистр. Также удалим слова которые состоят только из чисел.

Построим гистограммы по названию групп характеристик.



Для наименования групп аномальными являются характеристики с длинной более 40 символов.

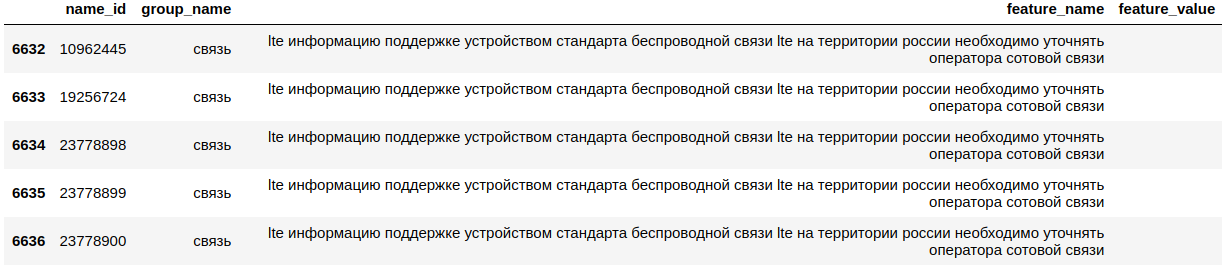


Среди этих товаров мы видим, что названия групп собраны неправильно, эти характеристики с ошибками, удалим их.

Построим гистограммы по названию характеристик.



Видим, что при длине более 50-ти символов начинаются аномалии. Посмотрим на эти характеристики.

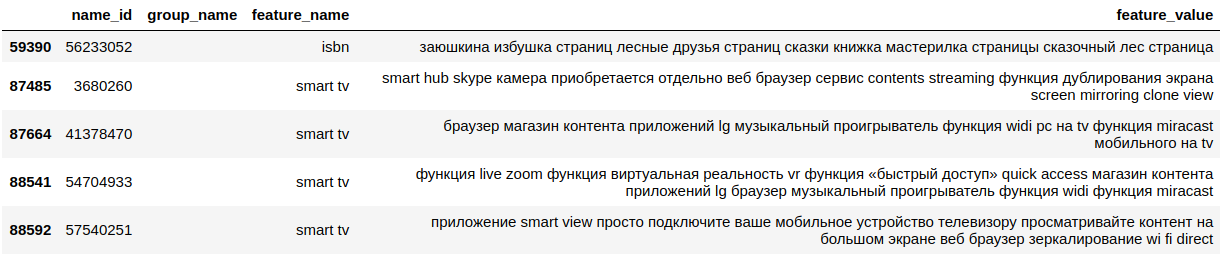


Видно, что у этих характеристик, название ошибочное. Эти характеристики были неправильно собранны с интернет-магазинов. Удалим эти характеристики.

Теперь посмотрим гистограмму по длине значения характеристик (поле feature\_value).

Аномальные характеристики, начинаются при длине более 40 символов. Посмотрим на них.





У этих товаров так же ошибки. Эти характеристики были собраны неправильно с интернет-магазинов.

Далее преобразуем все характеристики в мешок слов для каждого товара. Т.е. возьмем группу характеристик склеим её со всеми наименованиями и значениями характеристик.

Затем объединим мешки характеристик с наименованиями и посмотрим на полученную выборку.



## 3.4. Анализ каталога

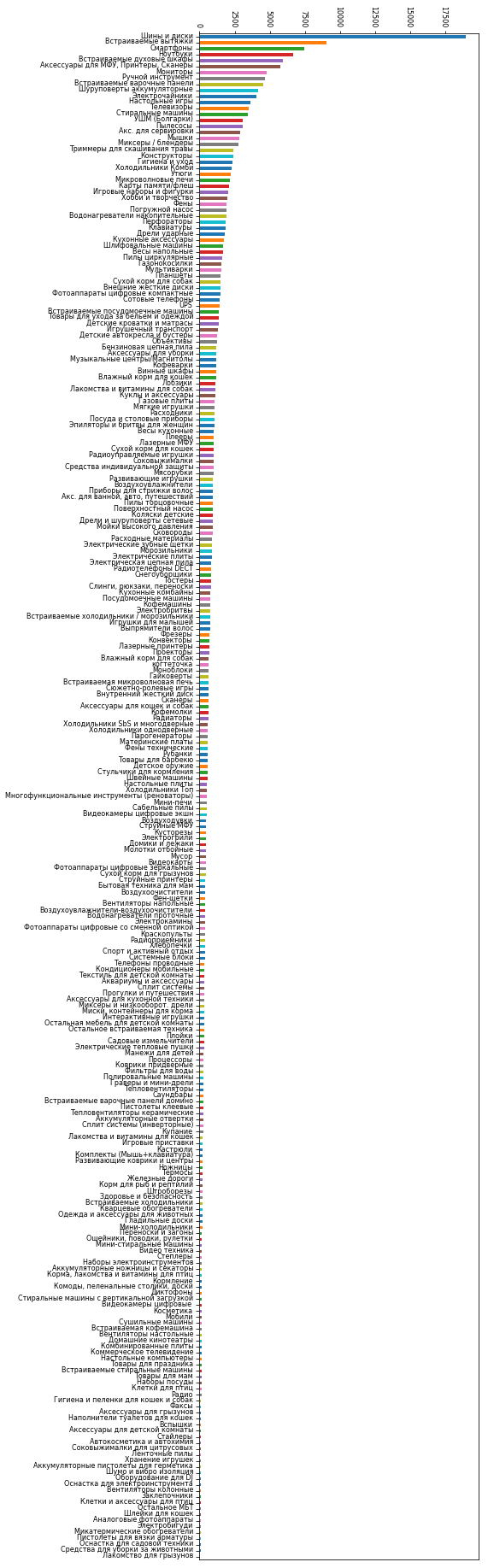
После визуального просмотра каталогов, было выявлено, что некоторые категории товаров сильно похожи друг на друга. Например категория 3-го уровня «Телевизоры» содержит подкатегории: «LCD/LED-телевизоры 15"-30"», «LCD/LED-телевизоры 32"-38"», «LCD/LED-телевизоры 39"-42"» и т. д. Т.е. эти категории содержат телевизоры с различной диаганалью. Было выявленно несколько аналагичных категорий. Эти категории трудоемко классифицировать методами машинного обучения. Поэтому было решено товары этих категорий классифицировать к 3-му уровню каталога. Соответственно, в обучающей выборке эти категории были заменены на категорию родителя.

После преобразования получилось всего 255 категорий в обучающей выборке.

Ниже представлен график количества категорий.

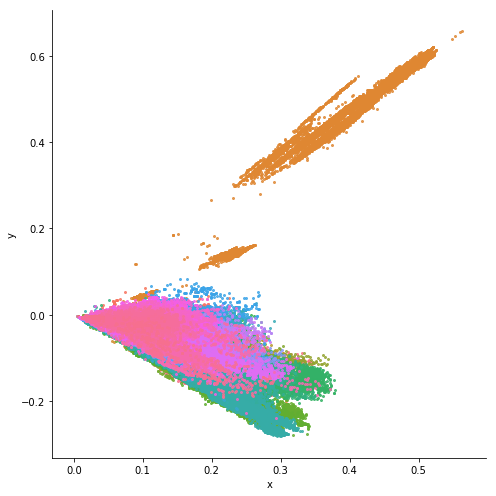
Из графика видно, что выборка несбалансированная. Товаров в категории «Шины и диск» гораздо больше чем остальных. К тому же, во многих категориях очень мало товаров, что может плохо сказаться на качестве классификации товаров. Для обучения таких категорий может просто не хватить данных.

Следующей проблемой являются пустые категории. Когда будет нужно классифицировать реальные товары, если попадутся товары из категорий, которых не было в обучающей выборке, то эти товары будут классифицироваться в категории, которые не были пустыми. Это проблема нецелостных данных. Например в обучающей выборке нет категории наушники. Но когда будет нужно классифицировать товары некоторого интернет-магазина, то там могут оказаться наушники, и тогда модель машинного обучения не сможет правильно определить его категорию.



Далее склеим наименование, текст категории и характеристики в одну строку, и преобразуем все признаковое пространство с помощью метрики TF-IDF.

Хотелось бы посмотреть на все объекты, в данном случае объекты — это товары, в двухмерном пространстве, и понять, есть ли какая лидо закономерность в классах. Можно ли их вообще отделить друг от друга с неким приемлимым качеством. Для этого воспользуемся сингулярным разложением, преобразуем выборку в двухмерное пространство.



На рисунке представлен график, где точками отмечены объекты. Для каждого категории используется свой цвет. Из графика видно, что некоторые категории хорошо отделяюся от других, некоторые плохо, но в целом есть некая закономерность. Этот график дает понять, что наши объекты можно классифицировать на различные категории с помощью машинного обучения.

# 4. Построение моделей машинного обучения

## 4.1. Постановка задачи машинного обучения

Требуется решать задачу многоклассовой классификации текстов. В качестве обучающей выборки используется множество товаров с информацией о наименовании товара, тексте категории и мешка слов из хактеристик товара. В качестве целевой метки будут использованы идентификаторы категории товара.

Для того чтобы преобразовать текстовую информацию в признаковое пространство, т.е. для того, чтобы каждый товар был представлен числовым вектором, будет применена мера TFIDF.

TFIDF - статистическая мера, используемая для оценки важности слова в контексте документа. Вес каждого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции. В качестве документа в данном случае представлена текстовая информация о товаре. В виде формулы она выглядит следующим образом:

здесь

где ni - число вхождений слова в документ. А в знаменателе общее число слов в документе

где |D| - количество документов в выборке, - количество документов, в которых встречается ti (когда ni≠0)

Большой вес в TFIDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Формальная постановка задачи будет выглядеть следующим образом.

Пусть X — множество описаний объектов, Y — конечное множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость — отображение , значения которой известны только на объектах конечной обучающей выборки

Требуется построить алгоритм , способный классифицировать произвольный объект .

Признаком называется отображение , где Df — множество допустимых значений признака. Если заданы признаки f1,…,fn, то вектор называется признаковым описанием объекта . В данном случае f1,…,fn  - меры TFIDF для каждого уникального слова в выборке по тексту товаров.

## 4.3. Метрики качества классификации

В качестве метрик классификиции были рассмотрены следующие:

* Доля верных ответов (Accuracy)
* Точность (Precision)
* Полнота (Recall)
* F – мера (F1)

Каждый из рассматриваемых метрик предназначен для бинарной классификации. Но в этой задаче 255 классов, и к тому же классы несбалансированны, поэтому будет рассматриваться взвешенная (по количеству объектов в классе) усредненная сумма оценок по каждому классу.

Доля верных ответов – самая простая метрика, представляет собой сумму верных ответов классификатора на общее число объектов в тестовой выборке. Она не подойдет для классификации из-за того, что плохо оценивает несбалансированную выборку.

Точность - это доля документов действительно принадлежащих данному классу относительно всех документов которые система отнесла к этому классу.

Полнота - это доля найденных классфикатором документов принадлежащих классу относительно всех документов этого класса в тестовой выборке.

Если представить ответы классификатора в виде матрицы, то матрица будет выглядеть следующим образом:



В таблице содержится информация сколько раз система приняла верное и сколько раз неверное решение по документам заданного класса. А именно:

TP — истино-положительное решение;

TN — истино-отрицательное решение;

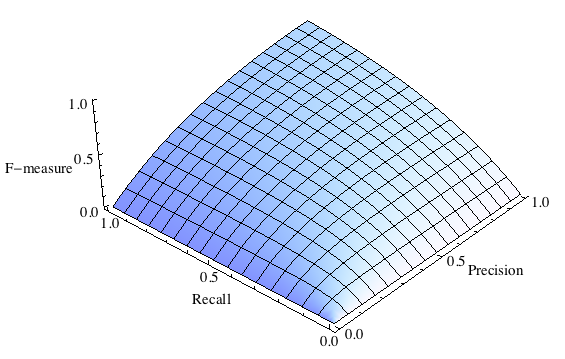
FP — ложно-положительное решение;

FN — ложно-отрицательное решение.

Тогда, точность и полнота определяются следующим образом:

Точность и полнота отражают разные качества классификатора. Высокая точность говорит о том что классификатор редко дает ошибочные положительные решения, но может пропускать верные ответы, и это никак не скажется на оценке. Высокая полнота наоборот, говорит о том, что классификатор пытается найти как можно больше верных положительных ответов, не никак не учитывает число ошибочных верных ответов. Т.е. максимальная полнота будет в случае положительного решения ко всем объектам.

Для верной оценки нужно как то обобщить эти 2 метрики. Что балансировать между этими двумя метриками существует F – мера. F-мера представляет собой гармоническое среднее между точностью и полнотой.



На рисунке представлен график зависимости F меры от точности и полноты. Видно что она стремится к нулю, если точность или полнота стремятся к нулю.

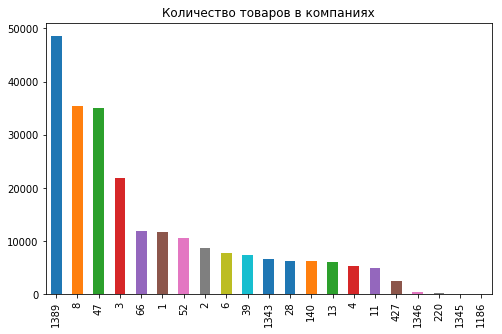
Для оценки классификации была выбрана F мера, поскольку она позволяет адекватно оценивать несбалансированную выборку.

## 4.4. Кросс-валидация

Для оценивания модели была применена кросс-валидация по блокам из компаний.

Кросс-валидация (Cross-validation) — она же «перекрёстная проверка» метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

В каждый блок были выбраны компании таким образом, чтобы в блок попадало как можно больше различных категорий, и чтобы блоки были приблизительно равного размера.



Если не проводить разбиение по блокам из компаний, а брать блоки из случайных объектов, то оценка может быть неправильной. Это может произойти из-за того, что если на обучение и тест попадут товары из одной компании и одной категории, то они могут иметь абсолютно одинаковый текст категории. А это не правильно, поскольку из-за этого модели могут обучиться некорректно, и переобучиться на тексте категорий.

На рисунке представлено количество товаров в компаниях. Для того чтобы не нарушать коммерческую тайну, компании заменены их идентификаторами. На рисунке видно, что есть компании с очень маленьким числом товаров в выборке, а есть наоборот.

## 4.5. Моделирование

Были рассмотрены следующие методы построения моделей машинного обучения:

1. Стохастический градиентный спуск с логистической функцией потерь
2. Метод опорных векторов
3. Градиентный бустинг над решающими деревьями