**Федеральное государственное автономное образовательное**

**учреждение высшего образования**

**«Национальный исследовательский университет**

**«Высшая школа экономики»**

**Факультет компьютерных наук**

**Основная образовательная программа**

**Прикладная математика и информатика**

**КУРСОВАЯ РАБОТА**

**на тему**

**«Задача кластеризации стандартизованных медицинских товаров на основе их текстового описания»**

**Выполнил студент группы БПМИ161, 3 курса,**

**Сумекенов Ахмад**

**Научный руководитель:**

**Доктор физико-математических наук,**

**профессор, Шаповал Александр Борисович**

**Консультанты:**

**Кандидат физико-математических наук, доцент, Ткаченко Андрей Викторович**

Table of Contents

[Абстракт 3](#_Toc9205437)

[Введение 3](#_Toc9205438)

[Подготовка данных 5](#_Toc9205439)

[Кластеризация товаров 8](#_Toc9205440)

[Исследования над кластерами 9](#_Toc9205441)

[Улучшения словаря 12](#_Toc9205442)

# Абстракт

# Введение

Успешная страна – эта страна, в которой государство эффективно выполняет свою роль. Государство выполняет много задач для функционирования общества и его совершенствования. Но есть те сферы жизни, где государство не может справиться в одиночку и должно сотрудничать с частным сектором, и в большинстве своем это государственные закупки. Государственные закупки затрагивают почти все аспекты общественности – инфраструктура строится за счет договора между государством и строительными компаниями, часто военная техника поставляется индустриальными частными компаниями. Так, например компания Boeing занимается не только построением гражданских самолетов, но также и созданием военной техники. Но государственные закупки также способствуют работоспособности одной из основных задач государства – здравоохранению.

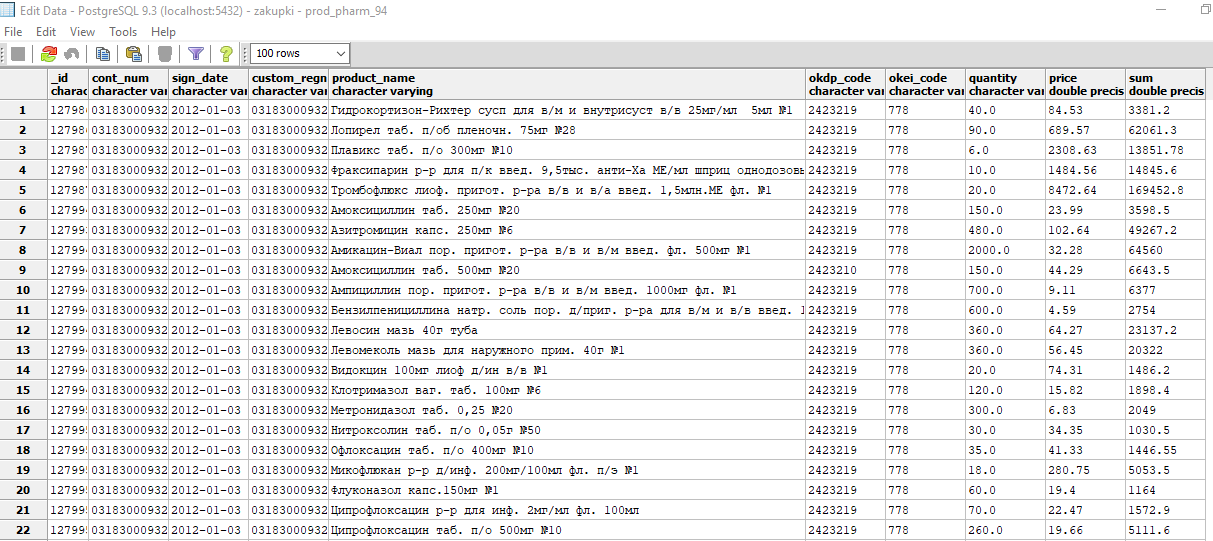
Чтобы местное правительство обеспечивало государственные больницы и другие медицинские учреждения лекарственными препаратами, они проводят конкурсы по покупке лекарств среди всех желающих поставщиков, так называемые тендеры. Основной критерий таких конкурсов, конечно же цена, по которой поставщик предлагает поставить требуемый товар. В нашем случае этот товар – медицинские лекарственные препараты. В разных регионах, городах и областях эти цены могут разниться – поэтому есть большой шанс того, что где-то закупки могут проводиться неэффективно, или наоборот, что в какой-то области есть малоизвестный поставщик, который готов поставлять товар по низким ценам. Естественно, это также зависит от самого тендера. Бюрократы, ответственные за проведение конкурса, могут злоупотреблять своим положением, таким образом закупая лекарства по завышенным и необоснованным ценам.

Нашей задачей является анализ государственных закупок в России по данных из государственного реестра. Сумев кластеризовать медицинские препараты, мы сможем сравнивать цены на одни и те же лекарства, купленные в разное время в разных местах или регионах. Таким образом мы проанализируем эффективность аукционов и сможем сделать выводы о том, какие аукционы были более эффективны и нашли более выгодного поставщика, а какие менее эффективны.

# Подготовка данных

Данные по государственным закупкам медицинских препаратов доступны на [ftp.zakupki.gov.ru](ftp://ftp.zakupki.gov.ru). На одном и том же аукционе могут быть проданы несколько лекарство разных видов и в разных количествах. В государственной базе данных лекарства разделены на классы по типу семизначного числа, которое означает целевое предназначение препарата, поэтому хоть эта классификация и помогает в разделении одних лекарств от других, она не решает проблемы кластеризации продуктов, так как лекарства могут быть от разных компаний и иметь разный состав.

Для первого этапа исследования я выбрал лекарства с кодом ОКДП 2423211-2423219, что означает лекарства, которые действуют « на центральную нервную систему». Всего лекарств с таким ОКДП насчиталось 127786 штук. Вот пример этой базы лекарств:



Данная база выглядит так: каждая строка — это покупка, и первый элемент в этой строке – ID покупателя. Следующим идет номер контракта, число контракта, код, полное название лекарства, код ОКДП, код ОКЕИ(Общероссийский классификатор единиц измерения), количество закупаемого препарата и его цена, а также сумма покупки(то есть количество, помноженное на цену). Самым главным образом нас интересуют только цена и название лекарства, однако другие показатели могут быть использованы косвенно.

Для анализа данных и последующей работы мы будем использовать среду, состоящую из языка программирования Python 3.6 и средства разработки Jupyter Notebook. Также мы будем использовать программные пакеты для Python. Об установке этой среды и более подробно об используемых пакетах будет рассказано в части «Приложение».

Чтобы работать с более структурированными данными, нам нужно извлечь из названия лекарства активное вещество, вид, то есть таблетка это или раствор, и массу. Для начала нужно вообще понять, как выглядят наши данные в Jupyter Notebook:



Наши данные – это DataFrame таблица, в которой есть столбцы “\_id”, “cont\_num”, “sign\_date”, “custom\_regnum”, “product\_name”, “okdp\_code”, “okei\_code”, “quantity”, “price”, “sum”. Для анализа этих данных и для наших целей нам не нужны столбцы “cont\_num”, “sign\_date”, “custom\_regnum”, “okdp\_code”, “okei\_code”, поэтому мы их удалим.

Для этого проделаем следующие шаги:

1. Переведем все буквы в строчные буквы.
2. Удалить символы № и вывести число после символа в отдельную колонку. Этот номер означает количество таблеток в упаковке.
3. Делаем «стоп-лист» слов, которые не несут значения: предлоги «для, в, на и т.д.».
4. Удаляем из названия лекарств все стоп-слова. Получившийся результат будет новой колонкой в таблице “name\_norm”.
5. Делаем словарь по типу «вид лекарства: слова, которые означают этот тип лекарства». Например: таблетки описываются словами «таб», «тб», «таблет», «капсулы», и т.д.. Растворы словами «рр», «раствор» и т.д. Так делаем для всех видов: таблетки, мази, растворы, порошок, капли и спреи. Точный маппинг будет таковым:
   1. 'tablet':['таблетки', 'таблетка', 'таб', 'таблет', 'тб', 'тбт', 'табл', 'драже', 'др', 'драж', 'капс', 'капсула', 'капсулы']
   2. 'cream':['мазь', 'мз', 'мзь', 'эмульс', 'эмульсия']
   3. 'powder':['пор', 'порошок']
   4. 'liquid':['рр','рас','раст', 'раств', 'раствор', 'концентрат', 'конц', 'сусп', 'сироп', 'сир']
   5. 'capsule':['амп', 'ампулы', 'ампула']
   6. 'spray': ['спрей', 'спр', 'аэр', 'аэрозоль', 'аэроз']
   7. 'dropley':['капли', 'каплях', 'капл', 'кап']

Ищем в названии каждого лекарства соответствующие слова и классифицируем тип лекарства, записываем в колонку ‘type’. Если ни одно такое слово в названии лекарства найдено не было, то в колонке ‘type’ будет None.

1. Как мы извлекаем вес: ищем в названии «мг» или «г», и переносим в колонку с весом то же число, если было «мг», то есть миллиграммы, и умножаем на 1000, если было «г», то есть граммы. Получившийся результат записываем в колонку ‘weight’
2. Удаляем из названия лекарства все символы пунктуации, тире, номера, числа, стоп-слова, также слова, обозначающие вид лекарства, и единицы измерения веса. Получившийся результат записываем в колонки ‘name\_del’ (список оставшихся слов) , а также ‘name\_del\_text’(то же самое, только в виде текста).
3. Создаем новый словарь. Теперь смотрим на каждую запись в данных, которые мы изменили: если ничего не осталось, мы удаляем эту запись, если что-то осталось, то мы токенизируем измененное название лекарства, и добавляем токены в словарь (если таких токенов еще нет в словаре). Получившийся словарь назовем med\_dic.

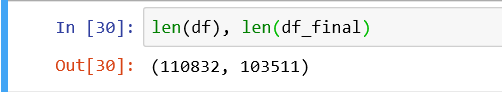
Получившийся словарь – словарь с названиями активных веществ.

# Кластеризация товаров

Как теперь можно кластеризовать товары? У нас есть медицинский словарь med\_dic, в котором имеются активные вещества. Будем относиться к этим активным веществам как к кластерам. Также, как я показал ранее, мы имеем колонку ‘name\_del’, из которых мы убрали единицы измерения, стоп-слова, тип лекарства. Остается логичным предположить, что то, что там осталось – это либо активное вещество, либо их список.

Поэтому мы отсекаем лекарства, в которых в колонке ‘name\_del’ не осталось ни одного слова, а в оставшихся ищем совпадения по словарю, и если находим совпадения, то классифицируем соответственно. Если в одном названии более одного активного вещества, то классифицируем по первому активному веществу по порядку.

В результате проведенных преобразований, по результату эксперимента нам удалось кластеризовать по активному веществу 103511 лекарственных препаратов из 110832 доступных.



# Исследования над кластерами

Первое, что мне интересно было узнать, это насколько и во сколько раз могут различаться цены на лекарства с одними и теми же активными веществами. Для этого я составил словарь (python dict()) с ключами, являющимися активными веществами, и значениями, являющимися отношением цены самого дорогого лекарства и самого дешевого лекарства с данным активном веществом.

Сначала я столкнулся с тем фактом, что в данных по ценам лекарств есть опечатки и просто бессмысленные данные. Отсортировав значения, я обнаружил такие значения:

1. ('галоперидол', -837.3280713091465)
2. ('феназепам', -837.3280713091465)
3. ('карбамазепин', -837.3280713091465)
4. ('неулептил', -837.3280713091465)
5. ('ламолеп', -837.3280713091465)]

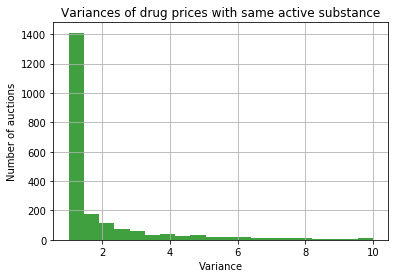
Очевидно, цены не могут быть отрицательными. Поэтому исследовав все записи с активным веществом «галоперидол», чтобы найти причину аномалий, я понял, что некоторые записи имеют опечатки. Так, была запись аукциона, где таблетки Галоперидола купили за минус 976 рублей. Чтобы не встречаться с такими данными, я удалил все записи, где цена являлась неположительной. Из 103511 записей таких «дефектных» оказалось только 123.

Проделаем аналогичные действия с данными, откуда мы убрали 123 записи с отрицательными ценами.

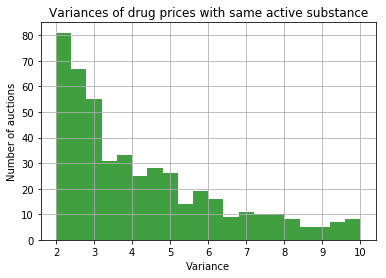
Оказалось, самые большие вариации наблюдаются с данными активными веществами:

1. ('действующих', 6369414.48),
2. ('медицинские', 10321471.23),
3. ('препаратов', 16039015.980392158),
4. ('медикаменты', 22405004.42),

Только дело в том, что это не активные вещества, а общие слова для лекарств, и попали они в med\_dic из-за не идеальности нашего метода составления словаря активных веществ. Ограничим наш список веществами, в которых разброс (наши отношения большая цена/меньшая цена) меньше 10. Из 2682 наименований в med\_dic, останется 2089. На них мы увидим такое распределение:



Как мы видим, в основном своем количестве большинстве, у лекарств цены разнятся не сильно. Интересно заметить, что если мы будем рассматривать кол-во лекарств с разбросом от 2 и больше, то зависимость становится более линейной:



Также заметим, что их намного меньше, чем тех, у которых разброс был в пределах 1.0 – 2.0 :

1. Разброс 1.0 – 2.0: 1621
2. Разброс 2.0 – 10.0: 468

# Улучшения словаря

В целях лучшего качества анализа данных по аукционам медицинских товаров, я решил улучшить наш словарь активных веществ med\_dic. У него есть несколько недостатков, из которых самых главных два:

1. Из-за отпечатков одно и то же название активного вещества могло быть добавлено в словарь активных веществ два и более раз.
2. Также из-за не идеальности метода туда могли попасть слова, которые не являются активными веществами, как мы видели ранее.

Чтобы решить первый недостаток, есть несколько методов. Наш недостаток является частью более общей проблемы «приближенного поиска совпадений строк» (*Approximate string matching, fuzzy pattern matching*). Один из способов называется «Locality-sensitive hashing» и его суть такова:

1. Хэшируем все строки такой хэш-функцией, что сильно похожие строки при хэшировании дают один и тот же результат, то есть попадают в одну и ту же «корзину»
2. Чтобы найти отпечатки, смотрим слова в одной «корзине», и каким-то способом, который устанавливает конкретный алгоритм, ищем отпечатки.

Вторым способом является использование дистанции Левенштейна:

***Определение 1.*** Расстояние Левенштейна (также редакционное расстояние или дистанция редактирования) между двумя строками в теории информации и компьютерной лингвистике — это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Например, чтобы превратить слово «строка» в слово «собака» надо «т» заменить на «о», «р» на «б» и «о» на «а», вставок и удалений не требуется — таким образом расстояние Левенштейна между этими двумя словами равно 3.

Расстояние Левенштейна и его обобщения активно применяется:

* для исправления ошибок в слове (в поисковых системах, базах данных, при вводе текста, при автоматическом распознавании отсканированого текста или речи).
* для сравнения текстовых файлов утилитой diff и ей подобными. Здесь роль «символов» играют строки, а роль «строк» — файлы.
* в биоинформатике для сравнения генов, хромосом и белков.

Источник: [<https://planetcalc.ru/1721/>]  
В нашей задаче мы будем использовать функция Левенштейна в следующем порядке:

1. У нас есть словарь med\_dic, в котором мы имеем 2682 наименования.
2. Сравним попарно все эти наименования: если дистанция Левенштейна между ними меньше либо равна 1, то будем считать это за опечатку.
3. При обнаружении опечатки мы удаляем это одно наименование из пары из med\_dic, таким образом оставляя только одно уникальное наименование.

При реализации данного подхода я столкнулся с несколькими проблемами. Также для наблюдения я посчитал количество опечаток для каждого активного вещества и составил список самых встречающихся опечаток по их частоте:

1. 'пирацетам', 12 опечаток
2. 'амитриптилин', 10 опечаток
3. 'феназепам', 8 опечаток
4. 'церебролизин', 8 опечаток
5. 'винпоцетин', 7 опечаток

Изначально, в med\_dic было 2682 наименования. После того, как мы убрали опечатки, осталось только 2141 наименования, то есть всего 541 опечаток, что является большой частью нашего словаря.

Второй недостаток нашего метода создания словаря med\_dic является тем фактом, что туда могли быть добавлены слова, не означающие активных веществ. Один из способов решить данную проблему это проверка на то, хранится ли данное слово в каком-либо известном корпусе русского языка.

***Определение 2.*** В лингвистике **корпус** — подобранная и обработанная по определённым правилам совокупность текстов, используемых в качестве базы для исследования языка. Они используются для статистического анализа и проверки статистических гипотез, подтверждения лингвистических правил в данном языке.

Логика заключается в том, что обычно в таких корпусах нет узкоспециализированных слов, как названия активных веществ медицинских препаратов. Так как в каждом корпусе есть функция проверки на то, хранится ли данное слово в корпусе, мы будем проверять слова в словаре med\_dic на принадлежность к корпусу русского языка. Если слово принадлежит корпусу, скорее всего это не активное вещество.

Существует множество корпусов русского языка, перечислим несколько из них:

1. Национальный Корпус Русского Языка – был разработан лингвистами и учеными из Москвы, Санкт-Петербурга и Воронежа как образовательный и научный проект. Один из самых первых в открытом доступе. Поддерживается крупной Интернет-компанией «Яндекс». Владелец – Институт русского языка имени В. В. Виноградова РАН. Размер – около 270 миллионов слов.
2. Генеральный Интернет-Корпус Русского Языка ГИКРЯ «Тайга» - корпус, собранный из российского сегмента Интернета. Согласно веб-сайту корпуса: «Генеральный Интернет-корпус Русского Языка (ГИКРЯ) – мегакорпус (более 15 млрд. слов), созданный при помощи полностью автоматической технологии сбора и разметки текстов из Рунета и основанный на современных достижениях компьютерной лингвистики.» Поддерживается крупной технологической компанией «ABBYY». Разрабатывается и поддерживается студентами и сотрудниками РГГУ, МФТИ, МГУ. Размер – почти 5 миллиардов слов.
3. Русская Википедия – корпус, основанный на русских статьях Википедии. Обычно для исследовательских целей совмещается с НКРЯ и используется дальше.

Итак, среди этих трех корпусов нам больше всего подходит НКРЯ, потому что там меньше всего вероятность обнаружить названия активных веществ медицинских препаратов.