**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

[Введение 3](#_Toc7814547)

[Подготовка данных 5](#_Toc7814548)

# 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.

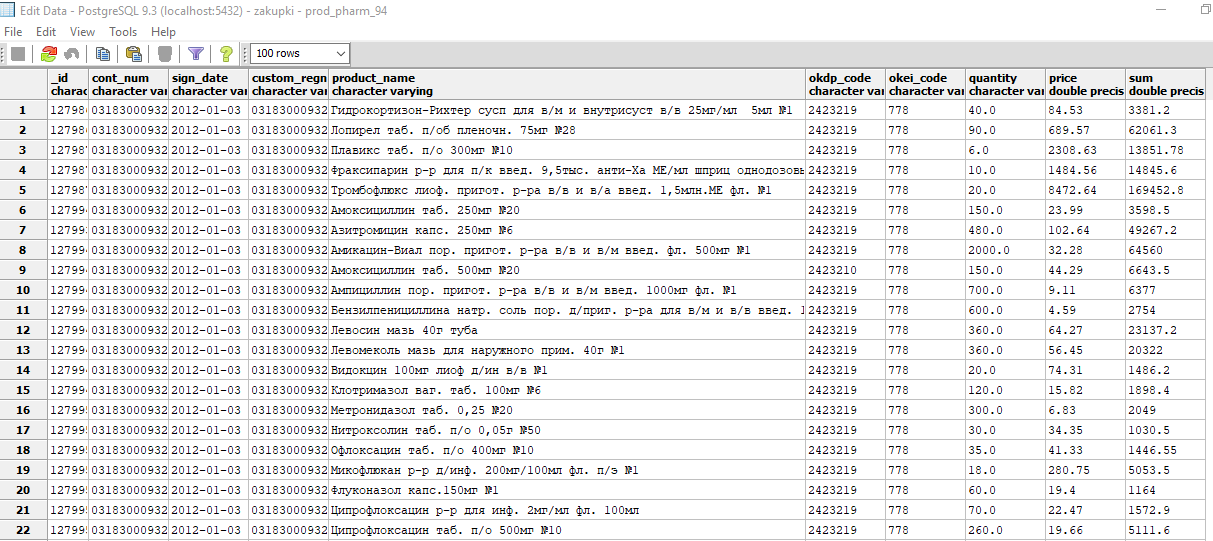
Чтобы местное правительство обеспечивало государственные больницы и другие медицинские учреждения лекарственными препаратами, они проводят конкурсы по покупке лекарств среди всех желающих поставщиков, так называемые тендеры. Основной критерий таких конкурсов, конечно же цена, по которой поставщик предлагает 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.

Чтобы работать с более структурированными данными, нам нужно извлечь из названия лекарства активное вещество, вид, то есть таблетка это или раствор, и массу. 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.
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.
6. 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.
7. Удаляем из названия лекарства все символы пунктуации, тире, номера, числа, стоп-слова, также слова, обозначающие вид лекарства, и единицы измерения веса.
8. 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.

Получившийся словарь – словарь с названиями активных веществ.