SULEYMAN DEMIREL UNIVERSITY

**ENGINEERING FACULTY**

**CSS 305 - Database Management Systems 2**

# Project documentation

**Made by: Ryspekov Ansar**

**Ismailova Dayana**

### 

### Kaskelen, 2020

**Structure:**

* Abstract
* Introduction
* Aim and objectives of research
* Background Review
* Methods and Materials
* Data and Results
* Discussion
* Conclusion
* References
* ***Abstract***

We live at a time where the “Coronavirus (COVID-19)” pandemic is shaping the economy, politics, social behavior, and the overall living of many people around the globe. The Republic of Kazakhstan was not the exception when our people went through this crisis, hoping for change. While the vaccine is developing by multiple countries and is passing through WHO’s rules and regulations, countless groups and companies offered their solutions to everyday problems of social distancing and lack of infrastructure or/and resources to face the pandemic. One of the people’s concerns was access to medicine. Some people bought a lot of penicillin and other necessities, which is added to the overall hype and created a shortage with big queues at drug stores.[[0]](https://rus.azattyq.org/a/kazakhstan-lack-of-medicaments-in-pharmacies-while-covid-19-cases-continue-to-increase/30701133.html)

***Key Terms:*** medicine, Oracle Database

* ***Introduction***

People need a solution where they can find medicines or share them when someone else is in need. All bought medicines need to be properly utilized or shared with others because there are ecological concerns over product consumption culture[[1]](https://recyclemag.ru/article/ekoinstruktsiya-delat-nenuzhnimi-prosrochennimi-lekarstvami).

We thought of a web application where people could find medicine in drug stores and also share their medicines.

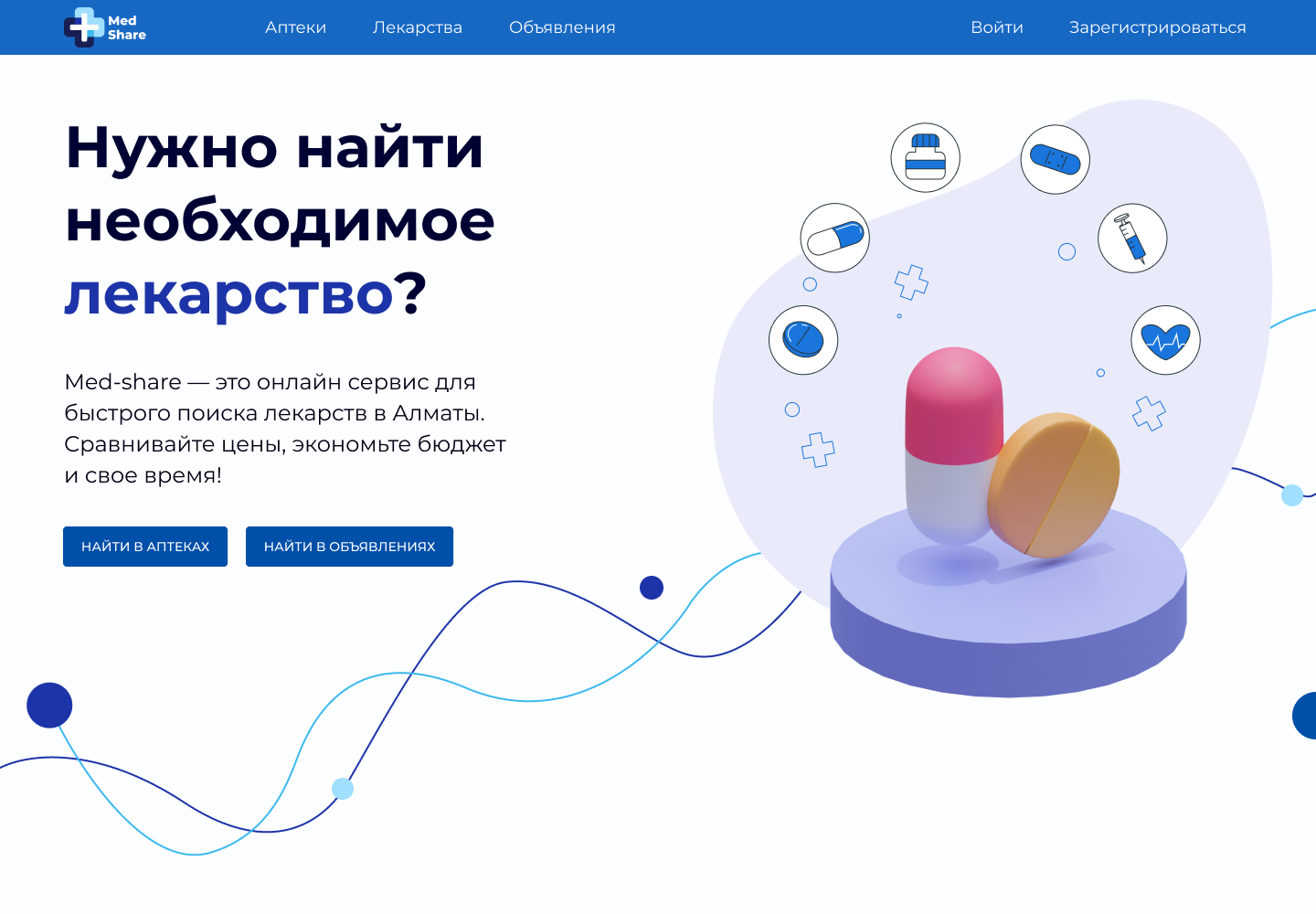


Figure 1. WEB APP. Main page

* ***Aim and objectives of research***

***Aim:*** To develop a web application that can help people find medicines and share their own with others.

***Objectives:***

Collect data on list of medicines

Build a web application

* ***Background review***

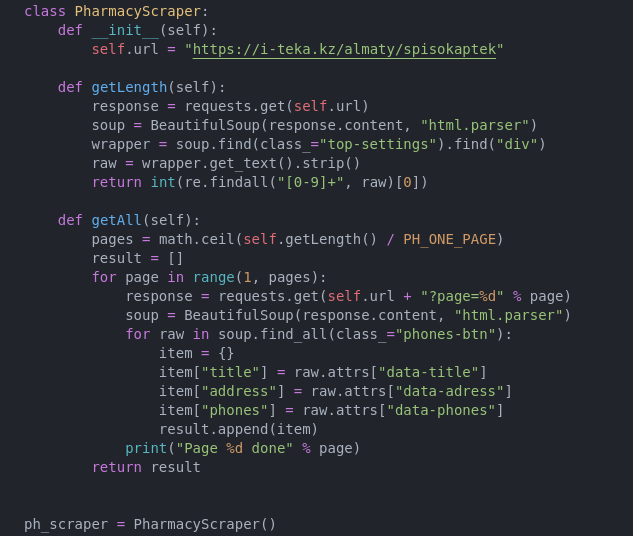
The list of medicines and drug stores they are contained in is found on multiple sources. We have chosen i-teka.kz[[2]](https://i-teka.kz/). It contains information about medicines, drug stores, their names, statuses and etc

* ***Methods and Materials***

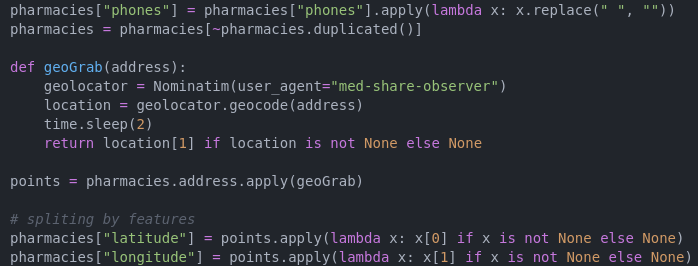
Unfortunately, we have not found any publicly available API that we could work with, therefore web scraping was our final solution for obtaining the actual data. We used python for its simplicity and libraries like Beautiful Soup and GeoPy. We needed GeoPy to find the exact coordinates of each drug store in Almaty since the source pages are provided with only the map.

Web Scraping is the process of obtaining the data from generated web pages by crawling through DOM and parsing the text returned. Beautiful Soup has a simple and straightforward API to work with. Since Beautiful Soup accepts HTML as a text and is not making any request itself, we used *the requests* python library to make a series of HTTP GET requests and pass the response to the soup object later on. But before getting our hands dirty, we opened the pages ourselves to inspect the generated DOM. By doing this we defined the steps to be done and what exactly target nodes with values. After passing the response and creating the Beautiful Soup object we are able to use its methods.

The type of data we need is drug stores, medicines, and the information joining those two: medicine availability in the particular drug store. We will start with drug stores.



We define the class that returns all the stores using the getAll() method. Information is contained in the multiple pages (48 drug stores each), so we first obtain the total number of available stores and then start a loop for each page. In the loop we make an HTTP GET request, create a soup object, and then start the inner loop that runs through all nodes containing the class *phones-btn* that each drug store container has. All the info we need is assigned to that node as an attribute, therefore we use *attrs* node property to obtain it. Each drug store is stored in the resulting array and returned from the method. Next is the cleaning and searching coordinates.



For the cleaning part, we need to remove all the whitespaces and also duplicated rows. We define the geoGrab function that uses GeoPy’s Nominatim class. It returns the coordinates (latitude and longitude) of the given address, we also delay the result since google API can classify the series of requests from the same source as a DDoS attack. Next are medicines.



There are approximately 16000 medicines on i-teka, but each page is loading too slow. To speed up the process we used a multiprocessing python library and parallelized tasks. Each task is grabbing a single page and returning all the medicines from it. In order to get information about drug stores we could find a particular medicine in, we also store its URL to run a third scraper operation.

The third operation was in fact the largest one. It is supposed to describe the relation between drug stores and medicines and in particular the list of drug stores where people could find the medicine in. It was the largest since in order to get information about each medicine’s list we needed to go by each of its own pages using previously saved URLs, which means we make requests for each ~16000 medicines. There were a number of troubles we encountered like the inconsistency of DOM structure from page to page (which was discovered at the near end of scraping), each full run went from 30 minutes to an hour to make, python specific issues with shared data across multiple processes, and etc. (We do not include the screenshot of the code piece since it would fit the whole page alone in order to save text quality and readability).

The last dataset we made was not scraped but built from what we had at that point. It describes the relation between medicines and their alternatives since some of them are having those.

* ***Data and Results***

We runned a sequence of scrapers that is chained together and obtained useful info to place in our project.

Pharmacies (467):

key - index, primary key

title - Drug store name, could be repeated in dataset (it has branches)

phones - Working phones separated by comma

latitude, longitude - geospatial coordinates of the drug store

Drugs (16 353):

key - index, primary key

title - Medicine title

status - Medicine’s recipe status

Drugs to Ph. (112 668):

key - index, primary key

pharmacy-id - Primary key of a pharmacy, the drug is located in

drug-id - Primary key of drug

price - The price of the drug in the specific pharmacy (if defined)

Drugs Analogs (34 947):

key - index, primary key

drug-id - Primary key of drug

analog-id - Primary key of analogous drug

We thought that that much data could potentially show us some statistics in it.



This graph shows the number of medicines not in stock in any drug store in Almaty

* ***Discussion***

The study has found very helpful data for building such systems as ours. It shows how the medicines are spread across the city and if added with more useful information it could be used to learn trends about medicine consumption and drug stores' performance.

* ***Conclusion***

Overall, such systems could be easily the most impactful in the day to day lives or even hard cases such as a global pandemic. We really hope that with proper resources and skills, our country will be able to implement much harder systems to dive into the world of open data that could be used for great, like learning people’s behaviors and correcting social policies and institutes towards the biggest future.

* ***References***

[0]<https://rus.azattyq.org/a/kazakhstan-lack-of-medicaments-in-pharmacies-while-covid-19-cases-continue-to-increase/30701133.html>

[1]<https://recyclemag.ru/article/ekoinstruktsiya-delat-nenuzhnimi-prosrochennimi-lekarstvami>

[2]<https://i-teka.kz/>