**IDS-560, Analytics Strategy Practice**

**Fall 2017**

**AHA SEEKER** *(KNOWING BETTER THAN EVER)*

**AMERICAN HOSPITAL ASSOCIATION**

**155 N UPPER WACKER DR**

**CHICAGO, IL 60606**

**Spurthi Bollina**

**Shubham Sirothia**

**Ashuthosh Gowda**

**Gaurav Chowdhury**

**Faculty Sponsor: Prof. Kyle Cheek**

**The University of Illinois at Chicago**

**INTRODUCTION**

American Hospital Association was originally formed in 1899 and formally became AHA in 1918. AHA is a non-for profit organization that is the voice of the hospital of the 6,500 hospitals in the US 4,500 of them are AHA members. AHA’s goal is to advance the health of individuals and communities. The AHA leads, represents and serves hospitals, health systems and other related organizations that are accountable to the community and committed to health improvement.

AHA ensures that the members' perspectives and needs are heard and addressed in national health policy development, legislative and regulatory debates, and judicial matters. Advocates the hospital legislative and executive branches in regulatory areas.

Health Forum is a strategic business enterprise of AHA which develops and delivers information and innovation services to help health care leaders achieve organizational performance excellence and sustainability.

**PROJECT PROPOSAL as of Oct 23rd, 2017**

**Primary Objective:**

AHA would like to know their US member hospitals which are 6000 in number by web scrapping the aggregator websites and extracting the hospital data and the key contacts in an organization categorized by specialty.

**Secondary objective:**

AHA requested for the similarities between their member hospitals over different domains such as hospital capacity, procedures used in a particular hospital, and the average rating over the different domains.

Initial project deliverables will consist of formatted data from web scrapping their member hospitals which includes contact list for each organization. Later deliverables include cluster visualization of these organizations based on the potential behavioral and demographic patterns.

The project deliverables will address the sponsor needs by December 5th, 2017.

**UPDATED PROJECT PROPOSAL as of Sep 15, 2017**

AHA would like to know their member hospitals better by web scrapping and extracting the key contacts for an organization, findings of member organizations/key contacts in the social media or news sources. In addition, AHA requested for the similarities between their member hospitals based on their mentions in the media, advertising patterns and the procedures used.

Initial project deliverables will consist of formatted data from web scrapping their member hospitals which includes contact list for each organization. Later deliverables include list of the key issues and major mentions of these member organizations in the social media/news, cluster visualization of these organizations based on the potential behavioral and demographic patterns.

The project deliverables will address the sponsor needs by December 5th, 2017

**PROBLEM STATEMNET**

**Context:**

AHA has 6000 hospital members whom it serves.

**Problem:**

1. Lack of Data
2. Providing a solution to every hospital individually might lead to a lot of redundancy in work and might lead to scaling issues.

**Solution:**

1. Web Crawler API
2. Providing a method to bucket Hospitals into clusters grouped by common issues to address, so a template can be developed where solutions can be shared and incrementally developed.

**OUR APPROACH**

**Web Crawler API:**

Developed a Web Crawler API that extracts hospital information from Healthgrades website given the hospital name and address into a JSON file. The extracted information includes patient reviews on a hospital, physician names and their specializations.

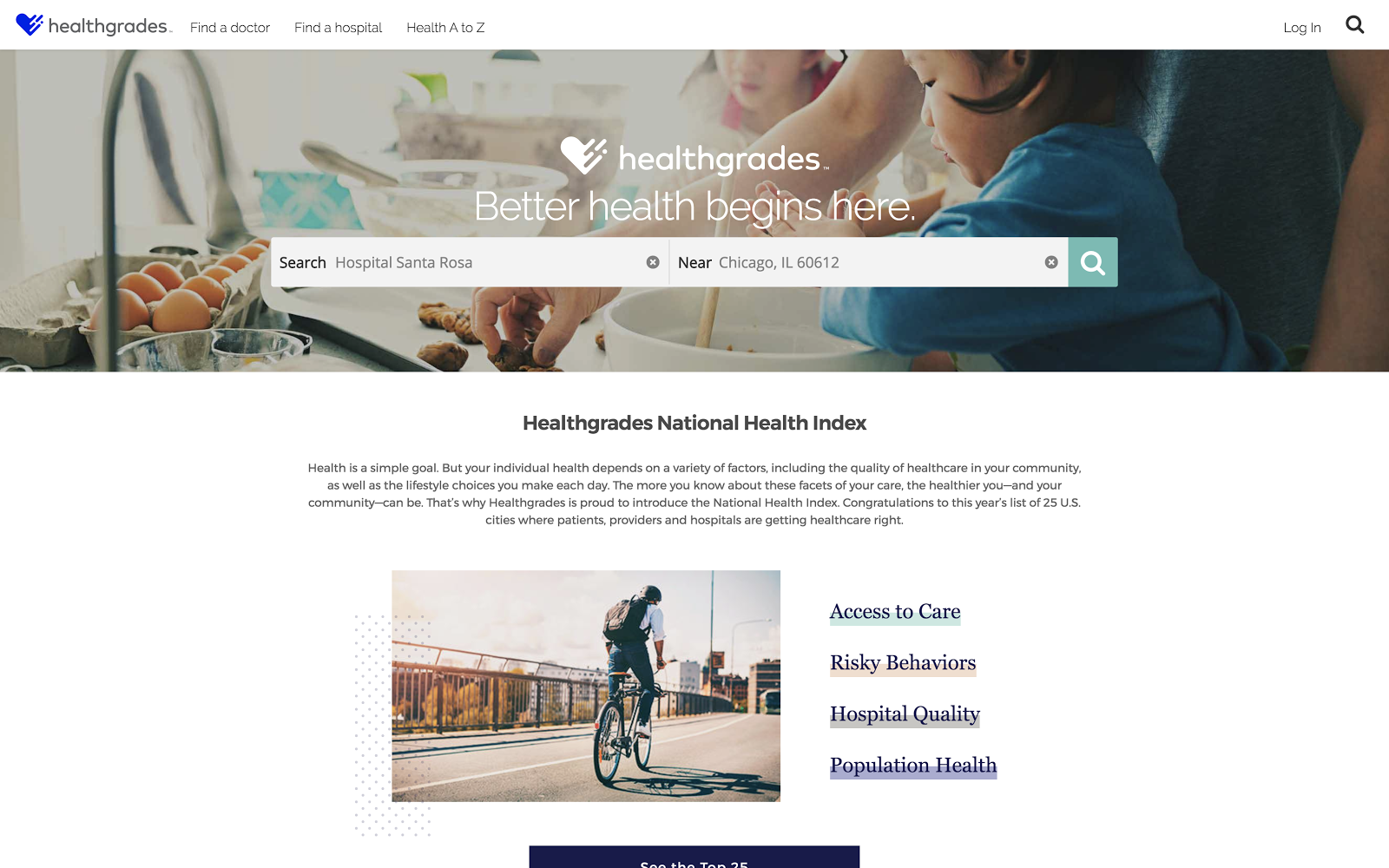
The necessary libraries used to develop the Web Crawler API are as follows:

* urllib
* re
* requests
* bs4 (Beautiful Soup)
* pandas
* json
* time
* robobrowser
* splinter
* selenium (webdriver)
* os
* unittest

**Work-Flow of the script:**

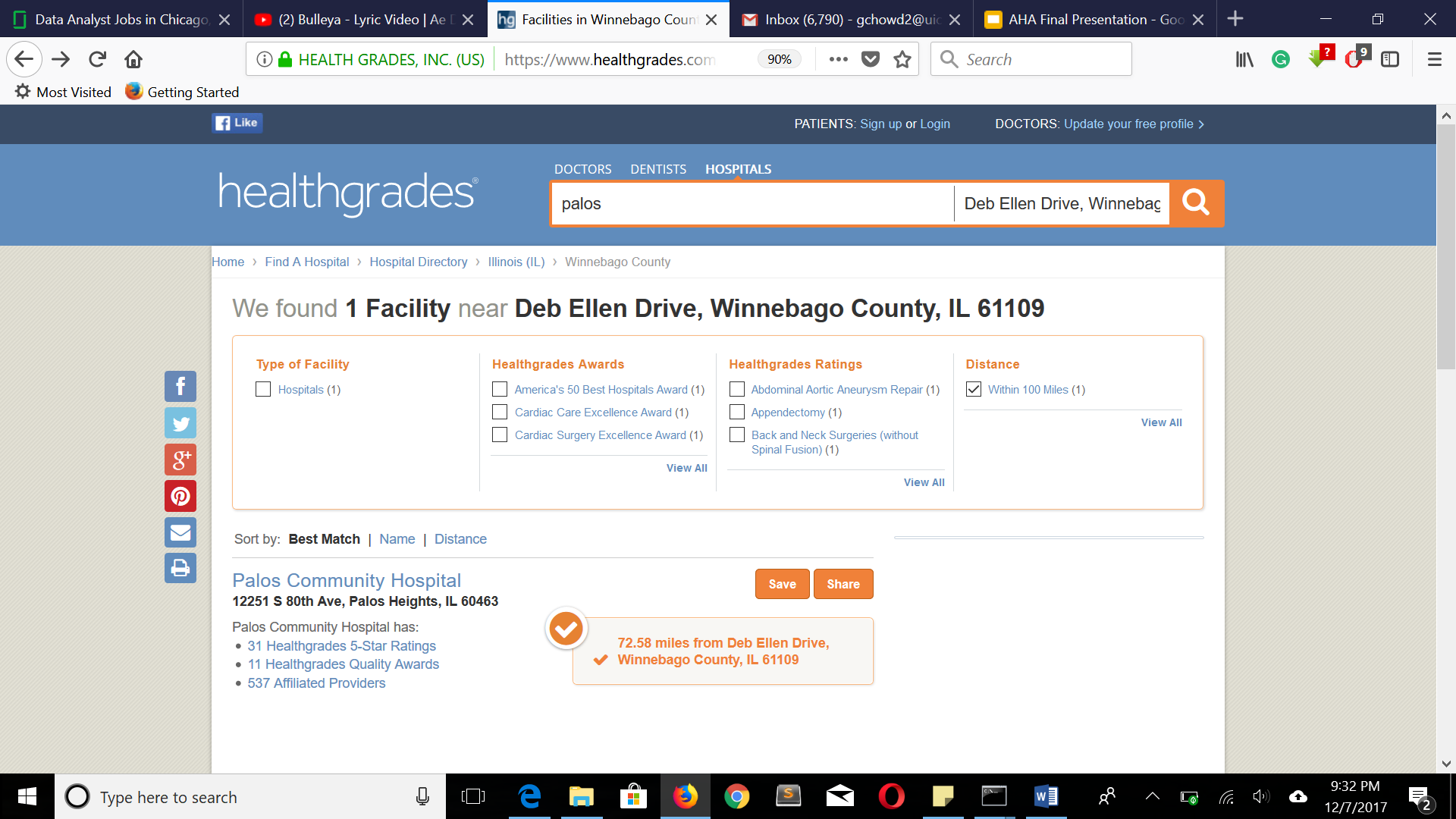
Step 1:

Input Data into the Search Box named Hospital Name and Hospital address & click the submit button of the Healthgrades Find your hospital HOME page.



Step 2:

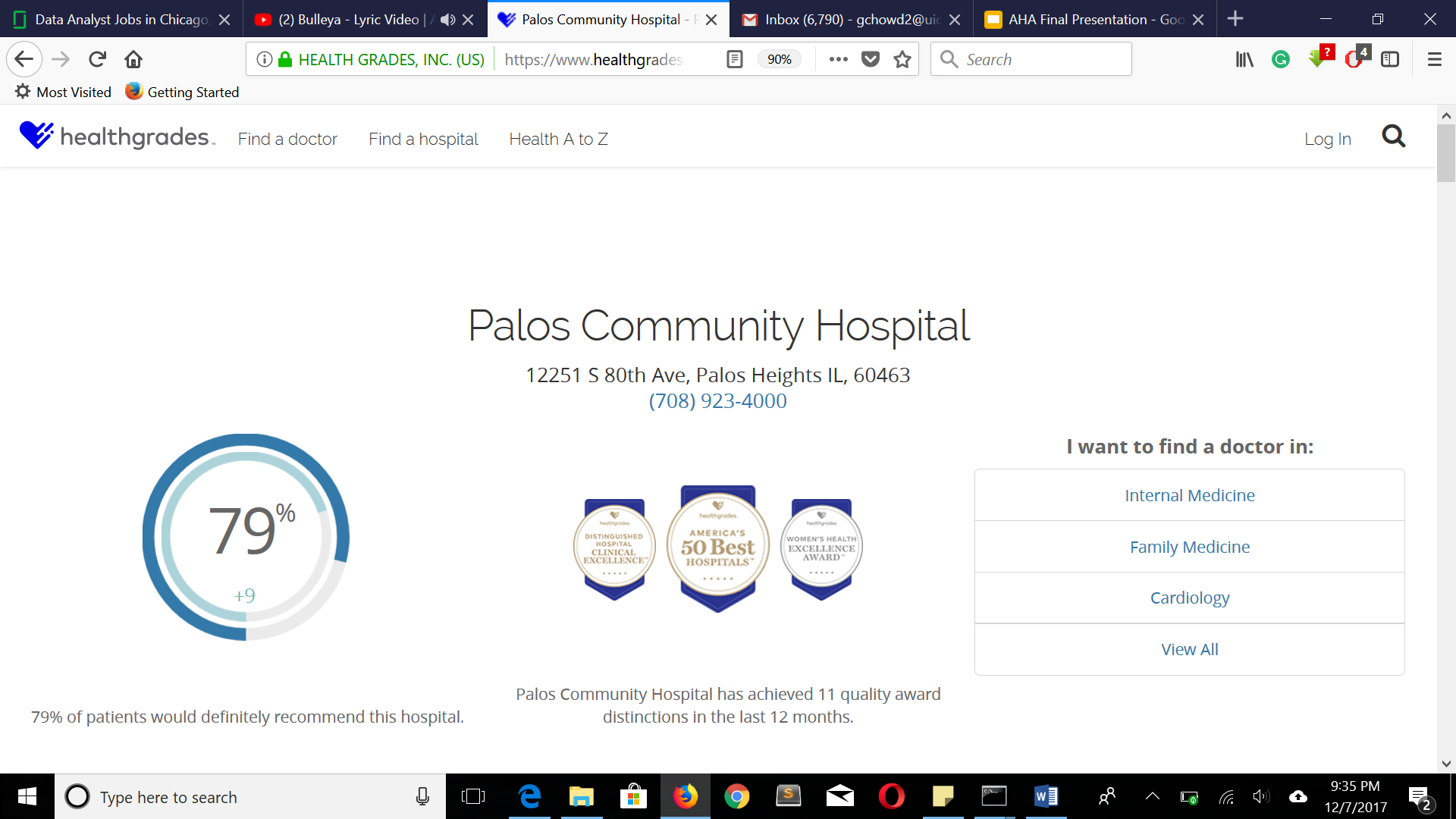
Lands at this page and compares the name of the hospital with all possible results and clicks on the best match hospital link to crawl to the particular hospital page.



Step 3:

Lands on the hospital page to collect required data fields from this page using the web scraping script

Data fields: Name, address, phone number from this section of this page.

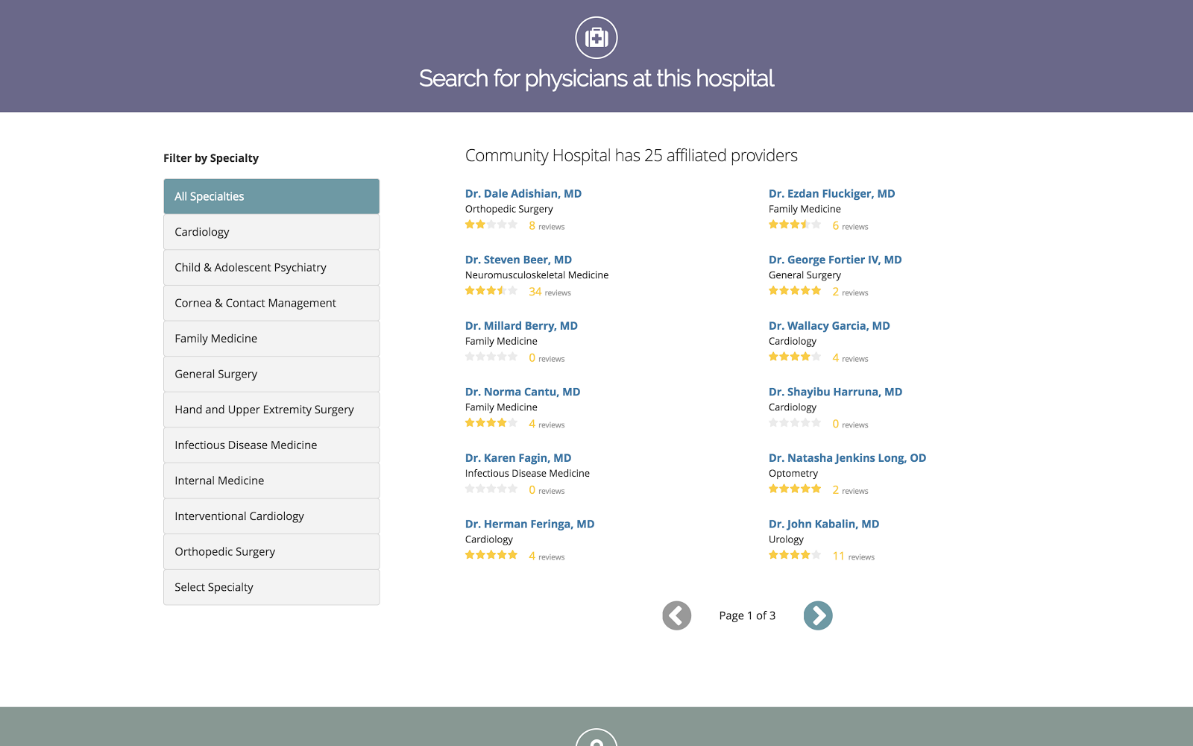


Data Fields:

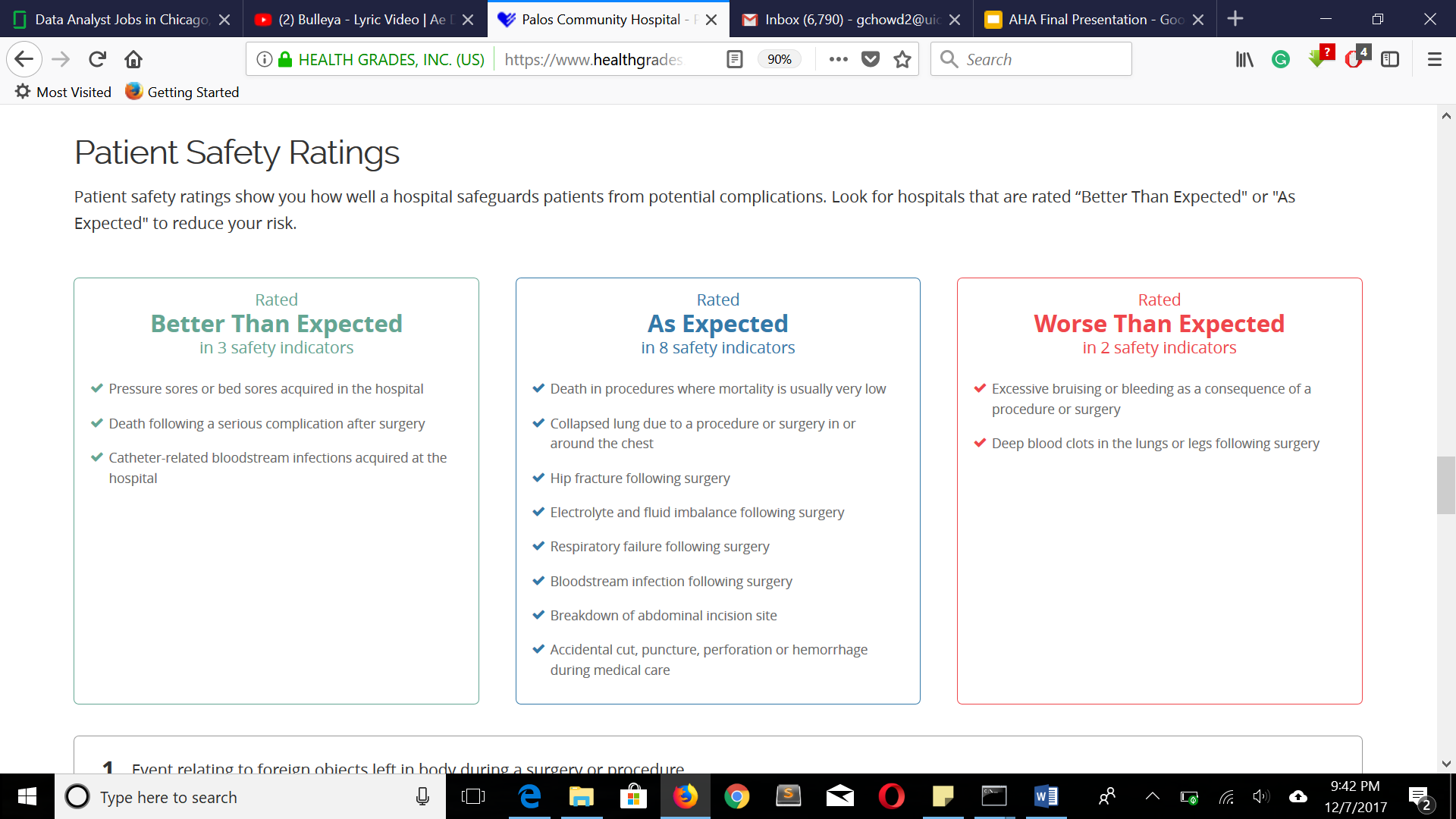
Patient reviews and ratings from this section of the page



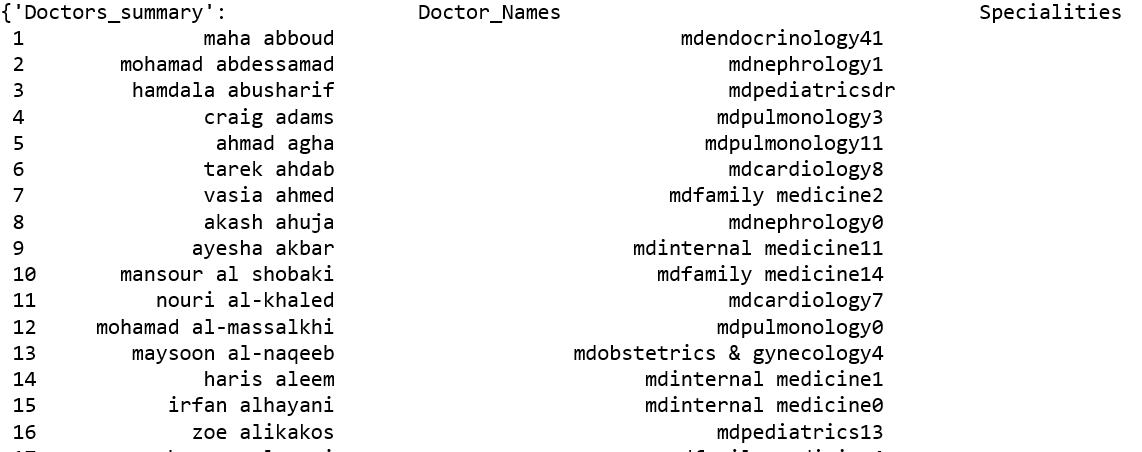
Data Fields: Doctor names & their specialty from this section of the Hospital page

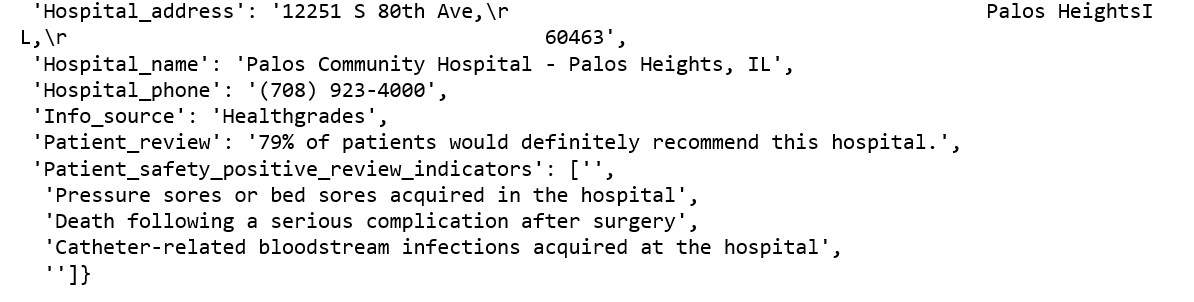


Data Fields:Patient Safety Positive Indicators from this section of the page



**Sample Output Data from the WEB CRAWLING API:**





**Clustering API**

We have developed four clustering solutions to answer the different set of business problems. Three of them cluster hospitals based on the specific set of variables to answer questions related to asset management, bed capacity and the staff of a particular hospital. Another cluster which is the aggregation of the above three clusters answers the generic questions related to hospitals.

Aggregated Clustering Solution

Bed Capacity

Staff

Management

Asset

Management

**USE CASES:**

1. **SCENARIO 1:**

There has been regulatory changes in the National Health Policy.

Create a platform so the AHA staff can use templates to frame their advice according to our Cluster Results.

**SOLUTION PIPELINE:**

Group by “Question to Ask” by using Variables from Domain expertise.

AHA members will document all advices and reasoning for a given template

Reiterate so this solution becomes scalable

1. **SCENARIO 2:**

AHA has product called AHA guide where all hospitals are listed

Provide a way to Group similar hospitals based on requirements

**SOLUTION PIPELINE:**

Ask user for a the “motive” behind the search and the hospital of interest.

Identify the cluster that is most relevant to the hospital.

Display the list

1. **SCENARIO 3**:

There has been regulatory changes in the National Health Policy in terms of Capacity of Hospital Care Units from the time that AHA collected survey data.

Assuming few hospitals affected by policy, AHA could send the survey to the hospitals that are likely affected by the regulatory change.

**SOLUTION PIPELINE**:

Group by “Bed Capacity Cluster”

Identify the cluster that is most affected by policy.

Send a survey to all the hospitals in the cluster

**DATA ACQUISITION & EXPLORATION:**

AHA has provided the hospital survey data (Hospital Data with URLs.csv), for 6251 hospitals. The survey data has 918 variables at the hospital level granularity. In addition, data dictionary was provided (AS2016layPRELIM.pdf) which detailed each variable.

Identified subset of variables related to assets, number of beds and staff for each hospital. Performed unsupervised clustering on these subsets to group hospitals in various aspects.

Imputed missing values with the mean and normalized the data before proceeding to the clustering.

**CLUSTERING RESULTS:**

Identified set of relevant variables for each clustering group,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Assets | Bed Capacity | Staff | Aggregated Cluster |
| # Variables | 5 | 9 | 14 | 28 |
| # Clusters | 4 | 7 | 4 | 5 |
| Clustering Method | Spectral Clustering | Agglomerative | KMeans | KMeans |

**Asset Management:**

Identified below variables related to assets and applied unsupervised learning to observe the patterns.

PLNTA - Property, plant and equipment at cost

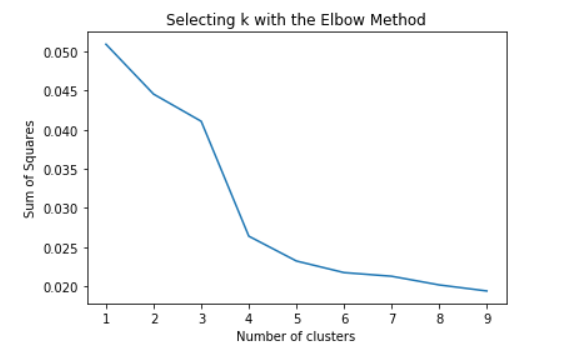
ADEPRA - Accumulated depreciation

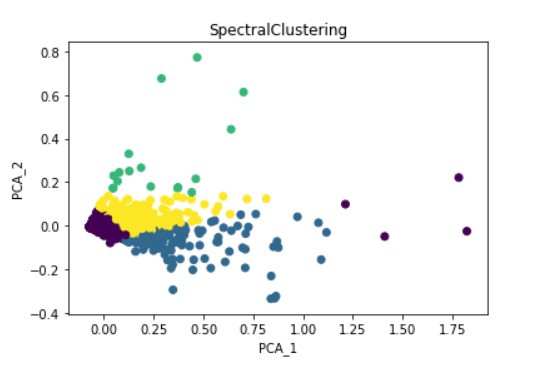
ASSNET - Net property, plant and equipment

GFEET - Total gross square feet of your physical plant

CEAMT - Total Capital Expenditures

From the elbow graph below, chose final number of clusters to be 4. Applied various clustering techniques and chose Spectral Clustering based on the patterns observed and the silhouette score.





**Cluster Characteristics in the order (0,1,2,3)**

* Low budget Hospitals (e.g. Health care clinics)
* High Profile with moderate Gross sq. feet Hospitals (Apollo, Fortis)
* Moderate Cost / valuation with High Gross Sq. Feet Hospitals (government hospitals)
* Moderate cost / valuation / gross sq. feet Hospitals (mid-size private hospitals)

**Bed Capacity**

Identified below variables related to the bed size.

BSC - Bed size code

GENBD - General medical and surgical (adult) beds

PEDBD - General medical and surgical (pediatric) beds

OBBD - Obstetric care beds

ACULTBD - Acute long-term care beds

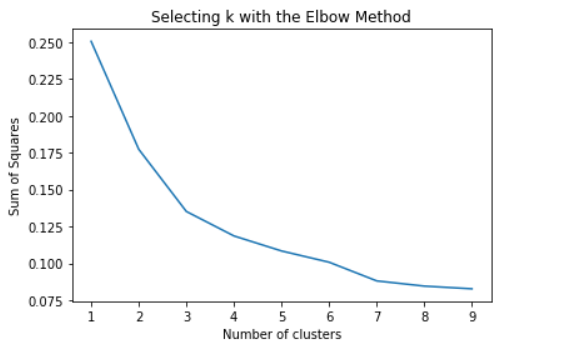
PSYBD - Psychiatric care beds

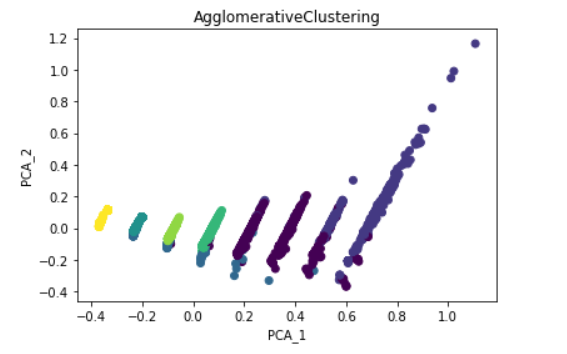
ALCHBD - Alcohol/drug abuse or dependency inpatient care beds

BRNBD - Burn care beds

MSICBD - Medical/surgical intensive care beds

From the below elbow graph, chose final number of clusters to be 7. Applied various clustering techniques and chose Agglomerative Clustering based on the patterns observed and the silhouette score.





**Cluster Characteristics in the order (1-7)**

* High Psychological and Drug Care bed capacity with low Acute long-term care bed capacity Hospitals
* High Pediatric Care bed capacity Hospitals with low in all other bed care facility
* Adult and Obstetric Bed Care Hospitals with moderate pediatric bed capacity
* Uniformly low Capacity Hospitals across all kind of beds
* Semi Uniform Capacity with moderate adult beds & obstetric care beds
* High Acute long-term care hospitals with low in all other bed care facility
* High Burn care bed capacity Hospitals with moderate Adult & Obstetric bed capacity

**Staff Management**

Identified below variables related to the staff.

FTMT – Total physicians engaged in arrangement with hospital

FTMDTF – Full-time Physicians and dentists

FTRES – Full-time medical and dental residents/interns

FTTRAN – Full-time other trainees

FTRNTF – Full-time registered nurses

FTLPNTF – Full-time licensed practical or vocational

FTAST – Full-time nursing assistive personnel

FTRAD – Full-time radiology technician

FTLAB – Full-time laboratory technician

FTPHR – Full-time pharmacists, licensed

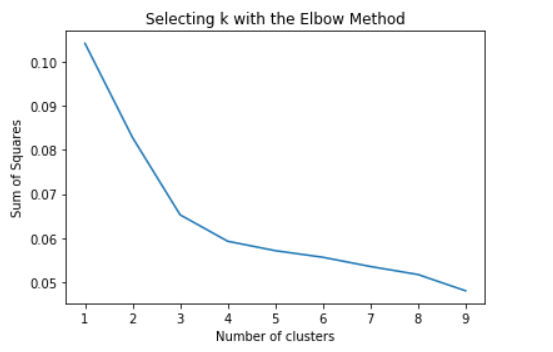
FTPHT – Full-time pharmacy technicians

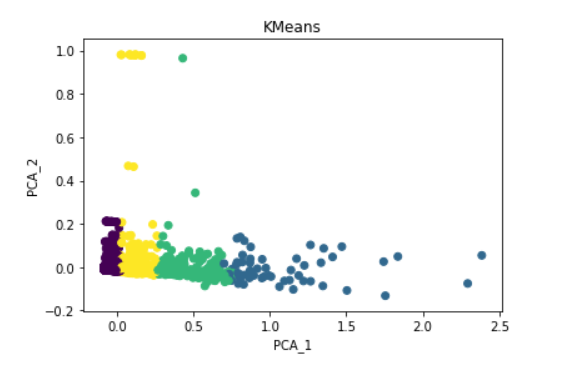
FTRESP – Full-time respiratory therapists

FTOTHTF – Full-time all other personnel

FTTOT – Full-time total personnel

From the below elbow graph, chose final number of clusters to be 4. Applied various clustering techniques and chose KMeans Clustering algorithm based on the patterns observed and the silhouette score.



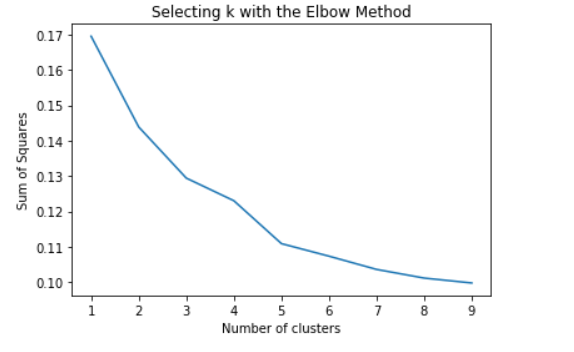


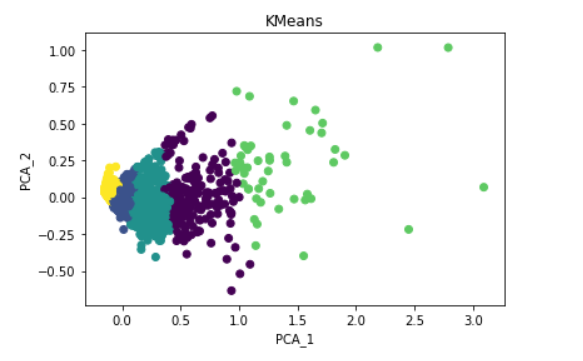
**Cluster Characteristics in the order (2,1,3,0)**

* High Man-Power- Hospitals
* Medium Man- Power Hospitals
* Low Man- Power Hospitals
* Very low Man – power hospitals

**Aggregated Clustering Solution**

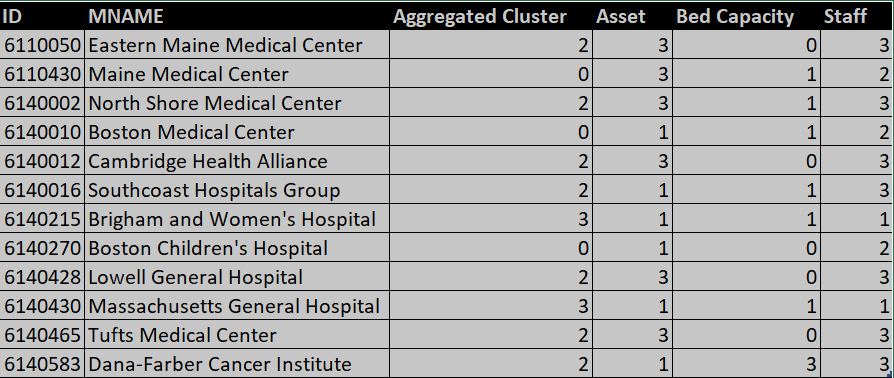
Combined features from all the above clusters and applied unsupervised learning on these 28 variables and obtained below results.





**FINAL RESULTS:**

Each hospital along with the unique identifier is joined with the clustering results as shown below.



**APPENDIX**

**Project Milestone List:**

|  |  |  |
| --- | --- | --- |
| Milestone | Description | Date |
| Defined Project Scope | * All requirements for AHA Seeker must be | 10/26/17 |
|  | determined to base design upon |  |
| Data Acquisition script, | * Handover the web scrapping script, * Preliminary cluster analysis results | 11/2/2017 |
| Initial Cluster Analysis |  |  |
|  |  |  |
| Complete Modeling and | * Hand over data collected from web scrapping * Improvised cluster Analysis results | 11/16/17 |
| Testing |  |  |
| Complete Model | * Handover the Model and documentation with visualizations of Clustering | 12/5/2017 |
| Deployment and Handoff |  |  |