

Thanks for showing interest in this role for Python/ElasticSearch developer with Datazymes. As the first round can you implement the below project in your local machine and upload the completed code to Github.

Important milestones

Send your participation acknowledgement by 1st June 2018
Questions and clarification to be sent by 2nd June 2018
Project completion date 6th June 2018

Business case:

Create a report with the physician level details from the New Jersey, US with the following metrics

1. As part of the project please scrap the data from the website; Write the scrap module so that it can be reused with other states in US. Use parameters as much as possible.

<https://health.usnews.com/doctors/city-index/new-jersey>

The website contains links for each city of new jersey (screenshot 1.0) and further each city has a set of specialty (screenshot 2.0) and further has the individual doctor (healthcare professional; screenshot 3.0 and 4.0) level detail in tabs; From the doctor page scrap the following details only

- a. Overview
- b. Full Name
- c. Years in practice
- d. Language
- e. Office location
- f. Hospital Affiliation
- g. Specialties and sub specialties
- h. Education and medical training
- i. Certification and licensure

Scrap all this data for all doctors across all specialties across all cities in New Jersey

2. Use python again to load the scrapped data from step 1 into elasticsearch; One document per doctor
3. Create the following summary report from the elasticsearch again using python.
 - a. Total number of doctors by city
 - b. Total number of doctors by specialty (element g of the scrapped elements)
 - c. Total number of doctors based on their experience range (experience range : 0 – 4 years, 5 – 10 years, 11 – 16 years, 17 – 20 years and 20 years above)
 - d. Total number of doctors by zipcode (The last five digit of the address; numeric field)
 - i. i.e., 222 New Rd, Linwood, NJ 08221 ← zipcode

Deliverables to include:

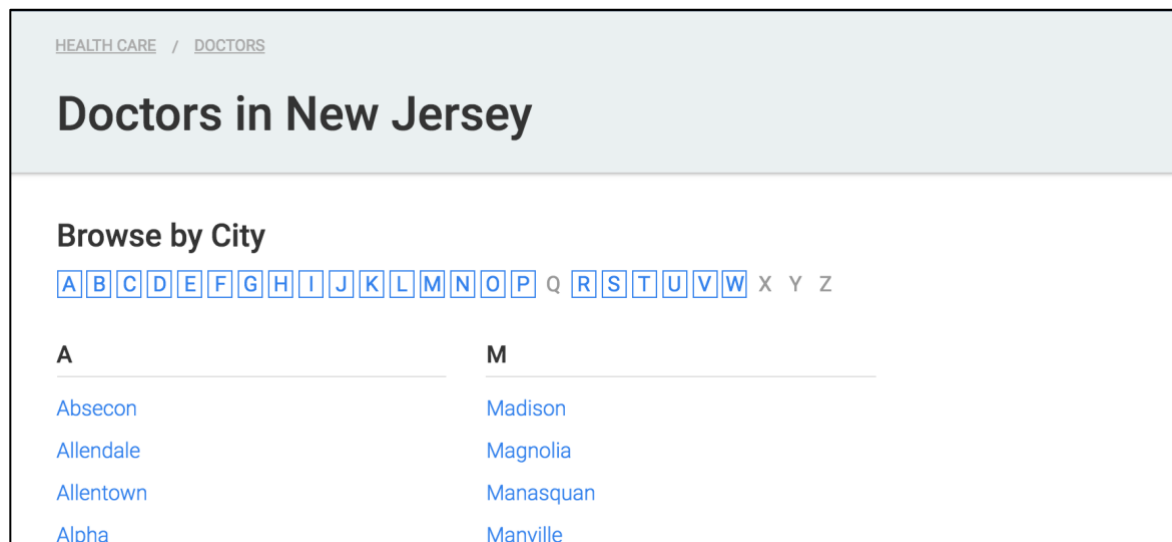
1. Working version of the code to be uploaded to github for review and also the requirements.txt
2. Summary report mentioned in business case point 3
3. We do not require the actual scrapped data but only the summary of the same

Points to note before starting with this project

1. Use only Python2.7 and Elasticsearch6.2 only
 2. All doctor profiles within each city and specialty should be scrapped and loaded into elasticsearch
 3. The final summary should be reported in a json format
 4. Build reusable python functions and break the functions into manageable pieces instead of a single monolithic functions
 5. Create your scrapping function using python modules like requests, urllib and beautifulsoup. Do not use third party scrapper for this exercise. Third party scrapper will include scrapy
 6. Do not overload the website with too many request within a small time frame. Optimize the scrapping frequency
 7. Implement multi-threading if possible
 8. elasticsearch==6.2.0 and elasticsearch-dsl==6.1.0 works great with python
- Use virtualenv to create the workspace
 - Comment your code blocks properly
 - Do not outsource this project to someone else as the second round of face 2 face will involve questions from this round
 - Have any questions reach out to Manikandan@datazymes.com before questions to be sent milestone

Screenshot for reference

Screenshot 1.0



Screenshot 2.0

Doctors near Absecon, New Jersey

Many doctors specialize in a particular area of medicine. Please select a specialty below.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

[Allergist-Immunologists](#)

[Allergist-Immunologists](#)

O

[Orthopedists](#)

[Occupational Medicine Specialists](#)

Screenshot 3.0

I'm Looking for...

☐ Hospitals

☐ Children's Hospitals

☒ Doctors

Doctor's Name

City, State or ZIP

All Distances

3 matches

SORT BY: Distance

Location: Absecon, NJ

Specialty: Allergy & Immunology

Clear All

Dr. Kenneth Better MD


Linwood, NJ | 7.42 miles from Absecon, NJ

Allergy & Immunology

Asthma & Allergic Conditions


Dr. Kenneth Better is an allergist-immunologist in Linwood, NJ, and has been in practice more than 20 years. [more](#)

21+ Years of Experience



Dr. Lawrence Schwartz DO

Somers Point, NJ | 10.14 miles from Absecon, NJ



Screenshot 4.0

Dr. Kenneth Better MD

Allergy & Immunology | Linwood, NJ

Overview

Contact

Insurance Accepted

Hospital Affiliation

Experience

Overview

Dr. Kenneth Better is an allergist-immunologist in Linwood, New Jersey and is affiliated with multiple hospitals in the area, including AtlantiCare Regional Medical Center and Shore Medical Center. He received his medical degree from Albany Medical College and has been in practice for more than 20 years. He is one of 3 doctors at [AtlantiCare Regional Medical Center](#) and one of 3 at [Shore Medical Center](#) who specialize in Allergy & Immunology.



Dr. Kenneth Better's Details

Phone Number

(609) 653-6676