Project: Python Application to Interpret Data.gov Data

Tools: Python(Pandas,Numpy,Request)

Project Participants

Thomas Sonderman

Lenore Montalbano

The purpose of our project is to explore a statistical analysis of regions and healthcare. Through the use of open API’s from Geonames and Data.gov we will obtain the raw data. With DCAT, Python and Pandas we will be able to combine the data into a usable dataset that we can explore. This data will be able to provide a way to see health care usage and anomalies per regions.

Do some regions of the United States have higher than average health care facilities? How does this relate to the population of an area? Are certain regions clustered with more incidents? These are all important questions that can be answered by combining datasets that exist on the web.

Python provides a foundation for building data intensive applications. With Python an application developer has access to a wide range of libraries that help in gathering data; one example of a Python library we will be using is Requests, this is an Http library that was built with ease of use in mind.

On top of being able to use Requests to grab the data, the data can then be loaded into data frames using Pandas and Numpy. Pandas and Numpy provide SQL type queries to the data frame that have been loaded. Python has the added advantage of using an easy to read syntax when working with a team on a project.

The use of a DCAT as the ontology will allow for the joining of disparate sources of data which will simplify the querying and should enable our team to produce the data store required for this project.

Several Datasets will be imported to use in the project. Data.gov has provided a universal dataset format DCAT that is available on most of the data released. With universal data the different datasets can be joined and manipulated to provide answers to our questions.

Hospital data can be merged with death data to provide some insight into the questions that were posed earlier in this paper. Using DCAT to normalize the disparate data sources should enable our team to more effectively query the data and produce results.

Data:

**Hospital Data**

<https://data.medicare.gov/api/views/xubh-q36u/rows.csv?accessType=DOWNLOAD>

This Dataset provides a look into all the hospital data that is available on data.gov. The data provides location, name, and affiliation information.

**Death Data**

<https://data.cdc.gov/api/views/qpap-3u8w/rows.csv?accessType=DOWNLOAD>

This Dataset provides access to a wealth of information on deaths with location and death type data.

**CHSI Indicators**

ftp://[ftp.cdc.gov/pub/Health\_Statistics/NCHS/Datasets/CHDI/chsi\_dataset.zip](http://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/CHDI/chsi_dataset.zip)

These indicators are provided by location to look at health risks.

The Data that will be going into the application comprises of several sources. The first set of data is the Hospital data that will be going in. The hospital data will give us a map of the healthcare availability per region/location.

Following the hospital data will be the death data that will map those same geographic locations to match deaths with the hospital availability.

The Third dataset that will be getting calculated is the CHSI indicators. The CHSI indicators can provide an insight into the risk levels that are associated with the geographic regions we have previously mapped.

The output of this data will answer some questions that we laid out. Are we able to look at a specific region and see if they offer decent hospital access but have a high number of deaths? If the deaths are high, is this consistent with the regions CHSI indicators? These questions if answered correctly could provide a useful correlation between the health data on data.gov and human behavior. If there are these risks, are they confined to specific regions of the United States?

One thing we will be looking at is whether there are any specific trends in the data. Do large metropolitan areas have higher health risks/deaths? With this data we can explore many different hypotheses about health and geographic locations.

The Semantic Web can improve searches if used correctly and efficiently. Whenever joining data there is always the factor that one can improperly join that data. The government's wide acceptance of DCAT can prevent these types of errors by specifying the exact columns that will be joined in the data represented on data.gov.

Whenever a technology to improve data aggregation is introduced it hinges on its success of acceptance. If a tool is rolled out that offers great advantages but relies too heavily on user participation then its usefulness will still correlate directly to how many people come on board.

Constantly we see many people pushing out these technologies that never gain traction. With Data.gov there is across the board consistency in the data that is getting pushed into RDF compliant datasets. If most of the data is in the RDF compliant format that will provide an “ease of use” for the data consumer. These principles of consistency for data access when widely adopted can only improve the access to data for a user/consumer.