CUNY MSDS Data 607

18 April 2019

Final Project Proposal

Team SPARC:

* Santosh **C**heruku
* **S**amantha Deokinanan
* **R**ajwant Mishra
* **P**riya Shaji

**Proposal #1: Recommendation System**

**Proposed Research Topic:** Establishing a Medicine Recommender System that can give a medication recommendation with an excellent efficiency and accuracy based on diagnosis and symptoms.

**Purpose:**

According to the World Health Organization, more than 42% medication errors are caused by doctors because experts write the prescription according to their experiences which are quite limited. With the advancement of technology and data analytical technique such as data mining and recommender systems, it is possibilities to explore potential knowledge from diagnosis history records and help doctors to prescribe medication correctly. A recommendation of right medicine based on diagnosis can targeted healing and decreasing trial-and-error when prescribing drugs. This can may further decrease undesirable drug side-effects. Therefore, the objective of this project is to build a medicine recommender system that can give a medication recommendation with an excellent efficiency and accuracy based on diagnosis and symptoms.

**Method:**

Team SPARC plans to design medicine recommender system that applies machine learning to the recommendation system. Datasets are retrieved from:

* Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125. DOI: [https://doi.org/10.1145/3194658.3194677]
  + These datasets provide patient reviews on specific drugs along with related conditions. Reviews and ratings are grouped into reports on the three aspects benefits, side effects and overall comment.
* USP Drug Classifications which provides therapeutic or pharmacologic groupings of FDA approved medications, consistent with current U.S. healthcare practices and standards of care.
  + This dataset has the hierarchical USP drug classifications which includes USP Categories, USP Classes in that category, the drugs (i.e. the general drug compound), and the medication or formulation you would buy of that drug. The dataset also has the individual KEGG pages for these drugs which leads to information the product and generic names, chemical formula, additional classes, ATC codes, biochemical information, other classifications, and links to the compound in other databases (e.g. PubChem, DrugBank, etc). [https://www.kaggle.com/danofer/usp-drug-classification]

These datasets are given its raw form; therefore, data cleaning and transformation will be conducted as necessary. Possible methods for building the recommender system will include collaborative filtering, SVM (Support Vector Machine), decision tree algorithm based on the diagnosis data and more. Any research questions will be tested and analysed using statistical methods and visualization. The data sets will be split into training and testing set to access and improve performance.

Limitations: Team SPARC is aware that this recommendation system will not be quite accurate since we are not specialist to say one medicine can indeed be recommended instead of another, in addition to the limitation on data such as cost of medicine, insurance company that cover specific type, etc which are not easily obtainable within the time frame for this project. Therefore, our model is thought as a basis in the right direction, and with more data, the better the recommendation can become.

**Tentative Bibliography:**

* Meisamshabanpoor & Mahdavi, Mehregan. (2012). Implementation of a recommender system on medical recognition and treatment. International Journal of E-Education, E-Business, E-Management and E-Learning. 2. 315-318.
* Kakulapati, Vijayalakshmi & Balaram, V V S S S & Reddy, Sheri. (2018). Collaborative Filtering Recommendation systems for personalized medicines. 5.

**Proposal #2: Trees Census and Air Quality**

**Proposed Research Topic:** The Effects Of Urban Trees On Air Quality

**Purpose: Why should be done in this analysis?**

President Ronald Reagan blamed trees for air pollution. Trees emit volatile organic gases such as the isoprene emitted from pine trees, into the atmosphere. These gases lead to the formation of ozone in our lower atmosphere when chemically reacted with nitric oxide derived naturally from soils and from various human sources, such as power plants and automobiles. However, trees remove a variety of air pollutants and particulates, including both ozone and nitrogen oxides, which reduces the ambient concentrations in the air we breathe. They help with temperature reduction and other microclimatic effects. Moreover, having urban trees along roadways can reduce the presence of fine particulate matter in the atmosphere within a few hundred yards of the roadside verge. Tree species also impact the how effective a tree is in to help improve air quality. For instance, greatest relative effect on lowering ozone were mulberry, cherry, linden, and honey locust. Therefore, the objective of this project is to investigate effects of tree location, number of trees and species with the exposure to primary air pollutants to determine their impacts.

**Method:**

In order to conduct a thorough analysis to answer the research questions, datasets are retrieved from:

* NYC Open Data makes public data generated by city agencies available for public use. More specifically, the catalogue over 500,000 sidewalk trees surveyed decennially in 1995, 2005, and 2015.
  + This dataset contains information on the trees such as status, health, measurements, location, species, signs of damage or problems, and more. It can be found at [https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh]
* United States Environmental Protection Agency provide daily outdoor air quality information for several pollutants.
  + The datasets that will be retrieve is the Air Quality Index Daily Values Report for each pollutants where there is a daily Air Quality Index values for the specified year and location. It can be found at [https://www.epa.gov/outdoor-air-quality-data/download-daily-data]

These datasets are given its raw form; therefore, data cleaning and transformation will be conducted as necessary. Possible analysis, in addition to diagnosis tests, will include hypothesis testing, regression analyses, geographic research, etc.

**Tentative Bibliography:**