FINAL PROJECT PROPOSAL

FINAL PROJECT PROPOSAL

Pollution Data Analysis and Machine Learning for prediction

Snigdha Joshi, Vipra Shah

Overview:

Pollution, we hear it every other day at school, college and read about it in newspapers but never had chance to analyze the situation. Pollutants are the key elements or components of pollution which are generally waste materials of different forms. Pollution disturbs our ecosystem and the balance in the environment. With modernization and development in our lives pollution has reached its peak; giving rise to global warming and human illness. The motive of this project is to analyze pollution data to deduce positive and negative changes on air quality over years.

Goals:

To provide pollution analysis which will help us to derive some inferences from dataset.

1. What are main causes of air pollution?
2. Which main pollutants are present in US?
3. What is city wise distribution of pollution?
4. What is trend of air pollution over the years?

Answer to such questions can give us some useful insights:

1. How much city wise distribution of pollutants vary.
2. Grouping of cities which has same type of pollutants.
3. What is source of these pollutants
4. Predict the pollution over next years
5. What actions should be taken required to restore air quality in cities

Use cases:

1. Estimate Air Quality Index for pollutants over next years which will help government to take preventive measures
2. Potential homebuyers can weigh their choices by considering air quality
3. Alert health care systems by predicting possible diseases due to pollutants (e.g. Asthma, Lung cancer)

Data:

1. A huge dataset is provide by United States Environmental Protection Agency(EPA): https://aqs.epa.gov/aqsweb/airdata/download_files.html

   The file download page contains seven types of files:

   a. Site Descriptions
   b. Monitor Descriptions
   c. Annual Summary Data
   d. Daily and Daily Summary Data
   e. Hourly Data

    f. 8-Hour Average Data
    g. Blanks Data
2. These files are summarized in one csv: https://www.kaggle.com/sogun3/uspollution

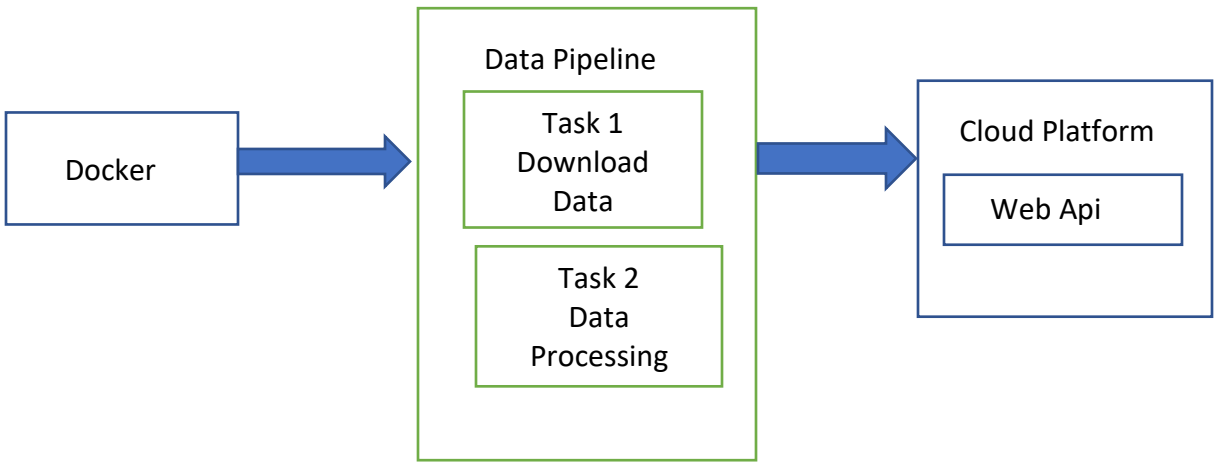
Process outline:

1. Data Preprocessing:
    - Data Cleaning, handling missing values
    - Extract required columns from data file
2. Exploratory Data Analysis
3. Study of Supervised approaches and select the best model for prediction
4. Study of Unsupervised approaches (Clustering and Associative rule mining) for recommendation
5. Design of a pipeline and system to implement this approach and discussion on the system's capabilities
6. Deploy the Model on Azure/AWS or Google Cloud Computing Platform
7. Build a web application to demonstrate the prediction and recommendation results.

Deployment Details:

1. Language: Python, R
2. Pipeline: Luigi
3. Container: Docker
4. Cloud Tools/Platforms: Microsoft Azure Machine Learning Studio, AWS (Amazon Web Services) EC2
5. Visualization: Tableau
6. Other Considerations: Google Cloud Platform

System Architecture Diagram:



User Interface: