INFO 7250 – Final Project Report
WENQING LIANG

PM2.5 Analysis of Five Chinese Cities

dataset link: http://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities#
python notebook link: https://github.com/aliang1994/INFO7250-BigDataAnalytics/blob/master/INFO7250-pm2.5Analysis.ipynb

Introduction

PM2.5—fine particles in the air with a diameter of 2.5 μm or less [Wikipedia]—is an important indicator of the severity of air pollution. In most recent years, China has undergone significant changes such as economic growth and technological development; however, the overall air pollution level has also surged alongside these activities. The dataset from UCI Machine Learning Repository contains PM2.5 data between 2010 and 2015 for five Chinese cities: Beijing, Shanghai, Guangzhou, Chengdu and Shenyang. Each row contains information for an hour in a day. Specifically, these are the five cities in which US Embassy is located and provides its own PM2.5 monitoring. In my final project, I would like to compare air pollution levels of these give cities by analyzing the PM2.5 value monitored by US Embassy ("PM_USPost" column in the dataset). I would only use data from 01/01/2013 to 12/31/2015 since some of the cities did not start monitoring PM2.5 earlier than the year of 2013.

Initial Analysis

According to the annual average set by the World Health Organization (WHO) [World Health Organization, 2006], the PM2.5 value for acceptable air quality (or clean air) is below 35 μg/m3. In my project, I use 3 ranges of PM 2.5 values to classify pollution level: "≤ 35 μg/m3" – low, "35 μg/m3<PM2.5<150 μg/m3" – polluted, and " >150 μg/m3" – severely polluted. Data preprocessing was done using Python Notebook (see link above). Counting total hours of pollution under each category for all the five cities, an initial sense of the pollution level was provided by the data frame in **Out[14]**:

	city	low	polluted	severlypolluted
0	Beijing	8130	12786	5054
1	Shanghai	10506	14025	979
2	Guangzhou	10715	13876	372
3	Chengdu	3934	17609	2948
4	Shenyang	6355	12943	2382

We can see that Beijing, Chengdu and Shenyang have relatively longer hours of "Severely Polluted", whereas Shanghai and Guangzhou have higher air quality. Visualization of data see **Out[15][16][17]**.

MapReduce Analysis

1) Counting hours of pollution for each month (counter)
   In order to get deeper sense of how the pollution level fluctuates, a MapReduce job with month as key and total hours of pollution under three level categories as value (PollutionLevelWritable) was performed for all five cities. Individual visualized result for each city see **Out[19][21][23][25][27**]; combined visualized result see **Out[28]**.
   We can see that most of the "Severely Polluted" data come from winter months for all the five cities. Shanghai and Guangzhou have more hours in "Low" level than the two polluted levels in summer and fall months; however, Beijing, Chengdu and Shenyang have longer hours in polluted levels than clean air level throughout the year. Beijing has longer hours in "severely polluted" than all other cities.

2) Min, Max, Aver PM2.5 per day (min/max/aver, top k)
   The dataset provides data for each hour in a 24-hour day. Part 2 is trying to find the minimum, maximum, and average hourly PM2.5 value for each day, and more specifically zoom into the outliers of these data– top 20 PM2.5 value for each city with date as unit instead of hour. A MapReduce job with date as key and min/max/aver as value (PollutionValueWritable) was created for analyzing daily PM2.5 values. Visualized time series plot see **Out[30][36][42][48][54]**. Lots of significant pikes can be observed for Beijing, fewer for Shenyang and Chengdu, and even less for Shanghai and Guangzhou. To zoom in for these pikes, a finding-top-20 MapReduce job was generated on top of the result of maximum daily PM2.5 values. A Heatmap combining five cities is generated for the result of top-20 (see **Out[73]**). It provides lots of insight to the severity of air pollution for these cities, and based on the color (darker = more pollution) we can infer the rank of pollution level – Beijing > Shenyang > Chengdu > Shanghai > Guangzhou – for the years 2013-2015.

Future Effort

Time series prediction of pollution levels using MapReduce can be made if more years of data are used in the analysis.