Zuha Ahmad

Dr. Jaume

DS-670-01-17SPSG: Capstone: Big Data & Bus Analytics

04 February 2017

**Assignment 3 – Expected Contribution and State of the Art**

In the Capstone research course, we have to select a dataset from the CityPulse EU Data Collection. The categorical dataset in which I am tackling on is the Pollution dataset. The Pollution dataset derives from two locations. The first is from Aarhus, Denmark and the second is Brasov, Romania. The timeline of the dataset is taken in the year 2014 for a few months for each country. The raw dataset is in the CSV which has 81,012,652 bytes. However, the annotated dataset is distributed in the gzipped file format which has 3,368,454,165 bytes. The structure of the dataset is structured because it has numerical fields and timestamps. The contributing variables in this dataset are ozone index levels, carbon monoxide, nitrogen dioxide, sulfur dioxide, and particulate matter. In addition, it also includes the latitude and longitude coordinates to track a traffic sensor in a particular location. The data is being calculated by using the Air Quality Index. By doing so, we can see the contents of the polluted air and how it will be done in the future.