PROJECT PROGRESS REPORT

Group Members: Anand Zupa, Ryan LaRocque

Subject: Million Song Dataset

For out project, we are employing the EPA’s annual air concentration by year (<https://aqs.epa.gov/aqsweb/airdata/download_files.html>) in order to create a machine learning model which can predict air quality based on wind data, temperature, location, meteorological data, and air toxicity. These features are broad and will not all correlate with air quality, but we hope that we will be able to predict air quality at a certain location based on unexpected features, for example wind and location data correlating to air quality because of forest fires. The downloadable data has 55 features and 51,000 instances, but this is only in the annual concentrations. We are also able to scrape finer data including daily data from each site, and we will compare to see if this is able to give us more accurate results. Additionally, we will make contact with a CU Boulder PhD student who is doing work sponsored by NOAA to see if they can contribute any other insights or offer any extra data. Our dataset as such is very dense, so a large portion of our work will involve finding which features give us usable results.

We are also researching scholarly articles on finding suitable features, optimizing our neural net, and extrapolating meaningful data. Articles we have consulted include *The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data* by Murat Kayri et. al., *Locating nearby sources of air pollution by nonparametric regressions of atmospheric concentrations on wind direction* by Ronald C. Henry et. al., and *Robust extraction of baseline signal of atmospheric trace species using local regression* by Andreas Ruckstuhl et. al.