CS5140 Data Mining
Aaron Goh & Ritesh Sharma
Data Collection

1. The data was collected from CapIQ where exporting data was a lot simpler and supported. The file was sent out as an excel file.

2. We have 5.5k companies with data on companies within the U.S. and the industry in which they are contained. We have the total enterprise value, total revenue, and EBITDA.

3. The data is stored as a CSV format (i.e. excel file). The data type being stored are in monetary value (or doubles) and string formatted for Industry Classifications and Geographic locations. However, the string formats are standardized in such a way that they are easily parsed.

4. No, we were able to get the data exported conveniently through CapIQ which supports doing analysis of this sort.

5. We hope to structure the data using distances to calculate regressions among different correlating fields of the companies values, industries, total revenue, and EBITDA. We will likely add more fields of data as time goes along.