CS838 Data Science project progress report - Stage 1

Jinman Zhao	Bin Guo	Di Wu
jzhao237@wisc.edu	bguo23@wisc.edu	dwu73@wisc.edu

● Goal of the project

1. Main idea: mining American Fortune 500 companies' stock information on Nasdaq, learn about the relationship between stock values and company locations, divisions, revenue and other facts.
2. Entry level goal: find out the relationship between absolute stock value and company's revenue.
3. Advanced level: Figure out the major impact factors of different company features on the stock values.

● The set of sources selected for the project

For this problem, we have three major sources
1, Nasdaq website for stock value information
2, Fortune 500 website to get a list of Fortune 500 company names
3, Get company synopsis from Wikipedia as plain text document.

● Method used to extract structured data from the data sources.

1. Download csv format data of company stock values and basic information from Nasdaq website.

2. Use "Scrapy" to crawl company information from Fortune 500 website. The information includes rank, employee number, revenue, profits, assets, locations and brandindex etc.

● Information extracted from the text documents

1. We want to extract the company's name from the text documents.
2. Synopsis of the first 100 companies on Fortune 500 list from Fortune 500 website, because only companies with higher rank have complete brief introduction. Other companies only have one sentence description, which is not suitable to use as the text documents.

3. Another 200 text documents are crawled from the company's Wikipedia page. We choose the first two paragraph as the company's text document.

● Open-source tools

- scrapy: web crawling
- numpy, scipy: scientific computation
- scikit-learn: machine learning
- Keras, Tensorflow/Theano: deep learning