Project report for step one

Hongyi Wang, Hao Fu, Miao Yang

Feb 5, 2017

1 Questions we are going to solve

1. **How's US university ranked among the world?**
   We explore rankings of US university among universities among the world. And try to tell the percentage of US universities in world's top 10, 20, 50 universities, and etc.

2. **What factors influence your income after graduation?**
   Incomes of people are largely influenced by lots of factors. However, which one influence people's income the most?

3. **Does Which university you attend influence your salary much?**
   Even we always do not want to tell, our educational background probably influences our life quite a lot. But we still want to know "how much" does my education background influence my future (say income)? Or will a top university education background guarantee an excellent income?

We are going to explore answer of all these foregoing questions in this course project using several fancy machine learning methods.

2 Data Sources

1. **World's university ranking by Shanghai Jiao Tong University**
   We wrote a web spider script using Scrapy lib in Python and extract data from world university ranking provided by Shanghai Jiao Tong University (http://www.shanghairanking.com/). Which contains 1000 universities all over the world.

2. **College Scorecard Data from US department of education**
   We also extract data from U.S. Department of Education (https://collegescorecard.ed.gov/data/) to see instances of people graduated from each university among the US and their future life after they graduated from college. This dataset contains more than 7000 cases of college graduations from US universities, it also contains their income, industries they working on.

3. **Comments of US universities**
   We wrote other crawl scripts to extract data from comments of people to each universities shown in our datasets from Google reviews and comments on Facebook official accounts of these Universities.

The world's university ranking datasets and college scorecard dataset are in csv format, and all the comments data are in text files. We create our own Github repository for this project: https://github.com/samfu1994/cs838webpage.