Career Path Prediction

Jeremiah Bill and Raayan Pillai
CSC 240

November 7, 2017

1 Introduction

1.1 Problem statement

With the excess of career opportunities in our society, it has become increasingly difficult to choose the right career path. This is dubbed the Paradox of Choice where having more options causes individuals to struggle in making decisions due to fear of making the wrong decision. With this in mind, we are seeking to limit this range of choices and give students and professionals the opportunity to better understand career paths based on the current and previous jobs they have had and the current job market. LinkedIn is very easy to scrape with the various frameworks availible for python and node.js and we have had experience doing this in the past. This data can then be stored in either a MongoDB or some SQL instance.

1.2 What to predict

What type of job would be a good fit for an individual's skill set? How does this predicted job compare to the current job market? Do they meet the qualifications for this job? How does someone's education play a role in the job they will receive?

1.3 Process

1. Scrape Linkedin Data
   - (a) Profile data
   - (b) Use current job position as ground truth and use work history as a means to predict current job
   - (c) Useful attributes: industry, salary, education, number of connections, featured skills and endorsements (numeric and categorical), interests, location
2. Gather data from other sources
   - (a) Indeed.com offers an anonymized resume db

    i. Can be used to analyze frequent patterns in job descriptions

(b) The WSJ offers a dataset with data from colleges within the U.S.

    i. Salary expectations
    ii. Type of university

(c) Job openings - market demand for jobs

    i. Analyze current market demand for jobs
    ii. View which industries offer jobs for a given individual
        A. How many jobs for new grads
        B. How many jobs for experienced people
        C. Can be matched with outcome of previous analysis from linkedin and resumes

1.4 Programming Languages and Tools

- *Python* Will be used for web scrapping and data analysis.
    - *requests* used retrieving HTML pages.
    - *beautifulsoup4* used for contextual mining from HTML pages.
    - *TensorFlow, scikit-learn, Numpy, Pandas* used for analysis.
    - *TensorBoard, Matplotlib, pyplot* used for visualization.
- *MongoDB, MySQL, PostgreSQL* Will be used for storing data.

1.5 Algorithms

1. Various clustering Algorithms to plot tuples (people) to compare and analyze job market.
2. Classification for LinkedIn and Resume Data
    (a) Naive Bayes - predict job category/industry
    (b) Create itemsets of users employment path ($Bank > TechCompany > QuantFirm...$)
    (c) Multivariate Regression to predict salary
3. Frequent Itemset Mining
    (a) Analyze job descriptions
    (b) Mine keywords

2 Data Sources

https://www.kaggle.com/ludobenistant/hr-analytics
https://www.kaggle.com/wsj/college-salaries/data
https://www.kaggle.com/datasets?sortBy=relevance&group=featured&search=job
https://www.indeed.com/resumes
https://linkedin.com