Web Data Mining&Analysis: University Career Guidebook

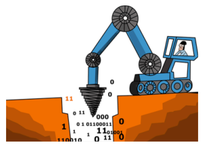
Nuochen Lyu, Zhao Chen, Minkang Yang
Department of ECE, Juniors, University of Illinois at Urbana-Champaign

INTRODUCTION

What are we Doing?

We design a scrapy to crawl career development related websites such as Linkedin to collect data about internships,connections, education experience of thousands of students in UIUC. We aim to provide you with an overview of UIUC students' career development status and strive to derive potential relationships between various aspects.

Checkpoint 1: Data Collection



The scrapy collects information of internships, number of connections, education experience in personal profiles and writes to a csv file for data analysis. The program is based on python with multiprocessing supported.

Checkpoint 2: Data Analysis

We analysis the data from thousands of students to find most popular internship companies, job titles, average connections, most common degree, school, looking for relationships between aspects such as number of internships and connections.

CONCERN

Will it takes too long?

Our scrapy supports multithreading and multiprocessing. In the graphical user interface, the user can specify how many pocesses to run. This feature enhances the efficiency of the process would be N times faster than single task scrapy.

Is it moral?

1. Our scrapy simulates normal human behavior to search for the results, which does not generate massive traffic to the target websites, so it will not cause pressure to the servers, and our data is only for research purpose.
2. No commercial usage.
3. Only run between 11 am to 6 am

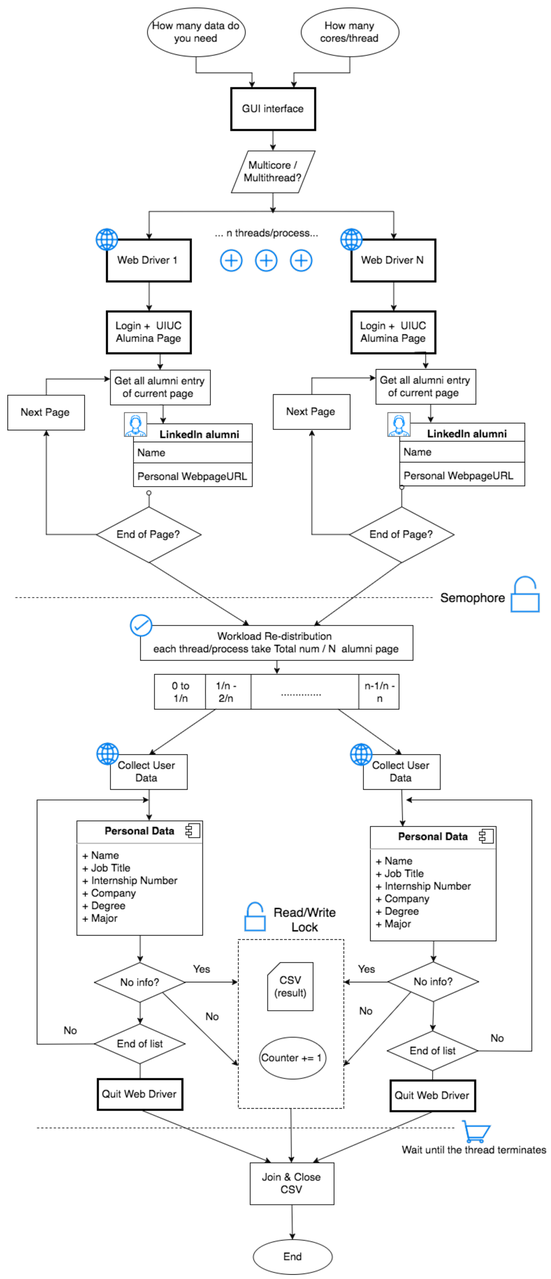
Is the data useful?

Websites such as Linkedin provide very limited analytical data about students' career information, through scrappy we can reorganize the date to find: the job demography, society trend, job market demand and new emerging industries.

How to deal with anti scraping techniques?

Our scrapy uses the python Selenium library to mimic real human actions in surfing Internet instead of sending raw requests or filling forms to interact with website server. Also we use multiple web drivers to avoid high frequency-browsing behavior. Therefore, usually it will not trigger some website defense mechanism against robot visitors.

METHOD



Single-Task < N-Thread < N-Process

1.0 x N 1.2 x N N x N

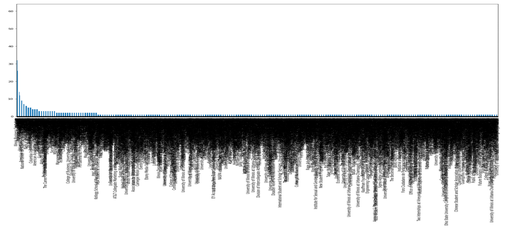
850 /H 1000 /H 3600 /H

*For i7(4 core 8 thread) Computer

RESULTS

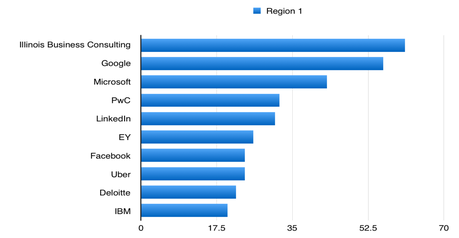
Total Data Size: 28,000
Takes 8 hours....

Power of data mining



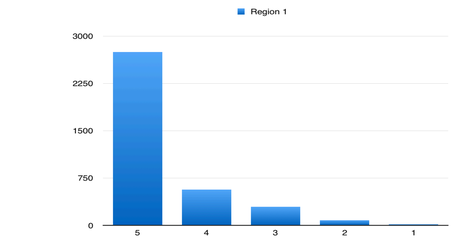
Company demography

In every 1000 students....



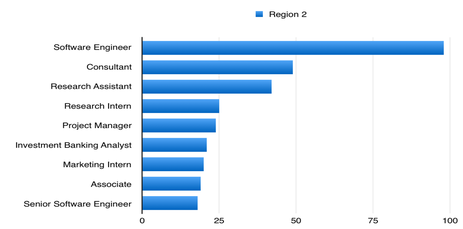
Number of interns

In every 4000 students....



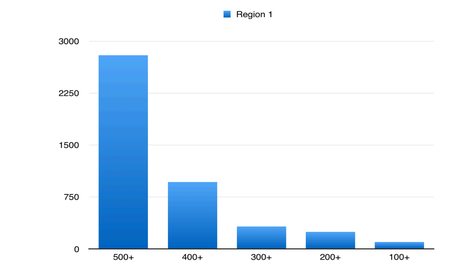
Titles

In every 1000 students....



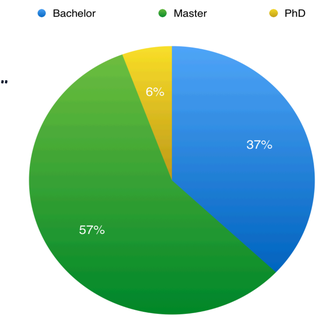
Number of friends

In every 4000 students....



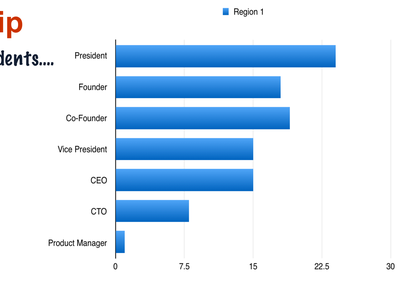
Degree

In every 1000 students....



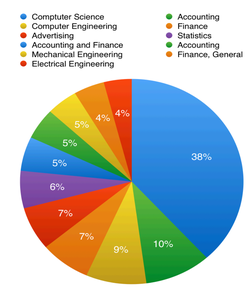
Leadership

In every 1000 students....



Majors

In every 1000 students....



Find on GitHub: scrapy_uiuc

ACKNOWLEDGEMENTS

pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN