Final Project

Yue Cao

Project Description:

Perform an analysis of "data scientist" jobs listed on job boards and on the employment pages of major companies. What are the most common skills that employers look for? What are the most unique skills that employers look for? Where are the types of companies that employ the most data scientists?

1. Introduction

As the modern society becomes more driven by digital activities and networks, data has become increasingly important in many industries. Besides software and Internet companies, even traditional industries, such as manufacturing and retail, are more dependent on high-quality data analysis. The demand for data scientists is rising accordingly.

Firstly, how to define a data scientist job? In general, Data Science is a field which studies how Information is gathered, what it conveys, and how that information can be converted into a valuable resource to generate meaningful insights for the betterment of a business as a whole[1]. Thus data scientists are the people who conduct these kinds of tasks in various industries.

To learn more about data scientist jobs, we perform a keyword-based skill search for data scientist jobs on two main online job boards: stackoverflow.com and dice.com. By extracting data scientist job information, such the job title, company name, skills, industry, etc., we can find out what are the most popular skills that employers preferred for data scientists and other unique characteristics.

2. Methods

2.1 Data Collection

In general, we collect data from online job boards using web scraping tools. This procedure contains a few key steps: job borads selection, web scraping and data cleaning.

There are many online job boards which we can search for data scientists jobs. We first tried indeed.com and monster.com. However, these two websites seem to have instable html nodes and varying structure of subpage when following the link corresponding to a job title, which cause much trouble when we do web scaping to extract data science job information. We notice that Stackoverflow.com and dice.com has nice and stable web structures for web scraping. Especially, dice.com returns more than 38,000 jobs when we search for data science jobs.

We mainly use `rvest` package by Hadley Wickham to srape information from webpages in R. To know the html nodes related to specific parts of a webpage, such as job title, company name, job description, we need to use the Chrome extension SelectorGadget. It is a web tool which returns the corresponding name of CSS or xpath when we click a specific part on a webpage.

2.2 Exploratory Data Analysis

Results

Discussion

Reference

[1]Shiv Shet, An Introduction to Data Science, Apr. 29, 2016 (https://dzone.com/articles/an-introduction-to-data-science, Accessed Oct. 9, 2017)