Predict 452 Section 55: Web and Network Analytics

Assignment 1: Automated Data Acquisition of Company Review Data

Jack Doig

Purpose of Study

The purpose of this automated data acquisition is to gather employee review data of their own company so as to be able to measure and understand the causes of employee attitudes towards their company and make organizational adjustments accordingly.

Management Problem

Understanding employee attitudes towards an organization is a vital management information problem for companies, particularly those competing for scarce human capital. If employee engagement and sentiment is understood, by location, over time and in a number of categories, management is much better placed to optimize decision making so as to retain a motivated workforce by adjusting work/life balance policies, compensation, investing in senior management training, and so on. Average attrition in highly skilled technology companies in the U.S is 22%, which is hugely expensive, so even marginal gains are very valuable.

Available Data

Glassdoor.com is in a dominant online position in the field of online employee company reviews. Glassdoor review data provides an anonymous channel for employees to review the company, provide feedback, and rate the company in a number of categories such as compensation and benefits, senior management and work/life balance. The rationale for gathering web data is that anonymous review sites like this afford employees an outlet without fear of recrimination from their employer.

The data acquisition process allowed me to gather the date of review, the pros, cons and employee's advice for management (all free text), as well as the ratings against categories (rating 1-5).

Scraping Process

The crawler I built accesses the initial page containing a company's reviews (this first page URL is configured), and then scrolls through a number subsequent of review pages, using the Glassdoor URL pattern, drilling down into each individual review and collecting the required data from the HTML using BeautifulSoup. This is then output to a csv text file for further analysis.

There were a number of challenges in scraping the Glassdoor site. Firstly there is a lot of Javascript generated content, this was overcome by using a JavaScript enabled browser within Python (Python Selenium). Without a JavaScript-enabled browser, Glassdoor.com would not return any content at all.

The second challenge was the requirement to be logged in, either with a validated email address or with social log in. This was handled by using Firefox as the browser and ensuring it was logged in to the site.

The slowness of the pages was another challenge, so it required sleep commands to be used while crawling the reviews.

This process was moderately successful in achieving its objective: my concern is about scalability. The process is slow, and Glassdoor.com have some protections against site scraping. A Captcha page is presented intermittently which interferes with the scraping process.

Next Steps

Given a large series of employee reviews over time, predictive and sentiment models can now be employed, augmenting the data with key company and macro-economic events as predictors, so as to infer the causes of positive and negative changes in employee engagement. Another interesting study could be to understand if there is a significant reactive element to reviews, or do strong positive reviews routinely follow strong negative reviews by employees who feel bound to defend their company.