A Scalable Pipeline for Analyzing Glassdoor Reviews: Insights into Employee Sentiment & Brand Personality Across Industries

Lab Group P1 - Team 4

Yap Ping 2101074 2101074@sit.singaporetech.edu.sg

Muhmmad Shahzad 2102060 2102060@sit.singaporetech.edu.sg

Eng Li Hao 2101704 2101704@sit.singaporetech.edu.sg

Mikhail Wee Ting Xu 2101138 2101138@sit.singaporetech.edu.sg

Tan Zong Wei 2100642 2100642@sit.singaporetech.edu.sg

Ong Yong Qi 2102646 2102646@sit.singaporetech.edu.sg

Abstract — Job seekers often use platforms like Glassdoor to view company branding and employee reviews. To date, few applications exist to scrape and analyze Glassdoor reviews. As such, we aimed to develop a data analysis pipeline that allows scraping and scalable analysis (using Hadoop MapReduce) of Glassdoor reviews. We analyzed a total of 243 companies across the 11 Global Industry Classification Standard (GICs). We successfully developed the pipeline and our findings indicate that the Airlines industry had the most positive average employee ratings on Glassdoor, while the Materials industry had the least positive. Overall, the development of our pipeline has provided valuable insights into the factors that influence employee satisfaction and sentiment. Additionally, this pipeline can be extended and used for larger scale studies.

Keywords - Python, Hadoop, Web-scraping, Glassdoor, Big-data, Scalable, Sentiment, GICs, MapReduce, Visualization, Data Analysis.

# Introduction

Glassdoor is a popular job search and career community platform that provides valuable insights about companies, job listings, and reviews by current and former employees. Glassdoor has become a go-to platform for job seekers who are looking for more than just job listings. It offers a variety of resources, including company ratings and reviews, and even photos and videos of the workplace. This information is provided by current and former employees, which helps job seekers get an inside look at a company’s culture and work environment. While Glassdoor may not necessarily be a social media platform, it allows its users to post reviews, rate the company and talk about their experience with a company, as current or former employees, similar to a social media platform. Job seekers would go to Glassdoor to find a company of interest to apply for. However there are many cases where a seemingly “good” company turns out to be a company that an individual would want to avoid applying for at all cost. Such cases happen because not enough research is done on the company or that the bad reviews did not surface at the top of the page; due to the way Glassdoor coded their website to only show the first 10 reviews on the page. Another problem about Glassdoor is due to its massive success as a job seeking/review platform. Glassdoor houses an estimated 50 million reviews to more than 1 million companies. This is one reason why job seekers would normally just scan through the page before applying for the company.

The main objective of this project is to develop a data analysis pipeline that incorporates scraping, analysis and visualization of Glassdoor company reviews. This pipeline will incorporate the use of Hadoop MapReduce to achieve scalability. Visualizing analyzed data from our pipeline will provide job seekers with an application that contains consolidated data of company reviews.

# Methods

The data collection, analysis and visualization can be broken down into 3 main steps (1) Scraping of Company information and reviews from Glassdoor (2) Analyzing reviews using Hadoop MapReduce and Python and (3) Visualizing analysis using Python-Flask framework. Link to all source codes and datasets used can be found at: [Yapping72/social-media-hadoop: A project for big data processing and analysis, focusing on social media data. (github.com](https://github.com/Yapping72/social-media-hadoop))

Text

Description automatically generated

Overview of the data-collection, analysis and visualization pipeline.

## Dataset collection

The glassdoor web scraper was developed in Python and utilizes Selenium and Beautiful soup to retrieve company information (Fig 2) and reviews (Fig 3) for each company. To bypass glassdoor anti-scraping mechanisms, each glassdoor scraper instance will login via Facebook before retrieving company information and reviews.

Graphical user interface, text, application

Description automatically generated

Sample of Company Information provided on Glassdoor. The scraper will attempt to retrieve all the information shown above.

Graphical user interface, text, application, email

Description automatically generated

Sample of 1 Glassdoor review, each review page holds 10 reviews, and the scraper manually navigates and stores all reviews for each company.

The web scraper works by generating URLs to retrieve both company information and reviews. Each company in Glassdoor is associated with a company name and company code.

https://www.glassdoor.sg/Overview/Working-at-Google-EI\_IE9079.11,17.htm. Here, *Google* will be the company name and *9079* will be the company's code.

Text

Description automatically generated

Using basic string manipulation, we can build URLs to retrieve company information using the company name and code .

URLs containing reviews follow the same logic. Each URL displays 10 reviews. To retrieve all reviews for a company, simply increment the ‘PX.htm’ in the URL.

<https://www.glassdoor.sg/Reviews/Google-Reviews-E9079_P2.htm?filter.iso3Language=eng>.

Text

Description automatically generated

Scraper will first retrieve total reviews displayed by glassdoor (not shown). It then builds a list of URLs by incrementing the PX.htm in the URL to navigate to all URLs containing reviews.

Once URLs have been obtained, the scraper will create instances of Google Chrome or Firefox browsers, navigate through URLs to extract HTML information using Selenium and BeautifulSoup4.

Both company information and reviews are retrieved and stored as dictionaries in JSON format.

Text

Description automatically generated

Sample of scraped company information stored using a dictionary.

To balance fault tolerance and performance (minimize I/O writes), the scraper will retrieve 1000 reviews (100 URLS) before dumping results to a JSON file.

Text

Description automatically generated

Sample of the JSON file containing reviews. Each JSON will contain up to 1000 reviews.

The scraped json files were then categorized based on their primary Global Industry Classification Standard  (GICS) and sorted into the following categories below.

*Text

Description automatically generated*

GICs industries, common classification method to categorize companies. Note that Institutions are not part of the GICs but were included. Institutions contain Singapore universities and polytechnics.

Graphical user interface, text

Description automatically generated

File structure for each individual company, each JSON file contains 1000 reviews.

## Hadoop MapReduce

Diagram

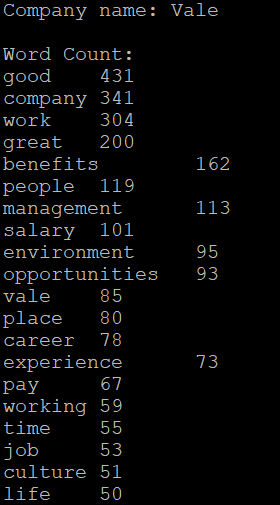
Description automatically generated

Hadoop MapReduce overall structure

[*https://lucid.app/lucidchart/dcd7b0b8-1a48-4129-89f7-2f6654d87b17/edit?viewport\_loc=-12060%2C-4648%2C20062%2C9228%2C0\_0&invitationId=inv\_c1ae095a-d38c-411d-9a46-50c52c000666*](https://lucid.app/lucidchart/dcd7b0b8-1a48-4129-89f7-2f6654d87b17/edit?viewport_loc=-12060%2C-4648%2C20062%2C9228%2C0_0&invitationId=inv_c1ae095a-d38c-411d-9a46-50c52c000666)

For the data pre-processing of this application. We used Hadoop mapreduce operation to create a word count operation. MapReduce is a programming model that allows for the distributed processing of large data sets across clusters of computers. Hence, since we need to process a large amount of data that we have scraped, MapReduce is the best way to do it. We compiled a .JAR file from our virtual environment using eclipse. From there, we send the .JAR file to the big hadoop cluster where the operation takes place. In this case, we are using MapReduce to process the JSON files that are found inside a folder with the company name. These companies folders are located within another folder which groups them up into the type of industry they belong to. The MapReduce job is divided into two phases: the map phase and the reduce phase. The mapper takes the input data and outputs key-value pairs, where the key is a word in the review and the value is the number 1. We have 3 different processes for MapReduce. In the first case, we are only interested in values from the “review title”, “pros” and “cons” keys.

The reducer then takes these key-value pairs and combines them by summing up the values for each key. The reducer then takes the top 20 words with the highest count and outputs them to a text file as shown in Figure 11. It is also worth noting that words such as “is”, “the”, ”and” and other filler words have been excluded. Special symbols such as “,” or “!” have also been filtered out. The word count is a sum of all the words found in all the JSON files in a single company folder. The output only contains words that are of importance.



Snippet of the output file for word count

In the next case, we are only interested in the ratings for each review. In particular, those reviews that have extreme ratings, which are either a “1.0” or “5.0”. Figure 4 below shows the output of the textfile calculating the number of “1.0” and “5.0” ratings for each company. From these 2 different types of data, we have information to analyze.

*Text

Description automatically generated*

Snippet of the output file for ratings analysis.

These JSON files were used for downstream analysis and visualization.

## Data analysis

A total of 243 companies listed on Glassdoor were scrapped for reviews, resulting in the retrieval of 1,255,054 reviews. The companies were categorized into nine industries using the GICs index - Information Technology, Financials, Consumer Discretionary, Industrials, Healthcare, Energy, Materials, Airlines, and Institutions.

The Information Technology industry had the highest number of companies scrapped (56), followed by Financials (30) and Consumer Discretionary (39). The Airlines and Institutions industries had the lowest number of companies scrapped, with 12 each.

When looking at the number of reviews found, the Information Technology industry also had the highest number (474,476), followed by Financials (270,397) and Consumer Discretionary (161,290). The Healthcare and Energy industries had the lowest number of reviews found, with 91,917 and 43,045, respectively.

The graph  below shows the number of companies scrapped per industry as a bar chart and the number of reviews found for each industry as a line chart. The bar chart clearly shows the differences in the number of companies scrapped for each industry, while the line chart provides a visual comparison of the number of reviews found for each industry. Overall, the data and graph provide insight into the distribution of companies and reviews by industry on Glassdoor.

Chart

Description automatically generated

Overview of companies and reviews scrapped.

To streamline downstream analysis, word count and review count obtained from Hadoop Analysis were consolidated using Python Pandas library.  Python was also used to compute overall industry information in other fields like percentage of X-star reviews and median reviews.

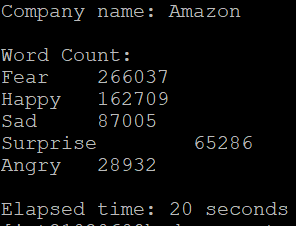
Text

Description automatically generated

Pandas library was used to consolidate industry information (left) and sample of compiled Google data-set (Right).

On top of generating word and reviews count, we also used MapReduce to perform sentiment analysis. Employee reviews were mapped to a comprehensive list of over 8000 words to determine the sentiment of the word. Each word in an employee review would be labelled as either ‘Angry’, ‘Fear’, ‘Happy’, ‘Sad’ or ‘Surprised’. By counting the occurrence of  words associated with these labels, we can develop a brand personality associated with each company.

The figure below, shows an example of the MapReduce sentiment analysis. Here we see that Amazon employee reviews are predominantly Fearful (266037) and Happy (162709). Both Fear and Happy appear to be the dominant emotions based on employee reviews.



Snippet of the output file for sentiment analysis.

With the data collected and processed, various analyses will be conducted via Hadoop and Python. ChartJS, Flask and Python will be used to plot various graphs depicting the relationship between elements such as industries and average ratings. Further details on the analysis insights and results will be explained in the later sections.

## Results and Insight

Following our analysis we observed that the highest rated industry was the Airlines industry, with an average rating of 3.926. Information Technology (3.877) and Institutions (3.859). Industrials and Healthcare were the second and third lowest rated industries, with average ratings of 3.599 and 3.616, respectively. Lastly, the lowest rated industry was Materials, with an average rating of 3.584.

Chart, line chart

Description automatically generated

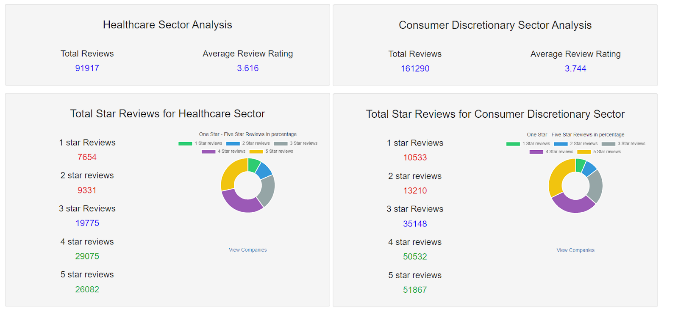
Average review rating by industry.

To better visualize our analysis, we developed a dashboard using the Python-Flask framework. The landing page (Figure 17) provides a stacked-bar chart that shows the different percentages of one- to five- star reviews, allowing for the composition of reviews to be visually displayed, enabling easy comparisons across industries.

The landing page also features information pertaining to the Energy and Airlines industry. For each industry, we have incorporated doughnut charts to provide a side-by-side analysis on each page, complete with color-coded review counts that enable easy visualization. The red hue denotes 1-2 star reviews, while green represents 4-5 star ratings, providing an intuitive way to interpret the data at a glance. This visualization strategy is deployed consistently across all industries, facilitating a comprehensive understanding of the review landscape for each industry.



Landing page of the dashboard, Shows overview of all industry as a stacked bar-chart. It also shows information specific to the Energy and Airline industry.



Total Star reviews for Healthcare Sector and Consumer Discretionary Sector

*Graphical user interface, application

Description automatically generated*

Total Star reviews for Materials Sector and Institutions Sector

*Graphical user interface, application

Description automatically generated*

Total Star reviews for IT Sector and Industrials Sector

*Graphical user interface, application

Description automatically generated*

Total Star reviews for Financial Sector

**Overall industry Analysis**

With five-star reviews comprising **41.815%** of their total reviews, the airline industry would be the favored industry to work at due to having the highest percentage of 5-star reviews amongst all industries. The large percentage of 5-star reviews indicates largely satisfied employees within the industry. As for the worst industry to work on, materials garner the most 1-star reviews at a total of 8.418%.

**Airlines**

Airlines have the highest percentage of 5-star ratings, indicating that employees generally have a positive experience with these companies. One reason why airlines may have a high level of employee satisfaction is that they often offer significant benefits to their employees. These benefits may include free or discounted air travel, health insurance, retirement plans, and other perks that can improve employees' quality of life.

**Financials**

Financial companies have the lowest percentage of 5-star ratings, indicating that their employees are comparatively less content with their jobs. There could be various reasons why financial employees may be less satisfied with their jobs. For instance, financial companies often have high-pressure work environments that demand long hours and have strict performance targets. Additionally, employees may feel underpaid, undervalued, or lack opportunities for career growth, which can lead to lower job satisfaction.

**Healthcare**

Healthcare and Industrials have the highest percentage of 1-star ratings, indicating that employees have a very negative experience with these industries. In the healthcare industry, several factors can contribute to low job satisfaction among employees. For instance, healthcare workers may face long working hours, high-stress levels, and frequent exposure to diseases, which can have a significant impact on their physical and mental well-being. Additionally, they may feel underpaid, undervalued, or lack opportunities for career growth, which can lead to lower job satisfaction.

**Consumer Discretionary**

The Consumer Discretionary industry has the highest percentage of 2-star ratings, suggesting that employees have a somewhat negative experience with these companies. This could be due to its fast-paced and constantly evolving nature, which can create a challenging work environment for employees. Additionally, companies in this industry may have to deal with rapid changes in consumer preferences and economic conditions, which can add to the pressure on employees.

**Energy and Materials Industry**

The Energy and Materials industries have the lowest percentage of 5-star ratings, indicating that employees are generally less satisfied with companies in these industries. This could be due to these industries having to deal with challenging and often hazardous working conditions, which can create physical and mental strain. Additionally, the cyclical nature of these industries can create a sense of job insecurity among employees, which can negatively impact their job satisfaction.

**Financials and Industrials**

The Financials and Industrials industries have the highest percentage of 3-star ratings, suggesting that users have a neutral or average experience with these companies. One possible reason for the neutral or average rating could be that employees in these industries have a balanced view of their work experience. They may find some aspects of their job to be positive, while other aspects may be negative. For instance, employees in the Financials industry may appreciate the competitive compensation packages offered by their companies, but may also feel pressure to meet sales targets and deadlines, which can create stress and impact work-life balance. Similarly, employees in the Industrials industry may enjoy the physical work involved in their jobs, but may also face challenges related to safety and long work hours.

**Information technology**

The Information Technology industry has the highest number of reviews and companies scrapped, indicating that there is a lot of activity and competition in this industry. One reason could be that the Information Technology industry is constantly evolving, with new technologies and software being developed and released regularly. This means that companies need to keep up with the latest trends and developments to stay competitive, leading to a high level of activity and innovation within the industry. The Information Technology industry is a highly sought-after field for job seekers, which creates a lot of competition among companies to attract and retain top talent. As a result, companies need to work hard to differentiate themselves from their competitors, leading to a higher volume of reviews and feedback from employees.

More in-depth analysis on each industry can be achieved via the ‘view companies’ button. Each industry page features a bar chart depicting a breakdown of one to five star reviews for each company within the industry, allowing for comparisons on which companies have better reviews in the industry. In addition, a word cloud and bar chart is shown, providing a visual summary of frequently used words, allowing for possible trend identification. (See figure 22).This provides a method of visualization that is simple to understand as word clouds are easy to read and is a visually appealing way to summarize textual data. Thus, allowing users to quickly grasp the most common sentiments expressed in the industry itself.

Chart

Description automatically generated

Deeper insights into the energy industry sector.

Lastly, if users wish to access detailed information on individual companies, they can select the company from the dropdown list. Which will reveal company specific information such as company's description, headquarters, mission statement, revenue etc. (Figure 23) This provides a quick method to view glaring red flags e.g., if there are a significant number of 1-star reviews, it may indicate an overall negative experience working at the company, thus enabling users to quickly gain a comprehensive and succinct overview of the company. Sentiment analysis for the particular company will also be displayed.

*Graphical user interface, application

Description automatically generated*

Summarized data for Amazon with sentiment analysis. Company information and Sentiment analysis for Amazon shown on the right.

**E. Conclusion**

     We have successfully developed a scalable Glassdoor review analysis pipeline that provides a solution for analyzing large volumes of company reviews and information. This application also allows for the communication of insights through data visualization. Based on our analysis, the airline industry appears to provide greater employment satisfaction, with the highest average rating and percentage of 5-star reviews, while the Materials industry has the lowest average rating and percentage of 5-star reviews.

In future work, it will be valuable to expand the scope of the analysis to include a larger sample of companies and industries, as well as to incorporate more detailed information on company-specific factors that can affect employee satisfaction and sentiment. This can be achieved by simply scraping more companies and passing it through our Hadoop and Python analysers. Overall, the development of our pipeline has provided valuable insights into the factors that influence employee satisfaction and sentiment, and has the potential to inform decision-making for both employers and employees.

|  |  |
| --- | --- |
| **Task Allocation & Contributions** | |
| Yap Ping | * Code for Glassdoor reviews scraper * Code for Glassdoor company information scraper * Code to merge Hadoop and python analysis * Scraping and Report |
| Eng Li Hao | * Scraping * Code to merge python analysis and report * Presentation Demo |
| Shahzad | * Code for MapReduce * Create Hadoop analysis * Scraping * Report |
| Zong Wei | * Scraping * Visualisation * Report |
| Mikhail | * Scraping * Visualisation * Report |
| Yong Qi | * Scraping. * Video * Report |

1. Task allocation and contributions

##### F. References

[1]<https://datareportal.com/global-digital-overview#:~:text=A%20total%20of%205.16%20billion,of%20the%20world's%20total%20population>.