Retrieving and Processing Information on Company Reviews

Information Processing and Retrieval - 2022/23

Bruno Rosendo
Faculty of Engineering of the University of Porto, Portugal
up201906334@fe.up.pt

João Mesquita
Faculty of Engineering of the University of Porto, Portugal
up201906682@fe.up.pt

Rui Alves
Faculty of Engineering of the University of Porto, Portugal
up201905853@fe.up.pt

ABSTRACT

In order to develop a functional search system, it's necessary to have a rich and structured dataset as its basis. In the theme of company reviews, the goal of the present work is to build a pipeline capable of collecting and preparing a dataset ready to be used in this context, starting from a dataset available on Kaggle [2]. To achieve that, we must first go through a data analysis phase to properly understand the data sources we're working with and how to improve them. Then, we can establish various small steps that the mentioned pipeline will perform, such as data cleaning, processing, and refinement.

KEYWORDS

data, datasets, company, companies, information, retrieval, processing, analysis, feup, search, Indeed, ratings

1 INTRODUCTION

The present work was made within the course of Information Processing and Retrieval, part of the Master's Degree in Informatics and Computing Engineering. This is the first part of a project whose goal is to develop a search system, powered by a rich and structured data collection.

The group chose to work with a dataset about company reviews, taken from Indeed, which was freely available on Kaggle [2]. The motivation is to develop a search system capable of providing information on a considerable number of companies.

The paper starts by describing the original dataset, followed by data characterization, the used processing pipeline, and a conceptual model representing the final collection, which is the result of this first part. Finally, it exposes some conclusions and the expected future work.

2 DATASET

The original dataset is composed of a large list of companies, accompanied by relevant information for people interested in working for them, including reviews from their employees, salary, and guidance on the interview process. This proved to be a very rich dataset with enough information for the project's goal, so there was no need to complement it with any additional data.

2.1 Company Reviews

The dataset contains more than 17 thousand companies (around 25MB), distributed across a large number of job fields and mostly located in the United States. The provided information was all taken from the Indeed website, a widely used job platform.

Each company review has around 20 fields characterizing them which can be arranged into three main groups:

Information about the company: Besides categorical fields like name and industry, we have textual fields, including a description of the company, the location of its headquarters, a revenue range, and the number of its employees. Some of these don't have the most appropriate format and would ideally be numerical fields.

Reviews: Each company has a given number of reviews associated, which are summarized by a set of textual fields. These include salary per role and ratings on different categories, like locations or CEO approval, as well as a general average rating. Most of them are stored as serialized documents with a great variety of categories, meaning that some data processing and exploration are needed in order to properly structure and understand these ratings.

Interviews: The companies also have reviews on their interview process. Here, we have two categorical fields exposing the difficulty and experience level of the interview, as well as a textual field describing its duration, which would ideally be a numeric field.

With this information, we should be able to implement a pipeline capable of cleaning the dataset and processing it into a more consistent and reliable format.

2.2 Data Source

The chosen dataset was collected from Kaggle [2] and is free to use, without any kind of authorization needed. The author stated that he built the dataset by scraping reviews from the Indeed website.

There are other people who have used this dataset for their own studies, which can be seen in the related notebooks on Kaggle's page [1] and shows that good projects can be done with this data source.

2.3 Data Characterization

In order to properly understand the dataset we're working with, it's essential to go through the process of data analysis and characterization. To start, we can see in Table 1 that most of the fields have a considerable fraction of missing values, so we cannot assume their presence when working with the dataset, especially the *happiness* field. More interestingly, around 2% of the companies do not have a name identifying them, which might not be very interesting when developing a search system. It's also important to note that the dataset doesn't contain any duplicate companies, all the names are unique.

Furthermore, we can calculate some statistical information regarding numerical fields, which is presented in Table 2. Note that some of these fields had to be previously processed in order to

Table 1: Number and ratio of missing values on each field of the dataset

field	missing	field	missing
name	338 (2%)	**ceo_approval**	5828 (34%)
rating	1434 (8%)	**interview_count**	5551 (33%)
reviews	1556 (9%)	**headquarters**	1817 (11%)
description	1 (0%)	**happiness**	12463 (73%)
industry	1846 (11%)	**roles**	9922 (58%)
employees	2027 (12%)	**salary**	5427 (32%)

retrieve their numerical values (e.g. the ratings on the companies' locations).

Table 2: Statistics for the dataset's numerical fields

field	mean	std	min	max	median
rating (0-5)	3.52	0.61	1.0	5.0	4.0
ceo_approval (%)	69.98	14.71	6.0	100.0	72.0
happiness (%)	61.91	9. 43	28.0	97.0	62.0
salary ($/hour)	12.43	5.72	1.77	83.47	11.71
locations (0-5)	3.96	0.62	1.52	5.0	4.0
ratings (0-5)	3.29	0.54	0.0	5.0	3.28

Even though the individual statistics of each field are useful, it's an even better idea to look at how they correlate with each other. [1] For starters, let's see two bar charts of the industries with the best and worse ratings in Figure 1 and Figure 2.

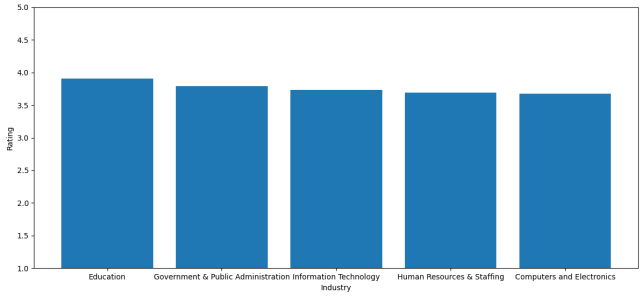

Figure 1: Industries with the best ratings

It's also interesting to see what the rating distribution looks like in the histogram of Figure 3. It bears a resemblance to a normal distribution around the 3.5 value!

One of the most important affairs of the dataset is understanding how the ratings are affected by the information of the companies. It's very interesting to see how well the CEO approval correlates with the respective company's rating in Figure 4's box plot. On the other hand, the revenue has almost no effect on the ratings, as we can see in Figure 5. Finally, we can use a heatmap to visualize how some of the rating factors correlate with each other, in Figure 6.

[1] All the plots created for this work are available in Google Drive, including some not present in the report.

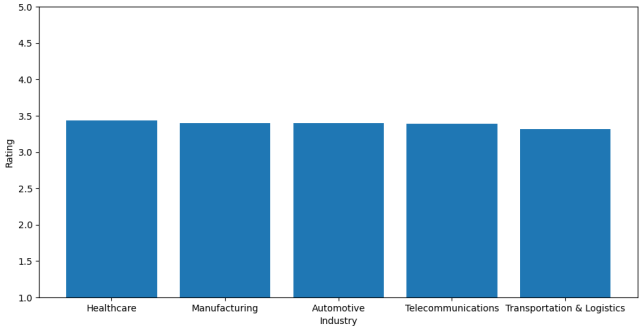

Figure 2: Industries with the worst ratings

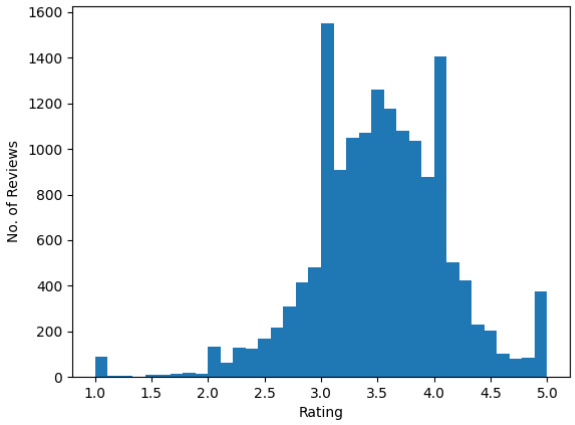

Figure 3: Distribution of the company ratings

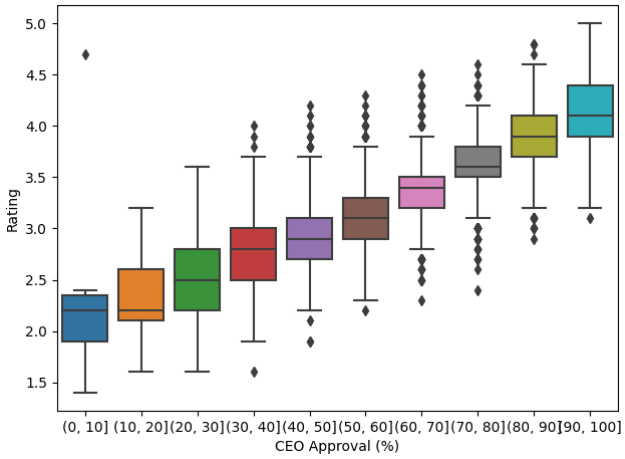

Figure 4: Correlation between the company ratings and the respective CEO approval

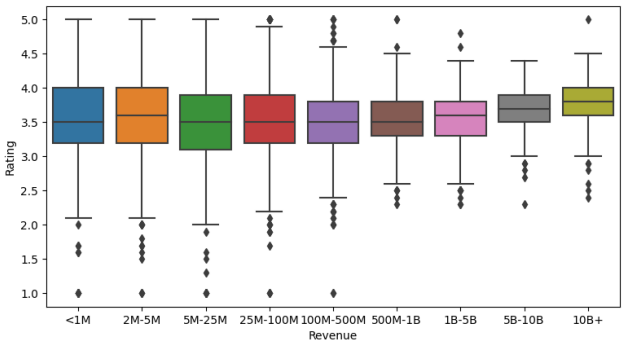

Figure 5: Correlation between the company ratings and the respective revenue

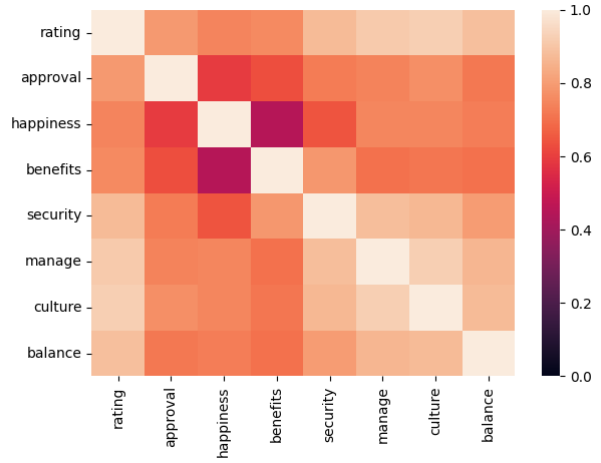

Figure 6: Correlation between the different rating factors

3 DATA PROCESSING PIPELINE

In order to obtain a ready-to-use dataset with relevant information and a well-defined structure, it is necessary to go through various steps that clean and format the original data, some of which depend on each other and should be executed subsequently. For this reason, it's important to define a pipeline where we can easily show and declare all the steps needed for the construction of the final dataset.

In Figure 7, we can see what the data pipeline looks like for this project. Each step is explained in the following sections.

3.1 Data Collection and Cleaning

As seen at the start of the pipeline and explained in Section 2, the dataset taken from Kaggle [2] was enough for the objectives of this work. Therefore, that is the only data source given as the pipeline's input.

After loading the dataset, the first step is to clean the data, which can be divided into incremental small tasks. First of all, the rows containing unidentified companies are removed. By unidentified, we mean companies missing the *name* field. At this time, the useless column *website* is also removed, since it only provides information about which websites the company possesses without giving any URLs.

After removing unwanted data, the next task is regarding the company's different ratings, given by the fields: *ratings*, *happiness*, *locations* and *roles*. As mentioned in Section 2, they are stored as documents whose values represent the rating of a given category, role or location. Although they resemble a JSON format, their keys are wrapped with single quotes which are invalid JSON syntax. Additionally, all their values contain numerical information stored as strings (text). Therefore, all of these fields are fixed, by replacing the single quotes with double quotes and converting the values to their respective types (either floating point or integer).

3.2 Data Processing

The next tasks of the pipeline involve processing the data. Contrary to data cleaning, they don't simply remove or quickly fix irrelevant or wrongly formatted data. Instead, they transform the data into a more structured and reliable format.

One of the first processing tasks is to transform the fields containing a numerical range that is described in a textual manner. The *revenue*, a monetary range indicating the revenue of the company in US dollars, and the *employees*, a range of the number of employees in the company, are the fields falling into this category. In these cases, the textual fields are replaced with categorical values, discriminating the range in which the company is described. For a more concrete example, the *revenue* field can be categorized by companies with less than a million USD, between 1 and 5 million, and so on, ending with a category for a revenue of more than 10 billion dollars.

Another important step is to fix some incorrectly typed fields. In the original dataset, the *reviews*, *ceo_count*, *ceo_approval* and *interview_count* fields hold numerical information. However, they are described in textual fields, so it's helpful to parse and convert them into simple numerical fields.

The *salary* field undergoes a similar treatment. The difference here is that the salary values (which are specified by role) are given by different time frames (i.e. salary per hour, month, year, etc.). To have a consistent and comparable salary rate, they're all converted into *dollars per hour*. In the case of the *happiness* field, a document of ratings related to personal topics, its percentual values are converted to a 0-5 score, similar to the other ratings.

After reformatting all of these fields, the pipeline creates a new field called *custom_rating*, a heuristic that tries to summarize the overall company rating, calculated with a weighted mean of all the other different ratings. The result of this step is then saved as a new CSV file.

Finally, the current CSV dataset is converted and saved in the JSON format, a document-based and well-structured format that is easy for humans to read and, more importantly, easy for machines to parse and generate. To improve the structure, the fields related to CEO and interview information are joined into their respective documents [2].

[2]The processed JSON dataset can be consulted in Google Drive.

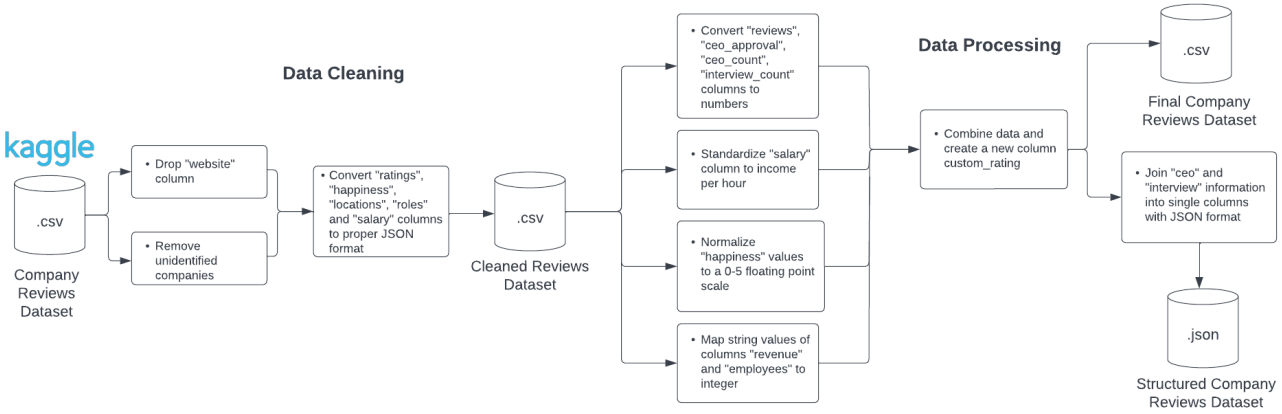

Figure 7: Data Processing Pipeline

3.3 Final Dataset

The result of the data pipeline is a JSON dataset stored in a *.json* file. It contains many different documents, each representing a company, all of its information and respective reviews and interview information. In order to visualize how these entities are related, the conceptual model in Figure 8 was built.

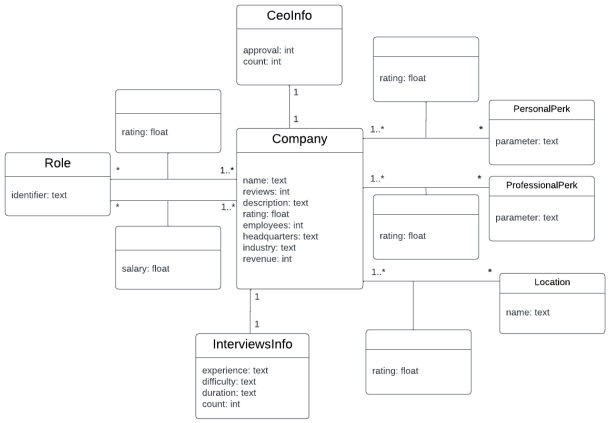

Figure 8: Conceptual Model of the Final Dataset

As shown in the figure, it is composed of the following entities:

- **Company**: a company has a few informative fields about itself, such as name, industry and revenue. Additionally, this entity contains its number of reviews and the average rating.
- **Role**: each role is associated with one or more companies and a given company can have as many roles as needed. Each of these relations has a rating and salary associated with it.
- **Location**: location of the company's different offices, associated with the respective rating. A location can be shared across different companies.
- **PersonalPerk**: a personal perk is a characteristic of the company that might increase or decrease the life quality of their staff, such as *Purpose* or *Flexibility*. Each relation between a company and a personal perk has an associated rating.
- **ProfessionalPerk**: a professional perk is a characteristic of the company that impacts the professional career of the staff, such as *Compensation* or *Job Security*. Each relation between a company and a professional perk has an associated rating.
- **InterviewsInfo**: information about the company's interviews, including the difficulty, expected experience and duration.
- **CeoInfo**: information about the approval of the company's CEO based on staff reviews.

4 CONCLUSIONS AND FUTURE WORK

This paper presented the process of collecting and preparing a structured dataset for the basis of a search system. We went through a data analysis phase that allowed us to better understand the original data and how it could be improved for our use case. With this information, a pipeline was developed to establish how the final dataset is generated and perform all the necessary tasks.

The following steps for the development of the search system are the use of information retrieval tools to obtain useful information from the dataset and the construction of an interface allowing for intuitive use of the system. The user should be able to search for companies that fit their needs, given certain parameters, as well as search for a specific company and see the opinion given by people who worked there, amongst other search scenarios.

REFERENCES

[1] Yael Man. 2021. Company review rating factors. https://www.kaggle.com/code/yaelman/company-review-rating-factors.
[2] Reza. 2020. Company reviews. https://www.kaggle.com/datasets/vaghefi/company-reviews.