COMP20008 Elements of Data Processing: Project 1

Ming Hui Tan

December 4, 2020

Introduction

The aim of this project is to crawl and extract information from a number of media reports on Rugby games contained in the web server http://comp20008-jh.eng.unimelb.edu.au:9889/main/ and use that information to improve our understanding of the teams' performance.

All codes written and used in this project can be found in the **submission.py** file.

1 Web-Scraping and Data Cleaning

1.1 Extracting URLs and Headlines

We begin by specifying the initial page (seed url), http://comp20008-jh.eng.unimelb.edu.au:9889/main/, from which we obtain the HTML text and begin our crawling. After obtaining the HTML text, we parse the HTML using a Python package called BeautifulSoup and find all available links in that HTML text. The available links are then joined with the seed url and appended to the `to_visit` list.

To scrap the headline of each article, we loop through each link (and deleting the link after visiting to avoid re-visiting the same link) in the `to_visit` list to obtain and parse the HTML text. We then find the headline tag ('h1') from the HTML text and extract all the text associated with the tag as the headline of the article. The headline in each link that we visited along with its URL are recorded and converted into a dataframe named `url_data`.

In total, we visited and scraped headlines and URLs from 147 links (excluding the initial page). The complete list can be found in **Task1.csv**.

1.2 Extracting Team Names and Match Scores

We first find all the team names that we want to match from **rugby.json** and store them in the list named `rugby_teams`. Then, we loop through each URL in `url_data` and obtain the body text in the article and join them with the headline of the article and store this resulting string in `content`.

To find the first team name in the article, we first compile a regular expression pattern named `search_team_re` where we join all the team names in `rugby_teams` separated by '|'. We will then use `search_team_re` to find all team names mentioned in the article that match the team names in `rugby_teams`. The main team of the article will be the first successful regex match that we found in the article. If there is no matching team name, we will discard the article from our dataset and proceed with the next URL.

To find the match-score of each article, we first use the regular expression r'\d+\-\d+' to match all possible match scores in each article and store the list of possible match scores in `scores`. We then define a function called `largest_score` where we compare each score in `scores` based on their weighted score using the following formula:

$$W_{score} = \alpha \times S_1 + (1 - \alpha) \times S_2$$

where S_1 and S_2 denote the first and the second score respectively, while α denotes the weight of the first score. In this case, we let $\alpha = 0.5$. The first and second score are extracted by searching and extracting the first group of text that matches the pattern r'(\d+)-' and r'-(\d+)' from each score in `scores`. The match score with the highest weighted score is selected and recorded as the most relevant match score in the article (with exception of scores that are

impossible to achieve such as '1994-96'). Similarly, we will also discard any article where zero match score is mentioned.

For each URL visited, we record the headline, URL, relevant team name and match score into a dataframe named `match_data`. After discarding irrelevant articles, we arrive at a total of 64 records in `match_data`. The complete list can be found in **task2.csv**.

1.3 Finding Absolute Game Difference

To find the average absolute game difference, we first extract the first score (S_1) and second score (S_2) of each record using the regex match patterns r'(\d+)-' and r'-(\d+)' respectively. Then, we calculate and record the absolute game difference of each record in a separate column in `match_data` using the following formula:

$$G = |S_1 - S_2|$$

Now, we group all the records by team names, calculate their respective average absolute game difference and store this result in `avg_game_diff`.

Overall, we observe that New Zealand has the highest average game difference at 30.5 whereas Scotland has the lowest average game difference at 5.0. The complete list of average game difference of each team can be found in **task3.csv**.

2 Results and Analysis

From `match_data` we count the number of records associated with each team, find the top 5 teams with the highest count and tabulate them in `team_article_data` (in descending order). We then plot the result of the top 5 teams that articles most frequently write about.

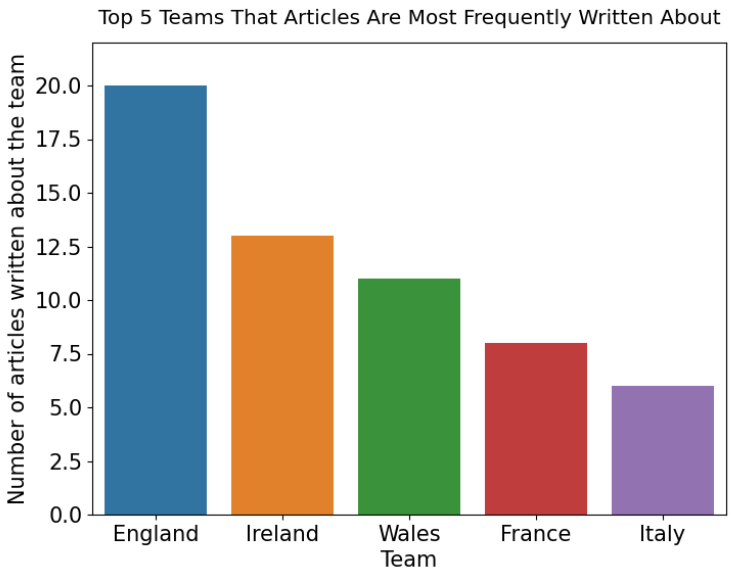

Figure 1: Top 5 teams that articles are most frequently written about.

Here, we observe that England is the most written about team followed by Ireland. Italy however is the least written about team among the top 5 teams. Let's see if there is any association between this and their average game difference.

We first retrieve the count of number of records associated with each team and store them in `full_team_data`. Then we will join `full_team_data` and `avg_game_diff` and store it in `join_data`. We then plot the result from `join_data` as follow.

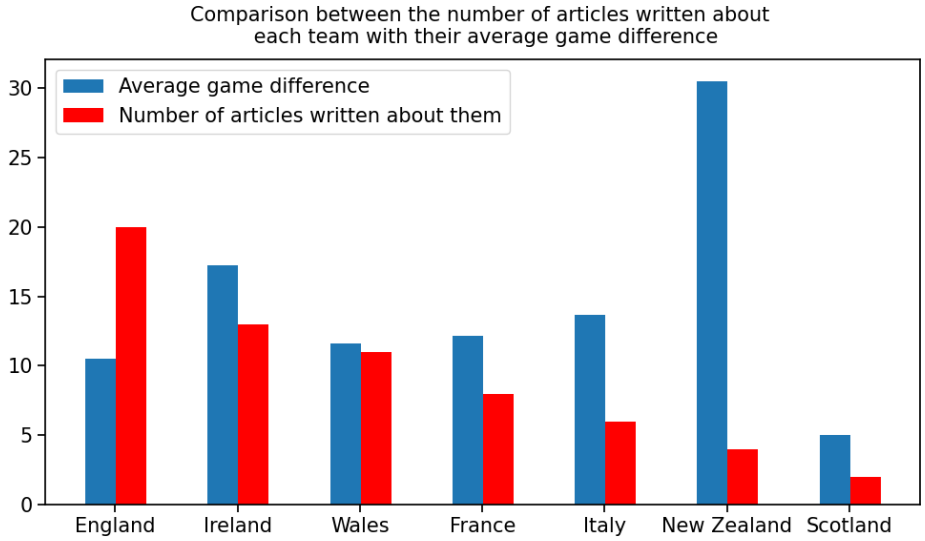

Figure 2: The number of articles written about each team against their average game difference.

Here we observe that New Zealand despite being one of the least written-about team, has the highest average game difference compared to the rest. England on the other hand has a relatively similar average game difference compared to most teams. On first glance, there seems to be some correlation in the middle section on the graph. However, by performing Pearson correlation test, we obtain the Pearson correlation coefficient, $\rho = -0.16097$, which indicates that there is very weak to none correlation between the two data.

3 Discussion

3.1 Appropriateness of Associating the First Named Team in the Article with the First Match Score

The method of associating the first name team in the article with the first match score may not be appropriate in this case because the rugby teams are named after countries. This may lead to wrong names being extracted especially when country names are often associated with the location of the sports events in sports-related articles. Besides, most articles mention the country's team in adjective form (eg. The French as the team from France) which we may not capture by just matching the first team name in the article.

On the other hand, extracting the first match score as the relevant match score is reasonable as relevant match score is mentioned early in most sports-related articles to attract the attention of the readers. However, there are cases where the match scores are mentioned in words instead of numbers which may lead to incorrect information being extracted.

Therefore, associating the first named team with the first match score in the article may only be accurate in certain articles where the team name is mentioned as noun early in the article and the match score is mentioned in numbers. In other cases, however, this method will fail to record correct information.

3.2 Suggestions of Method to Determine The Result of the Match in the Article

3.2.1 Sentiment Analysis

We can use Machine Learning (ML) based sentiment analysis where we train a ML model to recognize the sentiment based on the words and their order using a sentiment-labelled training set. To implement this method, we will need to extract the relevant team name in the article and then gauge the polarity of the sentiment in the body of the article using Natural Language Processing tools where positive sentiment is associated with winning and negative sentiment is associated with losing.

The advantage of using this method is that it is efficient at analyzing a large number of articles accurately provided a large labelled training dataset is available. The disadvantage of this method, however, is that we need to

create a large labelled training dataset to train the ML model for this method to be effective. Otherwise, the model will easily misclassify the result of the article.

3.2.2 String Match "Win" and "Loss" Synonyms in Articles

We can also implement the string matching method where we first tokenize the article followed by matching words that are synonyms of "win" and "loss" such as "victory" and "defeat". We will also need to keep count of the number of occurrence of "win" and "loss" synonyms in the article and compare them at the end. The word which occur more frequently in the article will determine the result of the match in the article.

The advantage of using this method is that it is easy to implement as we are just comparing the frequency of "win" and "loss" related words in each article. The disadvantage, however, is that if there are articles with equal number of both "win" and "loss" related words, this method will not work.

3.3 Other Information to Gauge Team's Performance

Using similar method described in 3.2.2, we can tokenize the article followed by finding the frequency of words that are associated with positive performance (e.g. "confident", "proud", "awesome") and negative performance (e.g. "struggling", "disappointing") in the article. Then we can create a scale to determine the performance of the team in the articles by the following formula:

$$P = \frac{N_p - N_n}{N}$$

Where P represents the performance of the team, N_p represents the number of occurrence of words associated with positive performance, N_n represents the number of occurrence of words associated with negative performance and N represents the sum of N_p and N_n. Thus, we can gauge team performance based on the P value where +1 indicates flawless performance, 0 indicates neutral performance and -1 indicates disappointing performance.

4 Conclusion

To summarize, from the result of webcrawling and webscraping, we found that England is the most written about team whereas Scotland is the least written about team. Other than that, New Zealand has the highest average game difference whereas Scotland has the lowest average game difference compared to the rest of the teams. However, these results may not be accurate due to the limitations of matching first named team and first match scores in each articles. For future studies, we should implement different methods of finding relevant team names and match scores to compare with the method mentioned in this paper in terms of the relative accuracy of the result obtained.