COMP20008 Elements of Data Processing Project 1 Report

**Crawling Method**

In this project the crawling method used was done with request library, You have your base\_url called 'http://comp20008-jh.eng.unimelb.edu.au:9889/main/' and you have a seed\_item which is the ‘index.html’, which is suffix of the actual website url, the seed\_url is the concatenation of base\_url and seed\_item. You then use the request library using requests.get(seed\_url) to contents of website pages.

**Brief summary of the output for Task 1**

The summary output for task1 was a csv file with the headers, urls and headline and then below the two headers contained the urls and the headline from the url article in sequential order.

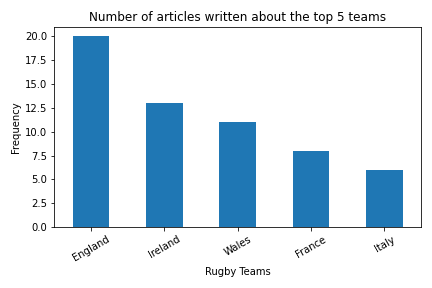
**Web scraping method**

The beautiful soup library allows you to extract data from the page requested. soup = BeautifulSoup(page.text, 'html.parser') used with html.parser as the data scrapped is written in html code. links = soup.findAll('a') allows to find all the html tags that include ‘a’ tag. soup.findAll('a',href=re.compile("^index.html")) allows us to find the the seed link which is the information before the index.html but after the base url. A list called to\_visit\_relative is made to visit each of the links. A list called to\_visit makes the absolute urls of website. After you get the all the urls. For task2 you also need to find the headers. The headers are inside the h1 tag of the html file. So using soup.findAll(‘h1’) and using a for loop you can append the text of h1 tag into a list. To find the first team name of each article page. You had to open the json file containing the team names and form a list of the relevant teams you wanted first. I would use soup.findAll find the main article text. I used a for loop and looped through each word and compared to if it was in the list of team names, if it was the first instance I appended it to a list called first\_team. If the team name was 2 words, e.g. New Zealand. I formed a list of list called new\_team, the first element in list of list containing the 1st word and 2nd element containing the 2nd word and 3rd element containing the whole word. If the new word was equal to 1st word and word after that was equal to 2nd word I would append the 3rd element which contains the team name. If the count was 0 by the end of the paragraph I would append None to list. To find the scores I used regular expressions re.findall(' \d{1,2}-\d{1,2}[, .?:;\|]',paragraph.text) to find all the scores. I appended list of scores to match\_score. I then wrote a function to find only the largest match score which was the sum of match. I then put all the of list into a data frame and stored it into a csv file.

**Brief summary of the output for Task 2**

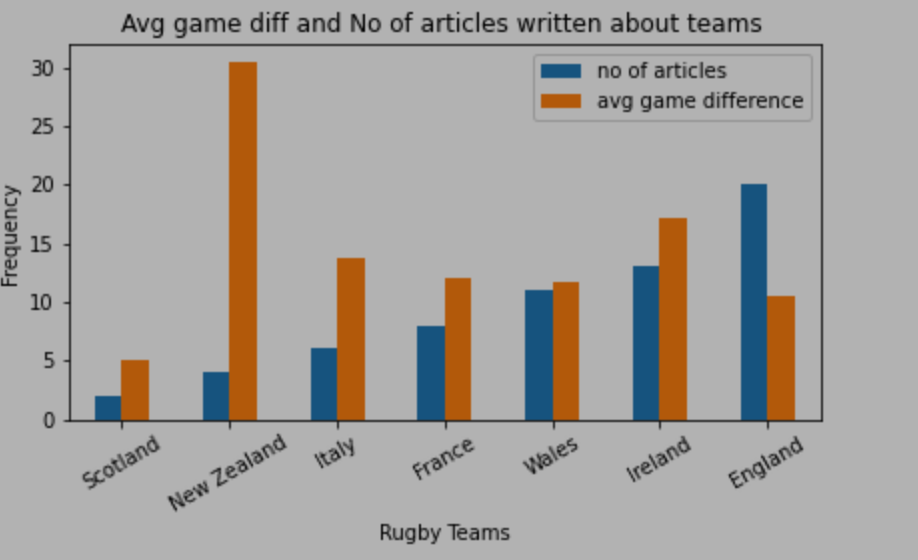
The summary output for task2 was a csv file with headers url, headline, team and score. Then below the 4 headers contained the data for each of the respective urls, if the urls had missing data it was removed from the csv file.

**Analysis of the information shown in the task4 plot**



The bar graph indicates the number of articles written about the top 5 teams, the data showed that England had the most number of articles written about them at 20, Ireland 2nd with 13 articles, Wales 3rd with 11 articles France 4th with 8 articles and Italy 5th place with 6 articles.

**Analysis of information shown in task5 plot**



The legend shows the blue bar graph representing the number of articles, whilst the orange indicates the average game difference. The average game difference is sum of differences divided by the total amount for each respective team. The teams are shown on x axis and y axis represents the frequency. The bar plot shows the New Zealand had the largest average game difference but had 2nd to last in terms of number of articles written about them. There is a correlation between the number of articles written about the team and their respective average game difference. Scotland with the lowest articles written about them had the smallest average game difference. As the number of articles increased the average game difference was similar. This shows the more articles written about each team the less fluctuation there was. Using the data from New Zealand and Scotland would be a false representation of their actual average game difference as not enough data is collected to give an accurate representation.

**A discussion of the appropriateness of associating the first named team in the article with the first match score.**

It is appropriate to associate the first named team in article with the first match score as the first match score generally be close by to the first mentioned team. It is best used when there is only one score in the whole article as it is not guaranteed that the first score would be referring to the first named team. You could also make a count to see how many words the score and the first named team.

**The methods for how you could whether the first named team won or lost the match being reported on from contents**

**1st method**

The first method would be to for each article find the names of rugby team countries and words such as defeat, failure, loss, lead, beat. Just using words that commonly suggest losing or someone winning. The list would be in chronologically ordered. If the word was between the two countries listed. Make a list of words represent winning and list of words that representing losing. If the word in winning list this suggest 1st country won and if in losing tells 1st country lost. Advantages of this method would be that it would work well for most cases, disadvantages it is possible to miss out a word, you would have to make sure list of words are stemmed.

**2nd method**

The 2nd method would be to find the first occurrences of 2 rugby team countries and 1st occurrence of score. If the score was in descending order the team on the left side would be the winner else wise if the score was in ascending order the team on the right side of the list would be the winner. Disadvantages would it is if the score was not between the 2 teams, or that the 1st score was not the actual score indicating the win. Advantages would be there would be no stemming or lemmatization needed.

**A discussion of what other information could be extracted from the articles to better understand team performance and a brief suggestion for how this could be done**.

Extracting the stadium from json file. This will help to indicate where they perform better given if the game is home or away. You could make an average difference for home games and away games and compare both of them.