Task 6

1)

In task 1, a web crawler is set up to extract all the URLs from the given website. The crawler initially gets the first hyperlink from the main URL and then other hyperlinks from each URL found in every iteration of crawling. The process then includes collecting the title of every article given by the URLs and then storing them along with their URLs in a csv file.

2)

In the first part of task 2, the data from each article is split based on sentence completion. This split data is then analyzed to check if it contains at least one team name mentioned in the rugby.json file. The first team which is mentioned in the article is stored whereas articles with no team name are discarded.

Regular expression is used in the second part of task 2 to find the valid scores. To avoid dates being considered into this expression, the list of data found using regex is then caped at the length of two to avoid the inclusion of years. The highest score from every article is stored. This process has some limitations as values such “50-50” which are not the scores of the match included in the article are also considered given the lack of information to set up flags differentiating such values. Any article without a valid score is discarded.

The output containing articles with valid team name and valid scores are stored in a csv file along with their URLs, article title, highest score, and team name.

3) a) Task 4

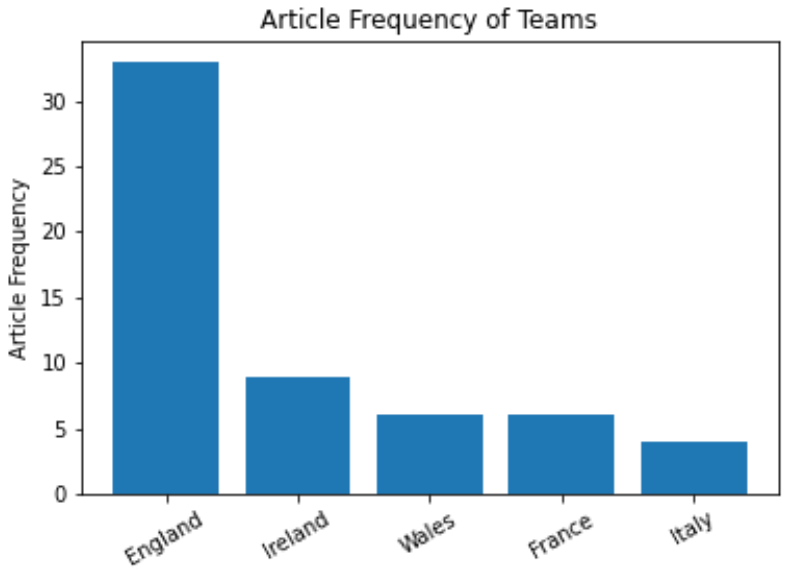

Figure 1 (Task 4)

The plot produced in task 4 (figure 1) displays the five teams about whom articles have been written most frequently. It is observed that articles have been written about England more frequently compared to all the other teams with Italy being least frequent among the five teams. A substantial difference can be seen in the frequencies between England and the rest of the teams.

The data used for this plot contains the articles in which the team names as well as valid scores were found. The articles with no valid scores were discarded and hence the plot produced can have some inconsistences as the articles discarded might have a valid team name.

b) Task 5

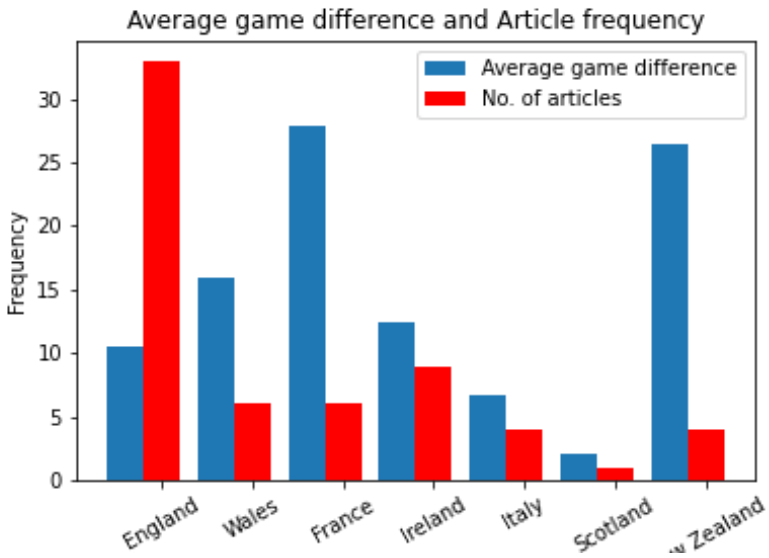

Figure 2 (Task 5)

The plot produced in task 5(figure 2) displays the teams with their average game difference and the number of articles written about them. As seen in the figure, France followed by New Zealand have largest score differences among all the teams whereas England has the largest frequency of articles with Scotland being the least among all the teams.

The game difference from the max score in every article has been associated to the team which has been mentioned first in the article. This creates discrepancies in the data as every score belongs to two teams involved in the match whereas only the first team has been considered here.

4)

The match score considered has been analyzed to be the max score of the match. Associating only the first team named in the article to this score creates a lot of inconsistencies in the data as which part of this score belongs to that team is not validated.

5)

The pattern search can be made more rigorous including words like defeat, win, loose, triumph, etc. which follow the scores. Another pattern can be included where the first team name is succeeded or preceded by words like beat, won etc. This can be used to differentiate whether the first team lost or won and then separate the scores accordingly.

The advantage of using this method is that it can give better results and match more precisely compared to the approach used in the assignment. The disadvantage is that it comes with language limitations as sentence construction can become very complex causing it to escape the specified pattern.

6)

Players' information for every team can be extracted from the articles. This can be used to observe the players from each team which are consistent or in a good form given the frequency of their name being mentioned in the article suggesting their good play.