Elements of Data Processing

Assignment 1 – Task 6

Submitted by:
Vinay Pinjani
1151832

A description of the crawling method and a brief summary the output for Task 1. (2 marks)

For the first task, a crawling method was required which makes use of all the available links on a particular webpage and uses those extracted links to access other pages if they previously have not been visited by the program. The program makes use of the BeautifulSoup library to extract all the anchor tags on a HTML file and appends them to a list unless they are the same as the current page.

As each of the link is visited by the program, the BeautifulSoup library is used to extract the H1 tags on the page, this should give us the headline for the article. Before the next link is accessed, both the link of the current article and the extracted headline are appended to a csv file. This loop ends when the all the links extracted from the page have been previously visited and no new links are available. In the end our output csv file contained the URL of each page visited and the headline of that page.

A description of how you scraped data from each page, including any regular expressions used for Task 2 and a brief summary of the output. (3 marks)

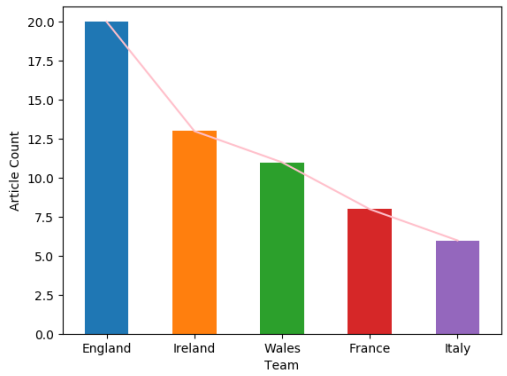
For this task, the provided JSON file was used extract the team names and were stored in a list. Then, using the CSV file from the previous task, each link was accessed and the BeautifulSoup library was again used to extract all the text on that page and store it in a variable. Then a regex pattern was created using the list of team names that was previously created and used the search method to extract the first matched team from the article and returned the name. For the scores, another regex pattern was created, which found all the strings that matched the pattern of a rugby score and stored them as a list of tuples containing the individual scores. This was done using the findall method. From that list of tuples, the score that had the highest sum was extracted. The score and the team name were appended to a CSV file along with the URL they came from. Out of the 147 links collected from the previous task, only 64 remained in this task. This may be due to the article not containing the team name or the program failing to recognise it, or it may not be containing a score or the program may have failed to recognise it.

An analysis of the information shown in the two plots produced for Tasks 4 & 5, including a brief summary of the data used. The plots are to be shown (included) along with your analysis. (4 marks)

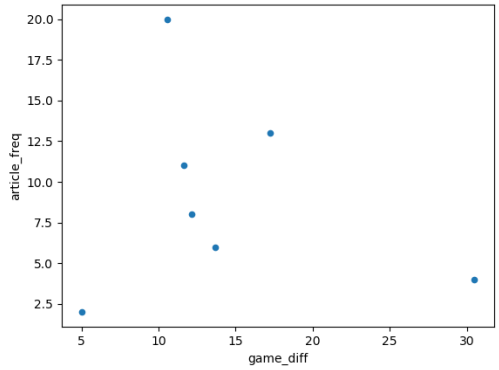
The data for task 4 and 5 was obtained by manipulating the CSV files obtained in the preceding tasks. For task 4, a dictionary was created for all of the teams that had been read as keys and the number of times the appeared as the value. This dictionary was used to create a Pandas series and a bar chart of the top 5 repeated values was plotted. The bar chart outlined a clear difference between the number of articles that were written for the teams. The highest was team England and the others followed.

For task 5, the dictionary obtained from the previous task was used, along with the game difference from the CSV file produced in task 3. A Pandas Dataframe was produced using the game difference and article count for each team. Since these were two different numerical variables, a scatterplot was used to represent them. The data showed no obvious trend, it had values on both extremes, this may be due to the small number of teams in our sample. From the given data, it is difficult to determine any relationship between the two variables.

Task 4 Output:



Task 5 Output:



A discussion of the appropriateness of associating the first named team in the article with the first match score. (2 marks)

It is important to understand that the articles do not follow a fixed writing pattern, it is impossible to achieve certainty with the data that we produce. However, it is likely that the first match team in the article is the team that is being talked about and the first matched score is associated with the team. It does make sense that out of all the scores in the article, the first one should be about the current team, and it makes sense to use that for our data.

At least two suggested methods for how you could figure out from the contents of the article whether the first named team won or lost the match being reported on and a comment on the advantages and disadvantages of each approach. (2 marks)

One method that we can use to analyse if the first named team won or lost is by defining keywords associated with winning or losing. This can be done by reading a few articles and understanding the language. The keywords such as wins, loses, defeat, triumphs etc. can be stored in two categories, winning and losing, we can write a program that looks for these words in the same sentence where the first named team is mentioned. Depending on which category the word belongs to, we can conclude if the team won or lost. This method can be unreliable sometimes as the words may be used in a different context or may not be used at all, in such cases our results might be misleading, if we find a way around this, this method may actually present accurate results.

Another method that can be used is by analysing the score itself. It is likely that the first named team is the home team, this can be confirmed by analysing the text, if the score extracted from the article is greater on the left side, for example, 6-4, this shows that the team won and vice versa. An advantage of this method is that it is easy to implement but it is also heavily dependant on the score that we extract from the article which may be incorrect and in that case our results will be false.

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. (1 mark)

Using the methods previously described, we could extract whether the team has won or lost in that particular article. This information could be stored and used to calculate the win-loss ratio for each team by dividing the number of wins per team and the number of losses per team. This way a direct comparison could be made between teams, and a team standing can be generated where the team on top will be one with the highest win-loss ratio.