COMP20008 Elements of Data Processing Assignment 1 Task 10

In Task 3, the regular expression used searched for two numbers of sizes one or two that are separated by a hyphen and enclosed by non-numeric characters. This is done so that the search would not recognise numbers larger than 99 that may potentially return the incorrect highest total number of goals. This regular expression may cause some problems such as when the articles contain dates, phone numbers or other similar numerical characters in the same format as the regular expression. For example, the expression '50-50' used to describe chance would be falsely recognised as a match score. Thus, resulting in the regular expression to identify a match that is not a score and may return an incorrect highest total number of goals in the article.

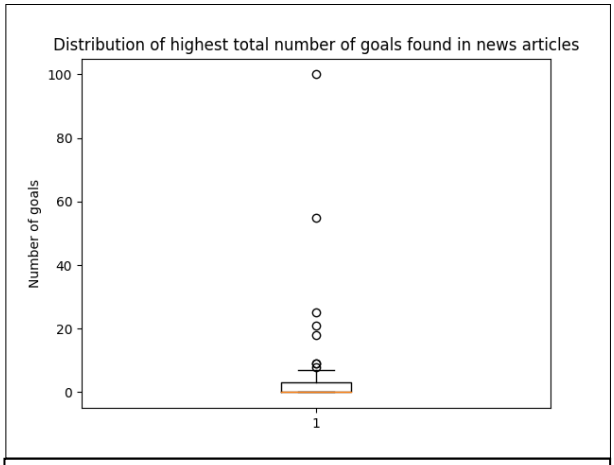

Figure 1: Task 4 boxplot of highest total number of goals

The following figure 1 is the boxplot produced from task 4. From this, it is implied that the median of the highest total number of goals in the news articles is equivalent to the lower inner fence and first quarter, which is 0. This demonstrates that more than half of the articles either had the highest match score of 0 or did not have a suitable score in the article. Through the boxplot, it can also be seen that there are extreme outliers around 100 and 55. This can imply that there are some problems within the implementation of calculating the highest total number of goals, as shown in the evaluation of the regular expression used in task 3.

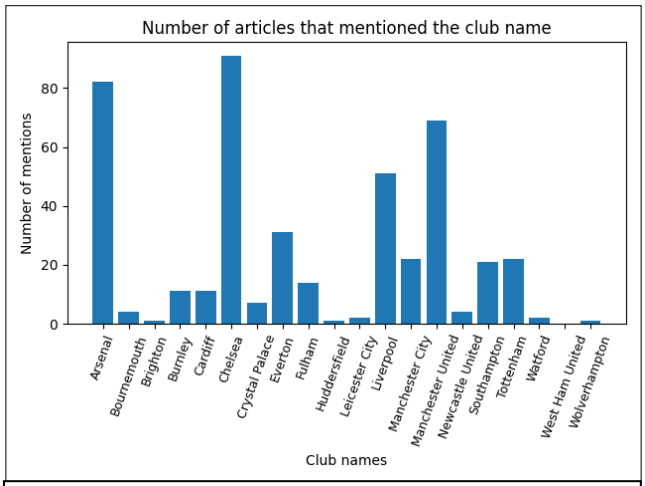

Figure 2: Task 5 bar chart of number of articles that mentioned a club

Figure 2 is the bar chart produced from task 5. Through this bar chart, it can be determined that Chelsea is the club that was mentioned the most within the articles, with around 90 mentions. The second and third most mentioned clubs were Arsenal and Manchester United, around 80 and 70 respectively. Additionally, it is evident that West Ham United was not mentioned at all within the articles.

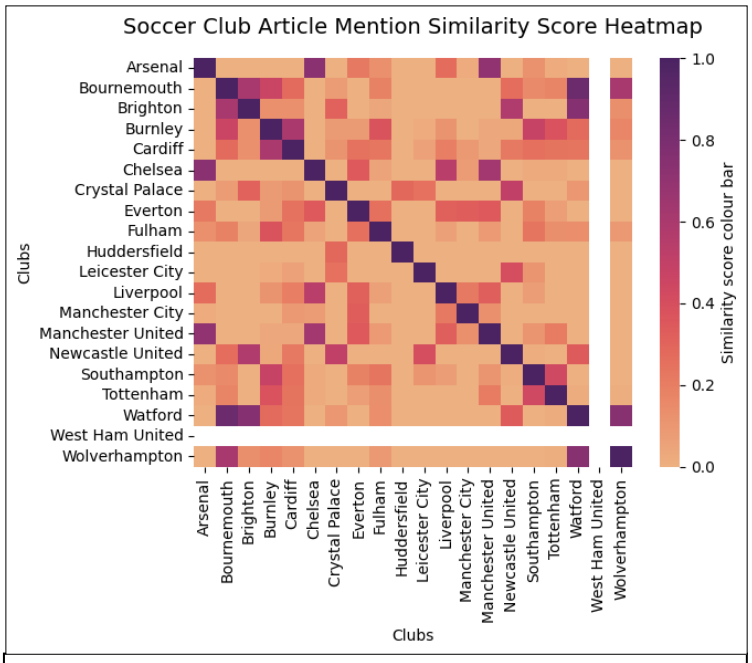

Figure 3: Task 6 heatmap of club mentions in article similarity score

Figure 3 is the heatmap produced from task 6. The heatmap shows that any club paired with West Ham United had a similarity score of 0. This is due to West Ham United having zero articles that mentioned their club's name, resulting in a similarity score of 0. Moreover, the diagram demonstrates that the two clubs who had the highest similarity score was Watford and Bournemouth, as they had the darkest representation of their similarity score according to the colour bar.

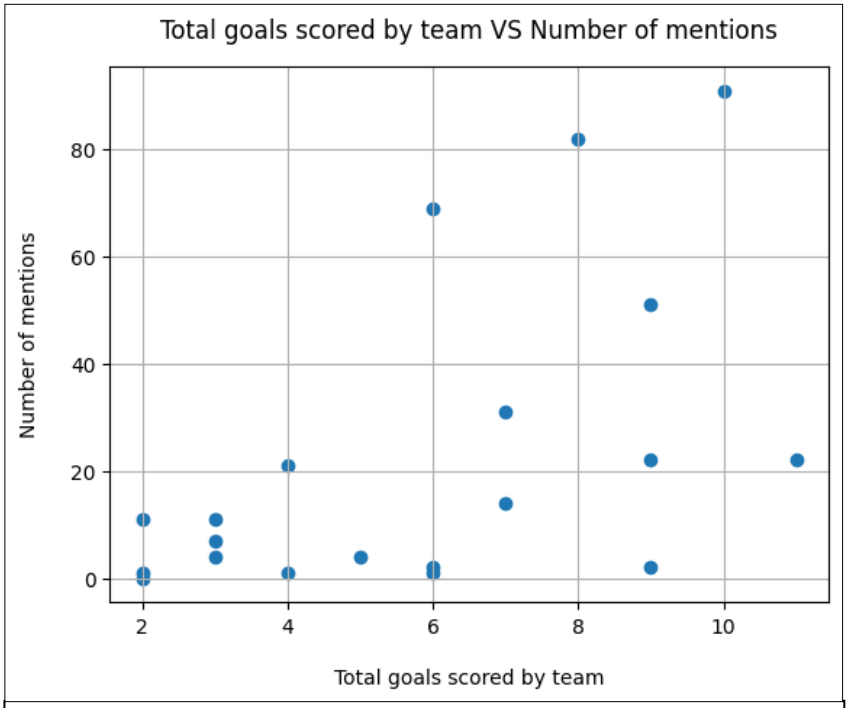

Figure 4: Task 7 scatterplot of the goals scored by each team

Figure 4 is the scatterplot produced from task 7. This scatterplot shows that majority of the teams that had a total of less than 6 goals scored, had less than 20 mentions in articles. The highest number of total goals scored were 11 and the highest number of mentions were around 90. The scatterplot shows a moderately positive trend, with the teams with the three highest number of mentions scoring 6 to 10 total goals. Therefore, it is shown that generally, teams with higher total goals scored have a higher number of mentions in articles.

Word count: 491