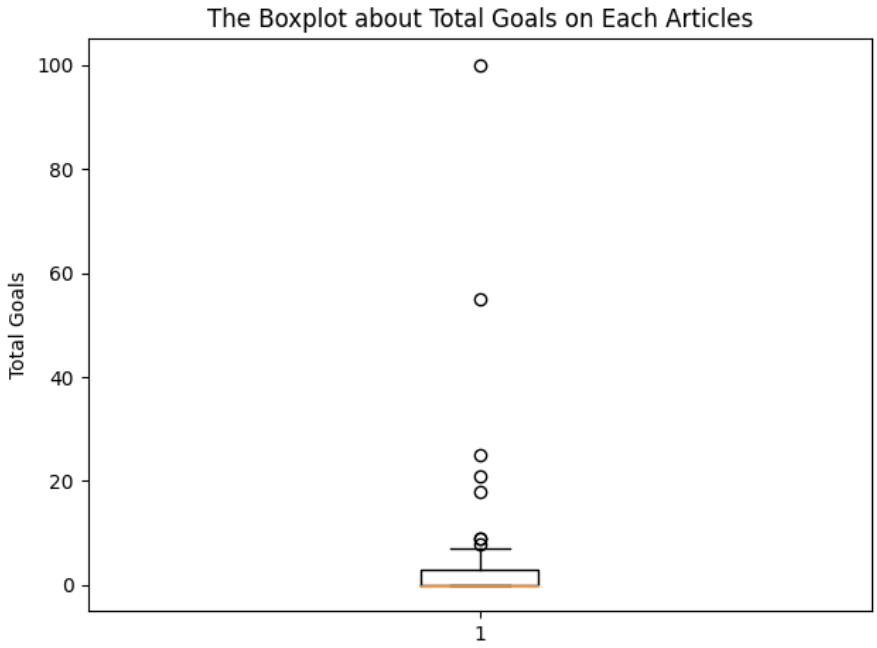
This project investigates some match results and news articles relating to soccer matches in the English Premier League provided. Many analysis methods, including using regular expression to find a particular pattern of data, text processing techniques and result visualisation, are implemented to analyse the data. This report will discuss the appropriation of the regular expression used in task 3, including both advantages and drawbacks, and then provide some interpretations based on the visualisation produced in task 4, 5, 6 and 7.

In task 3, regular expression is used to extract the total score of all matches in each article and find the largest one. Regular expression matching is to a large extent appropriate for this task because scores are always provided in a constant pattern in each article. However, regular expression is not the perfect method for this task. For example, the method will perform poorly when score data appears in a different pattern instead of number-number pattern. For example, number : number pattern, number~number pattern or even using English words to replace digits may occur in articles to represent scores. In this case, it will be hard for regular expression matching to find all scores. In addition, some data which have the same pattern of scores but not scores itself, like date data, may be wrongly interpreted as scores. In this case, the use of regular expression will disrupt the extracted data. Therefore, although regular expression matching is a solid method, it performs poorly when dealing with other patterns of scores and score-like data which are not scores.

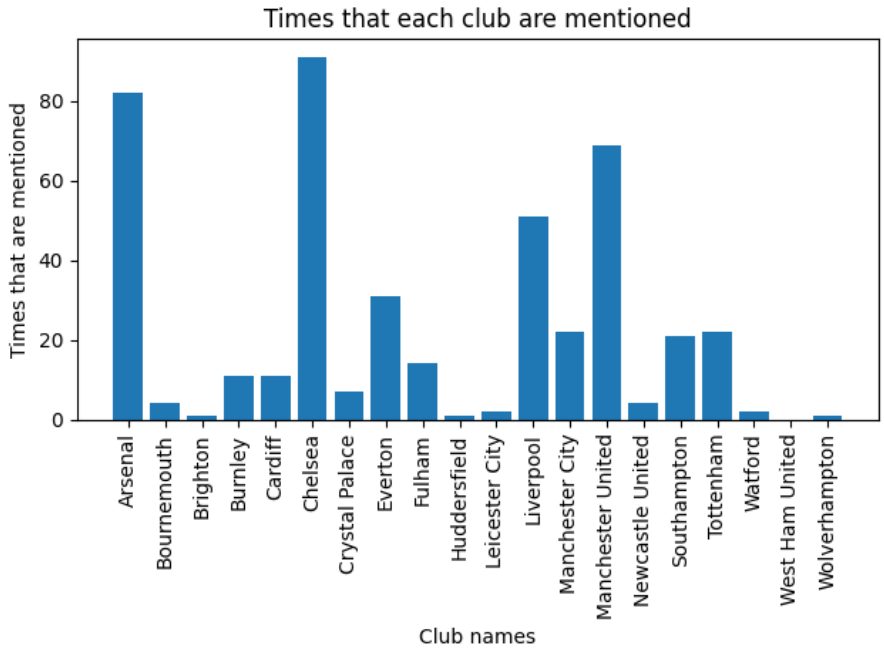
The project also retrieves all articles and allows important information to be displayed visually. Firstly, the maximum score in each article is calculated and stored in a boxplot, generated to visualise the result.



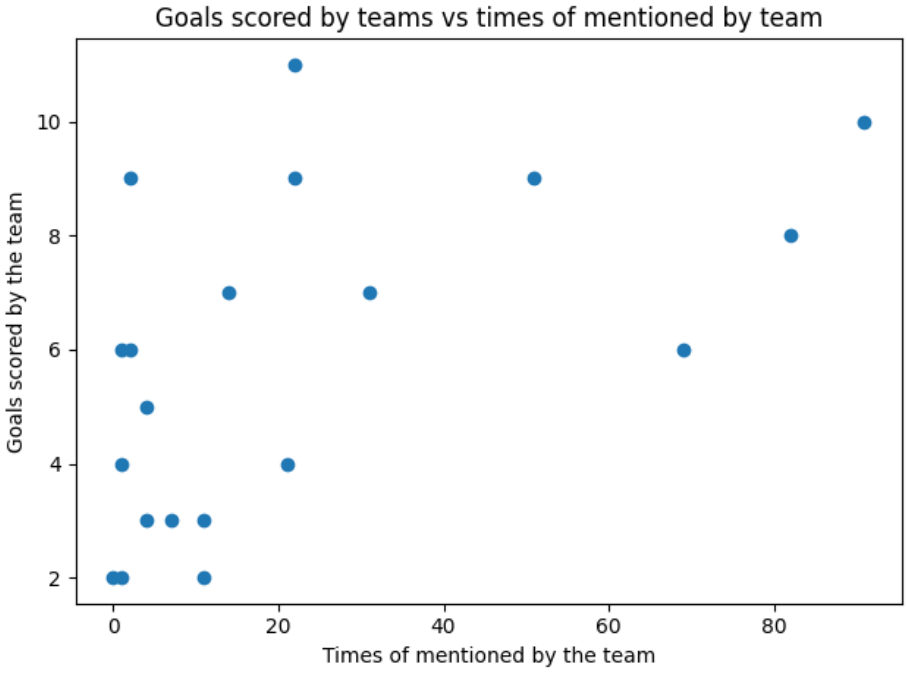
From the plot, it is easy to see that most of the maximum scores are located between 0 and 10, which is common for most football games, except extreme cases. It is interesting to find a score around 60 and another one around 100, so we can probably deduce that they are not

scores from real games.

Besides scores, a bar chart, which indicates how often each team is mentioned by articles is generated as shown below:

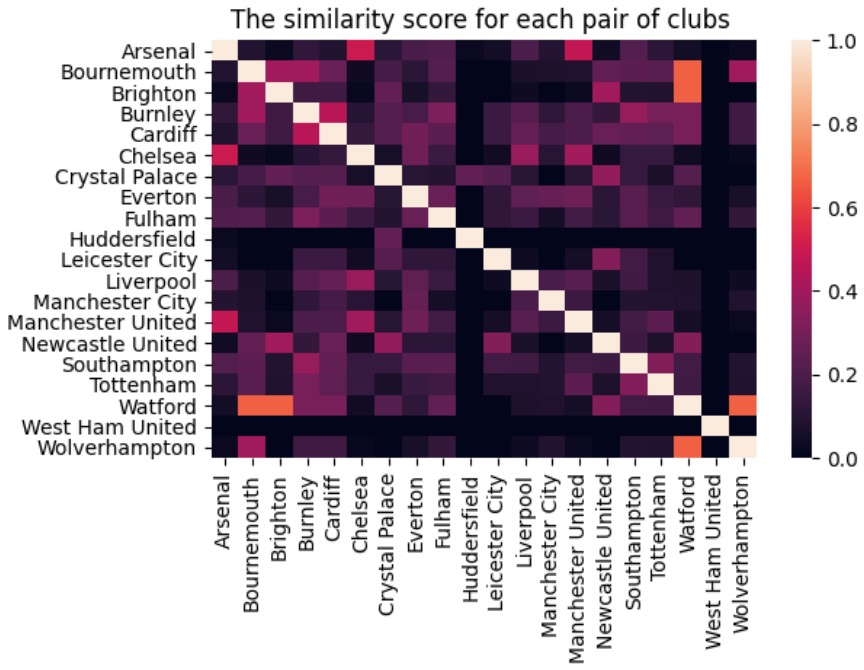


We can observe that teams are mentioned unevenly, with a huge gap between the maximum and minimum. A reasonable guess is that the performance of the team is positively correlated to the frequency of it being mentioned. To answer this question, the project also includes a scatter map to show the relationship between the number of articles and the total number of goals scored by each team, which represents the performance of each team:



The scatter map demonstrates that the performance of a team is somehow represented by the number of times the team has been mentioned although there is no strong relationship between these two.

Last but not least, the project visually displays the similarity for each pair of teams through the use of a heatmap:



From the heatmap, it is possible to say that two clubs may have a close relationship as the similarity score goes up. For example, it is likely that Arsenal and Chelsea may have similarity in strength, or there is a rivalry between them because they have a high similarity score.