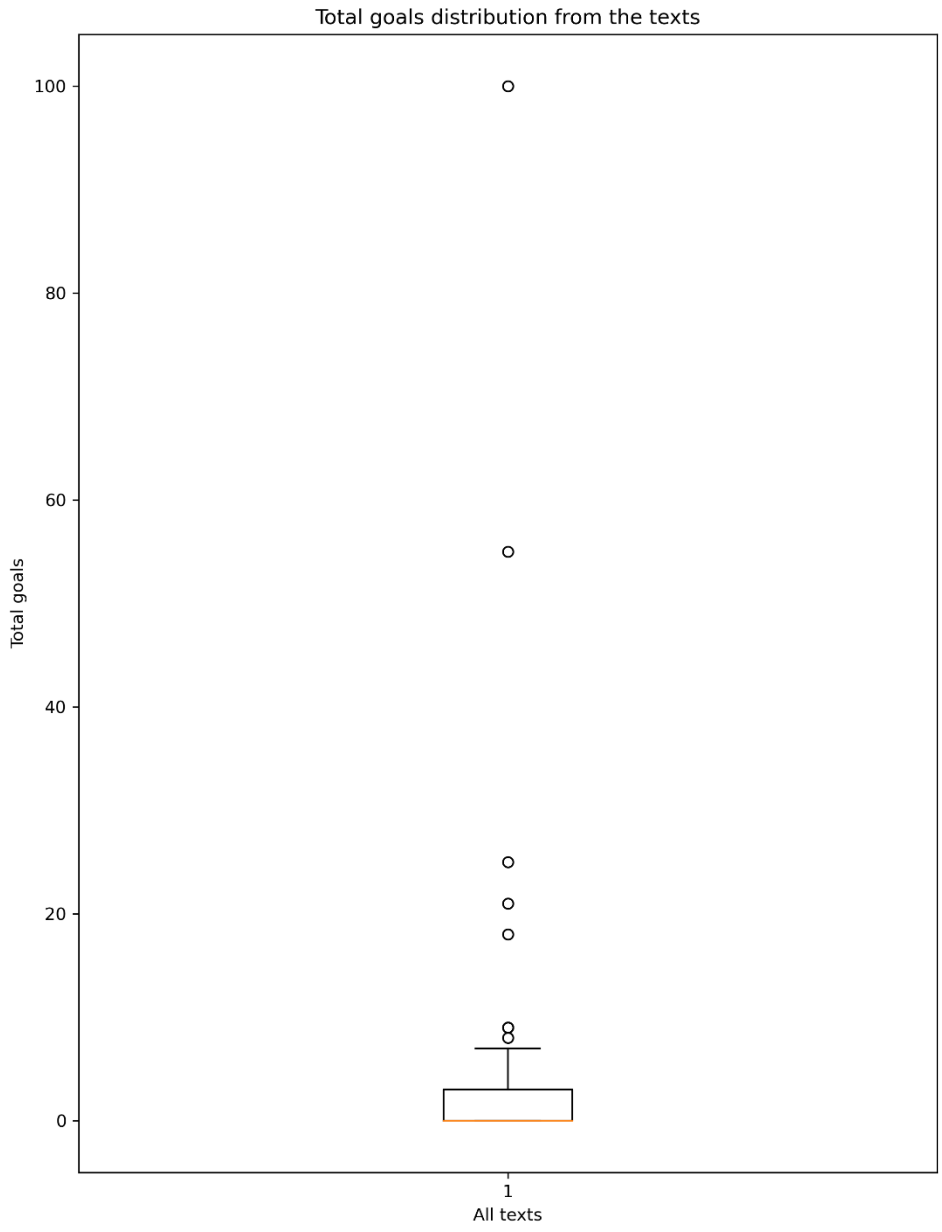
Assignment 1 report

Yuhao Tong

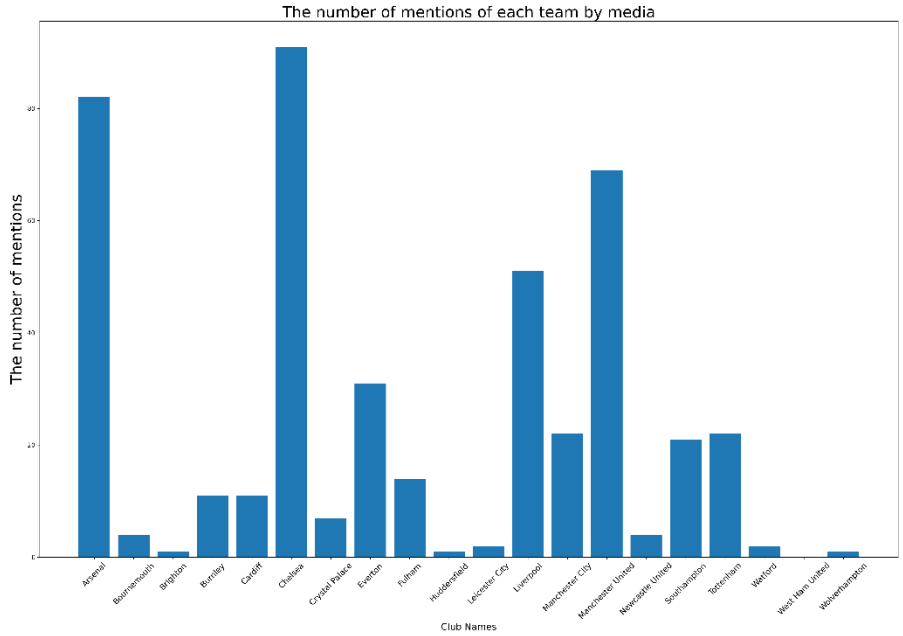
This report will cover following parts: The appropriateness of the usage of regular expression in task3, the interpretation of the visualization produced from task4 to task7.

The algorithm used in task 3 iterates through the articles and find the regular expression matches based on the pattern '\D[0-9]{1,2}\-[0-9]{1,2}' indicating the style of scores. The output suggests that the majority of articles produce a regular range of sum of scores, while with some unexpected outliers. For example, the result from 025.txt produces the highest record of all, 100 scores, which is impossible to be reached in a single game. Thus, the numbers which are not true soccer scores are also involved. For instance, in the 212.txt article the match '18' is actually sum of 2 digits in the context "6-12 months". Hence, the search using regular expression might be inappropriate and more specifications are needed.

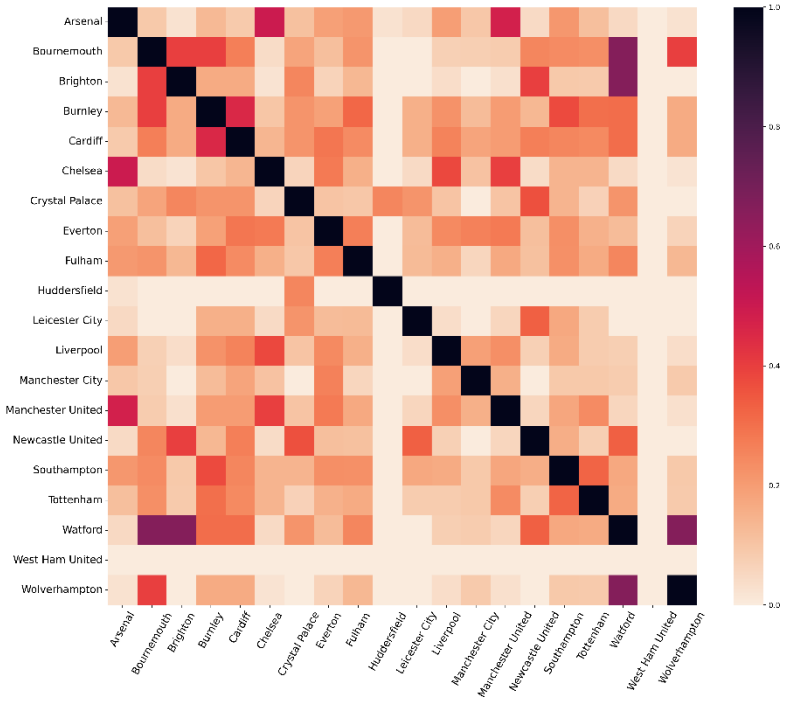
Secondly, visualizations from 4 tasks are analyzed. The task4 visualizes the total goal's distribution, which is tightly grouped and skewed down. The median and the quantiles are all located at the bottom, with the majority of data between 0 and 3. Furthermore, 3 articles output the suspected outliers and 6 produced true outliers, indicating mismatched results from the output of regular expression search used in task 3.



The visualization of task 5 counts the number of articles mentioning each club's name. Among 20 clubs, Chelsea owns the highest number of mentions, about 91 times, indicating a high level of focus by media. Arsenal, Manchester United and Liverpool have slightly less mentions, ranging from 51 to 82. Several clubs including the Leicester City and Wolverhampton, are less popular in the League, the West Ham United is even not mentioned by the media in the data, suggesting a lower frequency of participation in matches.

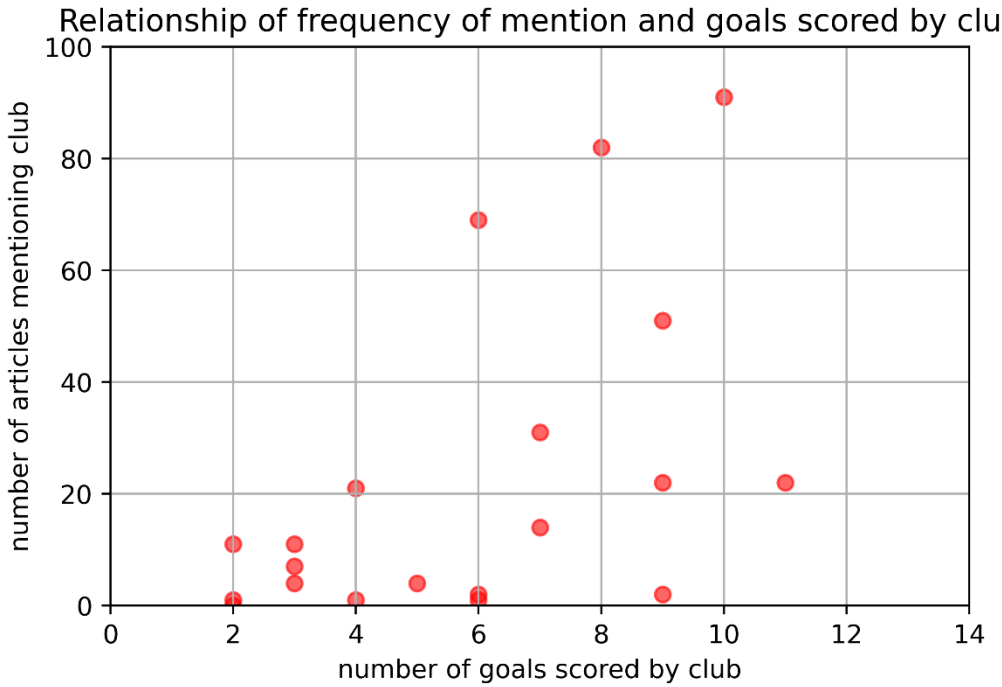


In task6, clubs co-existence in articles is illustrated by a heatmap based on the similarity scores. The West Ham United seems to have no games paired with other clubs, which is explained by the absence of media mentions. Several hotspots are found between Watford and other three teams: Bournemouth, Brighton and Wolverhampton, these teams have frequent co-existence in the report.



Finally, in task 7 relationship between frequency of mentions and goals scored by each club is explored using a scatterplot, the plot demonstrates a general trend that the clubs with higher scores

might be reported in higher frequency, suggesting a strong correlation between lower scores by club and the focus on them by media, whereas the correlation becomes weaker when goals increase.



In conclusion, the task3 pattern search needs more specification, because of the unexpected matches. Visualization from task 4 suggests a tight and down-skewed distribution of sum scores from the data, task 5 indicates a ranking of popularity of clubs, with Chelsea the most popular club. The heatmap in task 6 illustrates that West Ham United has no matches with other clubs, and task 7 verifies that higher scores generally raises higher focus for the club.