Analysis report for assignment 1

Discussion on the appropriateness of the regex in Task 3

The overall performance for the regex, “r' (\d{1,2})-(\d{1,2})'”, is not perfectly appropriate but acceptably effective. This regex is effective to extract the strings well fitted into the specific soccer scoring format, namely, ‘[single/double-digit number]-[single/double-digit number]’. Besides, it uses ‘capturing groups’ by parentheses to capture and return the goals scored by each side as a pair. This is convenient for us to perform the following calculation, namely, adding up the goals of each side as the total goals.

This regex also has limitations. Firstly, it cannot exclude the strings that can fit into the wanted pattern but do not represent the soccer goals, which means it doesn’t consider the context. This shortcoming can be overcome to some extent by manually examining the outliers. Secondly, if some goals are mentioned in other formats, for example, ‘two-three’, the regex will ignore them.

Some examples expected to perform poorly are, 1) -50’ in article ‘025.txt’ refers to probabilities instead of match goals. 2) ‘27-28’ in article ‘125.txt’ refers to dates. 3) ‘20-1’ in article ‘206.txt’ refers to odds for bets. 4) ‘6-12’ in article ‘212.txt’ refers to a span of months.

Analysis of the visualizations in Task 4, 5, 6, 7

Visualization in Task 4

As shown in **Fig.1**, we can see the 7 outliers are colored in red. All the goals greater than or equal to 9 are identified as outliers. By manual examination, the two outliers with a value of 9 are real total goals, though rare. This means they are valuable for the following study and we should not exclude them without careful consideration. All the other outliers are with values greater than or equal to 18. They are found as obvious artifacts.

We can conclude that most total goals in the provided articles are not above 9. Besides, in this case, 1.5 times the interquartile range above Q3 is relatively appropriate as the threshold, but not perfect.

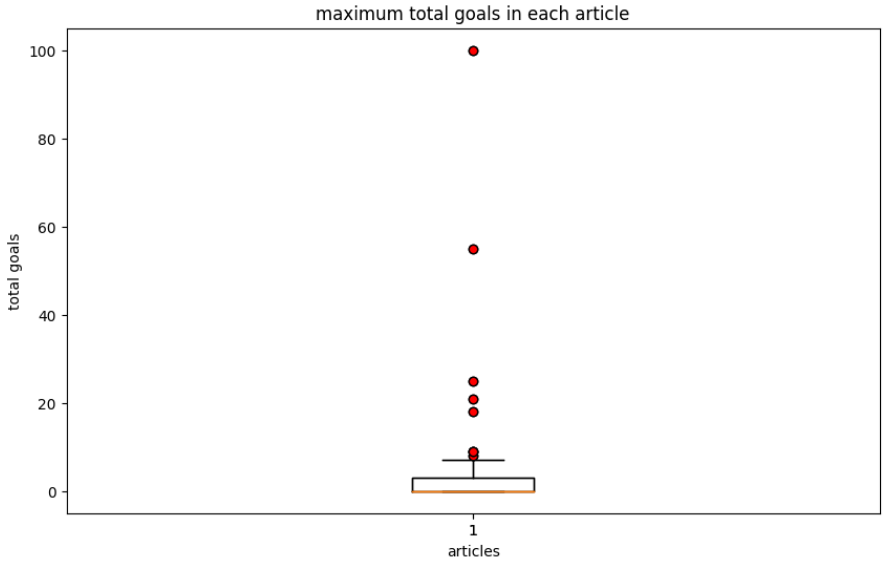

Fig.1. Visualization in Task 4

Visualization in Task 5

As shown in **Fig.2**, we can see the mentioned times vary considerably across clubs, with the maximum greater than 80 and the minimum equal to 0. Arsenal, Chelsea, and Manchester United are the top 3 most mentioned clubs by the articles, while West Ham United, Huddersfield, Brighton and Wolverhampton are the four least mentioned clubs.

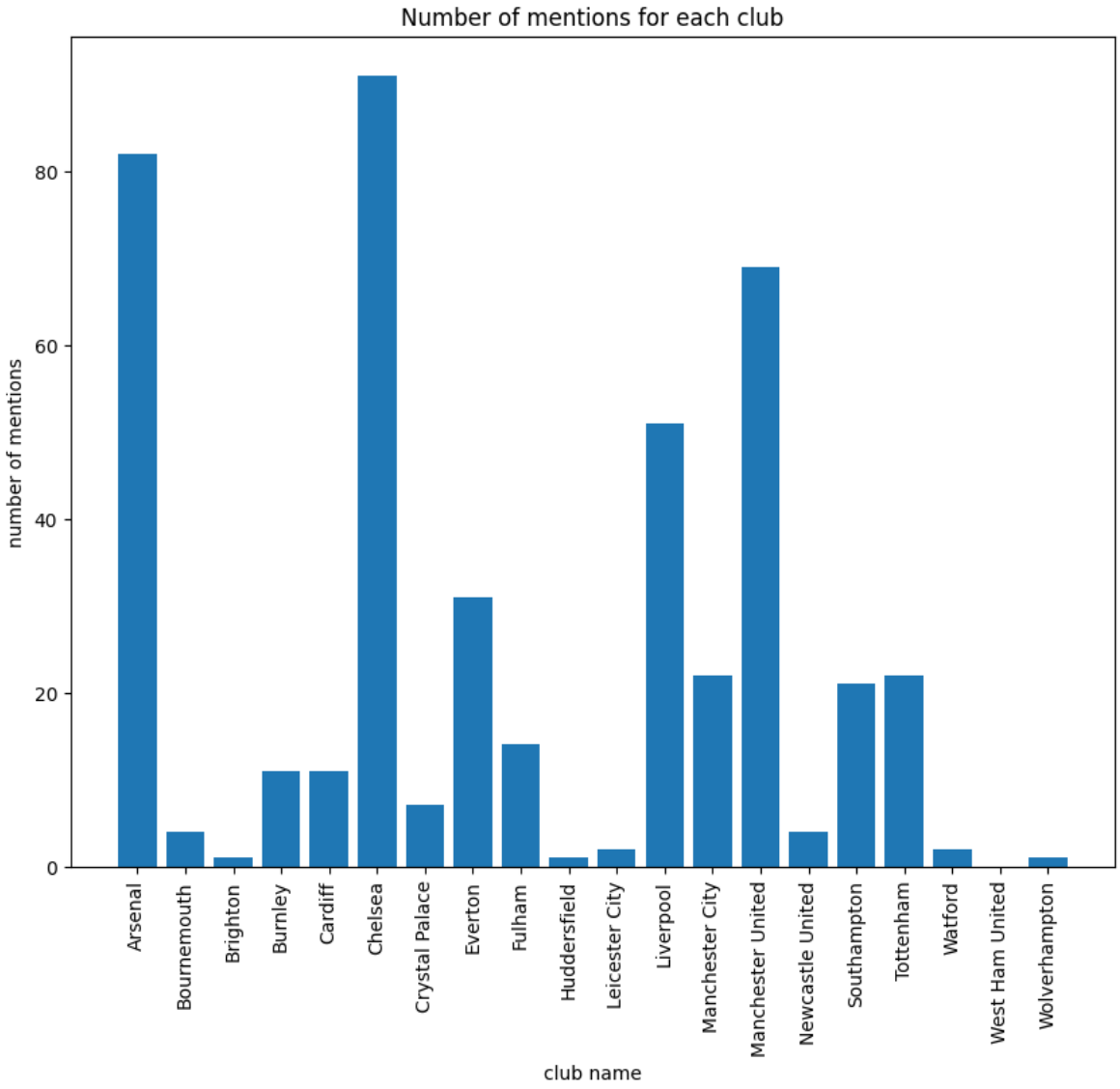

Fig.2. Visualization in Task 5

Visualization in Task 6

As shown in **Fig.3**, we can see the similarity scores (darkness of the color) vary considerably among different pairs of clubs, with the maximum about 0.65 and the minimum equal to 0. Watford and Wolverhampton, Watford and Bournemouth, and Watford and Brighton are the 3 pairs of clubs that are the most similar to each other.

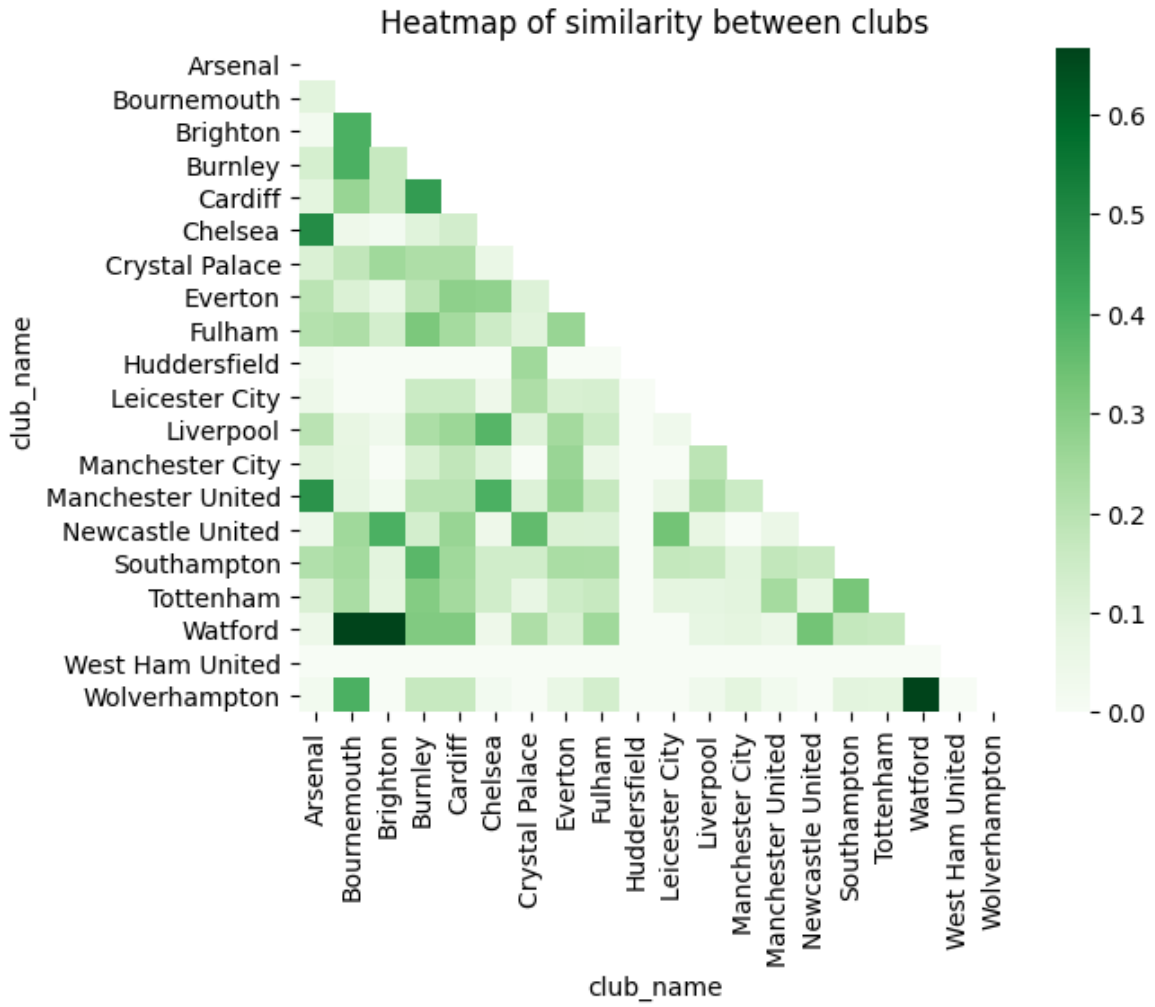

Fig.3. Visualization in Task 6

Visualization in Task 7

As shown in **Fig.4**, we can see there is no strong correlation between the mentioned times of clubs and their performance, since the data points are randomly spread out in this plot.

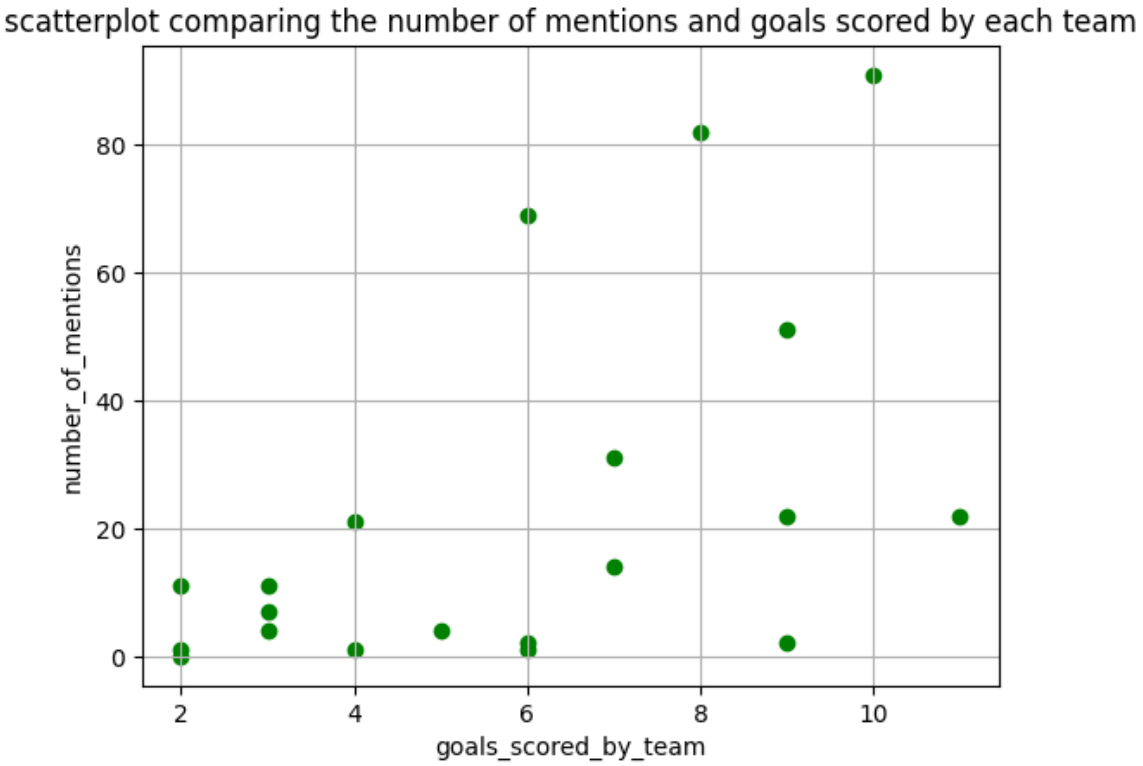

Fig4. Visualization in Task 7