Task 10: Report

Junfei Zhang (Fiona)

The University of Melbourne

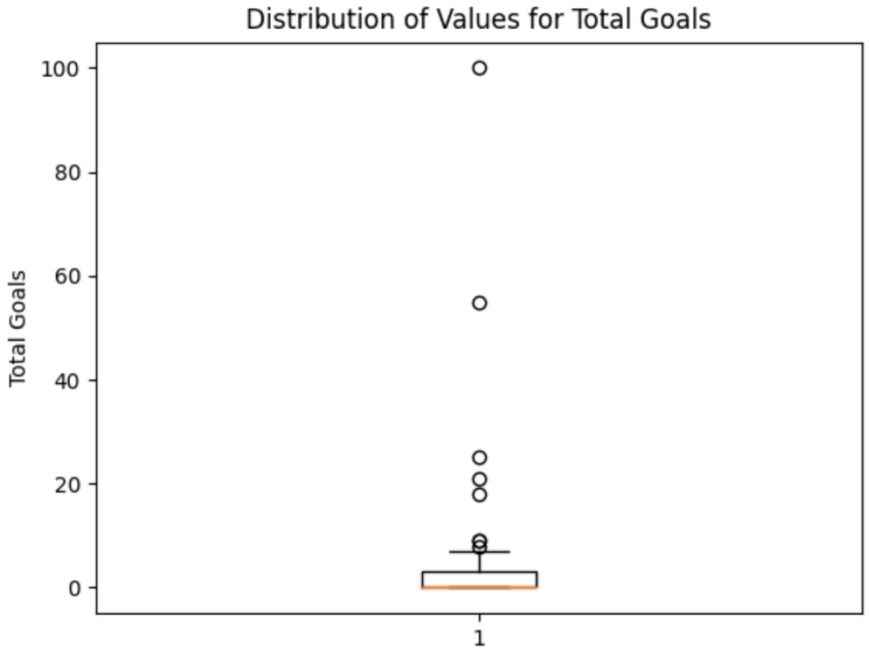
COMP20008 Assignment 1

September 9, 2021

Task 3: Appropriateness of Regular Expression

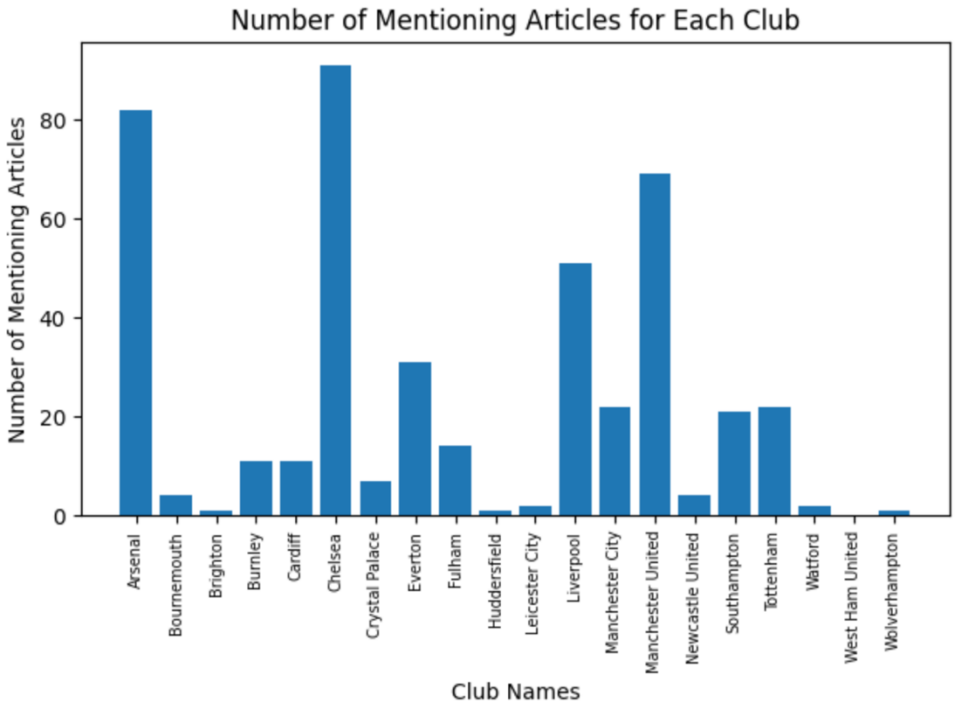
The regular expression used is '[^\d]\d{1,2}-\d{1,2}[^\d]', referring to any sequence that starts with a non-digit symbol, followed by one or two digits, a hyphen, another one or two digits, and a non-digit at the end. This expression can detect any match scores either in punctuation marks, such as “(15-16)”, or simply surrounded by white spaces. It can also avoid matching numbers such as 1998-1997 (98-19). However, the regular expression might fail if it encounters numbers with a similar format but has a different contextual meaning, such as “There are 10-11 players on the court”. This regex might also fail if the score is represented by the format XX:XX, or XX – XX (whitespaces surrounding hyphen).

Task 4: Boxplot



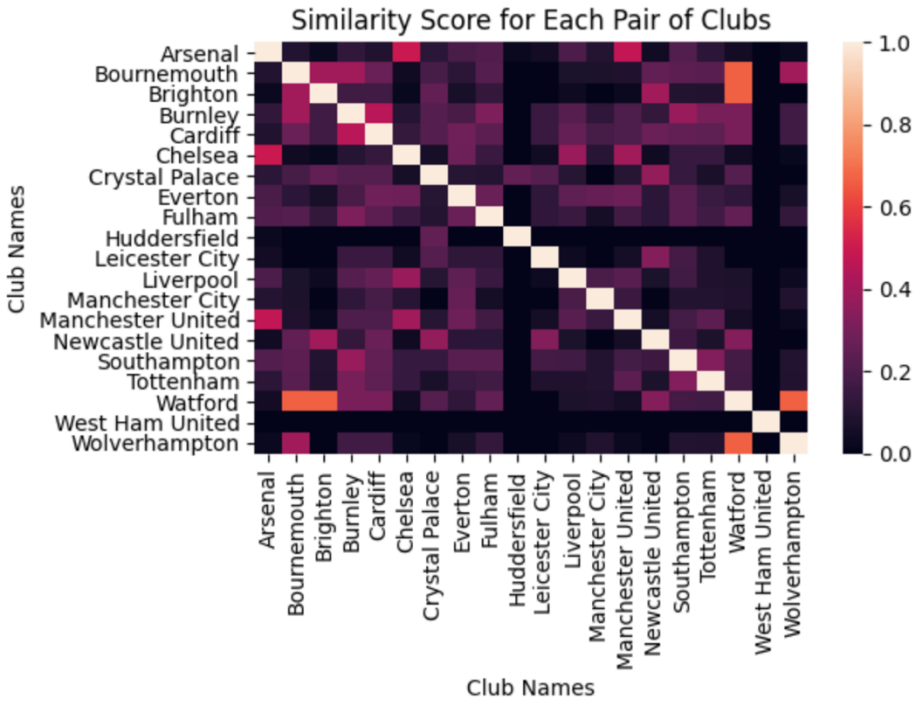
The box plot shows that 75% of the articles have a total goal below 3. There are 7 outliers shown on the graph. According to Task3, the maximum is ‘025.txt’ with 100 goals, followed by ‘125.txt’ with 55 goals. Both goals are not matching goals, but have different contextual meanings, with one referring to 50-50 chances and the other one refereeing to dates (27-28 November). Other outliers may also be affected by the same reason. This is where the regex perform poorly, as analyzed above.

Task 5: Bar Chart



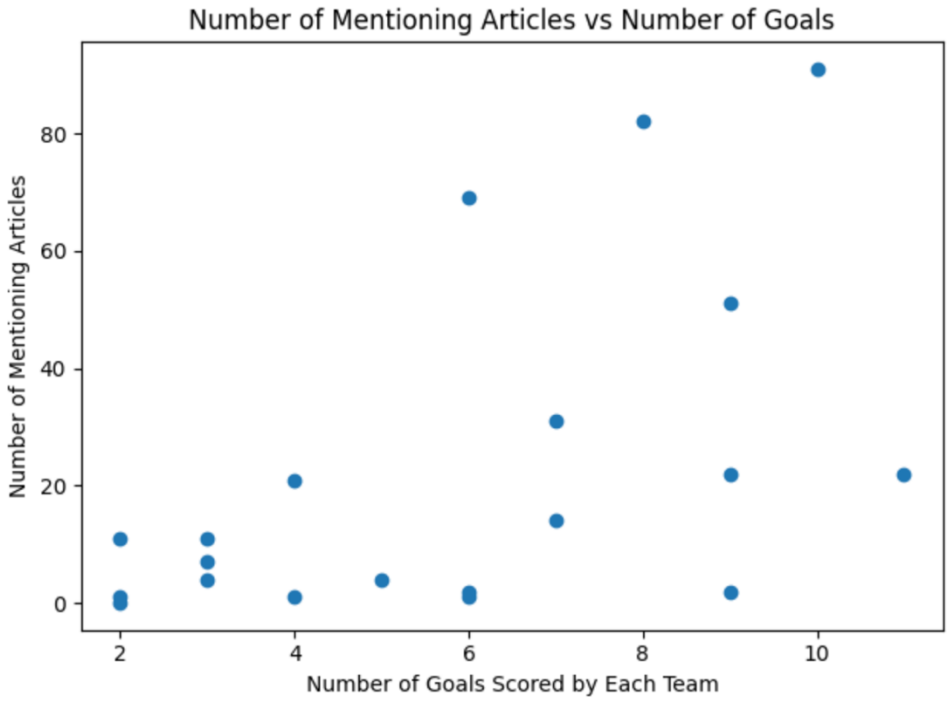
The graph shows that the clubs Chelsea, Arsenal, Manchester United, and Liverpool have a relatively higher number of mentions, whereas the clubs West Ham United, Wolverhampton, Huddersfield, and Brighton have a relatively lower number of mentions. The difference in mentions may be due to the popularity or performance of each club, which will be analyzed in Task 7.

Task 6: Heat Map



A brighter color on the graph indicates a higher similarity score, whereas a darker color indicates a lower score. A black block means that no articles are mentioning both clubs. The similarity score for the same club is decided to be 1 since a club is the same as itself. As the map shows, clubs that have a relatively higher score include Arsenal and Chelsea, Arsenal and Manchester United, and Burnley and Cardiff, meaning that those pairs are commonly mentioned together in the same articles. Those pairs are commonly mentioned together is maybe because they have many games against each other.

Task 7: Scatter Plot



The plot shows that the two traits have a weak positive correlation, meaning that the teams that scored more goals tend to be mentioned more. Some teams have relatively higher goals and a very low number of mentions. The high number of mentions in the middle (goals 6-8) is maybe because some teams are popular, and so they are mentioned by many articles even if they do not perform the best. The low points on the right (goals 6-10) are maybe because those clubs are the clubs that always lose but are always mentioned together with their winning opponents. However, there are no clubs that perform poorly but have a very high number of mentions, and the correlation is strong if we take the uppermost points.