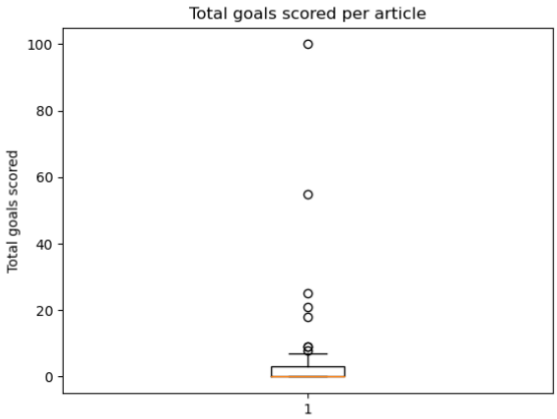
COMP20008 Assignment 1 Report

Assignment 1 dealt with many techniques for processing text, manipulating data frames, and creating graphical interpretations of the data. One of these text processing techniques included using regular expressions to find certain text elements, specifically the score written in each article.

Regular Expressions

My regular expression I employed was; “\s(\d{1,2})-(\d{1,2})[^\d]”. This was chosen since as per the assignment outline, the assumed maximum number of goals a team can score was 99, thus the regular expression can have only either one or two digits.
The “[^\d]” component is to avoid use cases of years such as “1965-1972” and ignore the “19” as a score. Since in regular English there would be whitespace before the score, the “\s” was used, however in text files with poor grammar there could be cases with no whitespace before the score, and thus the regular expression would fail to record this.
Examples include, but are not limited to; “the score was6-2”, or “Final score:3-0”

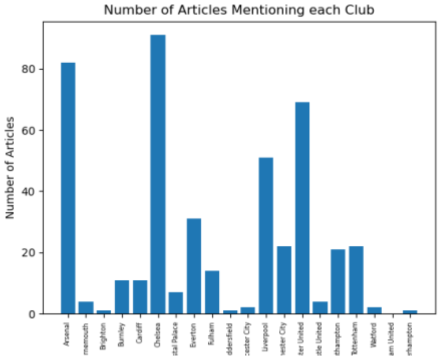
Graphical Data Analysis

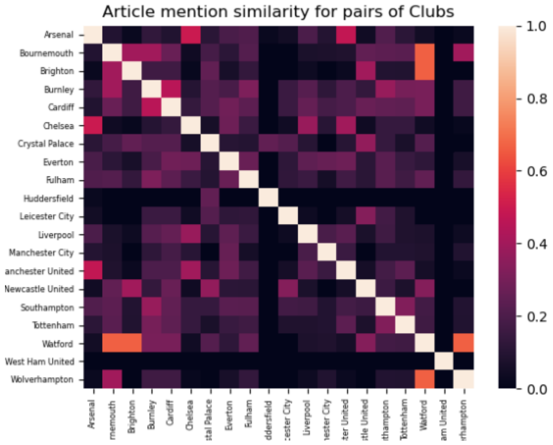


The task 4 boxplot (pictured left) shows the goals scored in each article, these being found using the regular expression previously referenced. The first key takeaway is the dense and small box, slightly extending above 0 goals, which can be interpreted as football being a low scoring game. This is true in practise since generally football games have low scoring with total goals being around 0 to 4.

The outlier dots being those except the two closest to the upper-inner fence are a reflection of the regular expressions correct identification of scores. Especially the dots at 100 and around 55, as these are far too unlikely score outcomes in a football match.

The task 5 bar chart represents the number of articles that mention each club, and from this data we can interpret which clubs are more popular and/or larger in fan-base/size. The four largest in order based on article mentions being; Arsenal, Chelsea, Manchester United, and Liverpool. These happen to be four of the largest clubs not only in England, but also the world, therefore this method of evaluation has merit.





The heatmap produced in task 6 compares pairs of clubs and quantifies their similarity based on the number of articles that mention both of them, and their total number of mentions.

This indicates that certain pairs, such as Watford and Bournemouth, and Watford and Brighton, as well as Watford and Wolverhampton, are mentioned together frequently and there are two factors that explain this.

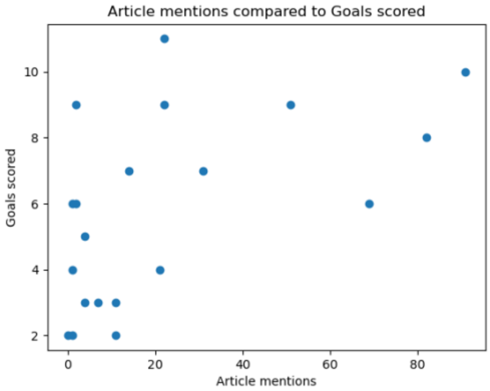
Firstly, is that all four clubs mentioned all have a very low number of articles mentioning them, which influences our similarity equation[1] less.

Secondly, this leads onto indicating that these four clubs have less general articles written about them and instead are only mentioned in match review articles between the clubs.

Task 7 produces a scatterplot that shows the goals scored by each team against the number of goals they have scored.

The main interpretation that can be made is the more articles a club is mentioned in generally means they score more goals, and the opposite is true. However there are still outliers such as one club with only 22 article mentions despite 11 goals being scored.

The data seems to follow a logarithmic pattern, especially since from task 4 we know that football is rarely high scoring.



[1] Sim(club1, club2) = (2 x number of articles mentioning both clubs) / (club1 mentions + club2 mentions)