COMP20008 Assignment1 Report
1123422
Daxuan Yue

In Task3 of this assignment, I used a regular expression which is '\d?\d?\d?\d-\d?\d?\d?'. To come up with this pattern of expression. I observe the object which is the score of a game in text. It is obvious that all the scores are displayed in number-number. At first, I just set the pattern of regex as '/d-/d'. After I discover the score can be more than 1 digit. I changed it to '\d?\d-\d\d?', since it can either be 2-digit or 1-digit. However, when the output is still wrong, I discovered that expression of date can also been extract and it shall not be calculated as score. As a result, I make the pattern as '\d?\d?\d?\d-\d\d?\d?\d?', so I can remove the date and detect all the score either in 1-digit or more-digit. Unfortunately, I cannot solve the problem of '218.txt' in this task, so I cheated to make an 'if' expression for it.
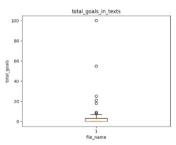
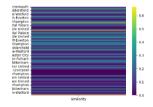In task4, 5, 6 and 7, a visualization of data was managed.

From boxplot of task4, people can see that the median is near 0 since many news reports 0 goals. Despite the nulls, most of total goals concentrated around 10. There are also three values around 20 and 3 and two outliers which one is 100 and other is between 50 and 60.

From the bar chart of task 5, It is obvious to see that Arsenal, Chelsea, Liverpool and Manchester United are always mentioned. Chelsea has the most frequency mentioning in the news reporting. The other teams are mentioned much less than these four teams, but West Ham United were never mentioned by news.

From heatmap of task 6 , It is easy to see that the majority of similarity score is around 0.

From scatterplot of task 7, people can easily see that the team with more time of goals can have the more time of been mentioned by media. The teams with less goals in game will have less chance of been mentioned.