COMP20008 A1 - Task 6: Analysis Report

An interpretation of the scatter plot in Task 3 and what conclusions we might draw from it (1 mark).

- There is no correlation between number of views and time since the publication of the article.
    - Excluding the top right outlier, the range is 0-0.6 * 1e6, with most values in a tighter range between 0-0.3 * 1e3, of which the articles are mostly evenly distributed throughout the period.
        - E.g., articles published at 0-2000-, 7000- and 13000-minute mark all have a max view count of 0.2 * 1e6, indicating time passed does not affect view count (null relationship)
- Additionally, the ‘cluster’ of articles at the 0–2000-minute period (compared to other periods where it is like a single ‘column’) can be explained by the conversion of the “when” values in task1.
    - E.g., if articles were published 5 days ago, it will convert to 7200 minutes, regardless of where during the 5th day it was published.
    - Whereas if multiple articles were published within 24hrs, the minute values will be more accurate/distributed due to the unit used to record the value.

An analysis of the suitability of text pre-processing steps used in Task 4 and suggestions for improvement (1 mark).

- The text pre-processing steps used are simple and somewhat effective in breaking down the titles into tokens.
    - Because the titles are properly formatted, we do not need to worry about format errors.
    - But because of this simplification, contractions such as ‘don’t’ map to “dont” which is not an English word.
    - Additionally, lower-case-folding means that “NOT GOOD” will be mapped to the same tokens as “not” and “good”, removing the emphasis and meaning of the capitalisation.
- Improvements include using lemmatizing or stemming and stop word removal to increase prevalence of content words.
    - This can be done via sklearn and nltk libraries, instead of using simple built-in methods.

An interpretation of the bar plot in Task 5, and what conclusions we might draw from it. You should consider the context in which the top five words appear in article titles to support your analysis (1 mark).

- From the bar-plot, we can conclude that the context of the most popular articles surrounds the persons “Watters”, “Kamala”, “Jesse”.
    - However, the dataset reveals that Jesse Watters is the same person, highlighting a drawback in the visualization as we could not have reached this conclusion without the original dataset.
- Additionally, the most popular articles are politics related as “liberal” is the term with the highest views.
- It is also important to note that being a common word does not correlate to belonging to the same topic, i.e., the context behind the “liberal” and “Jesse” articles could be different.
- To address these issues, we could apply clustering techniques to find similar topics and themes.