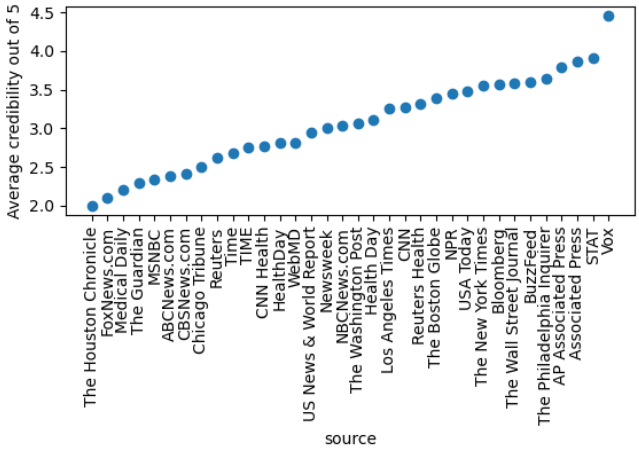
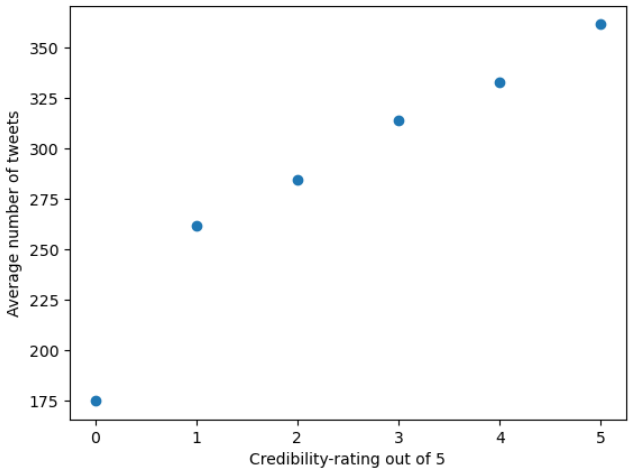
Task 8 - Analysis Report
Henry Routson

Plots

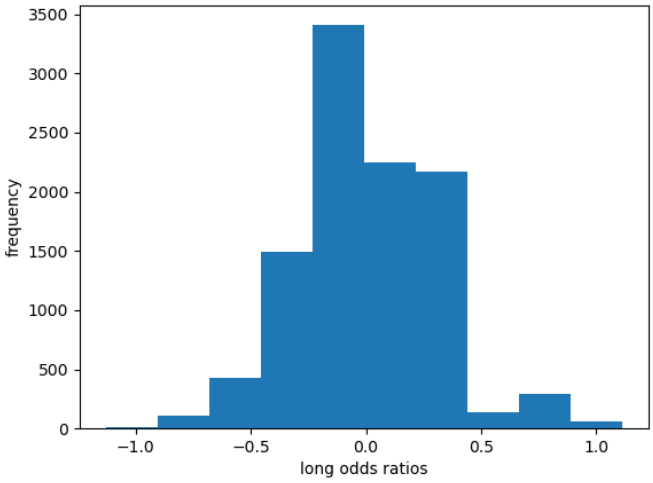
Task 4 b



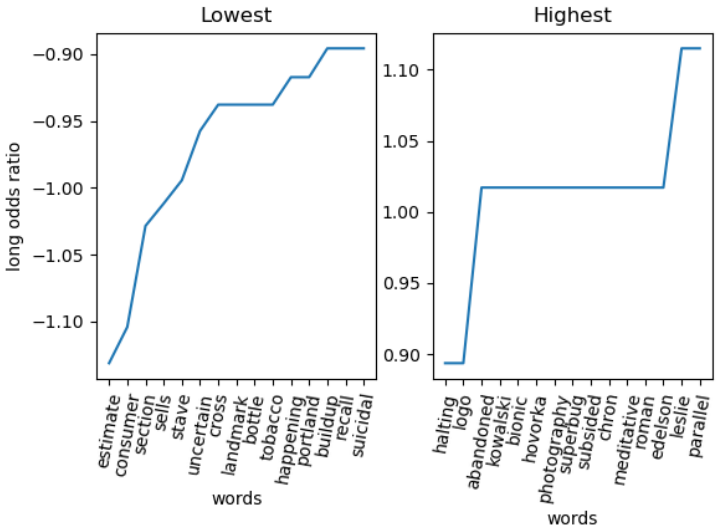
Task5.png



Task7b.png



Task7c.png



Data sources

The data for the graphical analysis above is sourced from the supplied
/course/data/a1/
folder containing tweets, articles and review data.

Plot description

Task4b -
A scatterplot with the news source on the x axis, sorted in ascending order of the average expert credibility rating out of 5

Task5 -
A scatterplot of the number of tweets for each integer expert credibility rating out of 5

Task7b -
A histogram with default matplotlib bin size of the distribution of long odds ratios, which is a ratio indicating the odds of a word being in a fake or real article.

Task7c -
Two line graphs with words with the highest and lowest and highest long odds ratios on the x axis and their ratios on the y axis

Credibility

News sources were overall credible, with the majority of news sources scored above an average article rating of 3, an arbitrary rating which separates real and ‘fake’ news in this analysis. Vox in this was a positive outlier, scoring nearly over half a point higher than any other news source.

Spread and Credibility

Against intuition, the graph task5.png seems to show that the more credible a news article is, the more it is retweeted.

Log odds ratios

Task7b shows a left skew bell curve, toward a negative odds ratio, which indicates that more than half of the words in the articles have better odds for being in a fake article. This could be due to fake articles being longer, but further analysis would be needed to see if this is the case.

“Fake” words

The Graph for task7.png shows the words which most indicate fake and real articles. In my personal experience these are not the words which I would use to identify fake articles, but they make sense in this for this analysis to detect them. I might use a semantics parts of the text like “flat earth” to determine if something is fake, but because not all fake articles would share content words, non content words like “uncertain” and “estimate” have been picked up. As for real article indicative words, the analysis seems to have picked up on non-controversial content words like “photography”, meaning the article that contains them is unlikely to be fake. Overall this analysis is interesting, but would not allow me to effectively detect illegitimate articles.

Limitations and future analysis

This dataset, like all others, is limited and does not allow for the perfect analysis. Having access to the content of tweets would be extremely useful, allowing us to look at engagement by their length, sentiment with sentiment analysis, and a number of other techniques. We could for example look at how closely expert and public opinion aligned, comparing expert ratings to some measure of public perception, like how often words like “fake” and “unrealistic” appear in article tweets.

Data processing

Much of the data processing used a number of pythons inbuilt data strucutres like dictionaries, and more advanced data structures like a pandas DataFrame, which allow the use of a number of high level methods.

There seems to be a typo in the task description, as there is no plot for task 6