**Assignment 4**

**Task 1**

**Frequency position of the words “applicant”, “and”, “attack”, “protein”, and “car”:**

[('and', 2), ('applicant', 448), ('attack', 512), ('car', 648), ('protein', 3167)]

**Task 2**

**five words with the largest regression coefficients**

[('mr', 141), ('applicant', 448), ('tribunal', 491), ('respondent', 921), ('appellant', 1035)]

As we can clearly see, at least 4 of the 5 words are definite legal terms and correlate very well with court cases. This shows that the model worked well in identifying important words in predicting Australian court cases.

Chart, histogram

Description automatically generated

**Task 3**

1. **F1 score obtained by the classifier.**

Unfortunately, I am getting a F1 score of 0 . I have 18347 True negatives (Wikipedia pages) and 377 false negatives (AU court cases classified as Wikipedia pages).

Since the precision is 0, the Fscore is 0.

1. Analysis of Three of the false positives

Actually, I did not have any false positives. All data was classified as negative (non- Australian court cases)

1)This result can be attributed to the heavily unbalanced nature of the dataset. Though there are good terms that have been identified as having highest weights for predicting Court cases, ultimately the imbalanced nature of the dataset caused the court cases text to be misclassified.

2) We did not filter for stop words. There are a lot of words in the dictionary that are common for both categories of text and they might have overwhelmed the overall classification.