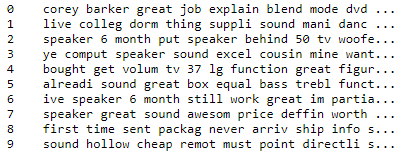
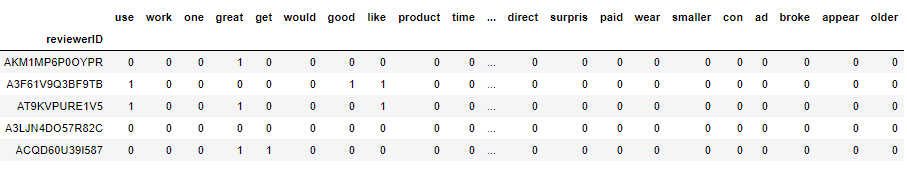
## Question 1 to 3:

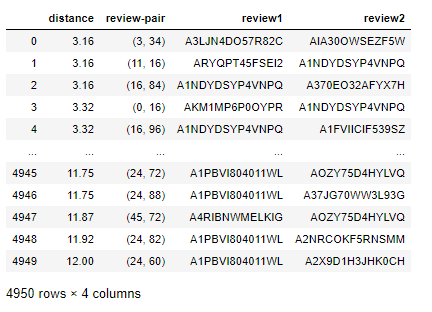
****The sample of output of converting to lower case and stemming after ignoring spellings and removing negations.

Question 4 : Frequency count, most frequent (500) and bag of word representation.

****A screenshot of a cell phone

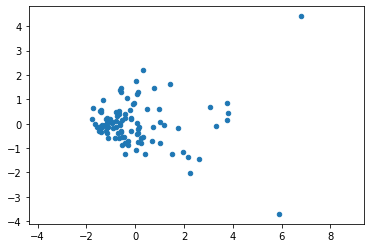
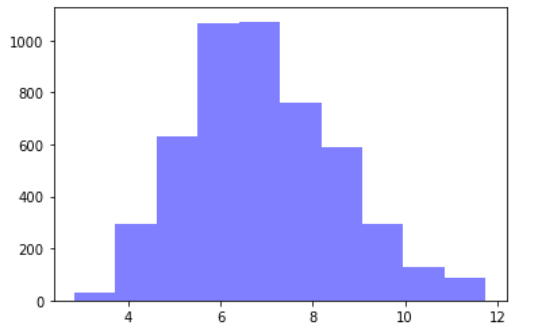
Description automatically generated

Question 5:

We look at the Euclidean distance (L2 norm) between the vectors to see how close the reviews are with respect to one another.

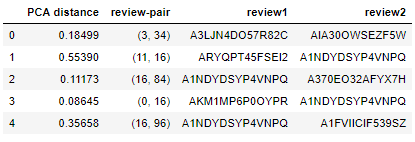
Question 6:

The PCA and graph the first two PCs for the first 100 reviews.

****

Firstly, from the histogram of BoW Euclidean distances, we can see that most vectors gather together, which also can be reflected by the scatter plot for the first two principal components.

Secondly, when we investigate the nearest pairs, the results of PCA distance match the result generated by Euclidean distance.

****

Question 7:

#### Is it possible to use logistic regression to predict the rating (5 levels) of a product? Why may it be advantageous to use a logistic regression to predict the 5-star rating over MNL with 5 categories?

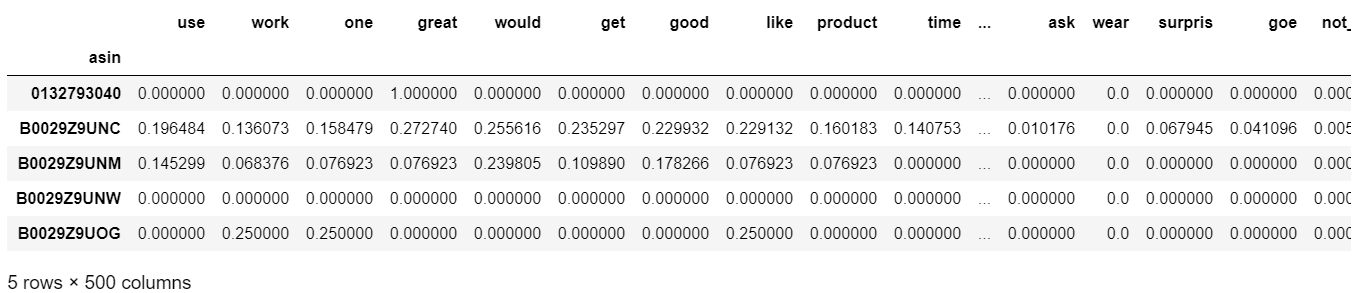
Yes, we can use logistic regression on ordinal variables.

It is advantageous to use logistic regression for that MNL does not retain the information of the order, i.e. it treats all-star ratings as separate but equal categories. While logistic regression could capture the order information when we curve the range of data.

Question 8: Lasso logistic regression and measure the out-of-sample accuracy

We use helpful\_start, helpful\_end, unixReviewTime from the original data, BoWs of reviewText and BoWs of reviewSummary as our dependent variables, and overall rating as the independent variable. The out of sample accuracy is 0.549, via a 3-fold CV lasso logistic regression.

Question 9: Method to aggregate reviews by product.



We average the BoWs of reviews grouped by product ID with a weight of ‘helpful’ proportion.

By giving the more helpful review a larger weight, the aggregated BoW might be able to map the product overall score with more accuracy.

Question 10: Method for measuring the distance between different products.

We apply Euclidean distance calculation on the ‘helpful proportion’ weighted average BoW, and the results suggest that the 2nd product and the 9th product are nearest, and hence most similar according to the reviews.

|  |  |  |
| --- | --- | --- |
| Review ID1 | Review ID2 | Distance |
| 7 | 9 | 5.08 |
| 2 | 9 | 5.12 |
| 1 | 6 | 5.25 |
| 5 | 9 | 5.27 |
| 4 | 9 | 5.33 |
| 8 | 9 | 5.37 |
| 3 | 9 | 5.53 |
| 6 | 9 | 5.61 |
| 0 | 1 | 5.62 |
| 0 | 9 | 6.08 |