# Homework 7: Text Bag-of-Words Search and Classification

**Page 1 Distribution graph (5 points)**

Show the distribution graph of words counts vs word rank.

**图片包含 屏幕截图

描述已自动生成**

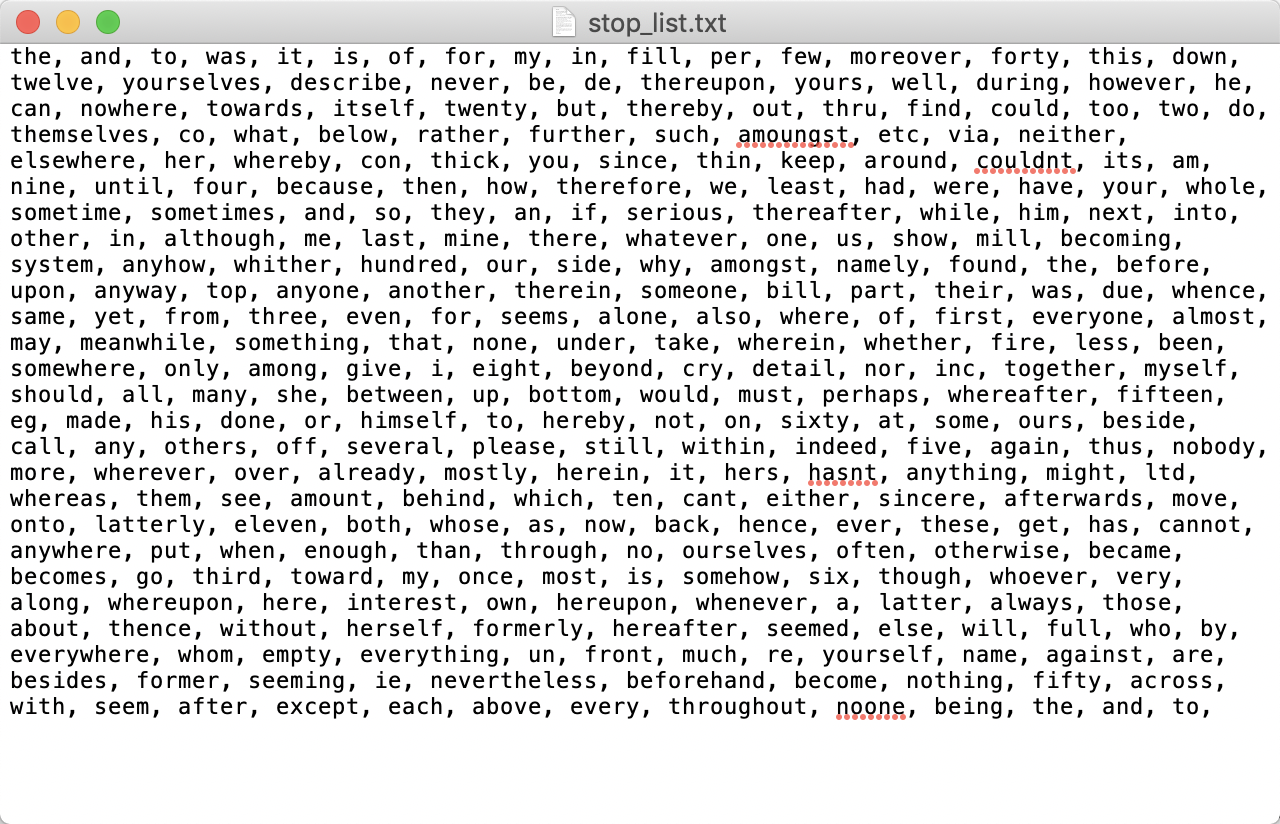
**Page 2 Identify the stop words (5 points)**

List the stop words you choose as well as the frequency threshold.

The max document frequency threshold is 5000.

The minimum word occurrence threshold is 10.

Stop words (331 words):



I choose Term Frequency-Inverse Document Frequency to decide the stop words and set threshold as 74.9.

And then add common English stop words and those words with word occurrence > 4000.

**Page 3 Distribution graph again (5 points)**

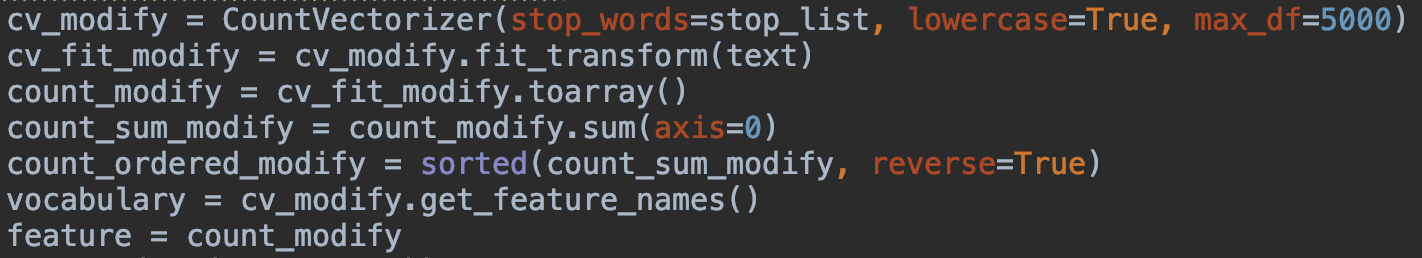
After chosing the stop words, show the distribution graph of words counts vs word rank.

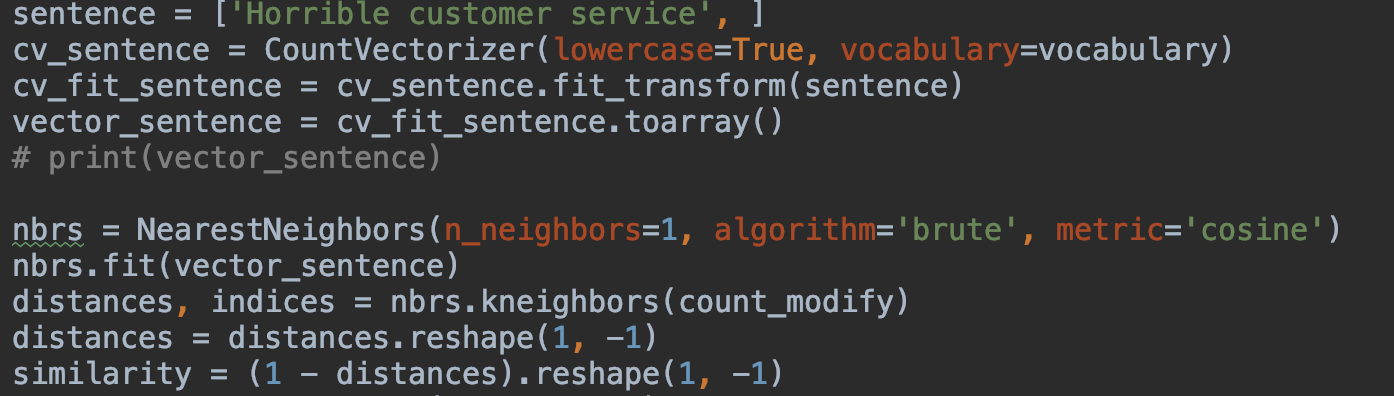
**图片包含 屏幕截图

描述已自动生成**

**Page 4 Code snippets (15 points)**

Show the snippet of your code that you convert all the reviews into bag-of-words formulation using your chosen stop words and your code for nearest-neighbours with cos-distance.

****

****

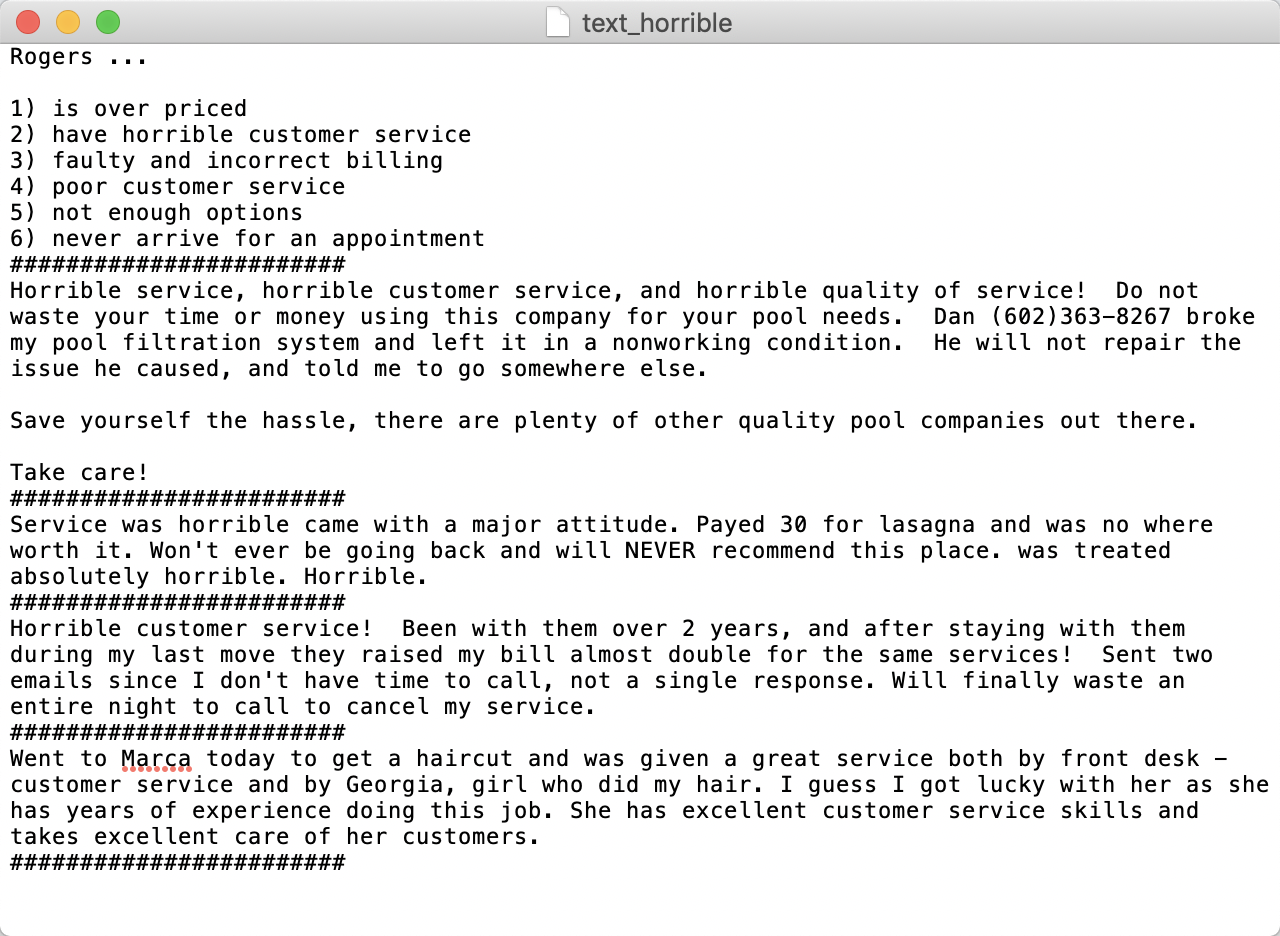
**Page 5 Reviews with score (10 points)**

Show the original reviews with the distance scores

The lowest 5 distance scores are:

0.22848325, 0.43408354, 0.49604737, 0.49604737, 0.51887478

The original reviews accordingly:



**Page 6 Query results (10 points)**

Show your document results and explain the reasons that you choose them.

图片包含 地图, 文字

描述已自动生成

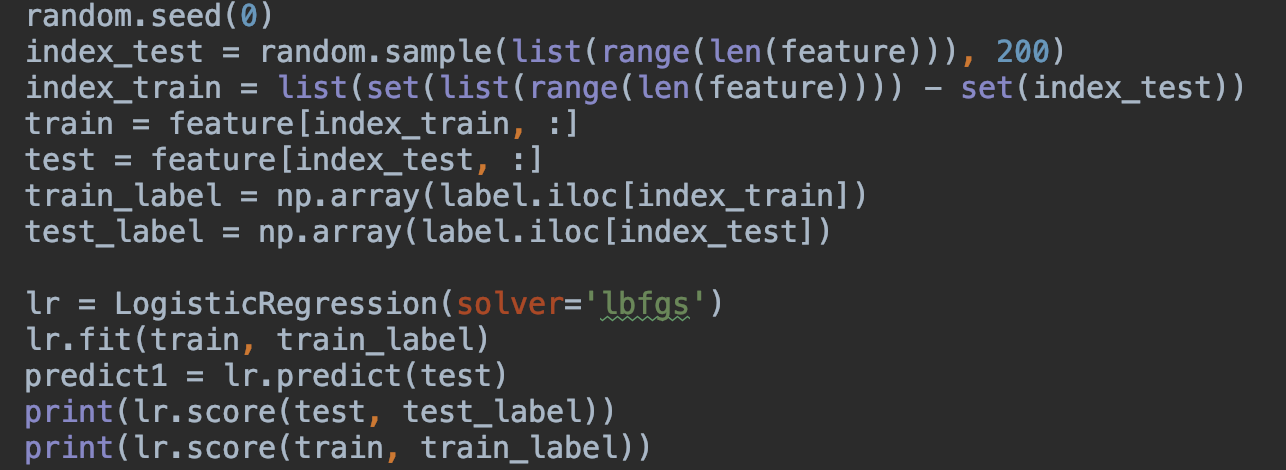
 I choose 0.6 as the cosine distance threshold (i.e. 0.4 as the similarity threshold) and there are totally 16 documents matching the result.

**Page 7 Accuracy with threshold 0.5 (10 points)**

Show your code for creating classifier. Report the accuracy on train and test dataset with threshold 0.5.

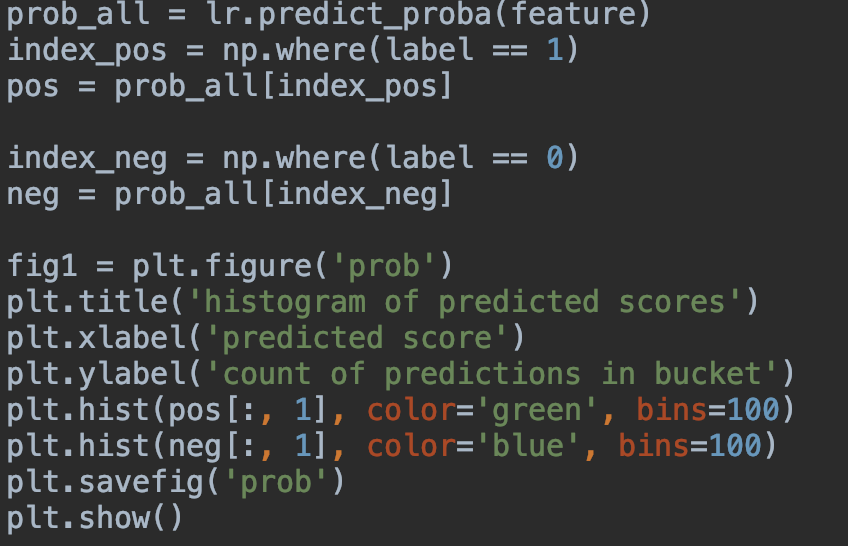
Accuracy of training set: 0.9988888888888889

Accuracy of test set: 0.945

****

**Page 8 Predicted scores (10 points)**

Show your code for plotting predicted scores and show the figure.

****

**图片包含 屏幕截图

描述已自动生成**

**Page 9 Accuracy again and curve (20 points)**

Report the accuracy on train and test dataset with a different threshold. Explain why you choose that threshold.

Plot the ROC curve.

When threshold=0.6, accuracy on train set is 0.9983333333333333 and accuracy on test set is 0.96.

The reason to choose 0.6 is that according to the histogram of predicted scores, threshold should be between 0.4 and 0.6 so that we could better tell different labels apart. And after trying different values, 0.6 gives the best accuracy result.

图片包含 地图, 文字

描述已自动生成

**Page 10 Best threshold (10 points)**

Choose the threshold that minimizes false positives while maximizing true positives. Explain your reason.

Threshold = 0.8789133271417255

False positive rate = 0.0380952380952381

True positive rate = 0.9578947368421052

The reason is that this result minimizes the distance between the corresponding point on the ROC curve and point (0,1) so that in this situation, it minimizes false positives and maximizes true positives.

