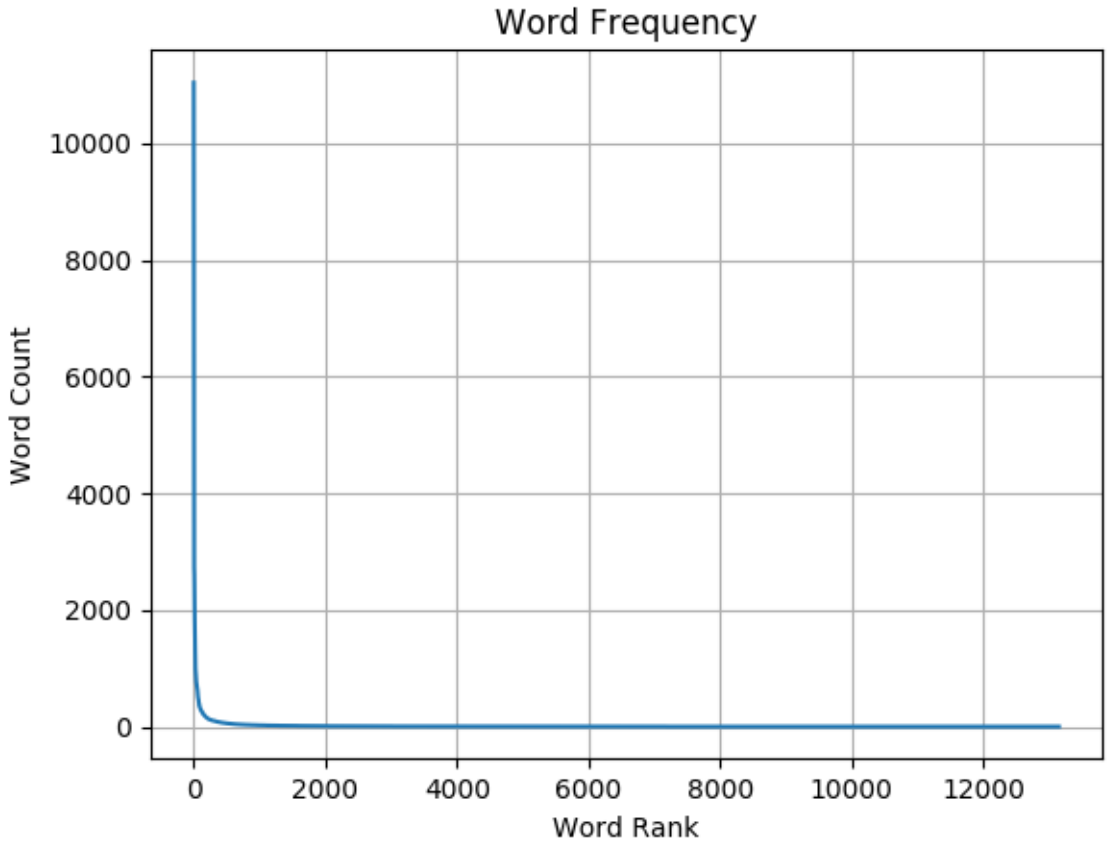
Page 1 Distribution graph (5 points)



Page 2 Identify the stop words (5 points)

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

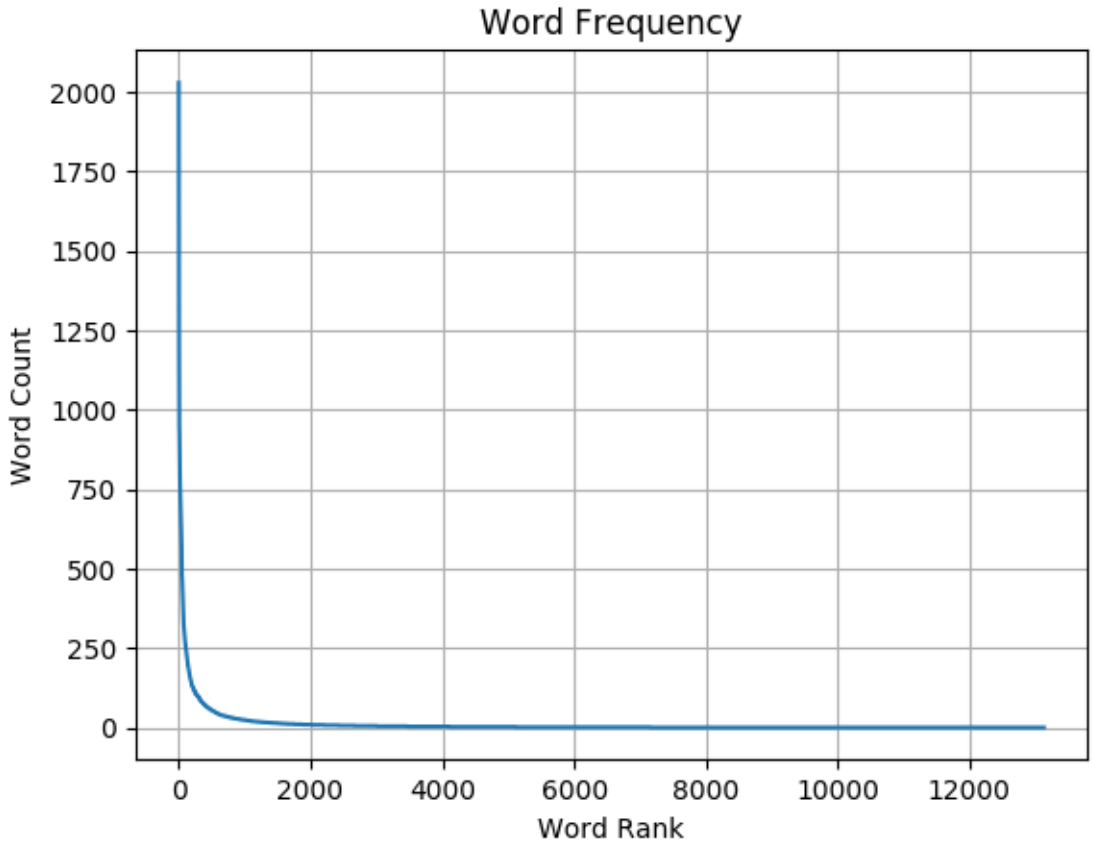
Stop words:

the
and
to
was
it
of
for
in
my
is
that
they
this
we
you
with
on
not
have
but
had
me
at
so
were
are
be
place
food
there
as
he
if
all
when
out
would

The max occurrence is 787 (which is the word "service"), and the min occurrence is 2.

Page 3 Distribution graph again (5 points)

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



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.

Convert all the reviews into bag-of0words formulation:

```
# By looking at the list of the document frequency,
# determine that the max occurrence is 787 (which contains "service")
# The min occurrence is 2
maxdf = float(787) / X.shape[0]
mindf = 1.0 / X.shape[0]
# Represent with bag of words using max_df and min_df
vectorizer = CountVectorizer(max_df=maxdf, min_df=mindf)
X = vectorizer.fit_transform(x_data).toarray()
# Get the word count vs word rank
```

Nearest-neighbours with cos-distance:

```
tfidf_transformer = TfidfTransformer()
X_train_tfidf = tfidf_transformer.fit_transform(X)
neigh = NearestNeighbors(n_neighbors=5, n_jobs=-1)
neigh.fit(X_train_tfidf)
X_test_counts = vectorizer.transform(['Horrible customer service'])
X_test_tfidf = tfidf_transformer.transform(X_test_counts)
res = neigh.kneighbors(X_test_tfidf)
print(x_data[res[0]])
np.savetxt("horrible_customer_service_reviews_Tfidf.txt", x_data[res[0]],
fmt="%s", delimiter='\t
```

Page 5 Reviews with score (10 points)

Review 1, score 1.075796:

Show the original reviews with the distance scores

service was horrible came with a major attitude. payed 30 for lasagna and was no where worth it. won't ever be going back and will never recommend this place. was treated absolutely horrible. horrible.

Review 2, score 1.097646:

rogers ...

1) is over priced

2) have horrible customer service

3) faulty and incorrect billing

4) poor customer service

5) not enough options

6) never arrive for an appointment

Review 3, score 1.154409:

horrible service, horrible customer service, and horrible quality of service! do not waste your time or money using this company for your pool needs. dan (602)363-8267 broke my pool filtration system and left it in a nonworking condition. he will not repair the issue he caused, and told me to go somewhere else.

save yourself the hassle, there are plenty of other quality pool companies out there.

take care!

Review 4, score 1.2149295:

horrible customer service. when you call you'll get transferred 3 times. no one knows what they're doing. was trying to buy a car and the dealer took forever to get back to us. not a good experience vs other places.

Review 5, score 1.2237004:

horrible customer service! been with them over 2 years, and after staying with them during my last move they raised my bill almost double for the same services! sent two emails since i don't have time to call, not a single response. will finally waste an entire night to call to cancel my service.

Page 6 Query results (10 points)

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

34 documents are good match, by looking at the score of all documents. As seen in below scores ranking from smallest to largest, at document 35, there is a larger difference between the score of document 34 and 35.

Also, the 35th document starts to have positive review:

“their flower arrangements are gorgeous! the attention to detail and excellent customer service is what keeps me coming back! highly recommend!!'”.

1.0757961633	1
1.0976463595	2
1.1544088518	3
1.2149295228	4
1.2237003714	5
1.2266540729	6
1.2270682799	7
1.234706347	8
1.2372764566	9
1.2388658106	10
1.2443750706	11
1.2467453589	12
1.2548146661	13
1.2588021859	14
1.2598877607	15
1.261833861	16
1.263626935	17
1.2659438416	18
1.2676944873	19
1.2746214597	20
1.2765978382	21
1.27868683	22
1.2787804836	23
1.2837202103	24
1.2838896926	25
1.2842600152	26
1.2845913381	27
1.2931751256	28
1.2936957273	29
1.2939073666	30
1.2958910187	31
1.2970740817	32
1.297189163	33
1.2978903998	34
1.3025038859	35
1.3026358561	36
1.3029975459	37
1.3033729954	38
1.3043988237	39

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.

```
x_train, x_test, y_train, y_test = train_test_split(X, y_data, test_size=0.1,
random_state=0)
logisticRegr = LogisticRegression()
logisticRegr.fit(x_train, y_train)
score_train = logisticRegr.score(x_train, y_train)
score_test = logisticRegr.score(x_test, y_test)
```

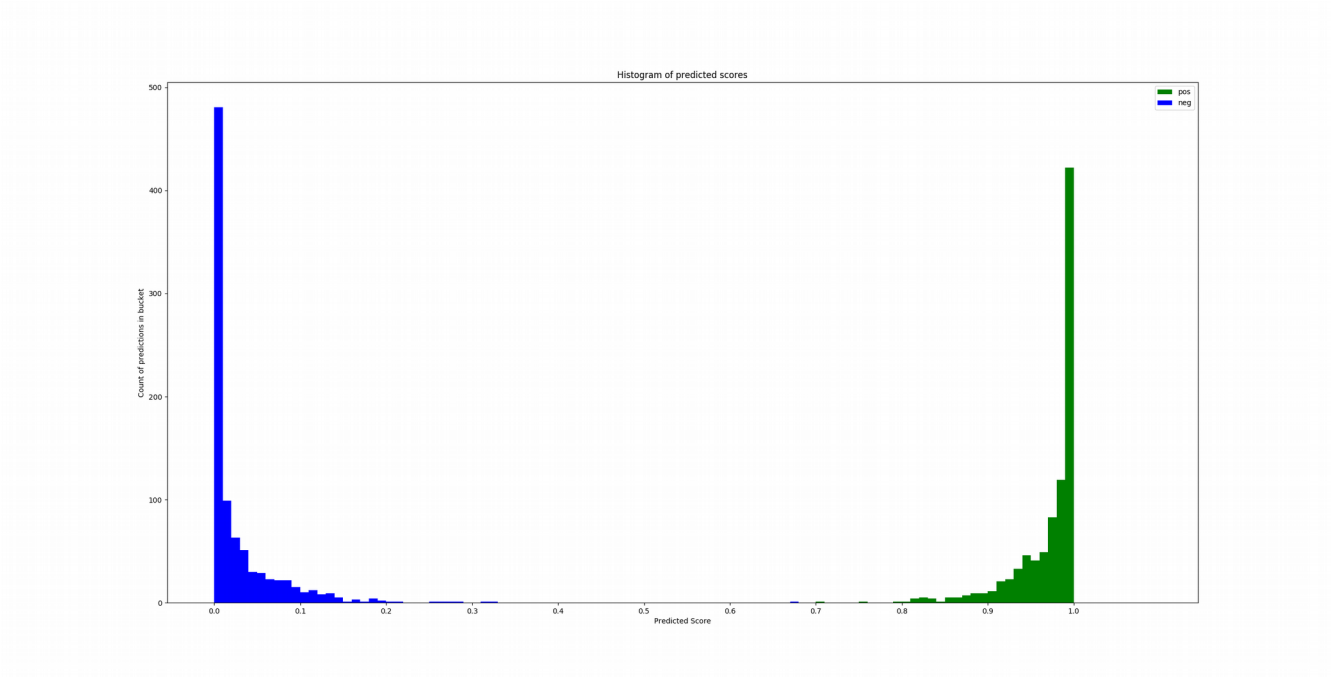
The accuracy on training data is: 0.99944

The accuracy on testing data is: 0.915

Page 8 Predicted scores (10 points)

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

```
prob_train = logisticRegr.predict_proba(x_train)
prob_positive_train = prob_train[:, 1]
index_pos = np.where(y_train==5)
index_neg = np.where(y_train==1)
plt.hist(prob_positive_train[index_pos], np.arange(0.0, 1.1, 0.01), color='green',
label='pos')
plt.hist(prob_positive_train[index_neg], np.arange(0.0, 1.1, 0.01), color='blue',
label='neg')
plt.ylabel('Count of predictions in bucket')
plt.xlabel('Predicted Score')
plt.title('Histogram of predicted scores')
plt.xticks(np.arange(0, 1.1, 0.1))
plt.legend()
```



Page 9 Accuracy again and curve (20 points)

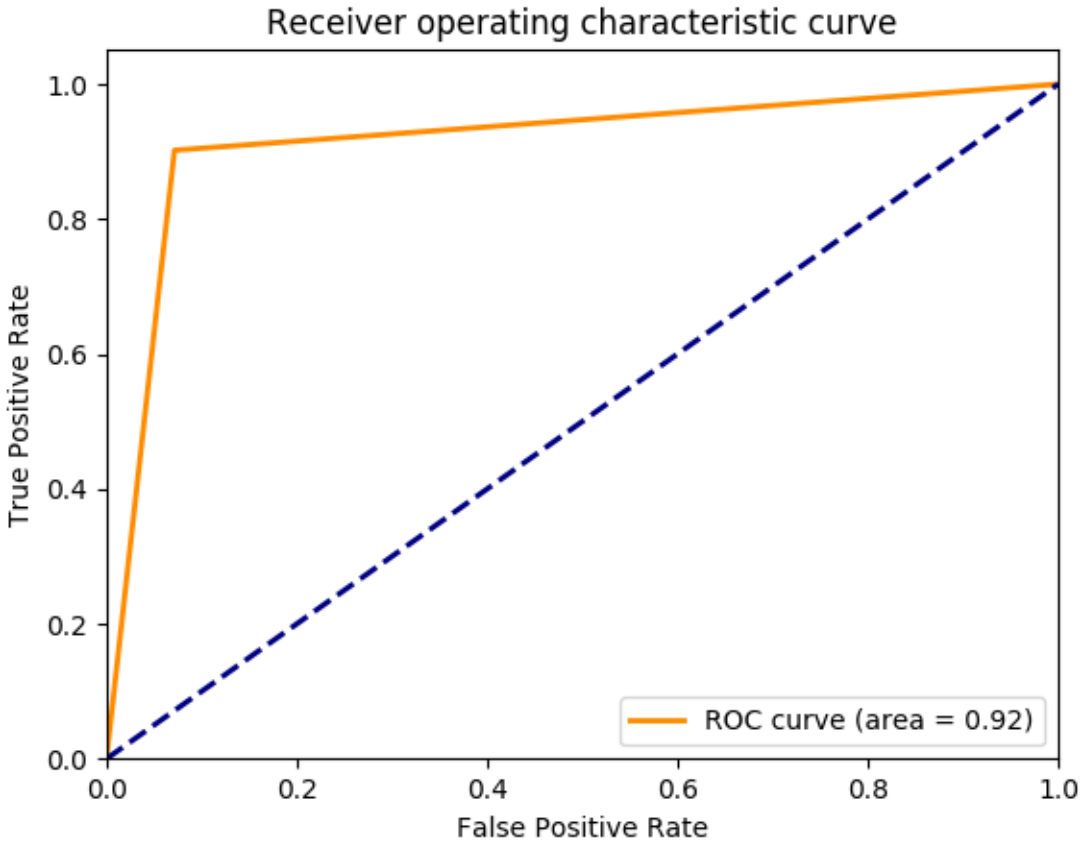
Report the accuracy on train and test dataset with a different threshold. Explain why you choose that threshold.

Plot the ROC curve.

By looking at the histogram of Page 8, we can see that several negative values with relatively high score (around 0.7). Therefore, we choose threshold value of 0.7, and rerun the codes. The below table shows the accuracy obtained with threshold value of 0.6, 0.7 and 0.8. As seen, although the training data accuracy can be increased with higher threshold value, the test data accuracy obtained are lower.

Threshold value	Train accuracy	Test Accuracy
0.6	0.99944	0.905
0.7	1	0.895
0.8	0.99833	0.86

ROC curve:



Page 10 Best threshold (10 points)

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

By looking at the ROC curve in Page 9, the point which minimizes the distance from (0,1) point is chosen. The false positive rate of this chosen point is 0.07142.