Assignment4.1

Part a:

The total number of unique (i.e., distinct) words that appear in the training set, the test set, and the entire dataset (training set + test set), respectively.

```
    unique word in train = 53975
    unique word in test = 47376
    unique word in all = 61188
```

The average and standard deviation of document length (in terms of number of words) in the training and test sets, respectively.

```
    average_Document Len_in_train = 245.3900
    std_Document Len_in_train = 499.3754
    average_Document Len_in_test = 239.4296
    std_Document Len_in_test = 473.9901
```

The total number of unique words that appear in the test set, but not in the training set.

```
    uw_only_test is 7213.
```

The 10 most frequent words and their number of appearances in the entire dataset

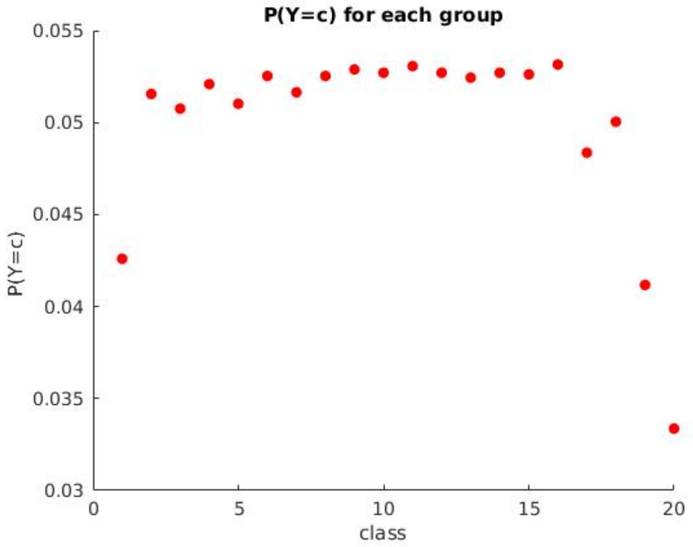
```
    freq_ten_word are:
        'the'   237369
        'to'    119324
        'of'    106116
        'and'   93719
        'in'    79824
        'is'    69167
        'that'  64897
        'it'    54650
        'you'   44312
        'for'   44014
```

The smallest number of times that a word appears in the entire dataset (training set + test set), and the number of words that appear these many times. Among these words, list all the ones that start with “od”.

```
    Smallest_time is 2 and 14786 words only appear 2 times;
    The word started with 'od' among them are:
        'odw'
        'oddibe'
        'odp'
        'odysseus'
        'odor'
        'oddballs'
        'odder'
        'odishe'
        'odiselidge'
        'oda'
```

Part b:

Attach a plot of the observed prior probabilities in the training set. Comment on what you obseve.



Most group has close appear probability, but class 1, 17, 19 and 20, has a relatively low probability of appearance.

Among the W × 20 estimated parameters, how many of them are zero?

200778

For some test documents, P(Y = c|x) = 0 for all c. What is the total number of such test documents? Can you explain why their probabilities are zero?

6958, If a word w_i appear in test but never shows in train, then the probability for all group is zero for all c, P(w_j|c) is zero.

The test CCR:

9.46%

The 20 × 20 confusion matrix:

		Real Lable																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Classified As	1	316	352	354	350	351	351	317	369	376	372	363	375	373	375	375	367	350	358	297	237
	2	0	26	3	3	2	2	4	0	0	0	0	1	1	0	0	0	0	1	0	0
	3	0	2	19	5	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	1	6	28	2	1	4	0	0	0	0	0	1	0	0	0	0	0	0	0
	5	0	1	3	2	22	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
	6	0	4	1	0	0	33	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	7	0	0	0	4	1	0	46	1	0	1	0	0	3	1	0	0	0	1	0	0
	8	0	0	0	0	0	0	2	19	1	0	0	0	1	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	2	2	20	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0	1	0	0	3	36	0	0	0	0	0	0	0	0	0
	12	0	0	1	0	0	0	0	1	0	0	0	17	3	1	0	0	0	0	0	0
	13	0	1	1	0	4	0	0	2	0	0	0	1	8	1	0	0	0	0	0	0
	14	0	0	2	0	1	1	0	0	0	0	0	0	1	14	0	1	0	0	1	1
	15	0	2	0	0	0	0	1	0	0	0	0	0	1	0	17	0	0	0	1	1
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	1	4
	17	0	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	12	1	2	0
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	15	0	0
	19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	8	2
	20	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	6

most missclassifying is caused by no probability and classified as group1.

Part c

The total number of non-zero estimated βw,c ’s. 200778

The test CCR. 11.14%

The total number of test documents where $P(Y = c|x) = 0$ for all c. If there are still such test documents even after removing the words that only appear in the test set, explain why.

6768 word with all zero probability. Because if some word never appear in one group and some word never apper in another group, the probability could be zero. It subpress all the information for other group and the information from new appear word.

Part d

The test CCR: 78.52%

The 20 × 20 confusion matrix. Make sure that you annotate the confusion matrix appropriately.

										Real Lable											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1	249	0	1	0	0	0	0	0	0	0	2	0	3	5	0	11	1	12	6	39
	2	0	286	33	11	17	54	7	3	1	0	0	3	20	7	8	2	1	1	1	3
	3	0	13	204	30	13	16	5	1	0	0	1	0	4	0	0	0	0	0	0	0
	4	0	14	57	277	30	6	32	2	1	1	0	3	25	3	1	0	0	1	0	0
	5	0	9	19	20	269	3	16	0	0	1	0	4	7	0	0	0	0	0	1	0
	6	1	22	21	1	0	285	1	0	0	0	0	1	4	0	3	2	1	0	1	0
	7	0	4	4	10	12	1	270	14	2	2	2	0	8	3	1	1	1	1	0	0
	8	0	1	2	2	2	1	17	331	27	1	1	0	11	5	0	0	2	2	0	0
	9	1	1	3	1	2	3	8	17	360	2	2	0	6	4	1	0	1	0	0	1
Classified As	10	0	0	0	0	0	0	1	0	0	352	4	1	0	1	0	0	1	2	0	1
	11	0	1	0	1	0	0	2	0	0	17	383	1	0	0	1	0	0	0	0	0
	12	2	11	12	4	3	5	0	1	0	0	0	362	21	1	4	0	4	2	5	1
	13	0	8	5	32	21	3	7	13	3	1	0	2	264	8	6	0	0	1	0	0
	14	3	6	10	1	8	6	4	0	1	3	0	2	9	320	5	2	5	0	10	2
	15	3	10	8	2	4	4	6	4	0	3	0	2	7	8	343	0	2	0	6	6
	16	24	1	3	0	0	0	0	2	0	5	1	0	1	7	3	362	1	6	2	27
	17	2	2	1	0	1	1	2	0	1	2	2	9	3	6	2	0	303	3	63	10
	18	3	0	0	0	0	1	1	0	1	1	0	0	0	5	1	1	5	326	6	3
	19	4	0	5	0	1	1	2	6	0	5	1	5	0	8	12	2	23	18	196	7
	20	26	0	3	0	0	0	1	1	0	1	0	0	0	2	1	15	13	1	13	151

The names of the 5 most confused class pairs and their degrees of confusion.

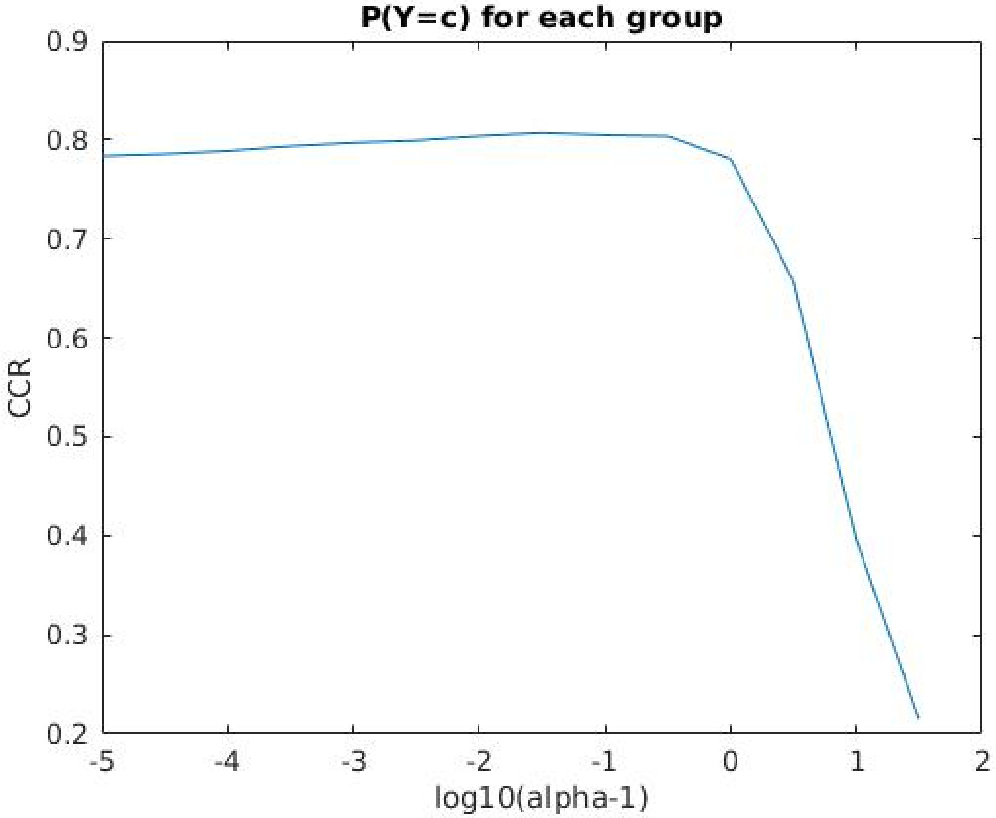
Through the matrix Find_conf we can know that, their are 5 pair has a total misclassifying number more than 55.

they are:

talk.religion.misc	alt.atheism	0.1186
comp.windows.x	comp.graphics	0.0975
comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	0.1112
sci.electronics	comp.sys.ibm.pc.hardware	0.0726
talk.politics.misc	talk.politics.guns	0.1332

Part e

Calculate the test error for different choices of α. Plot the test CCRs (y-axis) as functions of (α − 1) (x-axis). Use a log-scale for (α − 1).



Part f

The total number of unique (i.e., distinct) words that appear in the training set, the test set, and the entire dataset (training set + test set), respectively, after removing the stop words.

uw_train = 53975 uw_test = 47376 uw_all = 61188
which is the same as before removed.

The average and standard deviation of document length (in terms of number of words) in the training and test sets respectively.

average_DL_train = 116.9846
std_DL_train = 253.0562
average_DL_test = 114.6227
std_DL_test = 262.5982

The total number of unique words that appear in the test set, but not in the training set.

uw_only_test = 7213

The smallest number of times that a word appears in the entire dataset, and the number of words that appear these many times. Among these words, list all the ones that start with "aero".

There are 14786 words only appear 2 times.
Among these words, the words that start with "aero" are:
'aeronautical'   'aerodynamic'   'aerosols'   'aerostat'
'aeroplanes'

The test CCR: 78.23% which shows that removing common word does not provid improvement on Naive Bayes classifier. I think the reason is that those word still provid some information between different class

The 20 × 20 confusion matrix.

										Real Lable										
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1	251	0	1	0	0	0	0	0	0	0	1	0	3	4	0	10	1	14	6	37
	2	0	286	33	11	19	49	8	3	1	0	0	2	21	7	8	2	1	1	0	2
	3	0	9	207	33	11	15	5	1	0	0	1	0	5	0	0	0	0	0	0	0
	4	0	14	56	270	28	6	37	2	1	1	0	3	26	3	1	0	0	1	0	0
	5	0	9	17	20	268	2	20	0	1	0	0	3	7	0	0	0	0	0	1	0
	6	1	21	21	1	1	291	1	0	0	0	0	1	3	0	5	2	1	0	1	0
	7	0	5	5	13	10	1	250	13	2	2	2	1	10	2	2	1	1	0	0	0
	8	0	1	1	1	1	1	17	333	27	1	0	0	9	6	0	0	2	2	0	0
	9	1	0	3	1	2	3	9	17	360	3	2	0	7	2	1	0	1	0	0	1
Classified As	10	0	0	0	0	0	0	1	0	0	352	3	0	0	2	0	0	1	2	0	1
	11	0	2	0	1	0	0	3	0	0	17	385	1	0	1	1	0	0	0	0	0
	12	2	12	13	5	3	5	1	2	0	0	0	361	24	1	2	0	4	2	5	1
	13	0	9	5	32	21	2	11	11	2	1	0	3	255	9	6	0	0	0	0	0
	14	3	5	10	1	11	7	7	0	1	3	1	2	10	322	5	2	5	0	12	2
	15	3	11	9	2	5	5	5	3	0	4	0	2	9	9	342	0	2	0	7	6
	16	24	1	4	0	0	0	0	2	0	4	2	0	0	7	3	362	1	6	3	26
	17	1	1	1	1	1	1	3	0	1	2	1	12	3	6	2	0	300	2	60	11
	18	4	0	0	0	0	1	1	0	1	1	0	0	0	4	1	1	5	327	6	3
	19	4	3	2	0	2	1	2	7	0	5	1	3	1	7	12	2	25	18	195	7
	20	24	0	3	0	0	0	1	1	0	1	0	1	0	1	1	16	14	1	14	154

The names of the 5 most confused class pairs and their degrees of confusion.

Through the matrix Find_conf we can know that, their are 5 pair has a total misclassifying number more than 55.
they are:

talk.religion.misc	alt.atheism	0.1114
comp.windows.x	comp.graphics	0.0898
comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	0.0739
sci.electronics	comp.sys.ibm.pc.hardware	0.1137
talk.politics.misc	talk.politics.guns	0.1311