What to submit

A zipped directory with one subdirectory: **code**, and one PDF file. Name your directory "A3-uniquename", where "uniquename" is replaced by your own uniquename. The **code** subdirectory contains the code, the given stop words file, the given ground truth file, and a README with enough information to run your code. Specifically, your program should take one argument: the AAAC directory containing training and test files, and it should print out the test accuracy, the confusion matrix, the feature ranking, and the test accuracy for top k features (k varying from 10 to the maximum in steps of 10). We will go through the code to make sure they are present. Please submit on Canvas, through the Assignments tool. If for any reason that doesn't work, submit it to the Drop Box, and send email.

Here are two sample commands for C++ and python:[9]

```
./mycode AAAC_problems/problemA/
```

```
python mycode.py AAAC_problems/problemA/
```

Now, please answer the following questions (they are graded).[10] Once you are done, save this PDF as "answers.pdf" under "A3-uniquename" directory, zip it, and upload on Canvas.

AAAC Problem A

- What is the test accuracy?[11]

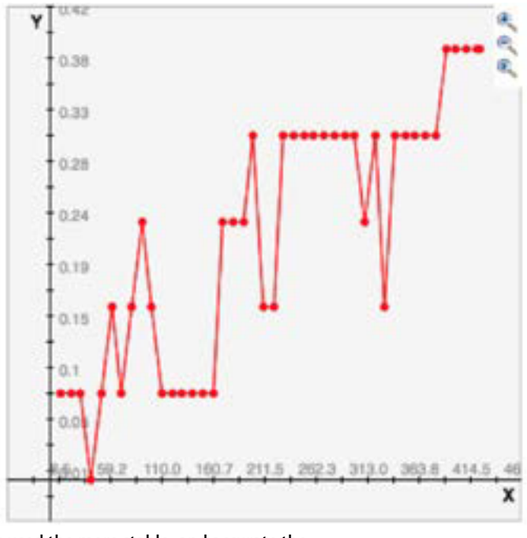
38.46%

- What is the confusion matrix?

- What is the feature ranking? Please list the top 20 features, along with their class-conditional entropy.

```
feature 1: going --- CCE: 0.512008
feature 2: against --- CCE: 0.510255
feature 3: everyone --- CCE: 0.510255
feature 4: under --- CCE: 0.506925
feature 5: find --- CCE: 0.506748
feature 6: always --- CCE: 0.503418
feature 7: need --- CCE: 0.503418
feature 8: greater --- CCE: 0.501842
feature 9: having --- CCE: 0.501842
feature 10: others --- CCE: 0.501665
feature 11: thought --- CCE: 0.500088
feature 12: until --- CCE: 0.500088
feature 13: get --- CCE: 0.500088
feature 14: high --- CCE: 0.496759
feature 15: areas --- CCE: 0.496759
feature 16: men --- CCE: 0.496759
feature 17: working --- CCE: 0.496759
feature 18: off --- CCE: 0.496759
feature 19: two --- CCE: 0.496582
feature 20: give --- CCE: 0.495678
```

- Please give the feature curve as described above (should be legible). What conclusion can be drawn from this curve?



The more features used the more stable, and accurate the graph-- and the program -- becomes. This will hold true for all feature curves.

AA AC Problem B

- What is the test accuracy?

30.76%

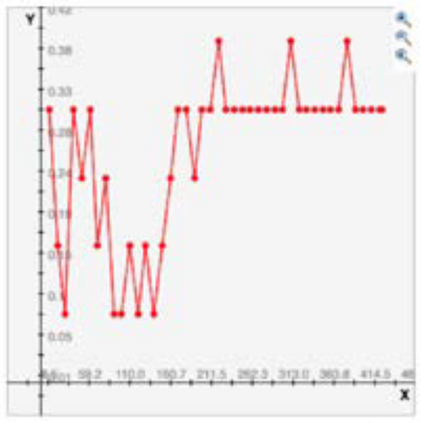
- What is the confusion matrix?

```
 0  1  2  3  4  5  6  7  8  9 10 11 12 13
 1  0  0  0  0  0  0  0  0  1  0  0  0  0
 2  0  0  0  0  0  0  0  0  0  1  0  0  0
 3  0  0  0  0  0  0  0  0  0  1  0  0  0
 4  0  0  1  0  0  0  0  0  0  0  0  0  0
 5  0  0  0  0  0  0  0  0  0  1  0  0  0
 6  0  0  0  0  0  0  1  0  0  0  0  0  0
 7  0  0  0  0  0  0  0  1  0  0  0  0  0
 8  0  0  0  0  0  0  0  0  0  0  0  1  0  0
 9  0  0  0  0  0  0  0  0  0  0  1  0  0  0
10  0  0  0  0  0  0  0  0  0  0  1  0  0  0
11  0  0  0  0  0  0  0  0  0  1  0  0  0  0
12  0  0  1  0  0  0  0  0  0  0  0  0  0  0
13  0  0  0  0  0  0  0  0  0  0  0  0  0  1
```

- What is the feature ranking? Please list the top 20 features, along with their class-conditional entropy.

```
feature 1: going --- CCE: 0.512008
feature 2: against --- CCE: 0.510255
feature 3: everyone --- CCE: 0.510255
feature 4: under --- CCE: 0.506925
feature 5: find --- CCE: 0.506748
feature 6: always --- CCE: 0.503418
feature 7: need --- CCE: 0.503418
feature 8: greater --- CCE: 0.501842
feature 9: having --- CCE: 0.501842
feature 10: others --- CCE: 0.501665
feature 11: thought --- CCE: 0.500088
feature 12: until --- CCE: 0.500088
feature 13: get --- CCE: 0.500088
feature 14: high --- CCE: 0.496759
feature 15: areas --- CCE: 0.496759
feature 16: men --- CCE: 0.496759
feature 17: working --- CCE: 0.496759
feature 18: off --- CCE: 0.496759
feature 19: two --- CCE: 0.496582
feature 20: give --- CCE: 0.495678
```

- Please give the feature curve as described above (should be legible). What conclusion can be drawn from this curve?



The stability with only a few features is very poor and therefore the accuracy is fairly random early on.

AAAC Problem C

- What is the test accuracy?

44.44%

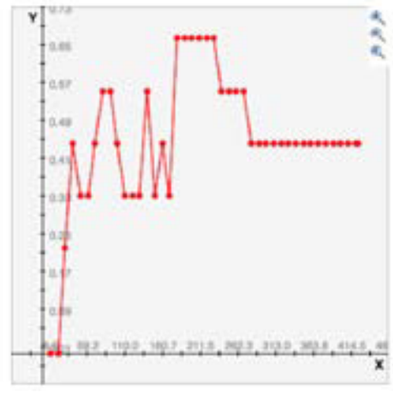
- What is the confusion matrix?

```
0 1 2 3 4 5
1 0 2 0 0 0
2 0 2 0 0 0
3 0 1 0 0 0
4 0 0 0 2 0
5 2 0 0 0 0
```

- What is the feature ranking? Please list the top 20 features, along with their class-conditional entropy.

```
feature 1: differently --- CCE: 0.525068
feature 2: yours --- CCE: 0.525068
feature 3: groups --- CCE: 0.518405
feature 4: gets --- CCE: 0.518405
feature 5: nowhere --- CCE: 0.513706
feature 6: cases --- CCE: 0.511741
feature 7: showing --- CCE: 0.511741
feature 8: ends --- CCE: 0.507043
feature 9: working --- CCE: 0.507043
feature 10: h --- CCE: 0.502129
feature 11: f --- CCE: 0.502129
feature 12: puts --- CCE: 0.495465
feature 13: wanting --- CCE: 0.495465
feature 14: finds --- CCE: 0.492516
feature 15: backs --- CCE: 0.490767
feature 16: goods --- CCE: 0.490767
feature 17: downs --- CCE: 0.490767
feature 18: m --- CCE: 0.490767
feature 19: mostly --- CCE: 0.490767
feature 20: d --- CCE: 0.486848
```

- Please give the feature curve as described above (should be legible). What conclusion can be drawn from this curve?



This plot seems to suggest that if a group of documents is being analysed with a group of stopwords that don't have very large CCE values, the usefulness of the procedure depreciates. The previous 2 documents all had top features with CCE's above .35

AAAC Problem G

- What is the test accuracy?

35%

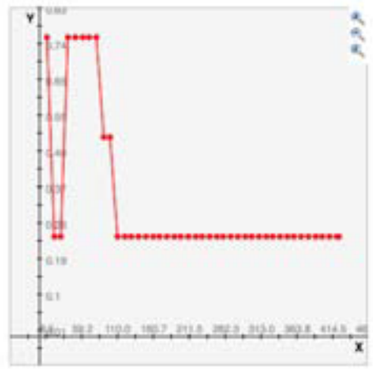
- What is the confusion matrix?

```
0 1 2
1 1 1
2 2 0
```

- What is the feature ranking? Please list the top 20 features, along with their class-conditional entropy.

```
feature 1: everyone --- CCE: 0.528771
feature 2: keeps --- CCE: 0.528771
feature 3: problems --- CCE: 0.528771
feature 4: somebody --- CCE: 0.528771
feature 5: wanting --- CCE: 0.528771
feature 6: differently --- CCE: 0.496578
feature 7: ordering --- CCE: 0.496578
feature 8: p --- CCE: 0.496578
feature 9: grouped --- CCE: 0.496578
feature 10: puts --- CCE: 0.496578
feature 11: s --- CCE: 0.496578
feature 12: cases --- CCE: 0.496578
feature 13: k --- CCE: 0.496578
feature 14: evenly --- CCE: 0.496578
feature 15: v --- CCE: 0.496578
feature 16: w --- CCE: 0.496578
feature 17: c --- CCE: 0.496578
feature 18: everybody --- CCE: 0.496578
feature 19: f --- CCE: 0.496578
feature 20: m --- CCE: 0.496578
```

- Please give the feature curve as described above (should be legible). What conclusion can be drawn from this curve?



This plot suggests that with a small sample size of documents, it is best to use fewer stopwords to analyze the data.

AAAC Problem H

- What is the test accuracy?

66.66%

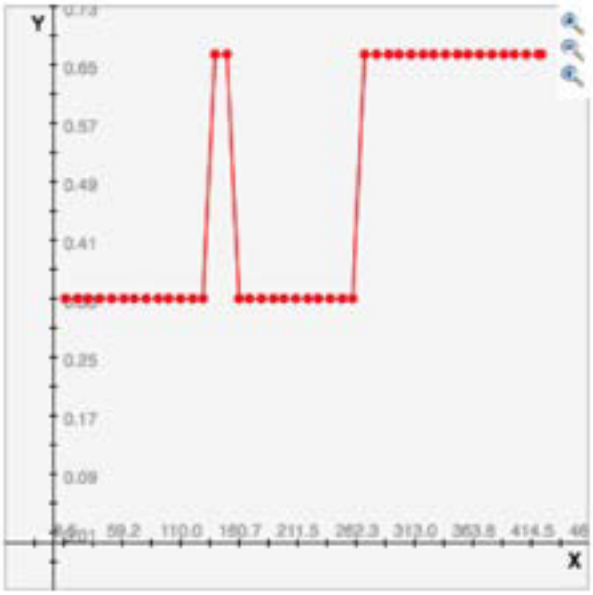
- What is the confusion matrix?

```
0 1 2 3
1 1 0 0
2 0 1 0
3 0 1 0
```

- What is the feature ranking? Please list the top 20 features, along with their class-conditional entropy.

```
feature 1: above --- CCE: 0.528321
feature 2: across --- CCE: 0.528321
feature 3: asking --- CCE: 0.528321
feature 4: asks --- CCE: 0.528321
feature 5: certain --- CCE: 0.528321
feature 6: ordered --- CCE: 0.528321
feature 7: ordering --- CCE: 0.528321
feature 8: orders --- CCE: 0.528321
feature 9: z --- CCE: 0.528321
feature 10: others --- CCE: 0.528321
feature 11: clear --- CCE: 0.528321
feature 12: goods --- CCE: 0.528321
feature 13: p --- CCE: 0.528321
feature 14: parted --- CCE: 0.528321
feature 15: parting --- CCE: 0.528321
feature 16: y --- CCE: 0.528321
feature 17: clearly --- CCE: 0.528321
feature 18: greatest --- CCE: 0.528321
feature 19: places --- CCE: 0.528321
feature 20: pointing --- CCE: 0.528321
```

- Please give the feature curve as described above (should be legible). What conclusion can be drawn from this curve?



What does this suggest? Who knows. Maybe that if all stopwords grant us the same amount of information, using more of them DOES work out in our favor.