ECE 448

MP3

Jiang Chenhuan

Qian Chen

Zhang Zheyuan

April 18, 2022

Section I:

The average accuracy is 0.8362. As we use multiprocessing, the train, test time cost was significantly decreased.

A screenshot of a computer

Description automatically generated

Figure X. CPU Utilization While Running our Program.

Figure X showed the confusion matrix. It showed that “5” had the lowest recognition rate, while “2” had the highest.

Graphical user interface, application

Description automatically generated

Figure X. Confusion Matrix of Test Results

Figure XX showed the visualization plots for the feature likelihood.

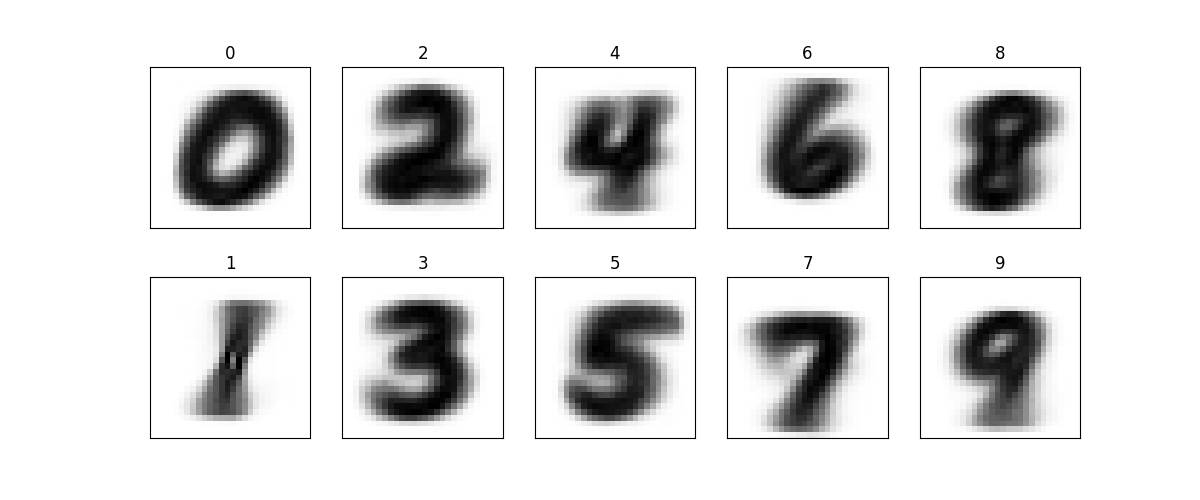
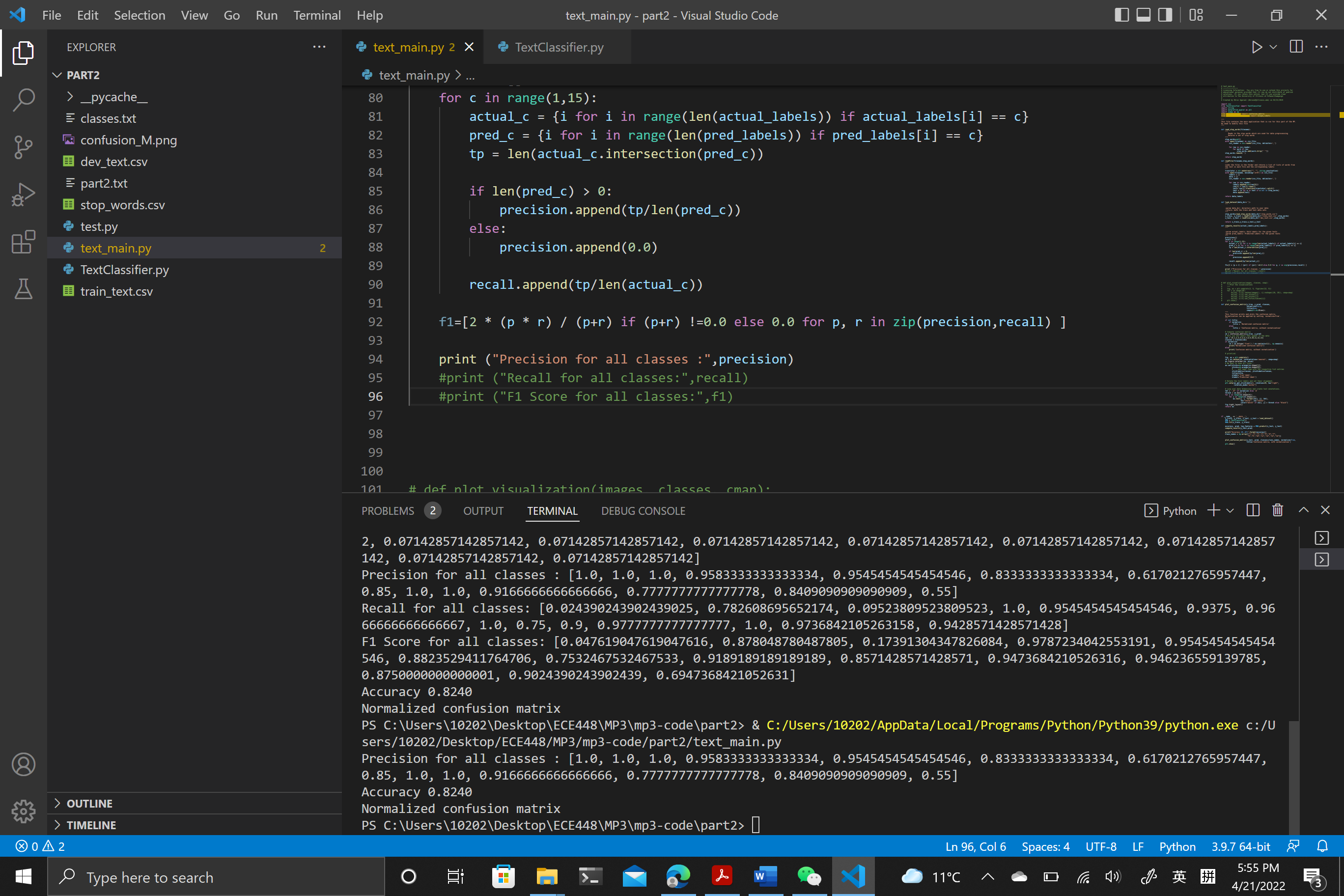
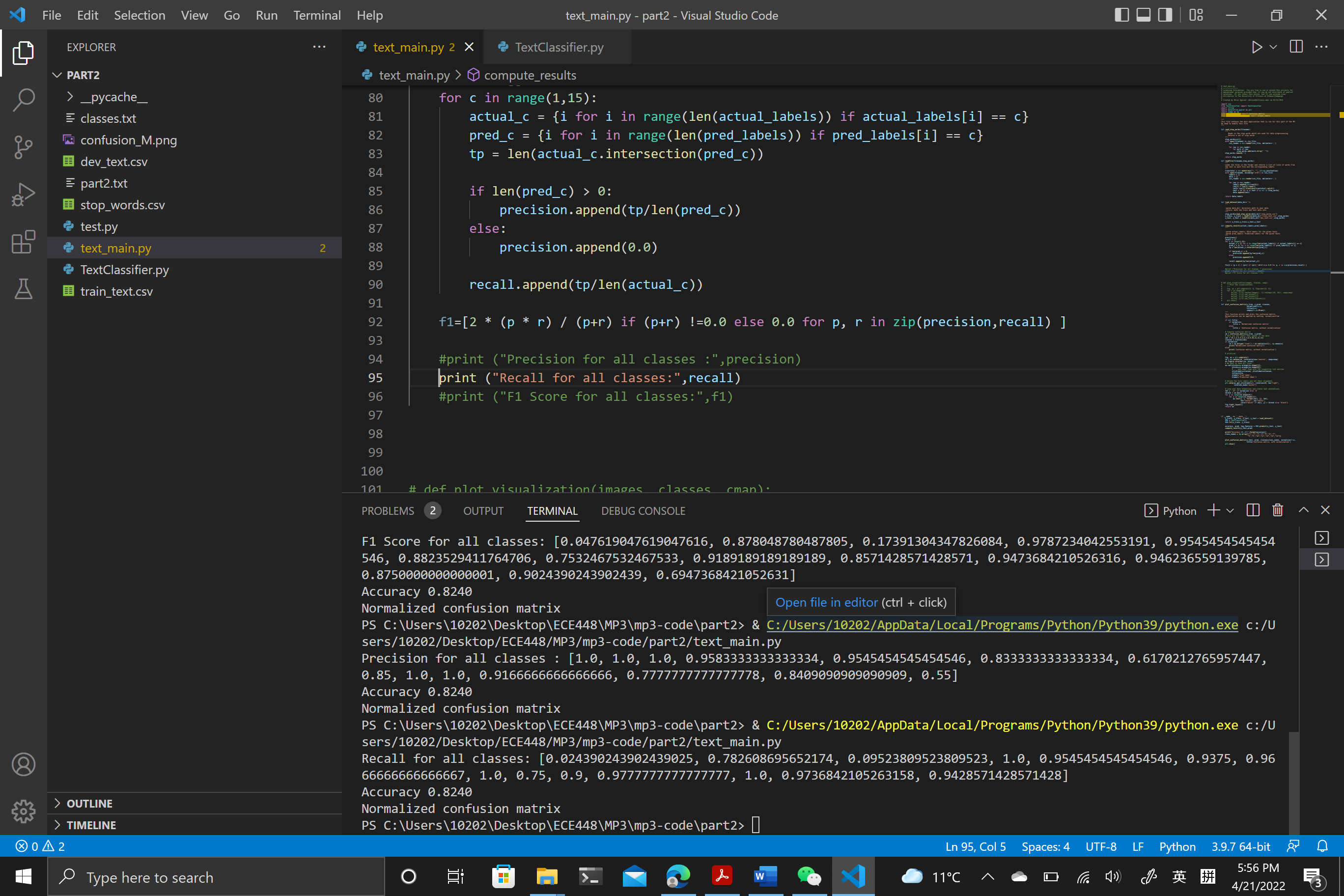


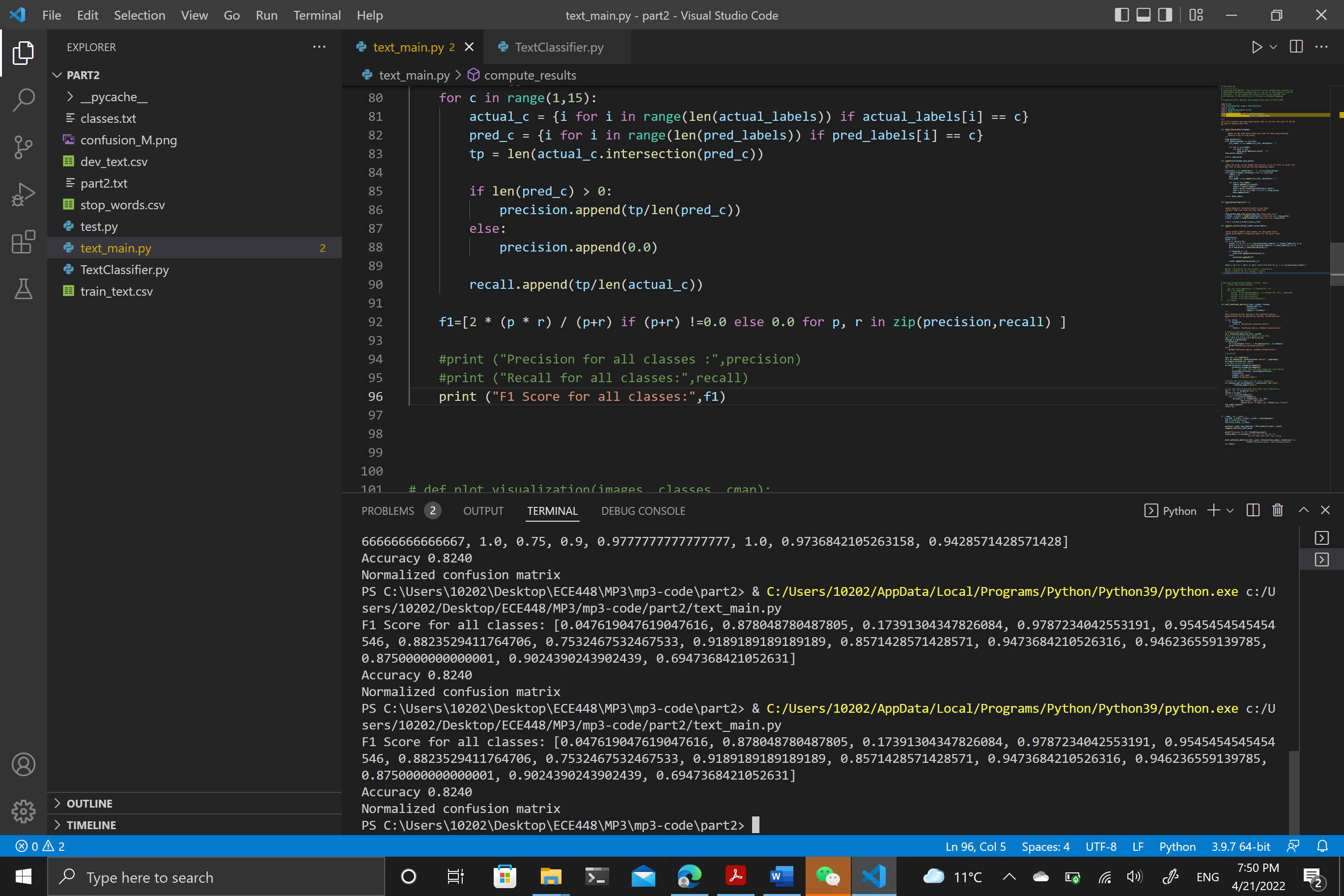
Figure XX. Visualization of Feature Likelihood of Each Number

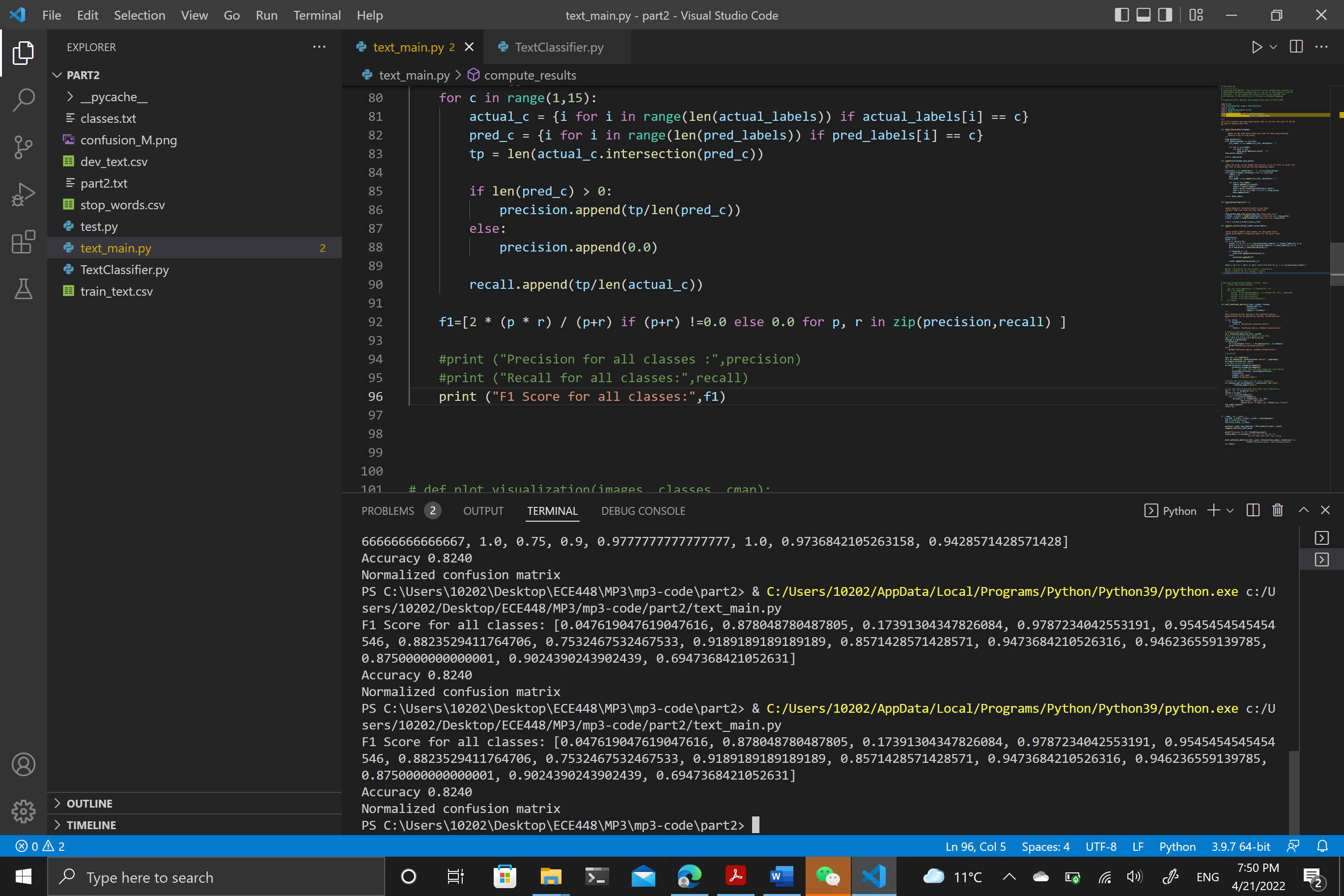
Section II:

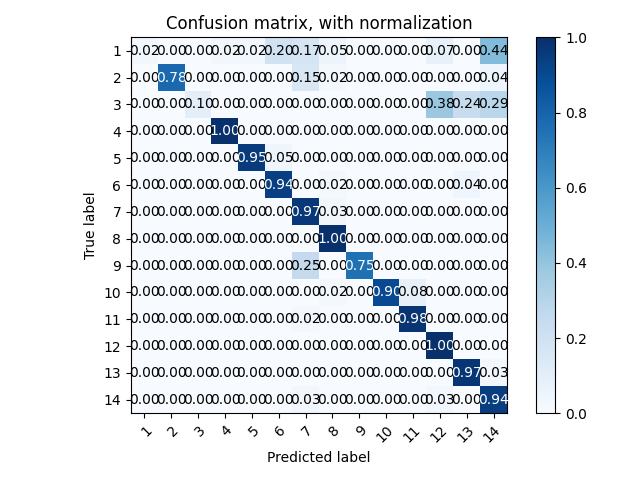
Classification result (with training data class prior):

**Precision:**

**Recall:**

**F1-Scores:**

**Accuracy:**

**Confusion Matrix:** 

**Top 20 features in the 14 class:**

The features are collected as (‘word’, # of occurance)

Class 1 (Company):

[('company', 28), ('based', 10), ('business', 9), ('founded', 8), ('record', 7), ('records', 7), ('bergen', 7), ('systems', 6), ('services', 6), ('products', 5), ('office', 5), ('buses', 4), ('located', 4), ('distribution', 4), ('national', 4), ('health', 4), ('virgin', 4), ('established', 4), ('also', 4), ('regional', 4)]

Class 2 (EducationalInstitution):

[('school', 231), ('high', 95), ('located', 53), ('university', 38), ('college', 36), ('public', 32), ('schools', 32), ('students', 24), ('education', 22), ('district', 19), ('county', 19), ('founded', 15), ('one', 15), ('new', 15), ('united', 14), ('established', 14), ('independent', 14), ('city', 14), ('part', 14), ('catholic', 13)]

Class 3 (Artist):

[('born', 67), ('american', 36), ('known', 28), ('new', 22), ('band', 17), ('writer', 16), ('best', 16), ('rock', 15), ('musician', 14), ('music', 14), ('work', 14), ('also', 12), ('singer', 12), ('york', 11), ('books', 10), ('author', 10), ('album', 10), ('university', 9), ('series', 9), ('united', 9)]

Class 4 (Athlete):

[('born', 124), ('football', 78), ('played', 70), ('league', 64), ('professional', 36), ('player', 36), ('plays', 36), ('footballer', 36), ('former', 34), ('national', 31), ('american', 28), ('also', 24), ('hockey', 24), ('currently', 24), ('rugby', 23), ('team', 22), ('australian', 21), ('november', 20), ('world', 19), ('new', 18)]

Class 5 (OfficeHolder):

[('born', 74), ('member', 69), ('district', 50), ('politician', 49), ('state', 34), ('senate', 32), ('democratic', 32), ('house', 32), ('party', 31), ('served', 30), ('former', 29), ('county', 26), ('since', 25), ('representatives', 24), ('republican', 21), ('united', 20), ('elected', 20), ('american', 19), ('representing', 18), ('national', 18)]

Class 6 (MeanOfTransportation):

[('navy', 133), ('built', 115), ('war', 82), ('ship', 71), ('uss', 64), ('united', 59), ('class', 58), ('aircraft', 58), ('world', 57), ('states', 55), ('launched', 54), ('service', 41), ('named', 40), ('first', 40), ('designed', 40), ('royal', 38), ('commissioned', 35), ('american', 34), ('ii', 33), ('us', 31)]

Class 7 (Building):

[('historic', 214), ('house', 166), ('built', 159), ('located', 129), ('church', 127), ('building', 117), ('national', 116), ('register', 96), ('places', 86), ('listed', 69), ('county', 62), ('street', 62), ('united', 52), ('known', 50), ('museum', 48), ('also', 48), ('states', 44), ('designed', 40), ('added', 39), ('hospital', 39)]

Class 8 (NaturalPlace):

[('river', 234), ('lake', 140), ('mountain', 91), ('located', 80), ('south', 62), ('km', 52), ('north', 50), ('county', 39), ('near', 38), ('tributary', 37), ('west', 36), ('range', 36), ('lies', 35), ('creek', 34), ('crater', 33), ('east', 33), ('ft', 32), ('state', 32), ('flows', 30), ('pass', 29)]

Class 9 (Village):

[('village', 95), ('district', 33), ('population', 23), ('province', 22), ('located', 21), ('census', 20), ('municipality', 16), ('nepal', 16), ('state', 14), ('india', 14), ('county', 12), ('km', 11), ('people', 11), ('within', 9), ('2010', 8), ('1991', 8), ('south', 7), ('township', 7), ('central', 7), ('southern', 7)]

Class 10 (Animal):

[('family', 335), ('species', 291), ('found', 170), ('genus', 164), ('moth', 98), ('gastropod', 62), ('sea', 60), ('known', 58), ('marine', 54), ('described', 53), ('tropical', 50), ('snail', 46), ('mollusk', 45), ('endemic', 44), ('subtropical', 41), ('habitat', 39), ('natural', 35), ('forests', 35), ('snails', 33), ('moist', 32)]

Class 11 (Plant):

[('species', 393), ('family', 263), ('plant', 231), ('genus', 220), ('native', 119), ('endemic', 117), ('flowering', 116), ('known', 113), ('found', 87), ('common', 75), ('plants', 69), ('leaves', 69), ('habitat', 66), ('tree', 59), ('grows', 52), ('name', 52), ('orchid', 49), ('south', 48), ('bulbophyllum', 44), ('perennial', 42)]

Class 12 (Album):

[('album', 950), ('released', 538), ('band', 175), ('records', 173), ('first', 150), ('studio', 144), ('american', 107), ('songs', 102), ('music', 101), ('second', 99), ('release', 96), ('recorded', 91), ('rock', 89), ('debut', 80), ('live', 80), ('tracks', 77), ('label', 74), ('albums', 71), ('new', 69), ('ep', 65)]

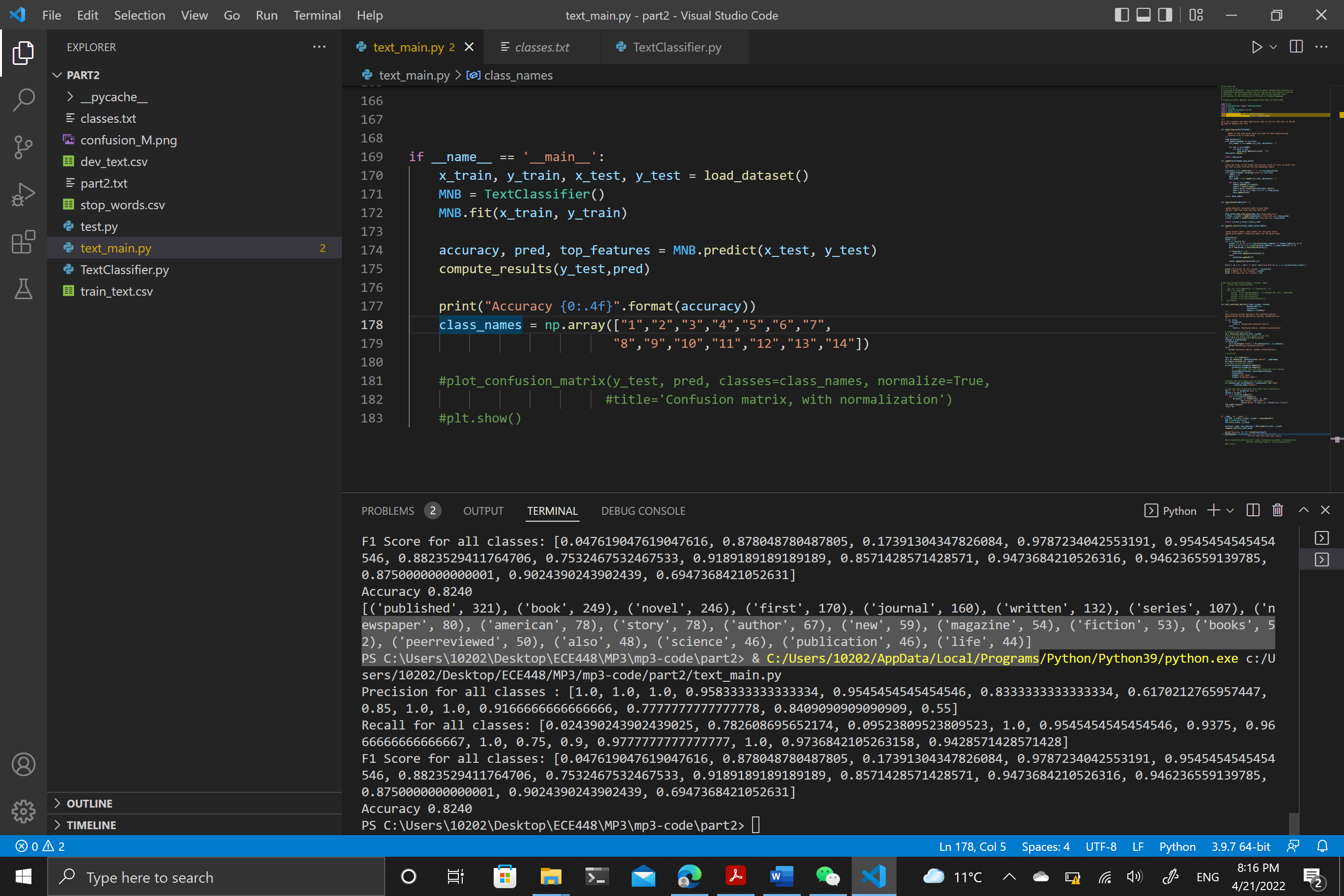
Class 13 (Film):

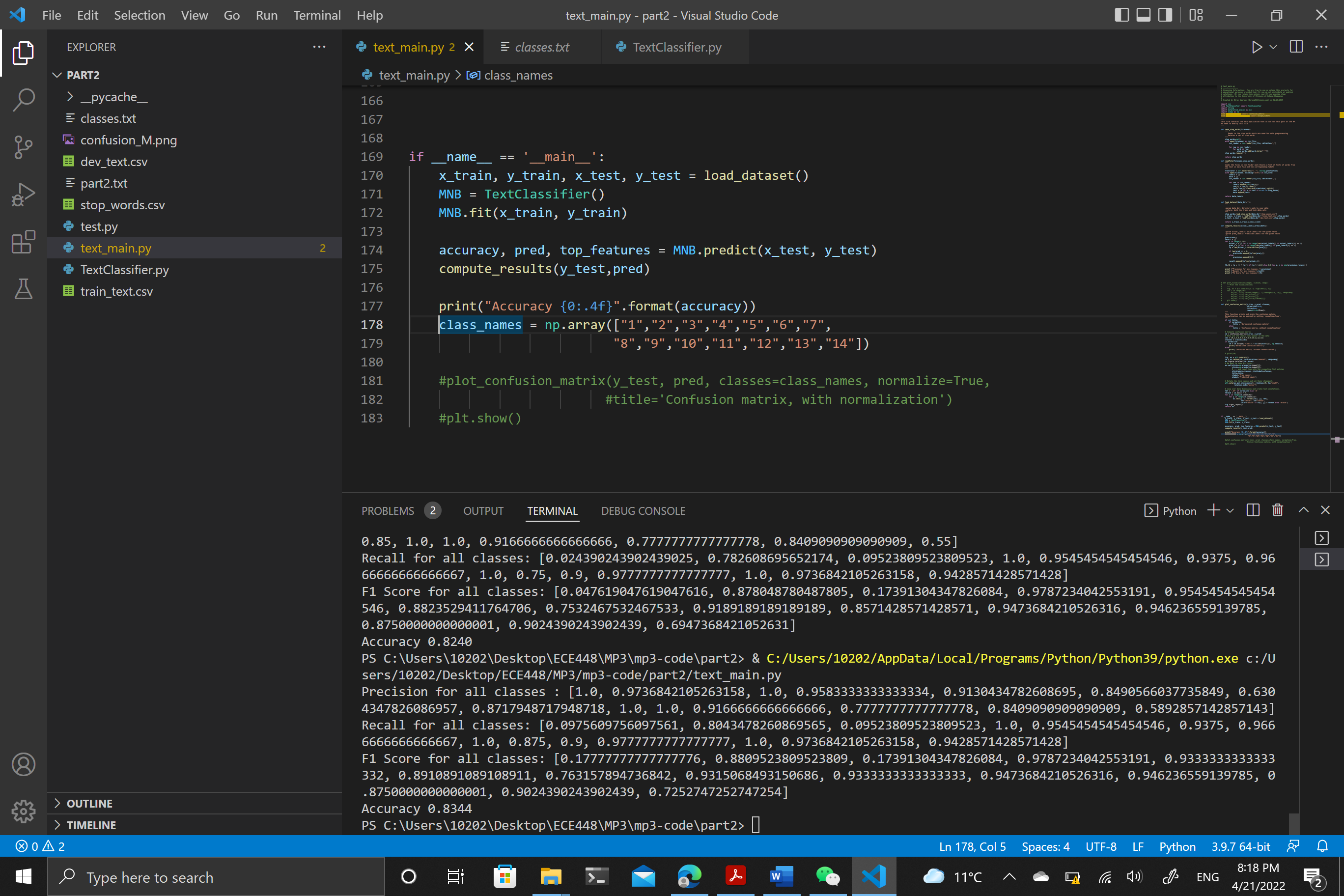
[('film', 870), ('directed', 354), ('starring', 139), ('american', 131), ('stars', 103), ('released', 101), ('written', 94), ('based', 87), ('drama', 85), ('comedy', 83), ('produced', 81), ('also', 81), ('films', 57), ('silent', 52), ('first', 52), ('movie', 51), ('roles', 43), ('novel', 42), ('name', 42), ('documentary', 41)]

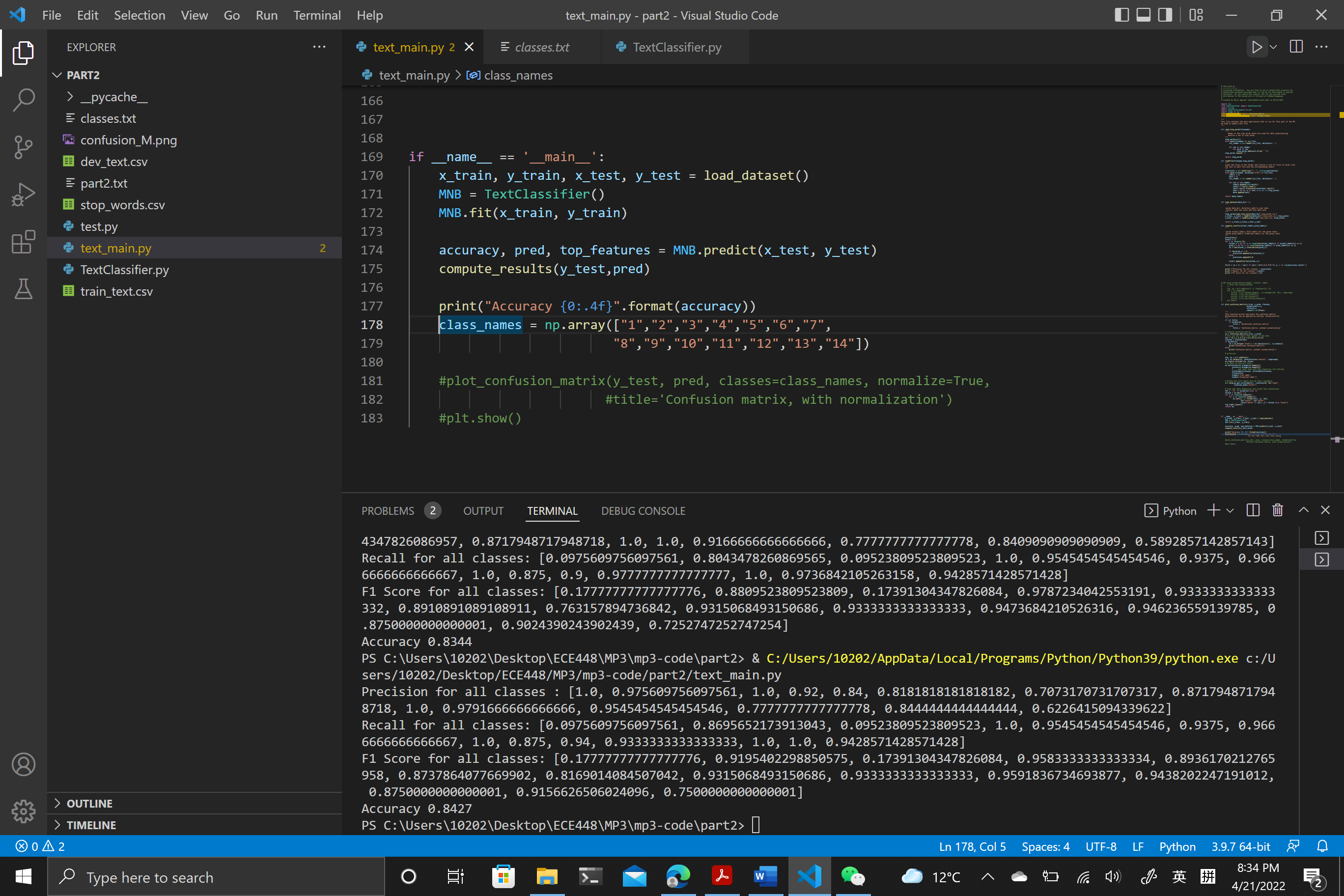
Class 14 (WrittenWork):

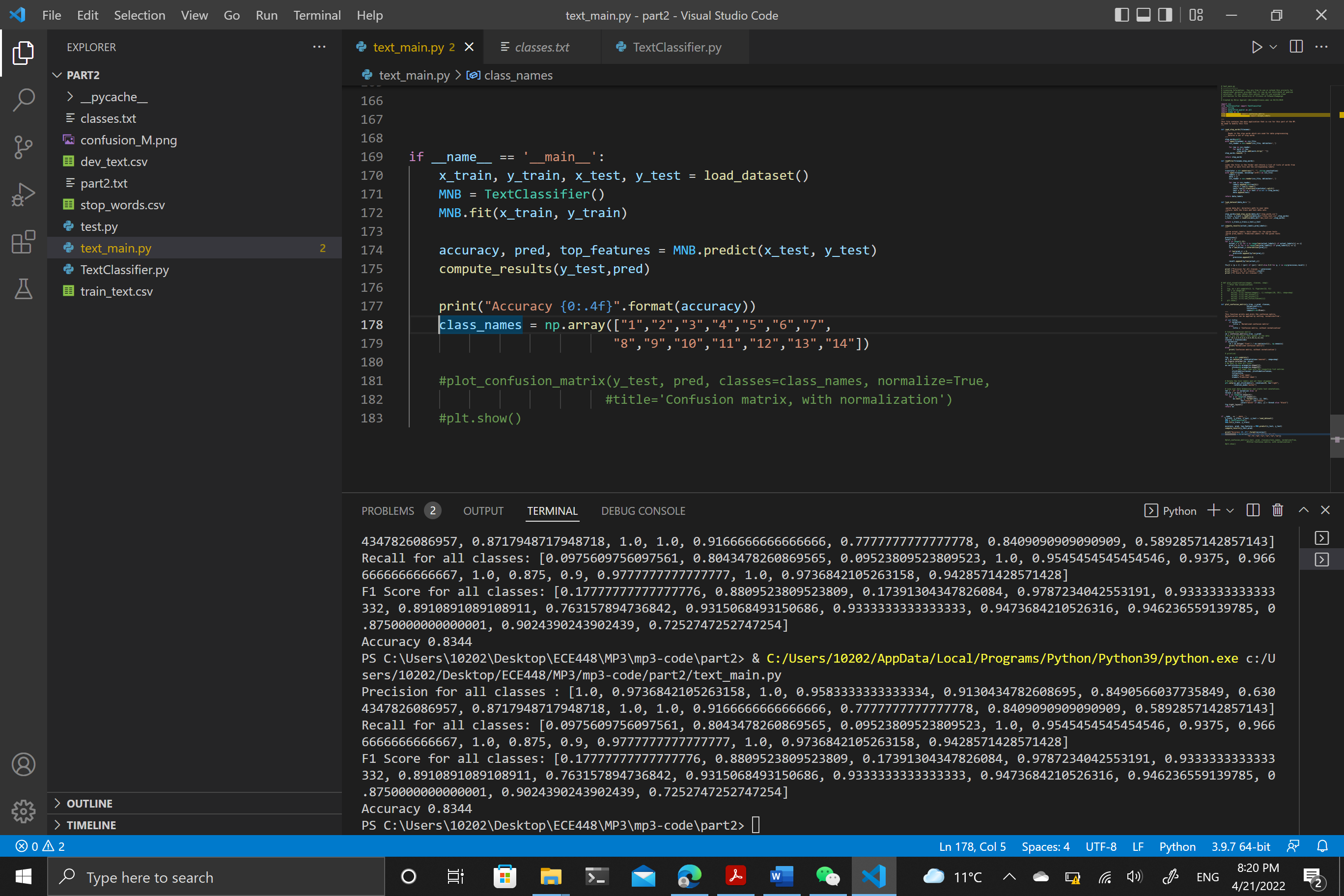
[('published', 321), ('book', 249), ('novel', 246), ('first', 170), ('journal', 160), ('written', 132), ('series', 107), ('newspaper', 80), ('american', 78), ('story', 78), ('author', 67), ('new', 59), ('magazine', 54), ('fiction', 53), ('books', 52), ('peerreviewed', 50), ('also', 48), ('science', 46), ('publication', 46), ('life', 44)]

**Different class prior distribution:**

1. Distribution of the training data:

2. Uniform Distribution

3. Normal Distribution

4. Without class prior

From the above 4 figures, it is clearly that if the class prior is normal distributed, the classifier will has the highest accuracy (among the four distrubutions), also, the accuracy would increase a little bit if we apply uniform distribution to the class prior or completely ignore the class prior (compared to the accuracy using the distribution of training data).

In our opinion, P(Class = type i) is more likely to be normal distribution since it is a relatively random possibility, so applying the normal distribution should have the highest accuracy since it described the P(Class = type i) more accurately. Except that, the uniform distribution and ignoring the class prior produced the same results, because in our computation, (log(P(Class = type i | Document = doc i)) ~ log(P(Document = doc i | Class = type i)) + log(P(Class = type i))) adding the same log(P(Class = type i)) is equal to add nothing.

Section III:

Average error rate of training is 0.4792000000000001, Average error rate of test set for your Logistic Regression model is 0.4801760000000002

Figure of the linear classifier.

图表, 散点图

描述已自动生成

Statement of Contribution:

Qian Chen and Chenhuan Jiang both try the task one. However, the task initially take a lot of time to run. Therefore, Qian Chen improve the classification rate. The task two is finished by Jiang Chenhuan and Zhang Zheyuan. The task 3 is written by Zhang Zheyuan and the other parts are written by three people. Three people work on the report together.