Assignment-3 (Report)

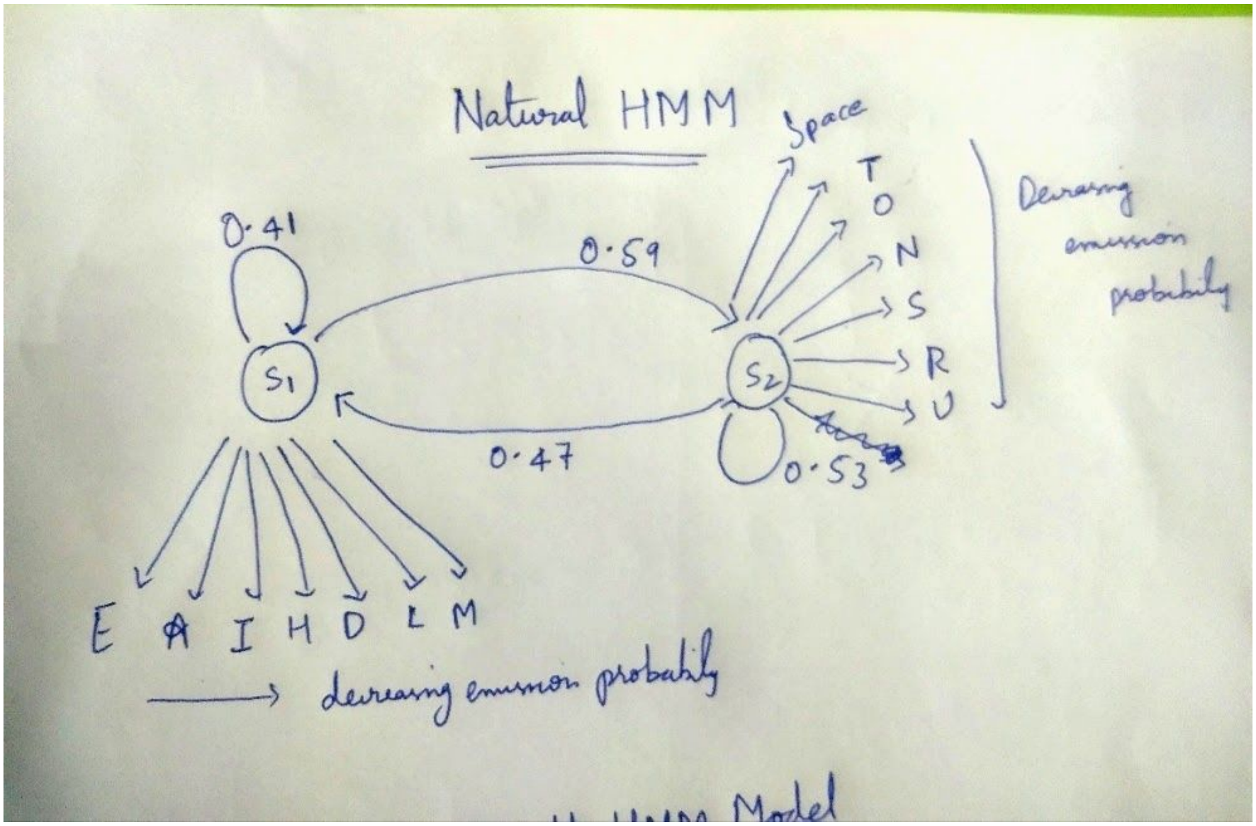
Task-1 :

I have preprocessed the text_file by removing all punctuations (as mentioned in the task) and made the test_file to uppercase alphabets by writing simple python code.

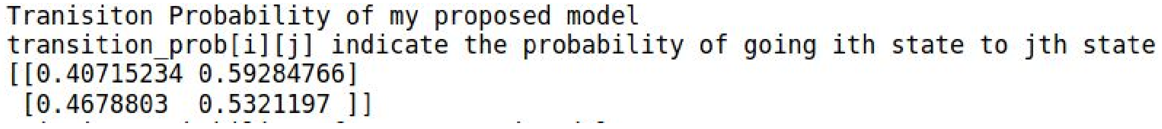
Task-2 :

Created two states s1,s2, and I have experimented with s1 having the first 13 alphabets and s2 having the later 13 alphabets + space. I wanted to have both states having the same number of characters so that's why I have created like that and also was interested to know its result. This is my natural hmm design

Handwritten hmm model

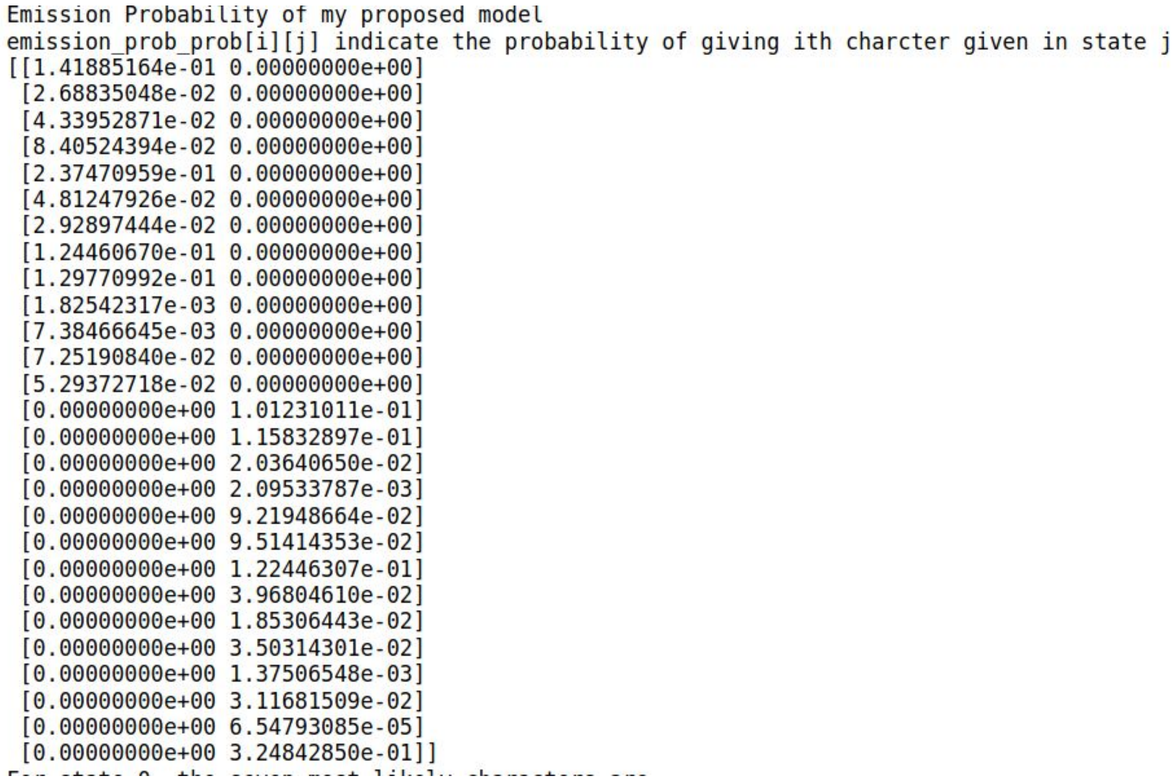


The calculated transition prob of natural hmm, done by reading the tect_file



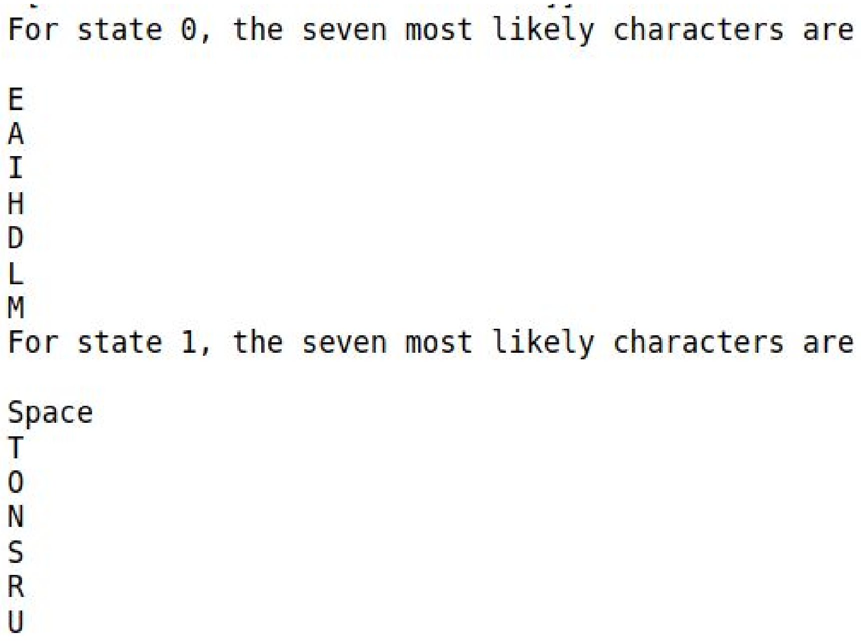
```
Tranisiton Probability of my proposed model
transition_prob[i][j] indicate the probability of going ith state to jth state
[[0.40715234 0.59284766]
 [0.4678803  0.5321197 ]]
```

The calculated emission prob of natural hmm, done by reading the tect_file



```
Emission Probability of my proposed model
emission_prob_prob[i][j] indicate the probability of giving ith charcter given in state j
[[1.41885164e-01 0.00000000e+00]
 [2.68835048e-02 0.00000000e+00]
 [4.33952871e-02 0.00000000e+00]
 [8.40524394e-02 0.00000000e+00]
 [2.37470959e-01 0.00000000e+00]
 [4.81247926e-02 0.00000000e+00]
 [2.92897444e-02 0.00000000e+00]
 [1.24460670e-01 0.00000000e+00]
 [1.29770992e-01 0.00000000e+00]
 [1.82542317e-03 0.00000000e+00]
 [7.38466645e-03 0.00000000e+00]
 [7.25190840e-02 0.00000000e+00]
 [5.29372718e-02 0.00000000e+00]
 [0.00000000e+00 1.01231011e-01]
 [0.00000000e+00 1.15832897e-01]
 [0.00000000e+00 2.03640650e-02]
 [0.00000000e+00 2.09533787e-03]
 [0.00000000e+00 9.21948664e-02]
 [0.00000000e+00 9.51414353e-02]
 [0.00000000e+00 1.22446307e-01]
 [0.00000000e+00 3.96804610e-02]
 [0.00000000e+00 1.85306443e-02]
 [0.00000000e+00 3.50314301e-02]
 [0.00000000e+00 1.37506548e-03]
 [0.00000000e+00 3.11681509e-02]
 [0.00000000e+00 6.54793085e-05]
 [0.00000000e+00 3.24842850e-01]]
```

The seven most likely words picked the ones which have the highest probability

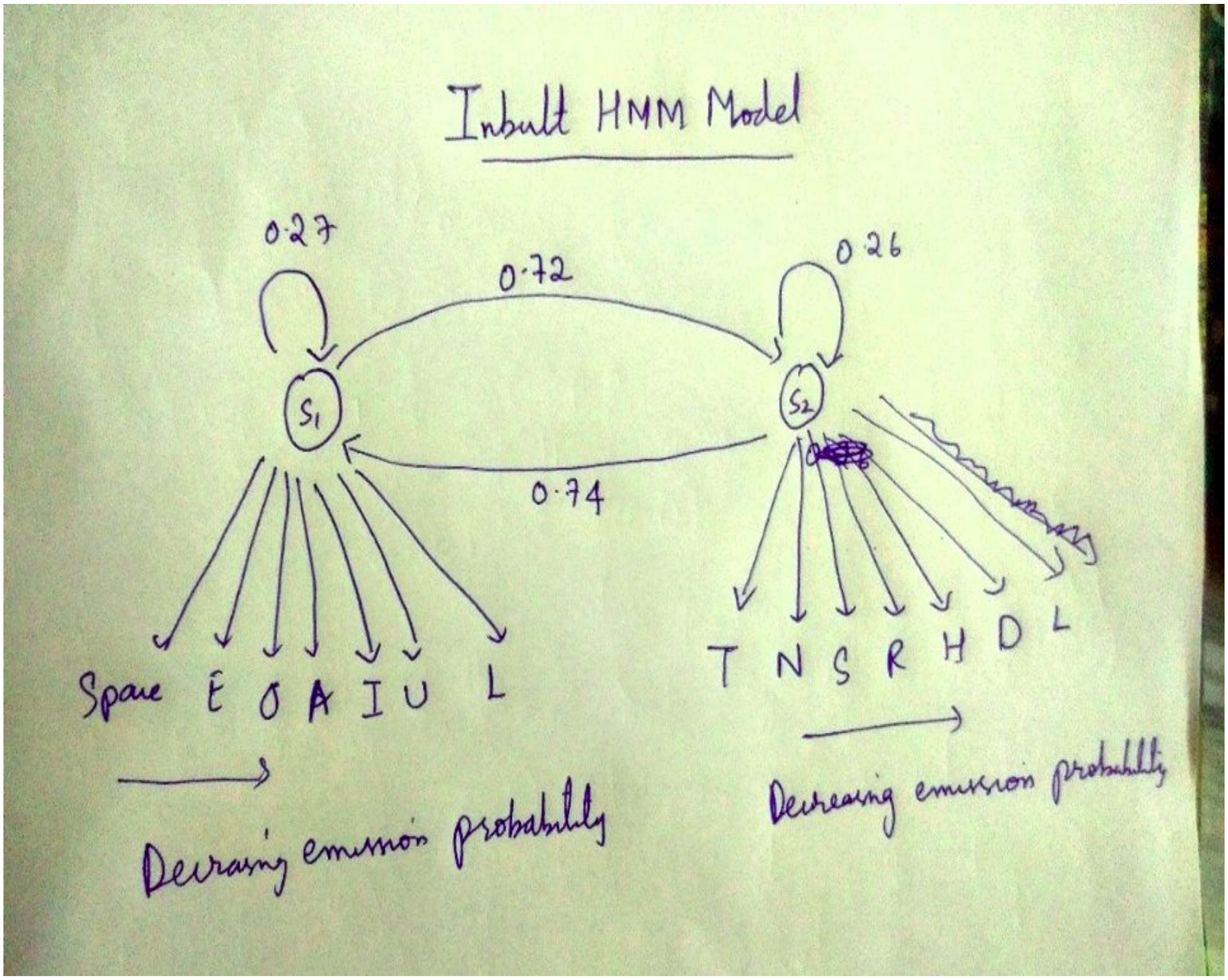


```
For state 0, the seven most likely characters are

E
A
I
H
D
L
M
For state 1, the seven most likely characters are

Space
T
O
N
S
R
U
```

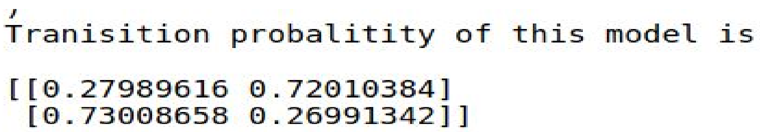
Task-3:

The final transition probability and emission results of the inbuilt library algorithm are given below along with the final result

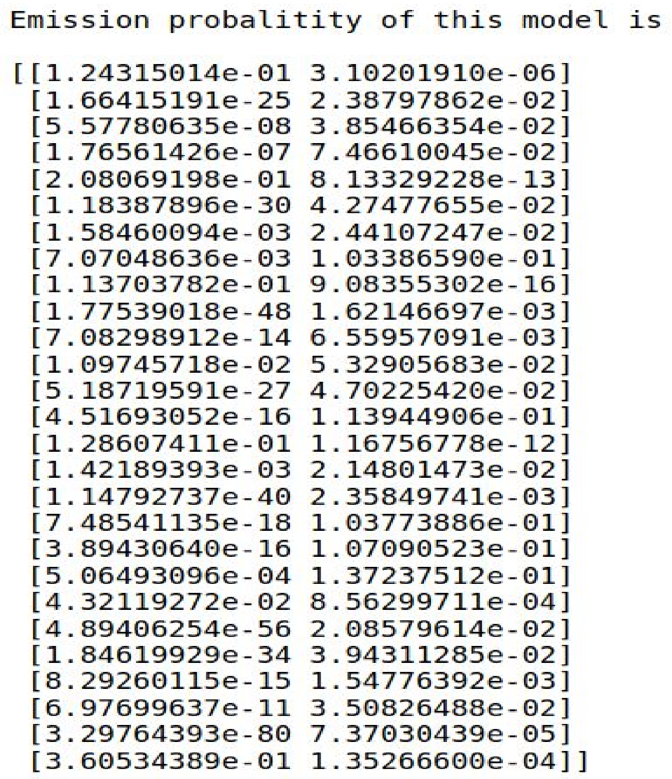
Handwritten hmm model



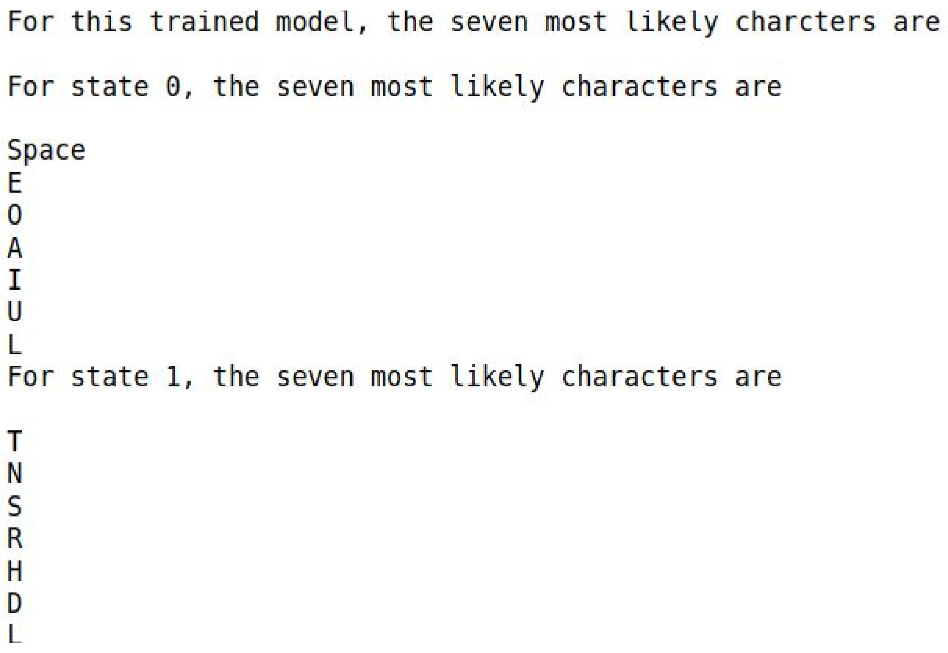
Transition probability



```
Tranisition probabilitity of this model is

[[0.27989616 0.72010384]
 [0.73008658 0.26991342]]
```



```
Emission probalitity of this model is

[[1.24315014e-01 3.10201910e-06]
 [1.66415191e-25 2.38797862e-02]
 [5.57780635e-08 3.85466354e-02]
 [1.76561426e-07 7.46610045e-02]
 [2.08069198e-01 8.13329228e-13]
 [1.18387896e-30 4.27477655e-02]
 [1.58460094e-03 2.44107247e-02]
 [7.07048636e-03 1.03386590e-01]
 [1.13703782e-01 9.08355302e-16]
 [1.77539018e-48 1.62146697e-03]
 [7.08298912e-14 6.55957091e-03]
 [1.09745718e-02 5.32905683e-02]
 [5.18719591e-27 4.70225420e-02]
 [4.51693052e-16 1.13944906e-01]
 [1.28607411e-01 1.16756778e-12]
 [1.42189393e-03 2.14801473e-02]
 [1.14792737e-40 2.35849741e-03]
 [7.48541135e-18 1.03773886e-01]
 [3.89430640e-16 1.07090523e-01]
 [5.06493096e-04 1.37237512e-01]
 [4.32119272e-02 8.56299711e-04]
 [4.89406254e-56 2.08579614e-02]
 [1.84619929e-34 3.94311285e-02]
 [8.29260115e-15 1.54776392e-03]
 [6.97699637e-11 3.50826488e-02]
 [3.29764393e-80 7.37030439e-05]
 [3.60534389e-01 1.35266600e-04]]
```

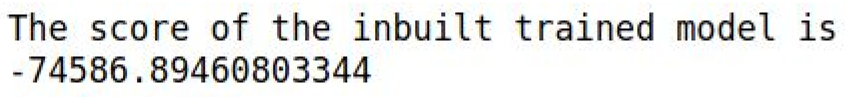


```
For this trained model, the seven most likely charcters are

For state 0, the seven most likely characters are

Space
E
O
A
I
U
L
For state 1, the seven most likely characters are

T
N
S
R
H
D
L
```

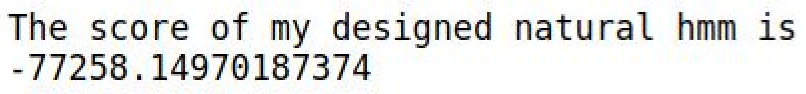
By seeing the most likely characters of this model, we can say that state0 contains mostly consonants and state1 contains all the vowels. By this observation, we can also design natural hmm having state0 with vowels and state1 with consonants

Task-4:

The score(**log probability under the model**) of the inbuilt hmm model was better than the natural hmm model.

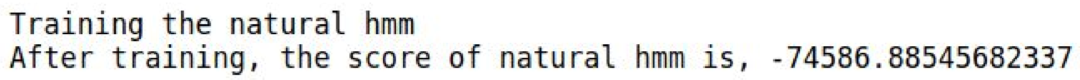


```
The score of the inbuilt trained model is
-74586.89460803344
```



```
The score of my designed natural hmm is
-77258.14970187374
```

After training the natural model, the score of the natural hmm increased, therefore we have succeeded in improving the solution



```
Training the natural hmm
After training, the score of natural hmm is, -74586.88545682337
```

The uninitialized inbuilt hmm model look fewer steps to converge whereas the initialized inbuilt hmm model took more steps to converge but the performance was good by the initialized one.

Therefore the best model is the initialized inbuilt model

Task-5:

On test data, all the model performance was similar, it concludes that there was a significant amount of **overfitting**