# Problem 1

## Part A

### i)

We have for some c

(note that can take two values -> 2 permutations)

(note that can each take two values -> 4 permutations)

As we can see if we continues this process we will get the following series to represent the permutations that must be considered:

We must do this for all c so altogether we have:

ii)

We know that

Each can take on 1 of 2 values

So each such calculation is

We need to do this of which there are

They are the same.

## Part B

With small N, I would expect naïve bayes to perform better on the testing set because will far more parameters than naïve, full bayes would almost definitely over fit drastically leading to a low training error but a higher testing error, whereas naïve would probably have a higher training error than full but a lower testing error.

## Part C

With large N, I would expect full bayes to out perform naïve bayes. This is because full bayes has more paramters which gives the model more flexibility to fit the underlying nature of the data, while a large N would ideally give a good representation of that nature to the point where the model would not suffer from overfitting.

## Part D

We want

Can be computed in since there are x’s

Can be computed in since we already have the value and we have a uniform class prior and it is a simple lookup.

We also have:

Which can be computed in since there are elements to the summation.

Altogether we have:

Computing the prediction for full bayes would also be a computation

We want

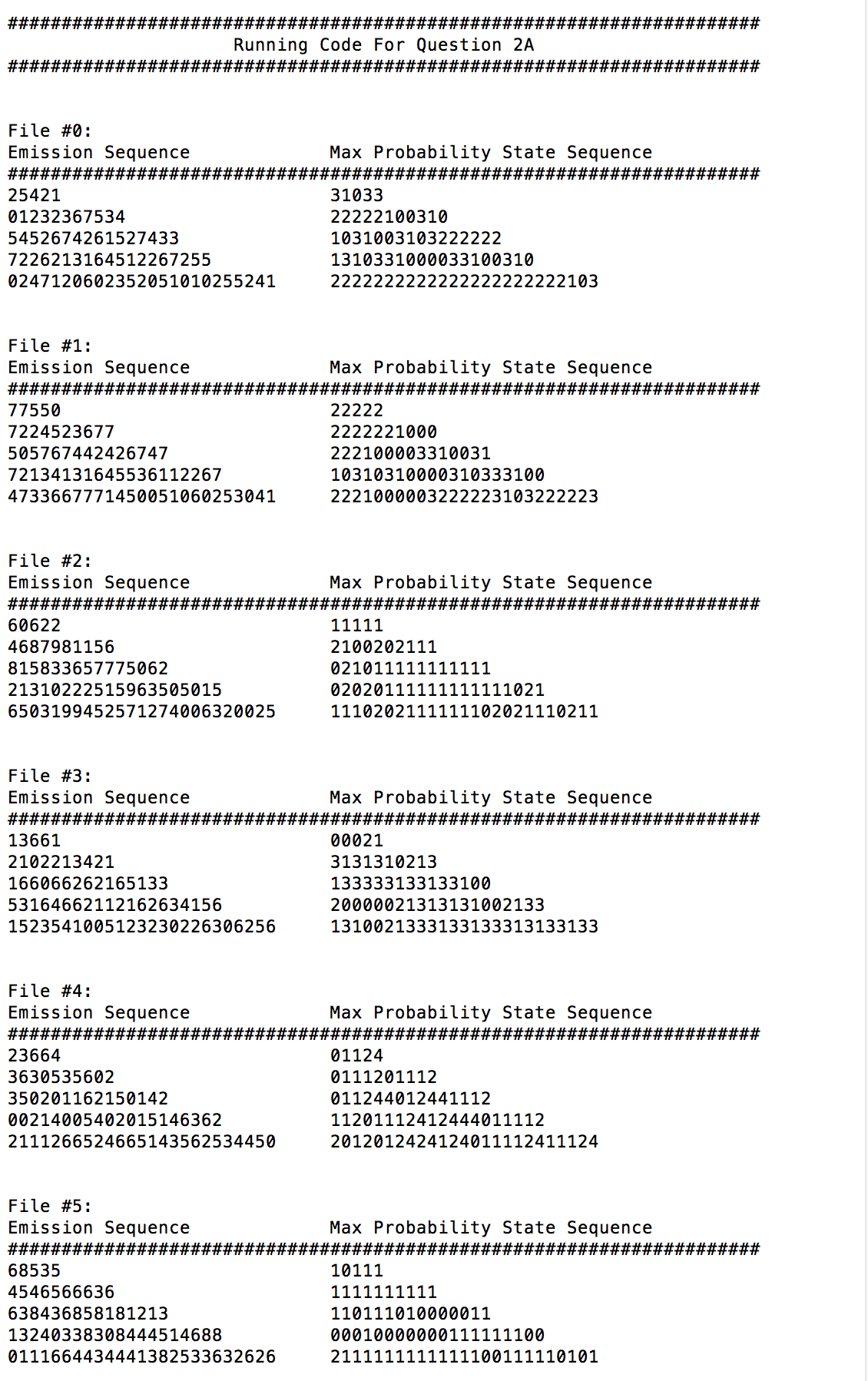
This information can be stored in a 3 nested array and you can get the first index as given in

To get probabilities for a given we need ,

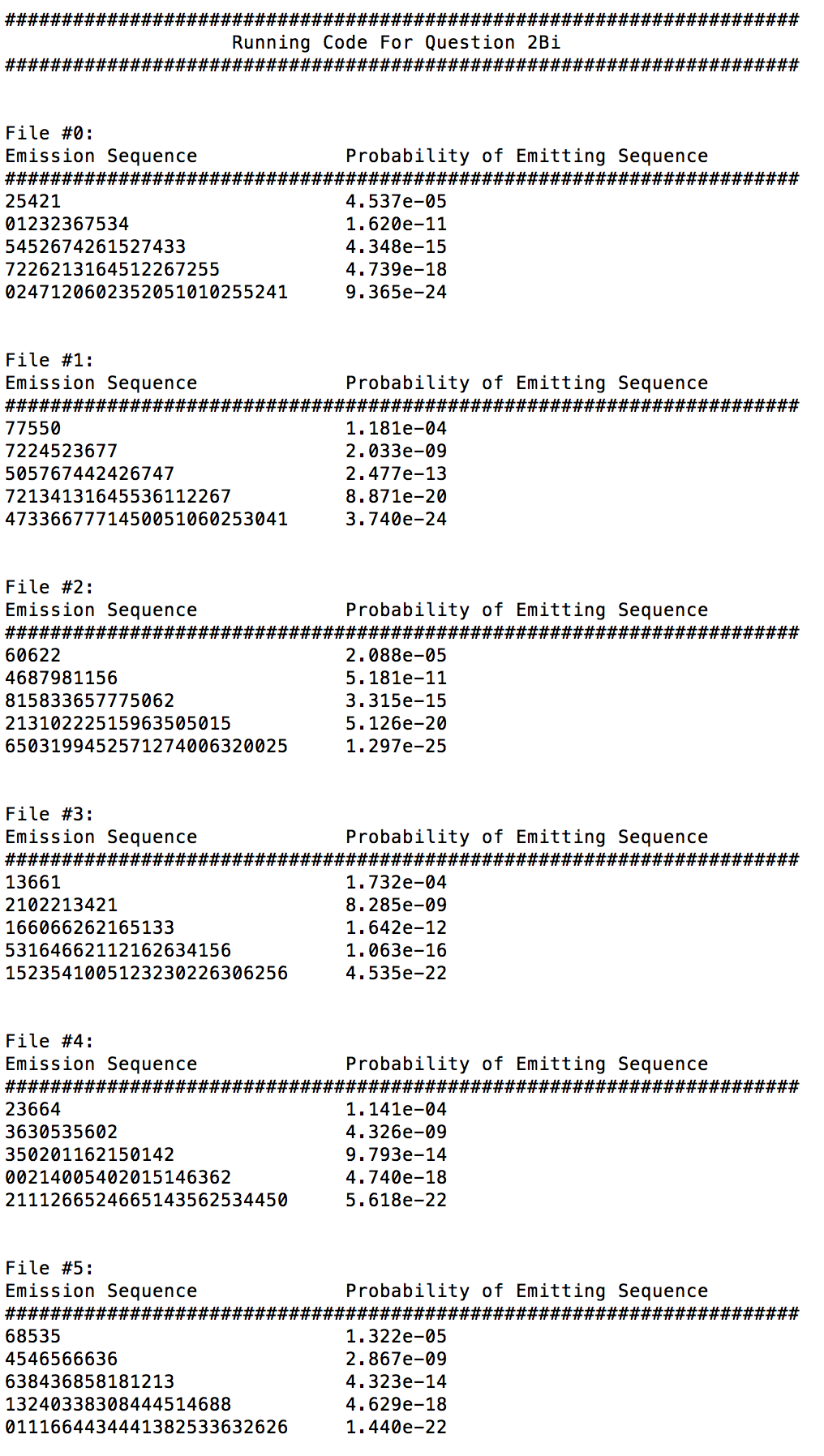
To do so for all possible y would give .

# Problem 2

## Part A



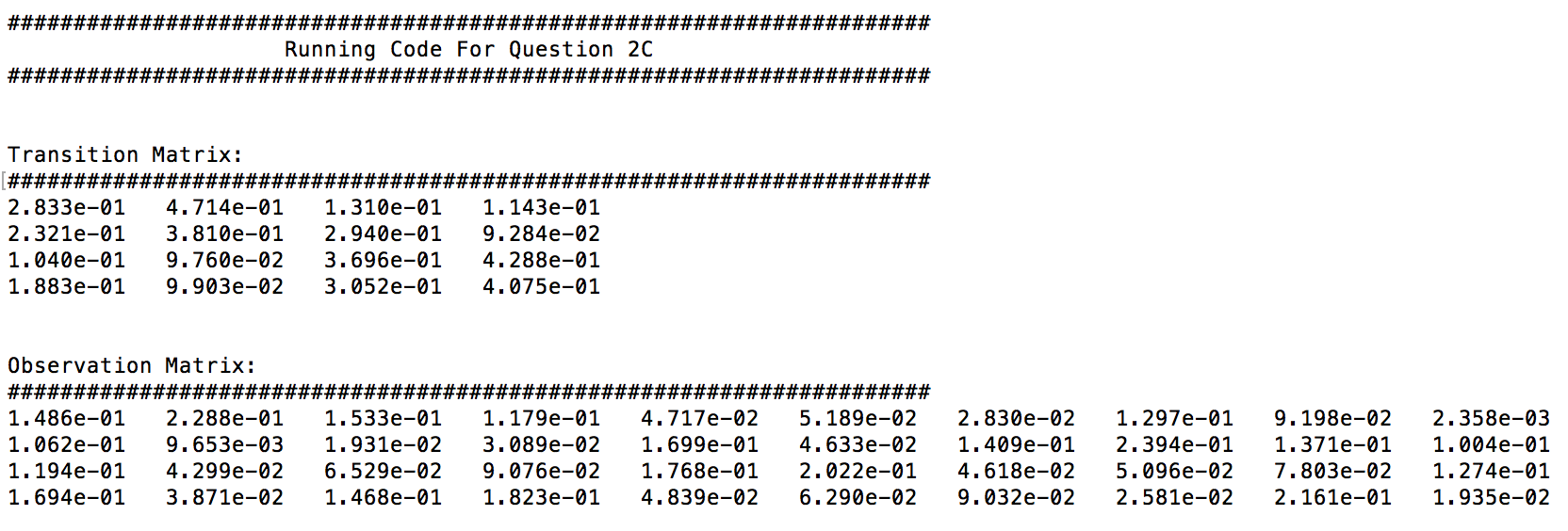
## Part Bi



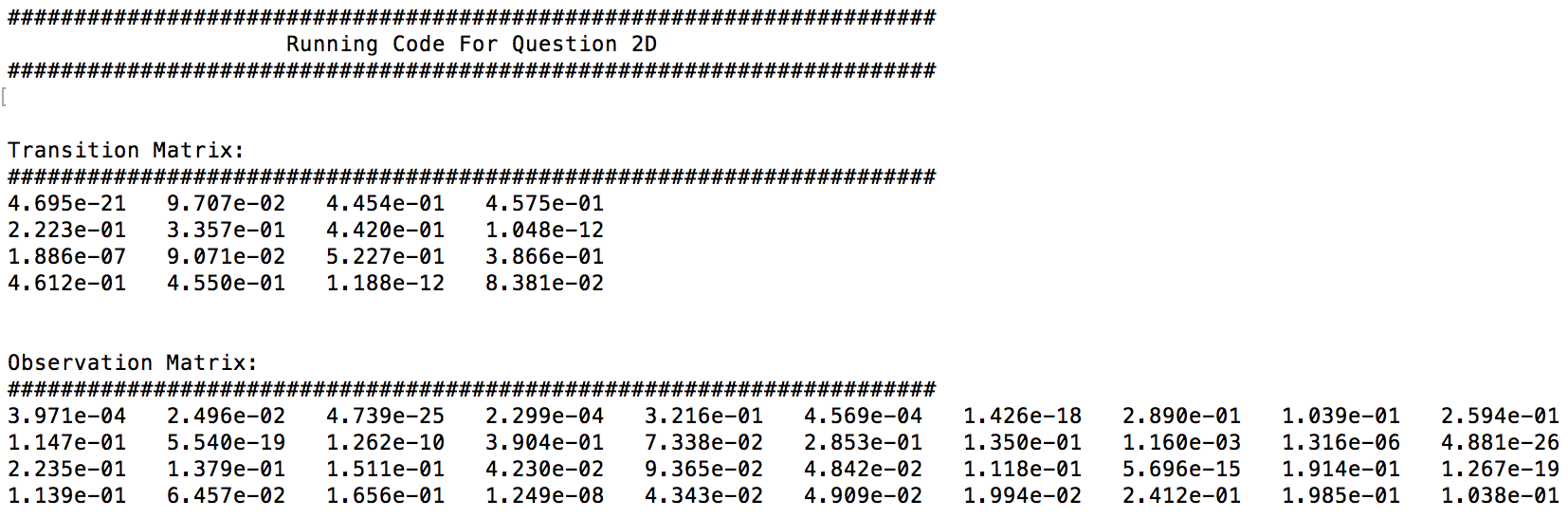
## Part Bii

## 

## Part C



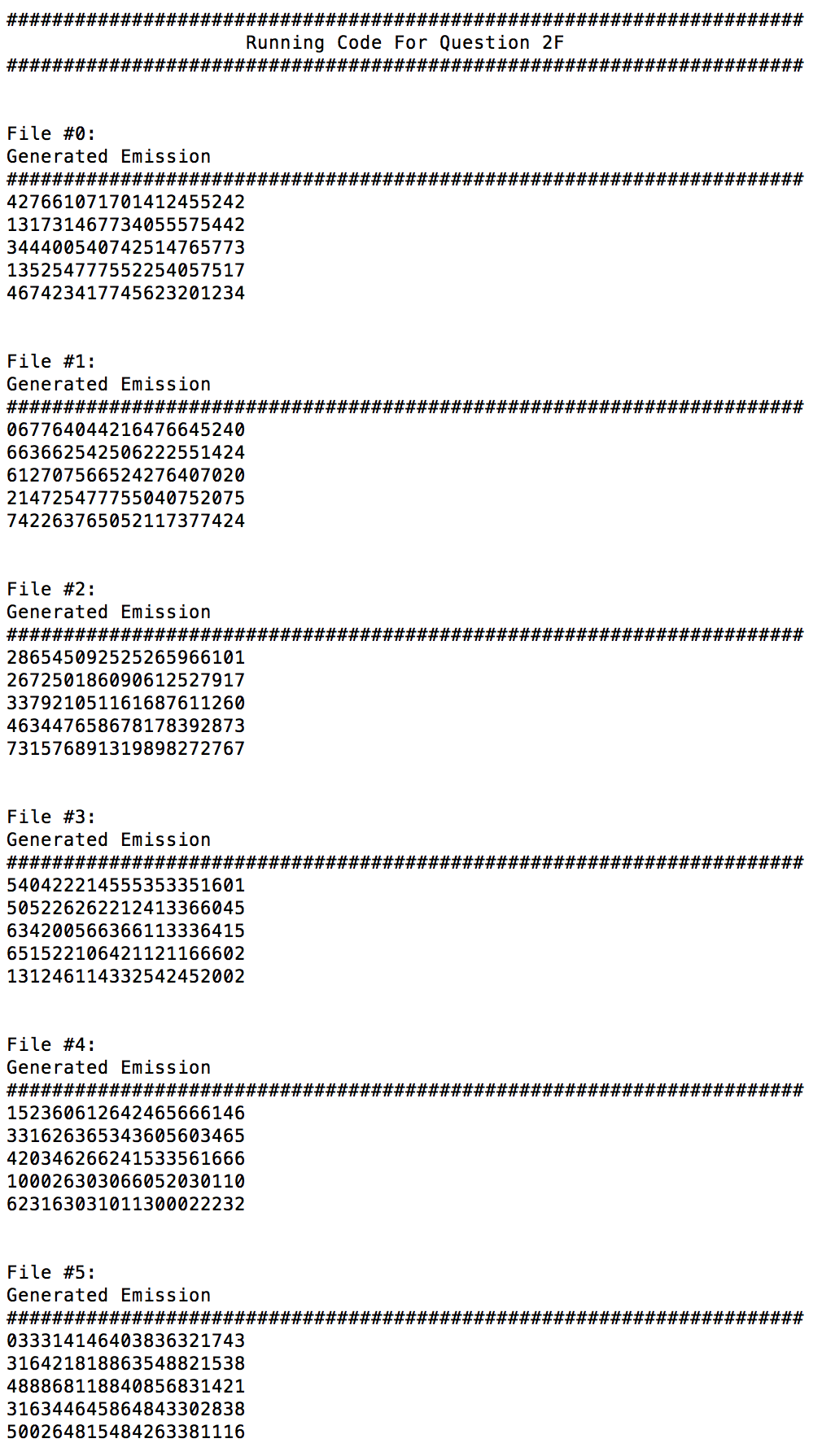
## Part D



## Part E

The transition matrices from 2C appear more accurate for a few reasons. For starters, it can’t be seen on the above images but I ran part D multiple times and the transition matrix values were highly variable which seems to indicate that the unsupervised algorithm doesn’t have as accurate convergence behavior and is encountering various local minima. Intuitively, it makes sense that the data with labels yields more consistent and accurate training.

## Part F

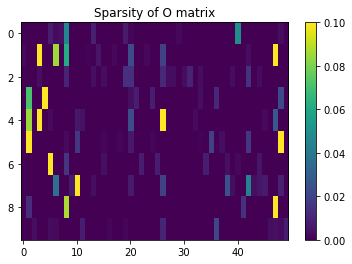
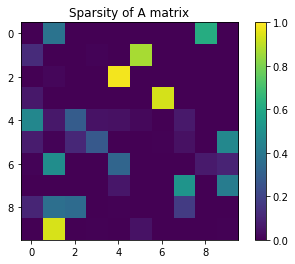


## Part G

The transition matrix is rather sparse with mostly values that are close to or at 0.

There are handful of entries that have high values which indicates a high probability of transition from one state to another. This indicates that the state transitions are fairly determined in that there isn’t a ton of variability in which state comes next given a current state.

We have somewhat similar sparsity in the O matrix. A strong majority of the entries appear to be 0. In each row there are generally no more than a few non-zero entries which indicates that a state is only liable to emit a very specific subset of emissions.



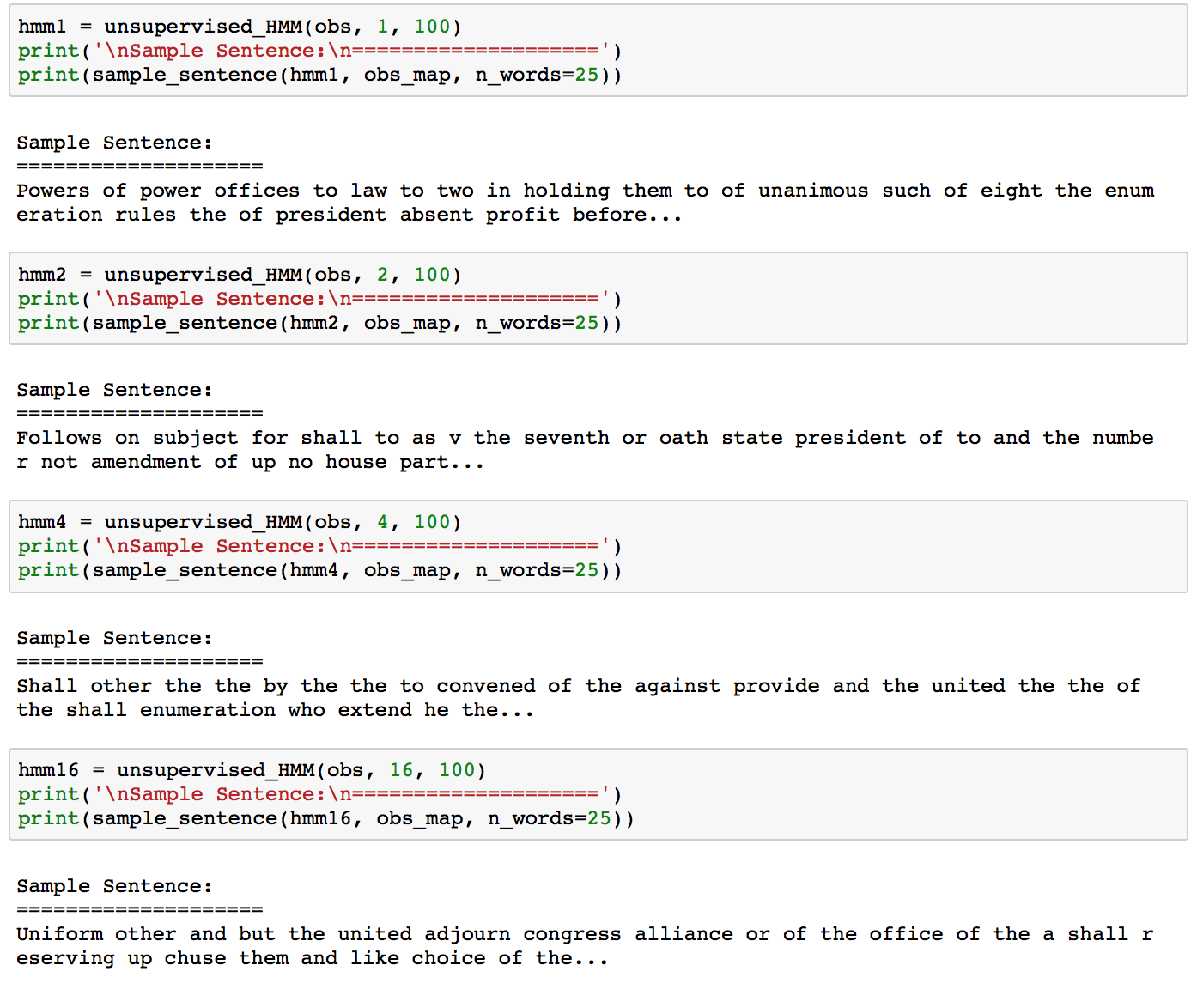
## Part H

In the special case of only 1 hidden state we know that we are essentially choosing words randomly. The transition matrix would be 1 by 1 and the observation matrix would be 1 by D.

Since there is only one state we can also say that the observation matrix would be uniform. Thus, this special case corresponds to choosing words at complete random.

We know that as we increase the number of hidden states we are increasing the likelihood of the training data since we are allowing more parameters with which the model can fit the training data.

We also notice an increase in subsets of the emission that are semblances of grammatically correct clauses in English. These occur more consistently with more hidden states.



## Part I

I find state 2 to be semantically meaningful in that the words have somewhat similar properties. The major components consist of words like Law, Power, Service, Public which are related in that they are nouns that do well to describe situations in public policies.

This differs from others in that a lot of the other word clouds in that the prominent words are related and could occur in similar contexts in sentences and they are also non-filler words.

