**HOMEWORK 7**

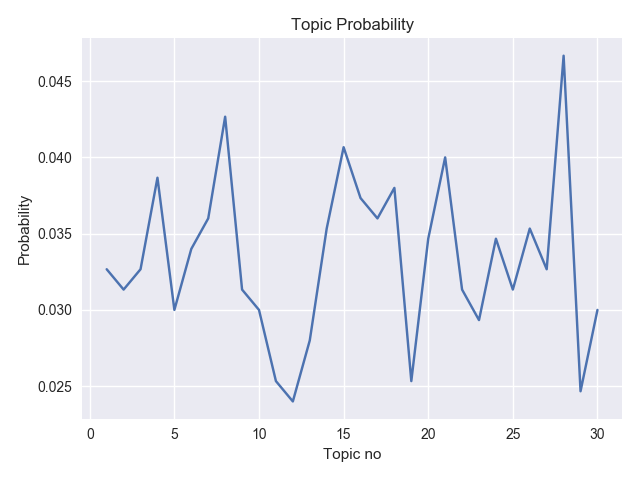
**Problem 1**

The aim of this problem is to use EM algorithm to perform soft clustering and create clusters of topics. The dataset we used in NIPS dataset which contained information about the word occurring in a document. We used this information to create a document vector array which has information about all the word count found in a document as a vector.

The next step was to have an initial cluster centroid to start from, for which we randomly assigned some documents as the cluster centroids and then every other document was assigned to one of these clusters, such that we have an initial value of probabilities. We created the count of all the words in every label and for those clusters that didn’t had any words we did some smoothening.

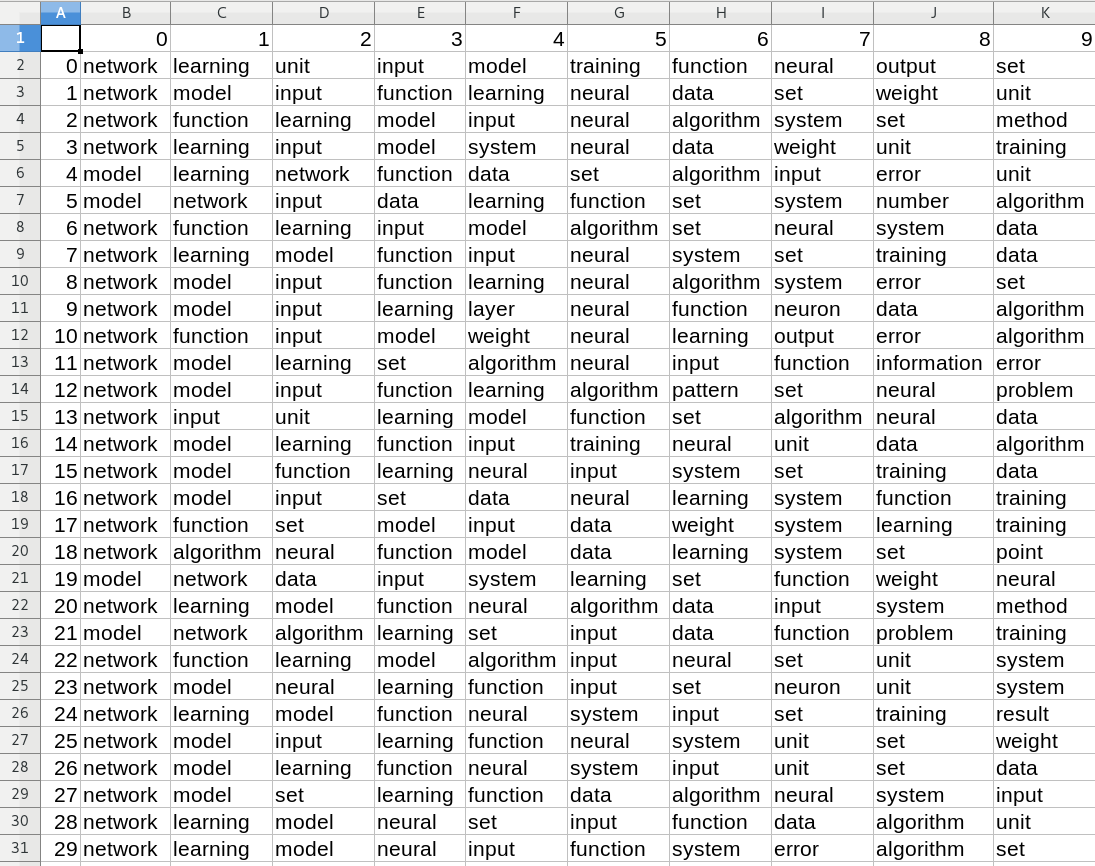
Going forward we perform the EM step as mentioned in the course book and checked for convergence of the algorithm.

We then produced a line graph showing the probability of each topic which is as follows:



There are some spikes around topic 13 and 27.

We also went ahead and generated the top 10 words for every topic, which are as follows:



***The assignment is done in python and you can find the code in problem1.py.***

**Problem 2**

The problem was to perform image segmentation using EM algorithm and assigning each pixel vector to its assigned group such that you represent an image by its no of segments.

We performed the above implemented EM algorithm to a set of 3 images and doing segmentation with 20,30,50 segments. The initial cluster centroids were computed using kmeans. Following is the output:

With 10 segment:



With 20 segments:



With 50 segments:



**Image 2nd**

With 10 segments:



With 20 segments:



With 50 segments:



**Image 3rd**

With 10 Segments:



With 20 Segments:



With 50 Segments:



Part 2:

We used sunset image with 20 segments and we ran the EM with different starting points to see if we could see much difference. We ran this with different seeds. The initial clusters were created using kmeans.

Following is the output:











**We didn’t see much variation in the images.**

***The solution is implemented in python and it can be found in problem2.py.***

**Citation:**

We looked at some of the numpy, misc operations out there in stack overflow. We also read some articles to better understand the topic and the implementation was mostly done as per the steps and formula discussed in course book.