HOMEWORK 7

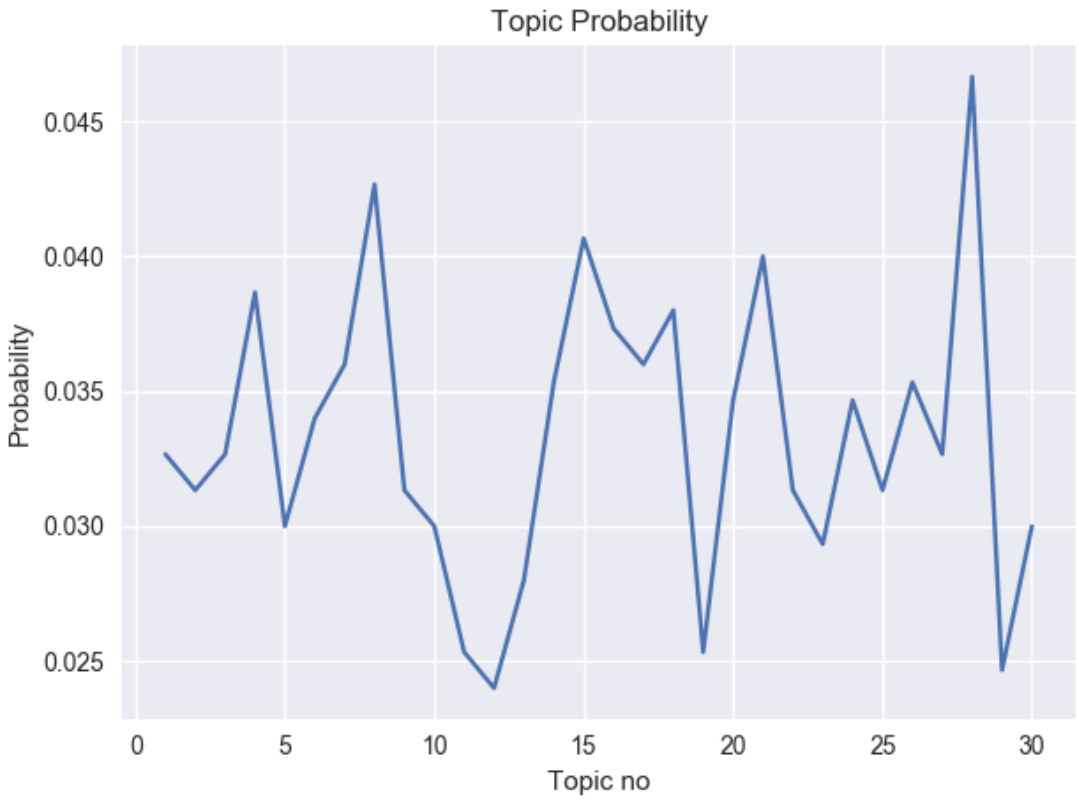
Problem 1

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:

	A	B	C	D	E	F	G	H	I	J	K
1		0	1	2	3	4	5	6	7	8	9
2	0	network	learning	unit	input	model	training	function	neural	output	set
3	1	network	model	input	function	learning	neural	data	set	weight	unit
4	2	network	function	learning	model	input	neural	algorithm	system	set	method
5	3	network	learning	input	model	system	neural	data	weight	unit	training
6	4	model	learning	network	function	data	set	algorithm	input	error	unit
7	5	model	network	input	data	learning	function	set	system	number	algorithm
8	6	network	function	learning	input	model	algorithm	set	neural	system	data
9	7	network	learning	model	function	input	neural	system	set	training	data
10	8	network	model	input	function	learning	neural	algorithm	system	error	set
11	9	network	model	input	learning	layer	neural	function	neuron	data	algorithm
12	10	network	function	input	model	weight	neural	learning	output	error	algorithm
13	11	network	model	learning	set	algorithm	neural	input	function	information	error
14	12	network	model	input	function	learning	algorithm	pattern	set	neural	problem
15	13	network	input	unit	learning	model	function	set	algorithm	neural	data
16	14	network	model	learning	function	input	training	neural	unit	data	algorithm
17	15	network	model	function	learning	neural	input	system	set	training	data
18	16	network	model	input	set	data	neural	learning	system	function	training
19	17	network	function	set	model	input	data	weight	system	learning	training
20	18	network	algorithm	neural	function	model	data	learning	system	set	point
21	19	model	network	data	input	system	learning	set	function	weight	neural
22	20	network	learning	model	function	neural	algorithm	data	input	system	method
23	21	model	network	algorithm	learning	set	input	data	function	problem	training
24	22	network	function	learning	model	algorithm	input	neural	set	unit	system
25	23	network	model	neural	learning	function	input	set	neuron	unit	system
26	24	network	learning	model	function	neural	system	input	set	training	result
27	25	network	model	input	learning	function	neural	system	unit	set	weight
28	26	network	model	learning	function	neural	system	input	unit	set	data
29	27	network	model	set	learning	function	data	algorithm	neural	system	input
30	28	network	learning	model	neural	set	input	function	data	algorithm	unit
31	29	network	learning	model	neural	input	function	system	error	algorithm	set

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

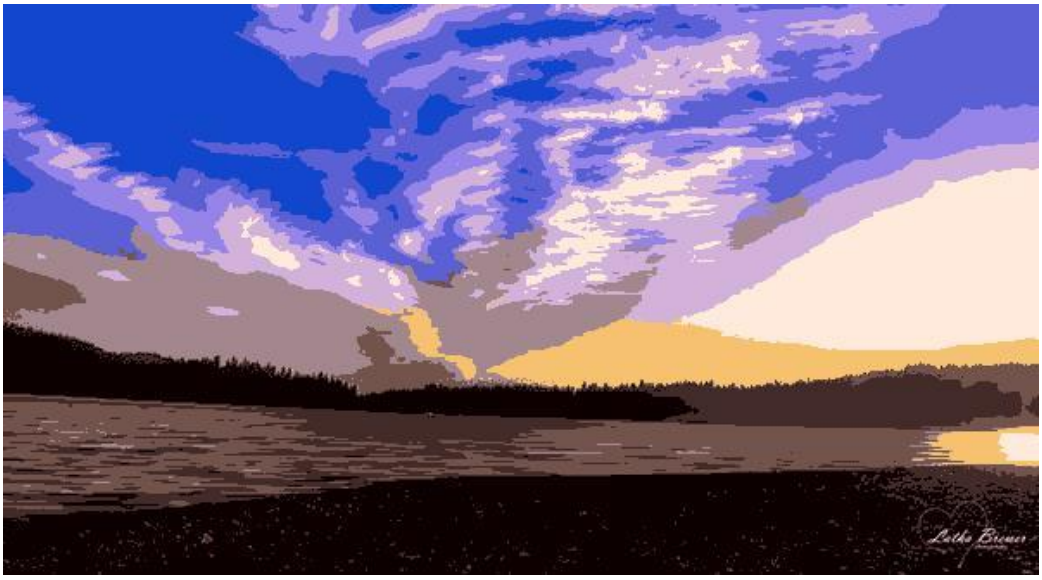
With 10 segments:

With 20 segments:

With 50 segments:

Image 3rd

With 10 Segments:



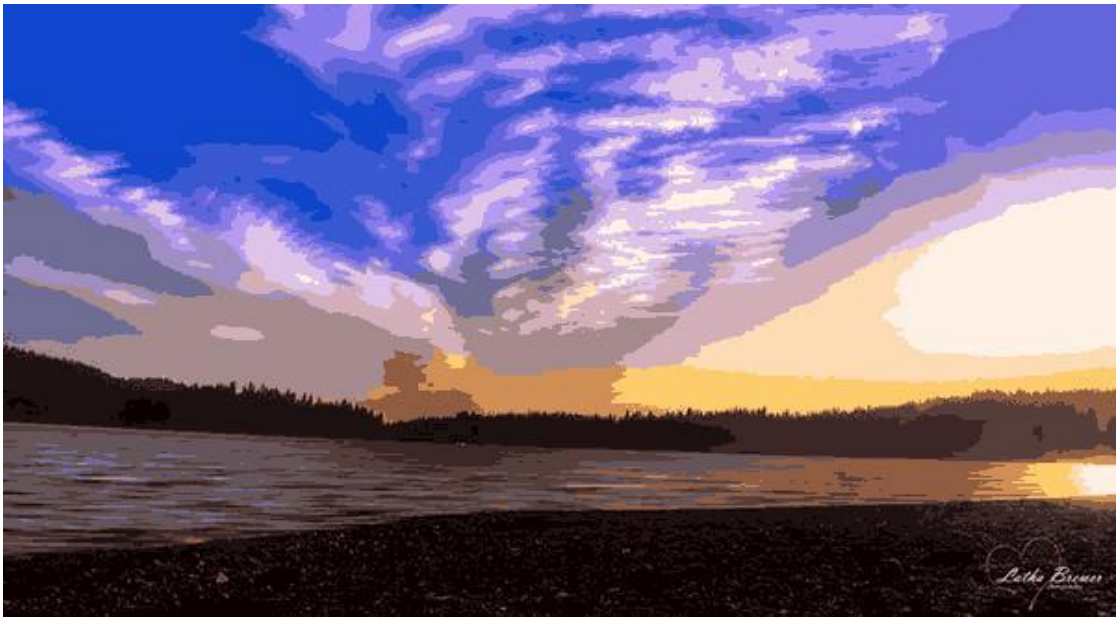
With 20 Segments:

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:


Latha Bremer


Latha Bremer

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.