**CS 498 AML HW7**

**Yuxi Gu(yuxigu2)**

**Songwei Feng(songwei3)**

**Xuanyi Zhu(xzhu42)**

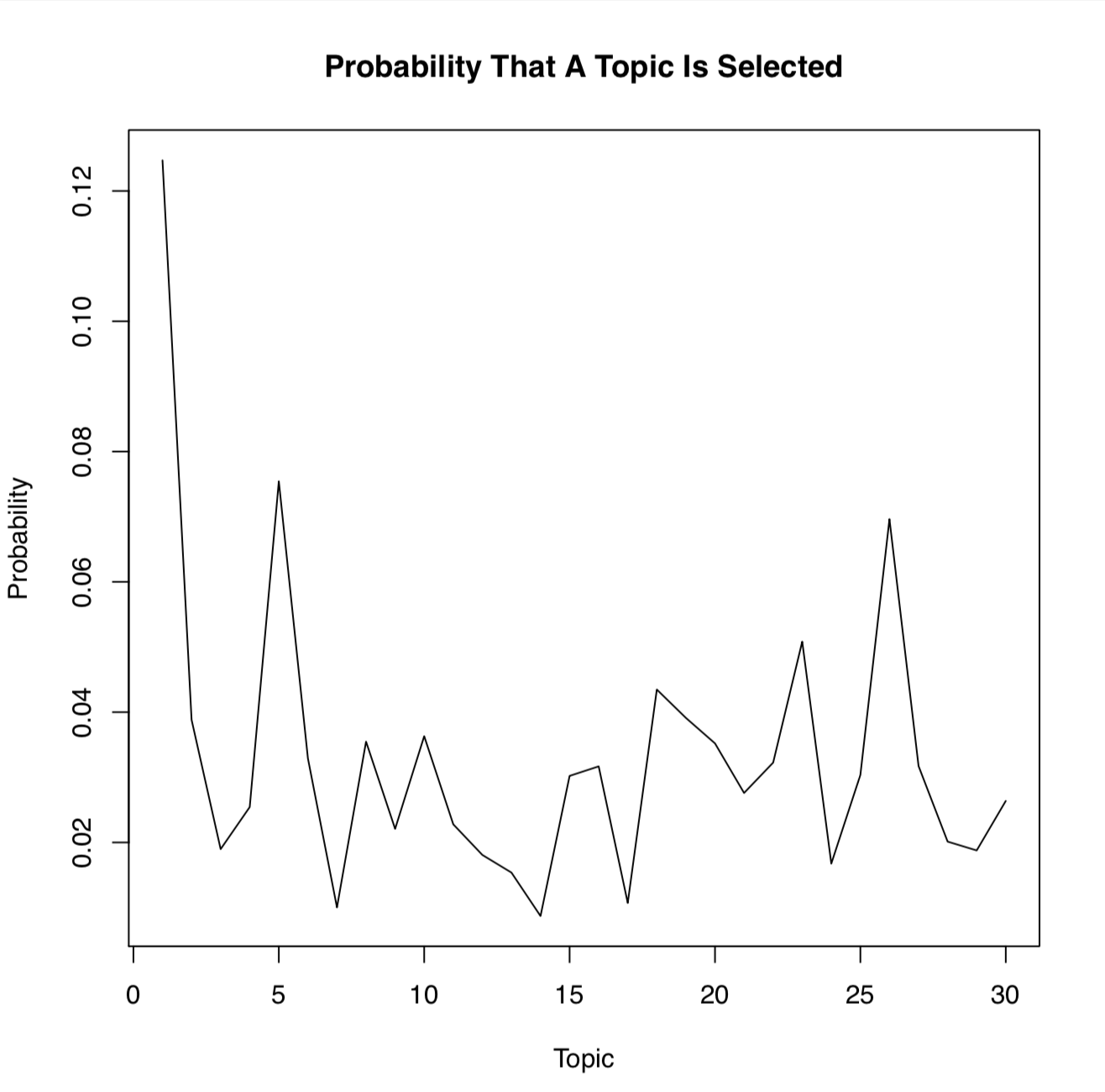
**EM Topic Model**

We used R as the programming language for this problem.

We used NIPS dataset from UCI machine learning repository for this problem. We first constructed a matrix from the dataset in docword.nips.txt. The row index for the matrix indicates the docID and the column index indicates the wordID. And each cell is the word count of the word (represented by wordID) in the document (represented by docID). Then we need to cluster 1500 documents into 30 different topics. And we used random distribution function, which was runif function in R, to get the initial value for each row of P\_j\_k, where each cell of P\_j\_k represents the probability of showing word k in topic j.

After data preparation, we did the EM steps stated in the textbook. And the result we got is showed as below.

**A graph showing, for each topic, the probability with which the topic is selected.**



**A table showing, for each topic, the 10 words with the highest probability for that topic.**

(For this question, we write the result to a csv file and below is a screenshot of the data in this csv file.)

