RQ 3.1: Based on your experiments in Problems 2 and 3, what is the relation between the number of clusters and the quality of a clustering? What would a "good" number of clusters be for this particular data set?

As I listed in the ipynb file, when the k tends to the number of categories, the k-means model performance would become better. In this case the "good" number of clusters should be 7, since the corresponding rand index is the largest.

RQ 3.2: Why is it important to run an LDA model for multiple passes, and not just one? Why is it important to monitor an LDA model for convergence (like you did in Problem 5) and not simply run it for, say, 1000 passes?

After running multiple passes, we can finally get an LDA model, with which we can get the maximum likelihood of all the documents. In other words, these documents can be categorized into appropriate topics as much as possible since we have already get the good parameters for the topic distribution.

RQ 3.3: What are the differences between k-means and LDA? When would you use one, when the other?

The most obvious difference is when using LDA, we can easily find the real topic of documents, since using "model.print_topics()" can give us the key words of a certain topic.

Another difference is deploying LDA is much harder than k-means.