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?

- by the rand index within the cluster numbers we experiemented, the more clusters we have the better the quality is.
- for this particular data set, 6 will be a good number of clusters be cause we have a prior knowledge that their are 6 categories.

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?

- we need to run for multiple passes to converge to a (hopefully global) optimal.
- we do the monitoring to save time. Too much iterations wont help yielding a better retsult pratically but only cosume unnecessary time.

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

- k-means is for clustering set of documents deterministically as a hard clustering, while LDA is for abstracting topics from set of documents probabilistically as a soft clustering.