We know that in NLP community, we often have collections of documents, such as blog posts or news articles or even a list of comments in social networks, that we'd like to divide into natural groups so that we can understand them separately.

So, how to summarize the corpus of documents by that way and perform this task fastly? Topic modeling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items even when we're not sure what we're looking for.

Another problem with topic modeling is that the number of topics is usually a latent variable that we cannot see intuitively from the whole corpus, so the question of how to estimate the optimal number of topic that most probably explain the diversity of the whole corpus becomes challenging.

The question for this first round: let's propose your methodology to approximate the optimal number of topics, generate the topic model corresponding with this optimal number and predict top 3 topics of a given document.

When you have any problem, Please contact: Dr. An Mai – an.mai@jvn.edu.vn - 0906.785.734