Application of LDA topic model in E-Mail Subject Classification

LDA (Latent Dirichlet Allocation) : **Dirichlet** distributions are most commonly used as the prior distribution of categorical variables

LDA + TF-IDF (to determine the reliability of topic classification)

What is LDA? Topic modeling providing methods for automatically organizing, and summarizing large electronic archives

- Discovering the hidden themes -> classifying the docs into the discovered themes -> using the classification to organize the documents
- The aim of LDA is to find topics a document belongs to, based on the words in it.
- Finds representative words for a topic
-

	Word1	word2	word3	word4	.....
Topic1	0.01	0.23	0.19	0.03	
Topic2	0.21	0.07	0.48	0.02	
Topic3	0.53	0.01	0.17	0.04	

Each topic contains a score for all the words in the corpus.

Finding Representative Words for a Topic

- We can **sort the words** with respect to their probability score.
  The top x words are chosen from each topic to represent the topic. If $x = 10$, we'll sort all the words in topic1 based on their score and take the top 10 words to represent the topic.
  This step may not always be necessary because if the corpus is small we can store all the words in sorted by their score.

e.. 2 topics that can be classified as CAT_Related and Dog_Related / Finding words such as milk, meow and kitten will have a higher probability in the Cat-Related, than dog-related.

How to measure probability of words belonging into a topic?

1. 여행이라는 단어를 제외하고, 여행에 연관된 단어만 extract -> 예, 해변, 바다 등 -> 이 단어의 수가 높다면 실제로 여행에 관련된 문서일 것이다.
2. 또한 이 해변이라는 단어가 다른 여행 연관된 문서들에서 얼마나 많이 나오는지를 판단해 해변과 여행이 상관관계가 있다는 것을 보여줌

Also uses stop words to rule-out useless words such a a, of, are is and so on from natural language tool kit (NLKT library).

TF-IDF measure the frequency of a specified word and IDF to measure the general importance of words

Choose top 30 key words of each topic in L -> LDA is used. The first five weighted words of each topic Compare these words if they match well it means LDA is doing great if not, switch topics and how it decides with words should belong to which topic.

5. Experimental Result

The matching accuracy of LDA topic model is changing with the increasing number of topics. The matching accuracy is the highest when the number of topics is 30. The result is shown in Fig 3.

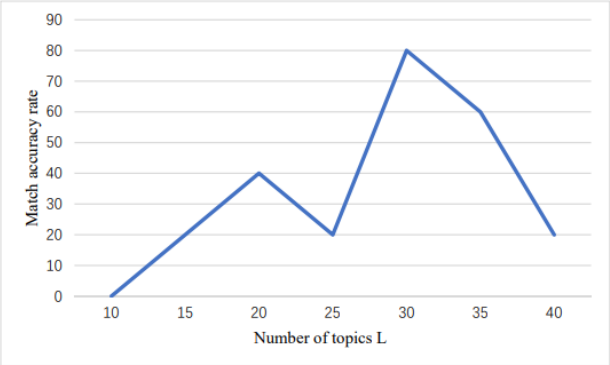

Fig. 3 matching accuracy of LDA topic model under different topic numbers

The proportion of value words in the LDA topic model is changing with the increase of the number of topics. The proportion of value words is the highest when the number of topics is 26. The result is shown in Figure 4.

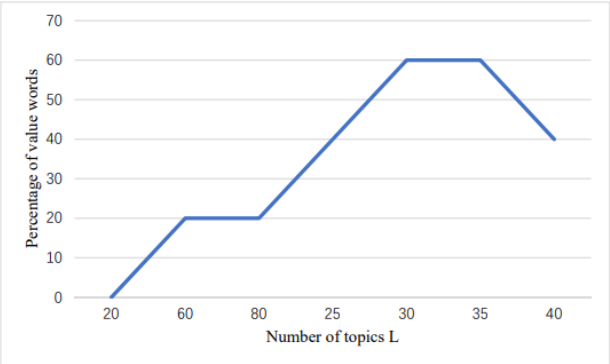

Fig. 4 the ratio of value words under different topic numbers in LDA topic model