Briefly summarize for Latent Dirichlet Allocation (LDA)

This is the age of information. It can be easy to find massive amounts of text content, such as various news articles, emails, blog posts and professional documents. Of course, humans can easily determine the topic of an article or two by reading. However, it is already an impossible task to read and judge by humans alone when we face massive amounts of text data. Therefore, there is the theme of this article, we focus on the topic model of text. Therefore, we will discuss and analysis the one of the topic models which is called Latent Dirichlet Allocation (LDA).

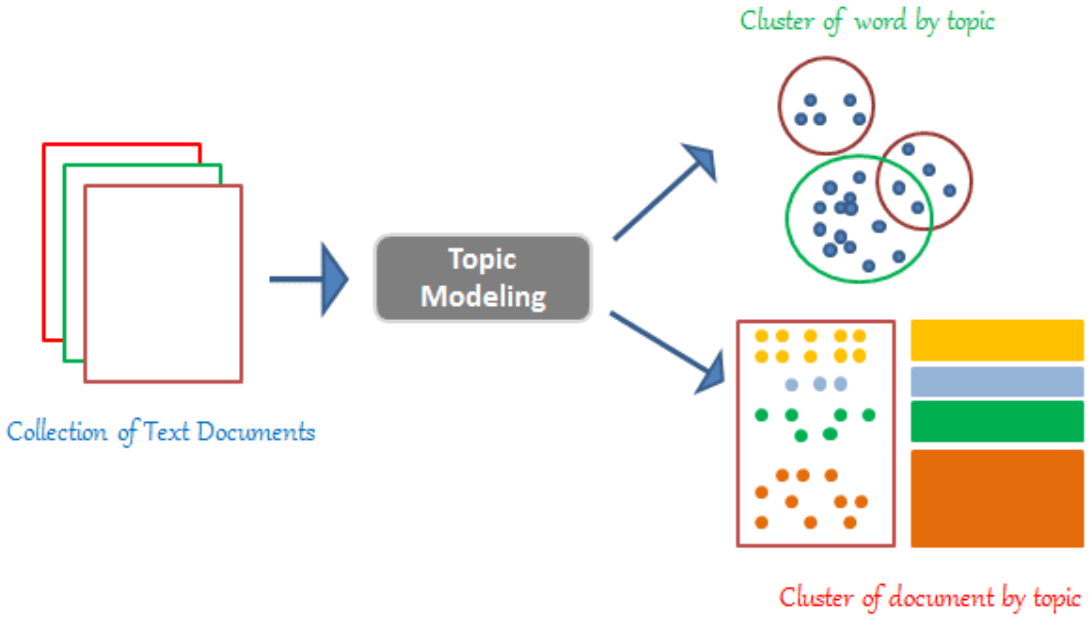

Figure 1. From https://medium.com/analytics-vidhya/latent-dirichelt-allocation-1ec8729589d4

LDA, Latent Dirichlet Allocation, is a generative statistical model which is usually used as topic classification. It is not only an example of a topic model, but also one of the most popular topic modeling methods in the Text information system. As mentioned above, in the face of massive data, we must use more efficient methods to classify the subject of the articles, posts or documents. By counting and classifying these massive text data, we can extract some important and useful information. it not only improves efficiency, but also provides data support and foundation for subsequent data analysis and reuse. Moreover, topic model is the subject of our group project. Therefore, LDA is also one of the candidate technologies for our project.

LDA is a form of unsupervised learning, and it views documents as bags of words. That means it doesn't care about the order of words. A plate diagram of an LDA model is shown below.

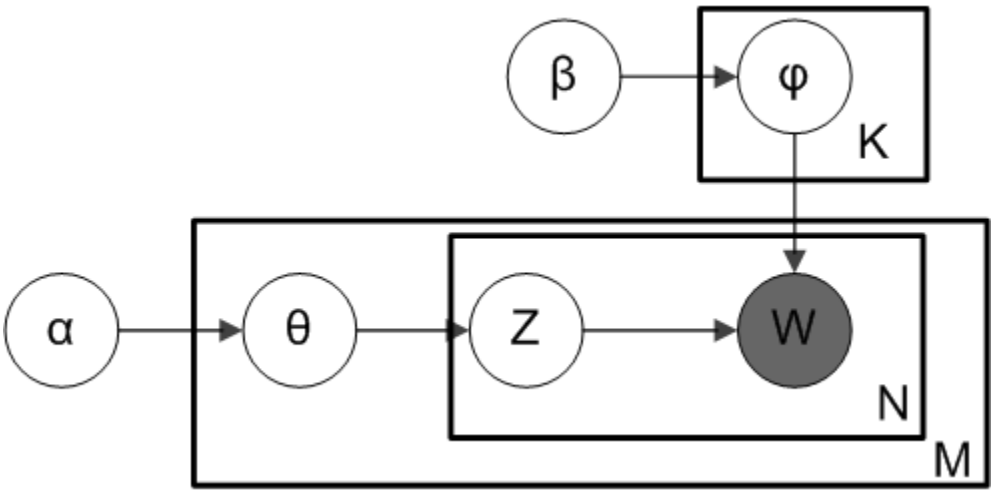

Figure 2. From https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

α - The per-document topic distributions.
β - The per-topic word distribution.
θ - The topic distribution for document m.
ϕ - The word distribution for topic k.
z - The topic for the n-th word in document m.
w - the specific word

First at all, LDA assumes that the way a document was generated was by picking a set of topics and then picking a set of words for each topic. Then, it reverse engineers this process for each document m.

1. The first assuming is the input documents have k topics.
2. Those k topics are distributed across document m. As the description of the LDA model plate diagram, we used α to present this distribution by assigning each word a topic. And α can be symmetric or asymmetric.
3. We assume topic of each word w in document m is wrong, but every other word is assigned the correct topic.
4. We consider what topics are in document m, and how many times word w has been assigned a particular topic across all the documents. This distribution is called β. Then, probabilistically assign word w a topic based on these two conditions.
5. For each document, repeat all steps above several times until all done.

The LDA process is not complex. But to get a good result, we need to pay attention to preprocessing before we use the LDA. Some typical preprocessing steps before performing LDA we can apply, including tokenization, punctuation, special character removal, stop word removal and lemmatized. Those steps can improve the precision of the results.

As mentioned before, The LDA process is not complex, but it is especially useful to resolve the problem in the real world. For example, a large law firm takes over a smaller law firm. To quickly take over the law firm's civil or criminal cases and other related documents. It is a satisfactory solution to apply topic classification based on the LDA algorithm.

Moreover, the second example is my final project for this course. We apply topic modeling to Coursera transcripts. Therefore, this technology can be linked to our project. Since Coursera has different topic contents. We compute the weight of various important chapter topics for each course which we focus on. Then, getting the important key word. The results of analysts could help users focus on the key knowledge points of the course. After using our model, users not only can more clearly understand the important theme of the course content, but also capture the similar knowledge points between different courses that the user has learned. The user can use the analysis result to get a clearer concept of the knowledge focus and structure of the topic they are studying.

Overall, the topic classification technology is increasingly important nowadays. The most popular technology in topic classification model, LDA, is very worthy of in-depth study and application of technology.

References

[1] Team, Great Learning. "Understanding Latent Dirichlet Allocation (LDA)." Great Learning Blog: Free Resources What Matters to Shape Your Career!, 16 Oct. 2020, www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation.

[2] Seth, Neha. "Topic Modeling and Latent Dirichlet Allocation (LDA) Using Gensim." Analytics Vidhya, 28 June 2021, www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn.

[3] "Latent Dirichlet Allocation - Wikipedia." Latent Dirichlet Allocation - Wikipedia, 1 Aug. 2017, en.wikipedia.org/wiki/Latent_Dirichlet_allocation.

[4] Bansal, Harsh. "Latent Dirichlet Allocation." Medium, 25 Nov. 2020, medium.com/analytics-vidhya/latent-dirichelt-allocation-1ec8729589d4.