A Survey on Latent Dirichlet Allocation for Text Classification

**1 Abstract**

Text classification plays a crucial role in various applications, such as sentiment analysis, topic modeling, and spam detection. One popular approach for text classification is the Latent Dirichlet Allocation (LDA) model.

This survey aims to provide an overview of the LDA model and its applications in text classification tasks. We discuss the fundamental concepts behind LDA, explore its probabilistic framework, and examine how it can be effectively utilized for text classification.

**2 Background**

In traditional text classification approaches, the Bag-of-Words model is often employed for feature representation, combined with classical machine learning algorithms such as Naive Bayes or Support Vector Machines for classification. However, these methods have certain limitations, such as the inability to capture the hidden thematic structure and semantic information within the text. LDA addresses this issue by introducing the concept of latent topics, enabling a better reflection of the inherent structure and semantic information of textual data.

**3 Introduction**

LDA model is primarily based on the following observations and assumptions[1]:

1. Documents consist of multiple latent topics: Each document can be regarded as a mixture of multiple latent topics. This implies that a document is not exclusively associated with a single category but rather composed of multiple topics.
2. Topics are represented by probability distributions over words: Each topic can be viewed as a probability distribution over words. This means that each topic has certain common words, and the probability of encountering these words is higher in documents related to that topic.

**4 Methods**

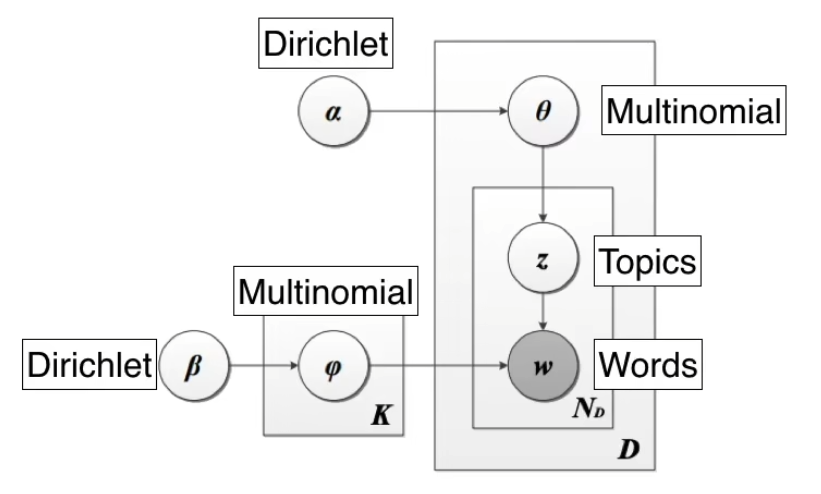


Figure 1. LDA structure[2]

At the first, the LDA model can be as a machine which can produce the document. In LDA, the process of generating documents is as follows[1]:

1. Initialize parameters (hyper-parameter):

K：The number of topics

α：doc-topic Dirichlet distribution

β：topic-word Dirichlet distribution

α represents the distribution that each article may be about each topic.

β represents the distribution of each word in each topic.

1. Generate multinomial distribution:

Θ is the distribution of topics of each document.

Φ is the distribution of words in every topic.

1. Generate some topics and words:

According to Θ distribution, there are n topics generated.

Then for each topics generated, there is one word will be selected according to Φ distribution.

Therefore, there is a document which contain n words.

LDA uses **Gibbs sampling** for training, which is an iterative random sampling method used to approximate the posterior of the probability distribution[2].

There are two properties:

1. Each document is as single topic as possible
2. Each word is as single topic as possible

Steps:

1. Set the number of topics K
2. Randomly assign a topic to each word in each document.
3. Count variables, including document topic count, topic word count, and the total number of vocabulary for each topic.
4. For each word w in document d:

For each topic t:

1. Count words with topic t in document d (document-topic count)
2. Count word w in topic t (topic-word count)
3. Calculate th
4. Update all counts and repeat step 4.

**5 Application**

* 1. Automatic Keyphrase extraction (TF-IDF):

TF: Term Frequency

IDF: Inverse Document Frequency

There are inevitably some stop words in the document, which appear in large numbers but have no practical meaning, such as “of”, “as” and so on. We need an importance adjustment coefficient to measure whether a word is a common word. If a word is relatively rare but appears multiple times in this article, it is likely to reflect the characteristics of the article, which is exactly the keyword we need.

Steps:

1. Calculate TF:

Where x is a word.

1. Calculate IDF:

Where x is a word and Document(x) represents the number of documents including word x.

1. Calculate TF-IDFL:

The higher the TF-IDF of a word, the more important it is.

The disadvantage is that measuring the importance of a word solely based on "word frequency" is not comprehensive enough, and sometimes important words may not appear frequently. Moreover, this algorithm cannot reflect the positional information of words, and words that appear at the front and those that appear at the back are considered equally important, which is incorrect.

* 1. Train LDA model:

1. Determine the number of topics K

Perplexity of the LDA model is a important measure which can be used to determine a good K.

1. Visualize the results and evaluate them

**Reference:**

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
2. Latent Dirichlet Allocation, luis serrano, 2020.

<https://www.youtube.com/watch?v=T05t-SqKArY>