Latent Dirichlet Allocation

112024504 鄭沛杰
112024509 吳振瑋
112024515 高童玄
112024519 孫利東
112024520 黃思緯

Applied Multivariate Analysis Final Project
[1]Institute of Statistics, National Tsing Hua Univsersity

June 16, 2025

目錄 Contents

Problem

1.1 From Simple Examples to Latent Dirichlet Allocation

First, let's refer to Figure 1.1 for a simple example. The figure shows three documents, each containing six words, with a total of four categories: "athlete," "galaxy," "experiment," and "forest." If we want to classify the documents based on their titles, we would identify the theme of the first document as "Sport" since "athlete" appears most frequently in it. Similarly, in the second document, "forest" appears most frequently, so we would identify its theme as "Nature." In the third document, "experiment" and "athlete" each account for half of the occurrences, leading us to classify it as "Sport Science."

Figure 1.1

However, this method is only feasible when the number of documents is

small, allowing us to classify them through visual inspection and judgment. When dealing with thousands or even tens of thousands of documents, how can we automatically identify topics in a large collection of documents? Especially considering that humans can understand the meaning of each word, while for machines, the differences between words may only be in their appearance. In this scenario, Latent Dirichlet Allocation (LDA) can effectively assist us in achieving this goal. LDA has the following features:

- LDA is a generative probabilistic model.
- It discovers hidden topics in large text corpora.
- It represents documents as mixtures of topics.

Intuitive Explanation of LDA

2.1 Introduction

In this chapter, we will not delve deeply into theoretical and mathematical details. To facilitate a better understanding of LDA, we will introduce it from an intuitive perspective.

2.2 What is LDA?

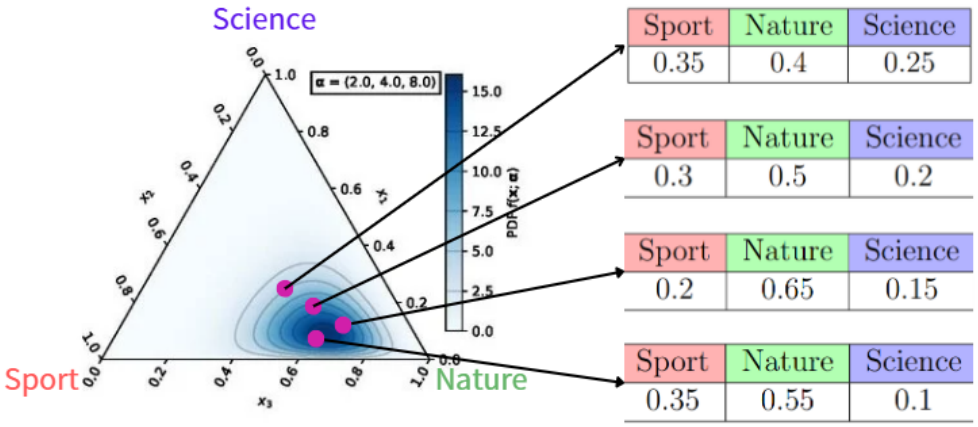
First, Equation (2.1) is a crucial formula for LDA, encapsulating its core principles, and many subsequent theories are built upon it. We can think of Equation (2.1) as a machine, where we feed two inputs, α and β, into the machine. The components $p(\theta \mid \alpha), \prod_{n=1}^{N} p(z_n \mid \theta), p(w_n \mid z_n, \beta)$ represent three gears within the machine. By following these steps, we can ultimately obtain a document. However, before introducing the functions of these "gears," we need to first understand the Dirichlet distribution.

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{2.1}$$

2.3 Dirichlet distribution

The Dirichlet distribution is typically used to describe the probability distribution of parameters for a multinomial distribution. The parameters of the Dirichlet distribution are a set of α vectors, which determine the shape of the

distribution. Using Figure 2.1 as an example, we can draw four points from a Dirichlet distribution with a specific α, representing the relationship between "Sport," "Nature," and "Science." Intuitively, these points represent "words," and the Dirichlet distribution describes the relationship between words and topics. We can use the Dirichlet distribution as the parameters for a multinomial distribution for further applications.



Sport	Nature	Science
0.35	0.4	0.25

Sport	Nature	Science
0.3	0.5	0.2

Sport	Nature	Science
0.2	0.65	0.15

Sport	Nature	Science
0.35	0.55	0.1

Figure 2.1

2.4 How LDA works?

With the concepts mentioned in the previous sections, we can now explain the significance of the three "gears" mentioned in Section 2.2. First, $p(\theta \mid \alpha)$ provides information about the topics, representing the Dirichlet distribution of words for the topics. Suppose we draw a point from this distribution, yielding proportions of 0.7, 0.1, and 0.2 for the topics "Sport," "Nature," and "Science," respectively. This information is then passed to $p(z_n \mid \theta)$, which represents the multinomial distribution of a topic. We can draw five points from this distribution, hypothetically resulting in "Sport," "Sport," "Nature," "Sport," and "Science." This information is then passed to $p(w_n \mid z_n, \beta)$, where z represents the topics we just drew, and β provides the distribution of words for different topics. Based on the topics drawn, we can extract words from the respective distributions, ultimately obtain-

ing "athlete," "athlete," "forest," "galaxy," and "experiment." This is the document generated by LDA, as illustrated in Figure 2.2.

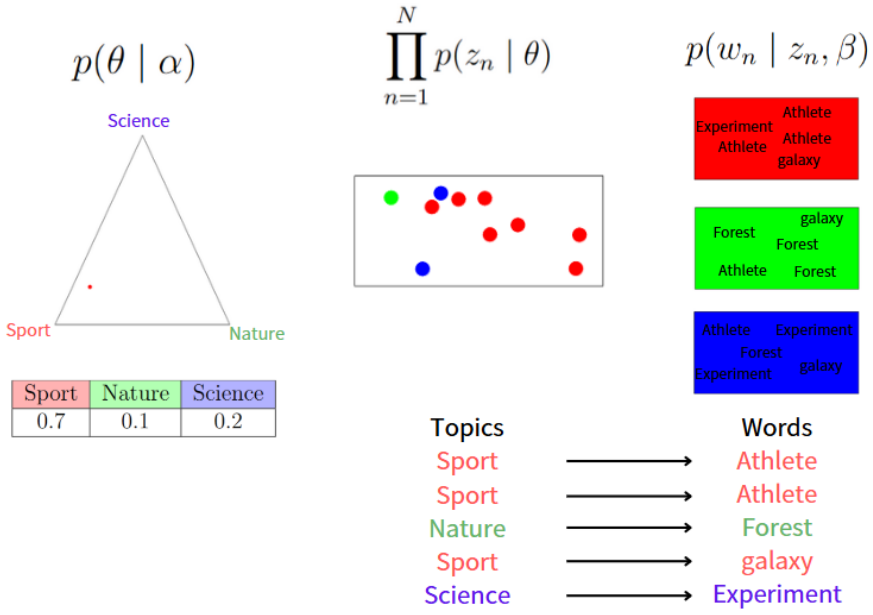


Figure 2.2

Model

3.1 Introduction

In this paper, we frequently use terms like "words," "documents," and "corpora" to describe text collections. This helps to simplify and clarify our explanations, particularly when discussing latent variables that represent abstract concepts like topics. It is important to note that LDA is not limited to text data; it can also be applied to various other data types, including collaborative filtering and bioinformatics.

Here are some key definitions:

- A **word** is a basic unit of data, represented as an item from a vocabulary indexed by $\{1, \ldots, V\}$. Each word is a unit-basis vector with one component equal to one and the rest equal to zero.
- A **document** is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$.
- A **corpus** is a collection of M documents denoted by $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

3.2 Latent Dirichlet Allocation

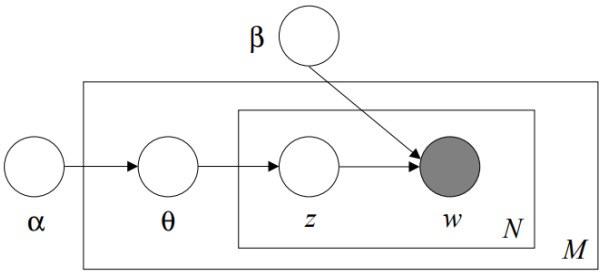
LDA is a generative model that represents documents as mixtures of topics, where each topic is a distribution over words. The generative process for each document w in a corpus D involves:

1. Choosing $N \sim \text{Poisson}(\xi)$.
2. Choosing $\theta \sim \text{Dir}(\alpha)$.

3. For each of the N words w_n:

   (a) Choosing a topic $z_n \sim$ Multinomial(θ).

   (b) Choosing a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.



In this simplified model, we assume the number of topics k is fixed and known. Word probabilities are parameterized by a $k \times V$ matrix β, which we estimate. The Poisson distribution is used for the number of words, but other distributions can also be applied.

3.2.1 Dirichlet Distribution

A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (a k-vector θ lies in the $(k-1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{3.1}$$

3.2.2 Model Inference

Based on the probability distribution assumptions, we have:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\mathbf{z}, \beta) \tag{3.2}$$

By the chain rule and conditional independence,

$$p(\mathbf{w}|\mathbf{z}, \beta) = p(w_1|z_1, \beta) \cdots p(w_N|z_N, \beta) \tag{3.3}$$

we can write the equation:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.4}$$

3.2.3 Posterior Distribution

Our goal is the posterior distribution of the topic proportions θ and the topic assignments z. The posterior distribution is given by:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{3.5}$$

However, why don't we just use this? Since we can now see the 'word' and the topics we selected after seeing the word, the topics may differ from the initial themes behind the word.

$$p(\mathbf{z}|\theta) \quad \text{where} \quad z_n \sim \text{Multinomial}(\theta) \tag{3.6}$$

The final form of $P(\mathbf{w}|\alpha, \beta)$ can be simplified as follows:

$$P(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left[\prod_{n=1}^{N} \sum_{z} P(w_n|z, \beta) P(z|\theta) \right] d\theta \tag{3.7}$$

Given the parameters α and β, the joint distribution is:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.8}$$

By combining these, we can derive the posterior distribution:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{3.9}$$

3.2.4 Marginal Likelihood

The marginal likelihood $p(w|\alpha, \beta)$ is difficult to compute directly due to the high-dimensional integrals and sums involved:

$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_{nj}} \right) d\theta \tag{3.10}$$

The equation is very important, as it relates to Gibbs sampling, which we will use later:

$$P(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha) \left[\prod_{n=1}^{N} \sum_{z} P(w_n|z,\beta) P(z|\theta) \right] d\theta \tag{3.11}$$

We use Gibbs sampling because we want to handle the problem of a word having multiple meanings.

Variational Inference

4.1 Introduction

Variational inference is a technique in Bayesian machine learning that approximates probability densities through optimization. It is particularly useful in scenarios where the true posterior distribution is computationally intractable to obtain directly. This report explores the principles and application of variational inference as presented in the Latent Dirichlet Allocation (LDA) model.

4.2 Variational Inference in LDA

The main idea of variational inference is to apply Jensen's inequality to derive a lower bound for the log likelihood of the observed data. Given a joint probability distribution $p(w, z, \theta|\alpha, \beta)$, the Evidence Lower Bound (ELBO) is defined as:

$$\ln p(w|\alpha, \beta) \geq \mathbb{E}_{q(z,\theta)} \left[\ln \frac{p(w, z, \theta|\alpha, \beta)}{q(z, \theta)} \right] = L(\alpha, \beta) \tag{4.1}$$

where $q(z, \theta)$ is a variational distribution that approximates the true posterior $p(z, \theta|w, \alpha, \beta)$. The objective is to find $q(z, \theta)$ that minimizes the Kullback-Leibler (KL) divergence to the true posterior.

4.2.1 ELBO and KL Divergence Decomposition

A key relationship in variational inference is the decomposition of the log probability of the data $\ln p(w|\alpha, \beta)$ into the ELBO and the KL divergence between the variational distribution and the true posterior. This can be derived as follows:

$$\begin{aligned}
\ln p(w|\alpha, \beta) &= \ln \int p(w, z, \theta|\alpha, \beta)\, dz\, d\theta && (4.2)\\
&= \ln \int \frac{p(w, z, \theta|\alpha, \beta)}{q(z, \theta)} q(z, \theta)\, dz\, d\theta && (4.3)\\
&= \ln \mathbb{E}_{q(z,\theta)} \left[\frac{p(w, z, \theta|\alpha, \beta)}{q(z, \theta)} \right] && (4.4)\\
&\geq \mathbb{E}_{q(z,\theta)} \left[\ln \frac{p(w, z, \theta|\alpha, \beta)}{q(z, \theta)} \right] && (4.5)\\
&= \mathbb{E}_{q(z,\theta)} \left[\ln p(w, z, \theta|\alpha, \beta) \right] - \mathbb{E}_{q(z,\theta)} \left[\ln q(z, \theta) \right] && (4.6)\\
&= \text{ELBO} && (4.7)
\end{aligned}$$

The inequality arises from Jensen's inequality applied to the logarithm of an expectation. We can also express the log probability as:

$$\begin{aligned}
\ln p(w|\alpha, \beta) &= \text{ELBO} + \text{KL}(q(z, \theta) \| p(z, \theta|w, \alpha, \beta)) && (4.8)\\
&= \mathbb{E}_{q(z,\theta)} \left[\ln \frac{p(w, z, \theta|\alpha, \beta)}{q(z, \theta)} \right] + \mathbb{E}_{q(z,\theta)} \left[\ln \frac{q(z, \theta)}{p(z, \theta|w, \alpha, \beta)} \right] && (4.9)
\end{aligned}$$

This shows that maximizing the ELBO is equivalent to minimizing the KL divergence, thereby making the variational distribution $q(z, \theta)$ a good approximation to the true posterior.

4.2.2 Variational Distribution

Assuming the independence of z given θ, the variational distribution is formulated as:

$$q(z, \theta|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{4.10}$$

where γ is the Dirichlet parameter and ϕ are the multinomial parameters for each z_n.

4.2.3 Optimization of ELBO

The ELBO can be further expanded as:

$$L(\gamma, \phi; \alpha, \beta) = \mathbb{E}_q[\ln p(w|z, \beta)] + \mathbb{E}_q[\ln p(z|\theta)] + \mathbb{E}_q[\ln p(\theta|\alpha)] - \mathbb{E}_q[\ln q(\theta)] - \mathbb{E}_q[\ln q(z)] \tag{4.11}$$

The optimization of ELBO involves updating the variational parameters γ and ϕ iteratively. This is achieved by setting the derivatives of the KL divergence to zero, yielding the update equations:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathbb{E}_q[\ln(\theta_i)|\gamma]\} \tag{4.12}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{4.13}$$

The update rules are derived from the fact that maximizing the lower bound $L(\gamma, \phi; \alpha, \beta)$ is equivalent to minimizing the KL divergence between the variational distribution and the true posterior.

4.3 EM Algorithm for Variational Inference

The Expectation-Maximization (EM) algorithm is a general framework for finding maximum likelihood estimates in models with latent variables. In the context of variational inference, the EM algorithm can be used to optimize the variational parameters and the model parameters iteratively.

4.3.1 E-Step (Expectation Step)

In the E-step, we update the variational parameters by maximizing the ELBO with respect to $q(z, \theta)$ while keeping the model parameters α and β fixed. This involves computing the expectations with respect to the variational distribution:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathbb{E}_q[\ln(\theta_i)|\gamma]\} \tag{4.14}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{4.15}$$

4.3.2 M-Step (Maximization Step)

In the M-step, we update the model parameters α and β by maximizing the ELBO with respect to these parameters while keeping the variational distribution fixed. This typically involves solving the following optimization problem:

$$\beta_{ij} \propto \sum_{d=1}^{D} \sum_{n=1}^{N} \phi_{dni} w_{dnj} \tag{4.16}$$

To update α, we can use the following fixed-point iteration method:

$$\alpha_{\text{new}} = \alpha_{\text{old}} + \frac{M \left(\psi \left(\sum_{k=1}^{K} \alpha_k \right) - \psi(\alpha_k) \right)}{\sum_{d=1}^{D} \left(\psi(\gamma_d^k) - \psi \left(\sum_{k=1}^{K} \gamma_d^k \right) \right)} \tag{4.17}$$

where ψ is the digamma function, M is the number of documents, K is the number of topics, and γ_d^k is the variational parameter for document d and topic k.

Perplexity

5.1 Perplexity in Language Modeling

Perplexity is a measurement commonly used in natural language processing and machine learning to evaluate the performance of probabilistic models, especially in language modeling. It measures how well a probability distribution or model predicts a sample. Lower perplexity indicates a better fit of the model to the data, implying better generalization performance.

5.1.1 Definition and Formula

In the context of Latent Dirichlet Allocation (LDA), perplexity is defined as:

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{5.1}$$

where $p(w_d)$ is the likelihood of the words in document d, M is the number of documents, and N_d is the number of words in document d .

5.1.2 Application and Interpretation

Perplexity is used by convention in language modeling and is monotonically decreasing with the likelihood of the test data. A lower perplexity score indicates better generalization performance of the model.

Example

6.1 Example: BBC Data

- Select 64 articles from BBC.
- Choose different article topics to increase the complexity of the information.
- Preliminary processing of data.

Most of the articles are from mid-May on the BBC. Each article ranges from 250 to 1200 words. Although forums like ptt were considered, we ultimately chose the BBC because English is easier to process and the variation in words or phrases is less frequent. Additionally, the source of the articles is easier to obtain.

Articles are selected from eight topics to increase data complexity:

- Sport: 15
- Business: 8
- Technology: 10
- Game: 5
- Travel: 6
- Music: 7
- Science: 6

- Political: 7

We hope to analyze the corresponding topics through the composition of the articles using LDA.

First, we exclude content unrelated to the text, such as reporters, authors, editors, publishers, advertisements, external links, publication time, punctuation marks, etc. After organizing the text of the same article, the words are represented as a vector.

6.2 Gibbs Sampling

Assume we decided to use 4 topics for analysis.

Properties:

1. The topic of the text within the same article should be consistent.
2. The subject of the same text should be as consistent as possible.

6.2.1 How Gibbs Sampling Works

Gibbs sampling uses Dirichlet distribution parameters α and β to iteratively refine topic assignments based on the properties mentioned.

To illustrate Gibbs sampling in LDA:

- Randomly assign a topic to each word in the document, usually using different color to represent pictures of different topic.
- According to the two properties, consider the first word "music" in the first article. If the first article has two red topics and two blue topics, and "music" appears red topic across all articles. Since events are probabilities, this procedure uses the product of probabilities for integration. Directly setting the unobserved topic to 0 is too extreme. Therefore, adding a positive value makes the process more flexible and yields more accurate results. α and β are the parameters of the Dirichlet distribution introduced earlier, set by the user and known values.
- From the sum of these products, randomly choose a topic for the word. Repeat this process for other words to get the final converged result, which is Gibbs sampling work in the LDA.

Since computers cannot distinguish the meaning of words like humans, they can only distinguish whether words are the same. Therefore, Requires manual or other tools to name topics(e.g., bank stocks usually appear in economic topics, while galaxy and comet appear in natural topics).

6.3 After LDA

- Assign topic names created by LDA (e.g., economy, technology, humanities).
- Make the conclusion based on the topic distribution.

Visualization: In low-dimensional space, each corner represents a topic, and each point represents an article's topic distribution. Most points should be near the corners because most articles have one to two specific topics. By observing the distribution, one can infer how many articles are in each topic.

Data Analytic Demo

7.1 Natural Language Processing

7.1.1 Original Sentences

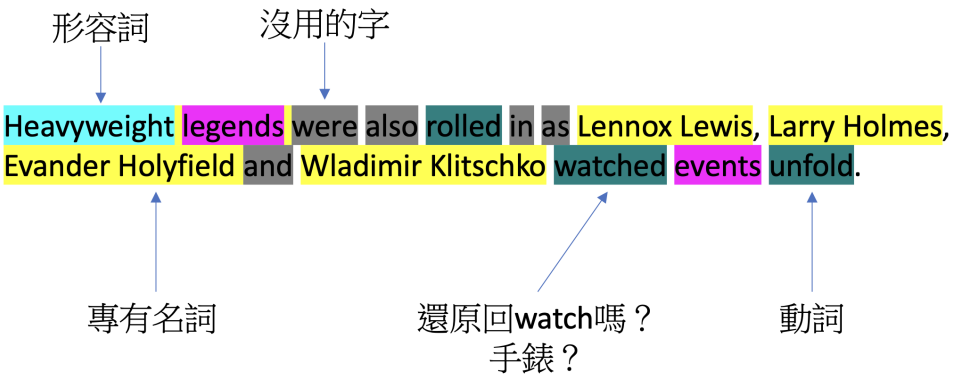


Figure 7.1: 英文句子拆分

From the above figure, we can observe several issues in the sentences, including different tenses, homophones with different meanings, and other ambiguities. Therefore, we employ the following steps to process the text.

7.1.2 Processing Steps

Step 1: Retain Nouns, Verbs, Adjectives, and Adverbs

- Extract nouns, verbs, adjectives, and adverbs from the text.

Step 2: Restore Original Forms

- Convert plural nouns to their singular form.
- Convert verbs to their present tense.
- Use techniques like Lemmatization.

Step 3: Remove Stop Words

- Remove common stop words (e.g., “is”, “the”, “at”, etc.) from the text.
- Increase the conciseness and processing efficiency of the text.
- Filter using a stop word list.

7.1.3 Processed Sentences

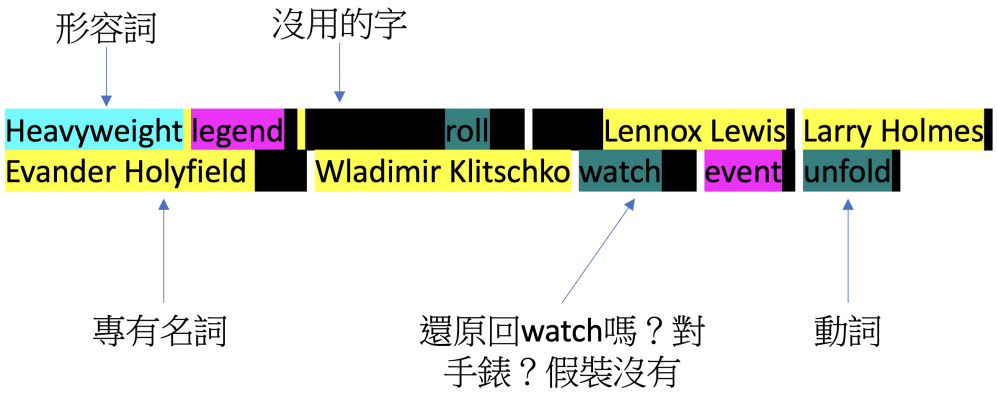


Figure 7.2: Processing Result

The above figure shows the sentences after processing. However, we can notice that the word ”watch” has both meanings of a wristwatch and the act of looking. Due to the limitations of LDA, this issue is not addressed.

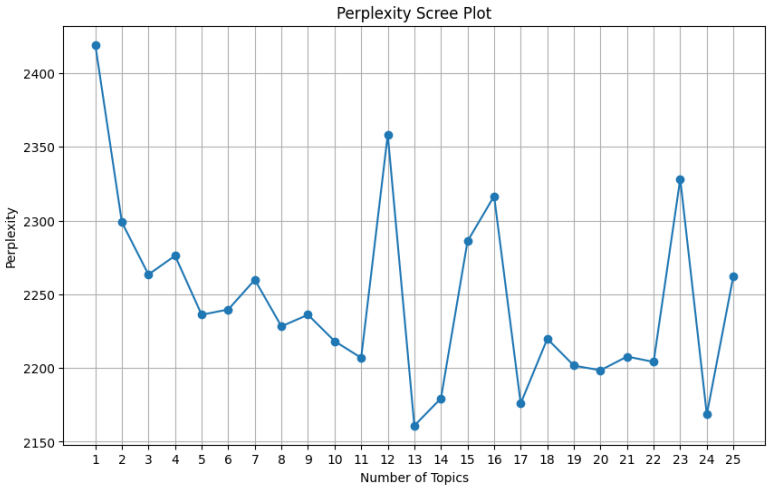


Figure 7.3: Perplexity Scree Plot

7.2 LDA: Model Selection

7.2.1 Number of Topics Selection

From the above figure (figure 7.3), we can observe a decreasing trend in perplexity. Theoretically, lower perplexity indicates better performance. However, selecting too many topics can lead to difficulties in interpretation. Therefore, we choose 5 topics, as the rate of decrease in perplexity diminishes at this point.

7.2.2 LDA: Will cross validation work?

From the cross validation log-likelihood performance, it is evident that LDA does not generalize well to unseen datasets. This makes it unsuitable for using CV to select the number of topics.

7.2.3 LDA: T-SNE visualization

From the T-SNE projection to the 2D plane, we can observe that our topics are clearly divided into 4 clusters. This does not entirely align with our division into 5 topics. However, a closer look reveals that one of the clusters is

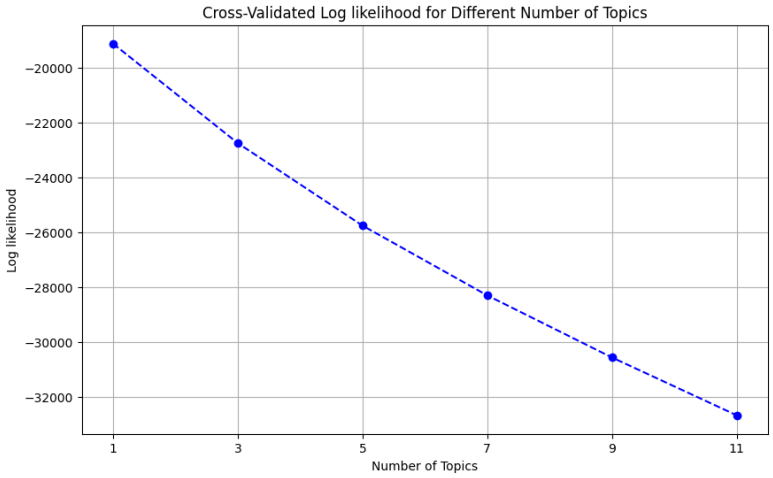


Figure 7.4: CV log-likelihood

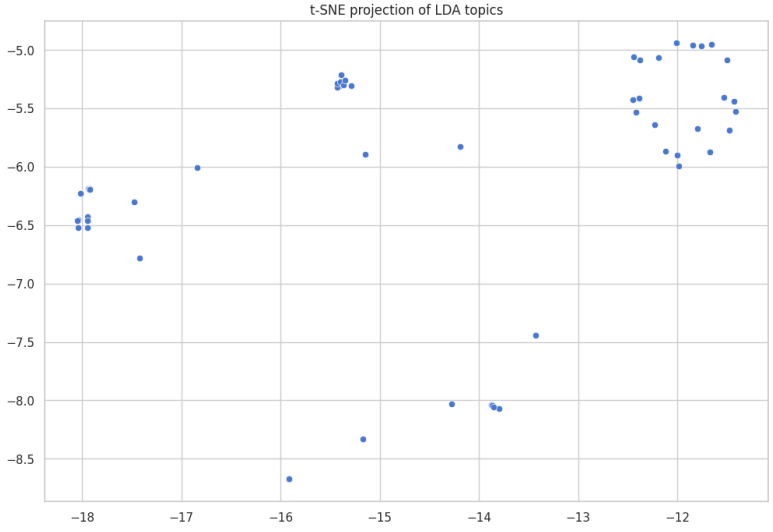


Figure 7.5: T-SNE plot

particularly elongated, suggesting that it might actually contain two distinct groups.

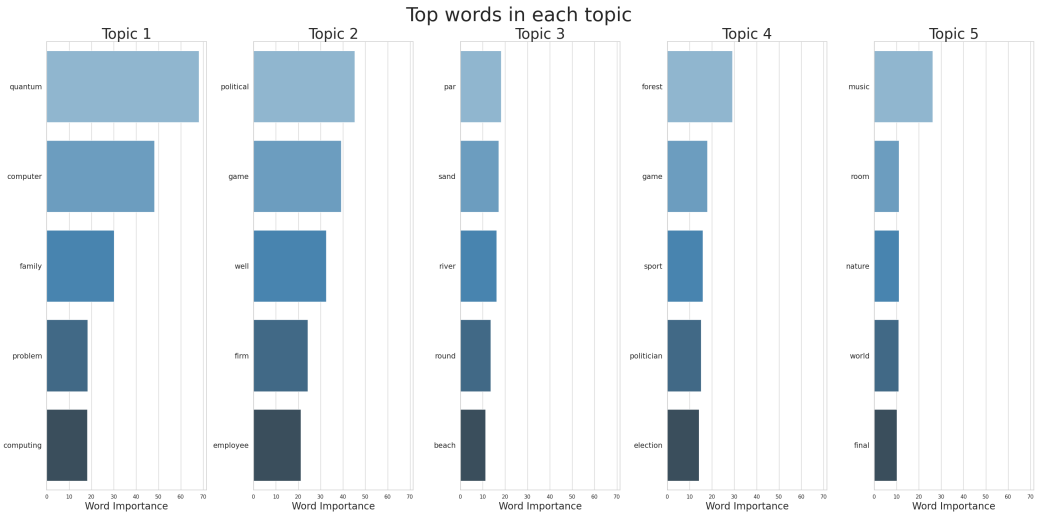


Figure 7.6: Topic

7.3 LDA: Topic Explanation

The above figure (figure 7.6) shows the top words in each of the five topics identified by the LDA model. Here is an interpretation of each topic:

- **Topic 1: Tech** - Words like "quantum", "computer", "computing", and "problem" suggest this topic is related to technological advancements and issues, particularly in computing and quantum technologies.
- **Topic 2: Politics** - Words such as "political", "game", "firm", and "employee" indicate this topic deals with political matters, possibly including political strategies and employment-related discussions.
- **Topic 3: Nature and Environment** - The presence of words like "par", "sand", "river", and "beach" suggests this topic focuses on natural environments and possibly discussions around conservation or outdoor activities.
- **Topic 4: Sports and Recreation** - With words like "forest", "game", "sport", and "politician", this topic seems to encompass recreational activities, sports, and possibly the involvement of political figures in these areas.

- **Topic 5: Music and Entertainment** - Words such as "music", "room", "nature", and "world" indicate that this topic revolves around music, entertainment, and perhaps cultural events.

7.3.1 Topic 1: Tech

We extracted two news articles from the tech topic as examples to evaluate the performance of LDA. The first article is clearly related to technology, as it mentions several aspects of quantum computing. However, the second article is more political, focusing on the actions of Humza Yousaf, which is more aligned with political news. These examples show that the performance of LDA is not always accurate.

Ex1. Quantum breakthrough could revolutionise computing

Scientists have come a step closer to making multi-tasking 'quantum' computers, far more powerful than even today's most advanced supercomputers....

Ex2. Humza Yousaf's decision follows on from SNP political time bombs

In his brief stint as Scotland's first minister, there is one moment for Humza Yousaf I will never forget.

Last October, Mr Yousaf was embarking on a political ritual - a round of interviews with political editors before his party's conference in Aberdeen. ...

Reference

- David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(2003), 993-1022.
- https://www.bbc.com/news