Latent Dirichlet Allocation

吳振瑋　孫利東　高童玄　黃思緯　鄭沛杰

Institute of Statistics National Tsing Hua University

2024/5/28

outline

Problem

Example

Doc 1	Doc 2	Doc 3
Athlete	Forest	Experiment
Athlete	Forest	Experiment
Athlete	Forest	Athlete
galaxy	Athlete	Athlete
Experiment	galaxy	Experiment
Athlete	Experiment	Athlete

Sport Nature Science

Problem

Question:

How can we automatically identify topics in a large collection of documents?

Solution: Latent Dirichlet Allocation (LDA)

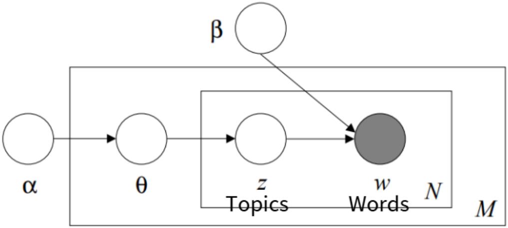
LDA

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

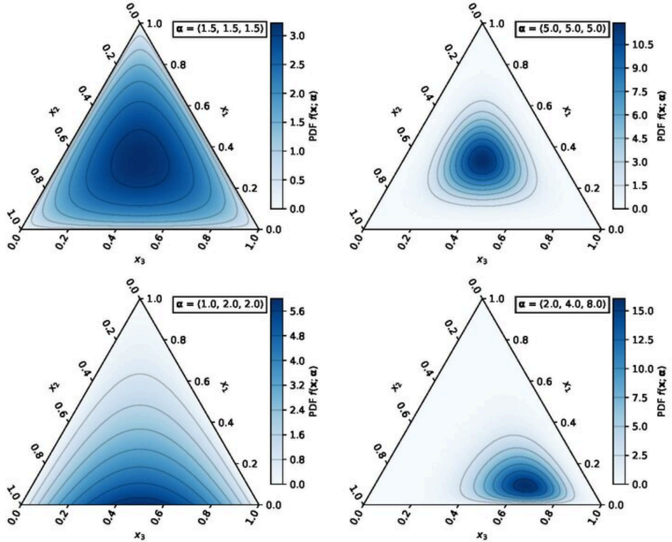
Intuitive Explanation of LDA

What is LDA

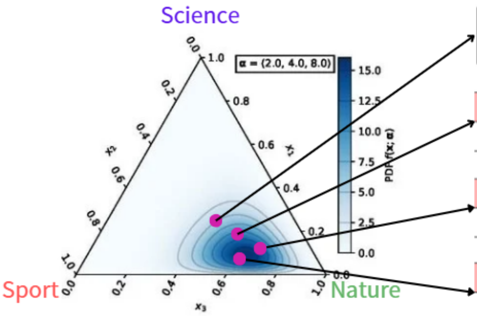
$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$



Dirichlet distribution



Dirichlet distribution (cont.)



Sport	Nature	Science
0.35	0.4	0.25

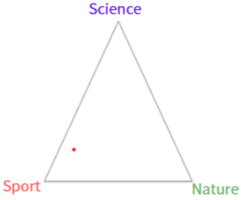
Sport	Nature	Science
0.3	0.5	0.2

Sport	Nature	Science
0.2	0.65	0.15

Sport	Nature	Science
0.35	0.55	0.1

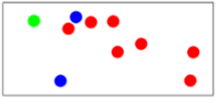
How LDA works?

$$p(\theta \mid \alpha)$$



Sport	Nature	Science
0.7	0.1	0.2

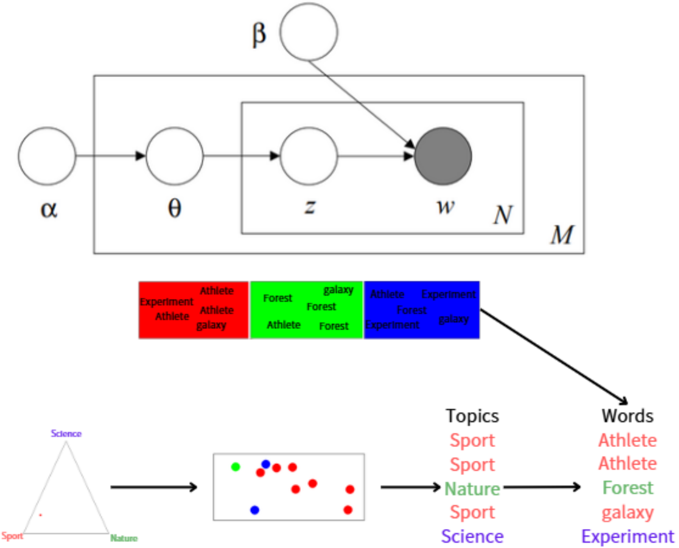
$$\prod_{n=1}^{N} p(z_n \mid \theta)$$



$$p(w_n \mid z_n, \beta)$$

Topics		Words
Sport	$\longrightarrow$	Athlete
Sport	$\longrightarrow$	Athlete
Nature	$\longrightarrow$	Forest
Sport	$\longrightarrow$	galaxy
Science	$\longrightarrow$	Experiment

How LDA works?



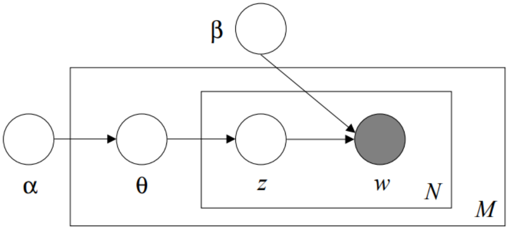
LDA

- **Word:** A word w is an element from a vocabulary indexed by {1, …, V} and represents the basic unit of text data.
- **Document:** A document is a sequence of N words denoted by $\mathbf{w} = (w_1, \ldots, w_N)$.
- **Topic:** Each word w_k in a document is associated with a topic z_k.
- **Corpus:** A corpus is a collection of M documents, denoted by $D = (\mathbf{w}_1, \ldots, \mathbf{w}_M)$.

LDA

For each document w in the corpus, **LDA** assumes the following generative process:

1. Choose the number of words $N \sim \text{Poisson}(\xi)$.
2. Choose a topic distribution $\theta \sim \text{Dir}(\alpha)$.
3. For each word w_n:
    1. Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
    2. Choose a word w_n from $p(w_n \mid z_n, \beta)$.



LDA

$$\theta \text{ (document topic dist.) } = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1k} \\ \vdots & \ddots & \vdots \\ \theta_{M1} & \cdots & \theta_{Mk} \end{bmatrix}_{M \times k}$$

$$\beta \text{ (topic word dist.) } = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1V} \\ \vdots & \ddots & \vdots \\ \beta_{k1} & \cdots & \beta_{kV} \end{bmatrix}_{k \times V}$$

- M : Number of documents
- k : Number of topics
- V : Number of words

Model

Model

Based on the probability distribution assumptions, we have:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) p(\mathbf{z} \mid \theta) p(\mathbf{w} \mid \mathbf{z}, \beta)$$

By the chain rule and conditional independence,

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

we can write the equation:

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

Model(conti)

Our goal is the posterior distribution of the topic proportions θ and the topic assignments $\mathbf{z}$. The posterior distribution is given by:

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

However, why don't we just use this?

$$p(\mathbf{z} \mid \theta)$$

where $z_n \sim \text{Multinomial}(\theta)$

Model(conti)

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

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

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

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

By combining these, we can derive the posterior distribution:

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

Model(conti)

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

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

Model(conti)

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

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

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

Variational Inference

Variational Inference

- The idea of the variational inference is to apply the Jensen's inequality to obtain a lower bound of $P(\mathbf{w}|\alpha, \beta)$.
- Let $q(\mathbf{z},\theta)$ be a joint probability density of $\mathbf{z}, \theta$, applying the idea of importance sampling yields.

$$
\begin{aligned}
ln\ [p(\mathbf{w}|\alpha, \beta)] &= ln \int \sum_{\mathbf{z}} p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) d\theta \\
&= ln \int \sum_{\mathbf{z}} \frac{p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta)}{q(\mathbf{z}, \theta)} q(\mathbf{z}, \theta) d\theta \\
&\geq \int \sum_{\mathbf{z}} q(\mathbf{z}, \theta) ln\ \frac{p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta)}{q(\mathbf{z}, \theta)} d\theta := L(\alpha, \beta)
\end{aligned}
$$

- $L(\alpha, \beta)$ is also referred to as Evidence Lower Bound(ELBO).

Variational Inference

- The Evidence Lower Bound (ELBO) can be further decomposed as follows :

$$
\begin{aligned}
L(\alpha,\beta) &= \int \sum_{\mathbf{z}} q(\mathbf{z},\theta) ln \frac{p(\mathbf{w},\mathbf{z},\theta|\alpha,\beta)}{q(\mathbf{z},\theta)} d\theta \\
&= \int \sum_{\mathbf{z}} q(\mathbf{z},\theta) ln \frac{p(\mathbf{z},\theta|\mathbf{w},\alpha,\beta)p(\mathbf{w}|\alpha,\beta)}{q(\mathbf{z},\theta)} d\theta \\
&= \int \sum_{\mathbf{z}} q(\mathbf{z},\theta) ln\, [p(\mathbf{w}|\alpha,\beta)] d\theta - \\
&\quad \int \sum_{\mathbf{z}} q(\mathbf{z},\theta) ln \frac{q(\mathbf{z},\theta)}{p(\mathbf{z},\theta|\mathbf{w},\alpha,\beta)} d\theta \\
&= ln\, [p(\mathbf{w}|\alpha,\beta)] - KL(q(\mathbf{z},\theta||p(\mathbf{z},\theta|\mathbf{w},\alpha,\beta))
\end{aligned}
$$

- Our goal is to find a $q(\mathbf{z},\theta)$ that has as small KL divergence to $p(\mathbf{z},\theta|\mathbf{w},\alpha,\beta)$ as possible.

Variational Inference

- Assume $\mathbf{z}_1, ..., \mathbf{z}_N$ are independent, we may have the following variational distribution :

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

where Dirichlet parameter γ and multinomial parameters $(\phi^1, ..., \phi^N)$ are the free variational parameters.

Variational Inference (cont.)

- The expansion of ELBO :

$$
\begin{aligned}
& L(\alpha, \beta) \\
=& L(\gamma, \phi; \alpha, \beta) \\
=& E_q[ln\, p(\mathbf{w}|\mathbf{z}, \beta)] + E_q[ln\, p(\mathbf{z}|\theta)] + E_q[ln\, p(\mathbf{w}|\mathbf{z}, \beta)] \\
& - E_q[ln\, q(\theta)] - E_q[ln\, q(\mathbf{z})]
\end{aligned}
$$

Variational EM Algorithm

- E-step :
  Given (α, β) , find the optimizing value of the variational parameters (γ, ϕ) that maximize $L(\gamma, \phi; \alpha, \beta)$.
- M-step :
  With (γ, ϕ) obtained from the E-step , find the optimizing value of the model parameters (α, β) that maximize $L(\gamma, \phi; \alpha, \beta)$.
- Repeat the two steps until the value of the ELBO converges.

Perplexity

Perplexity

- The perplexity , used by convention in language modeling, is monotonically decreasing in the likelihood of the test data.
- A lower perplexity score indicates better generalization performance.

$$perplexity(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} ln\ p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\}$$

Example

Example:Processing data

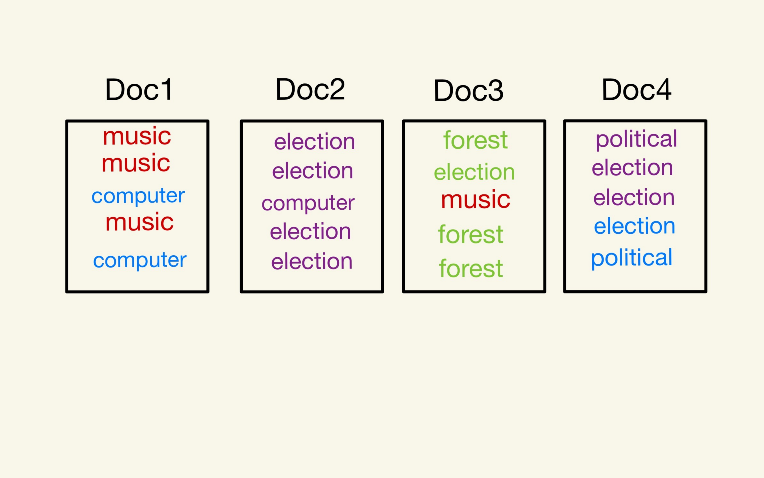
- Select 64 articles from BBC
- Choose different article topics to increase the complexity of the information
- Preliminary processing of data

Example:LDA program(rough)

- **Gibb-sampling**
- Determine the number of topics (BY perplexity)
- Plot data in low-dimensional space

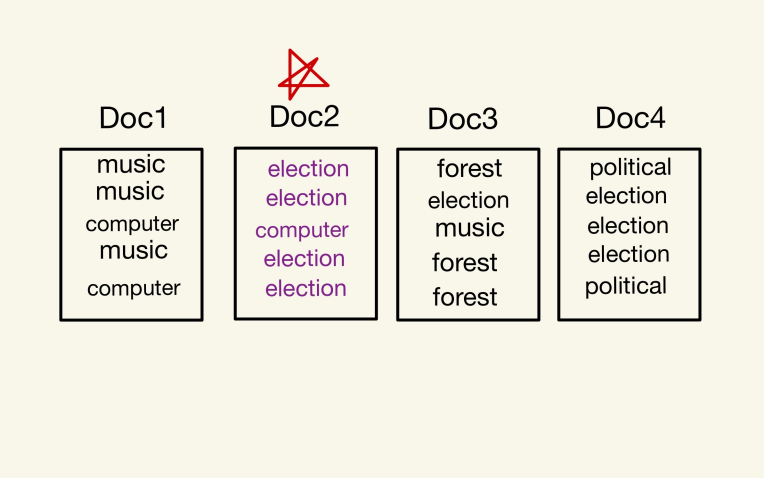
Gibb-Sampling

Assume we decided to use 4 topic for analysis.



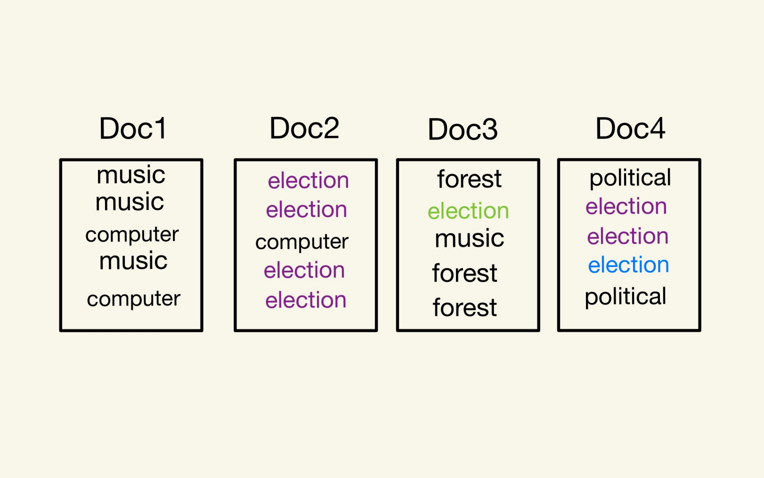
Gibb-Samplimg

Property1: The topic of the text within the same article should be single

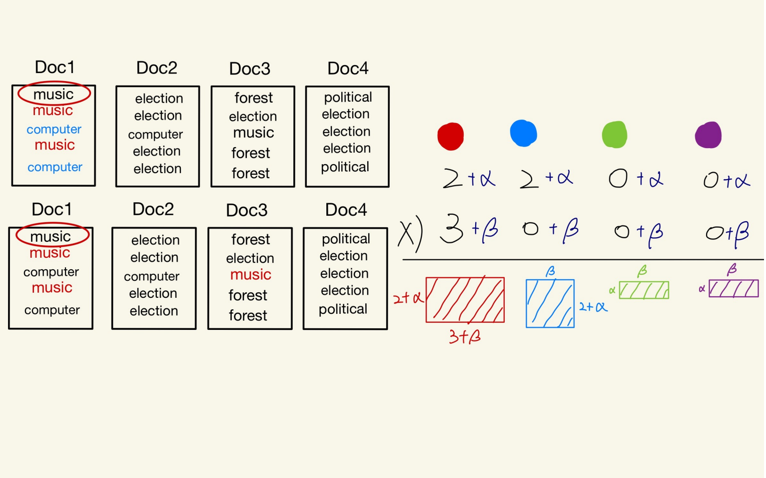


Gibb-Sampling

Property2:The subject of the same text should be as single as possible

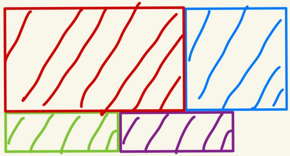


Gibb-Sampling



α, β is Dirichlet distribution

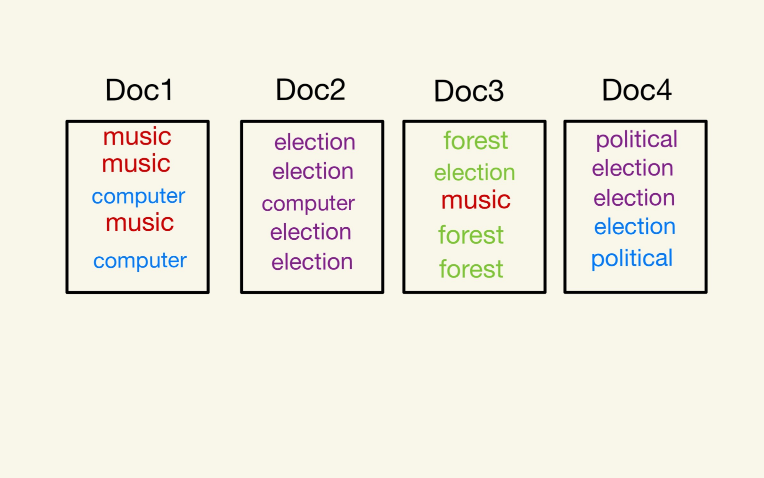
Gibb-Sampling



Example:After LDA

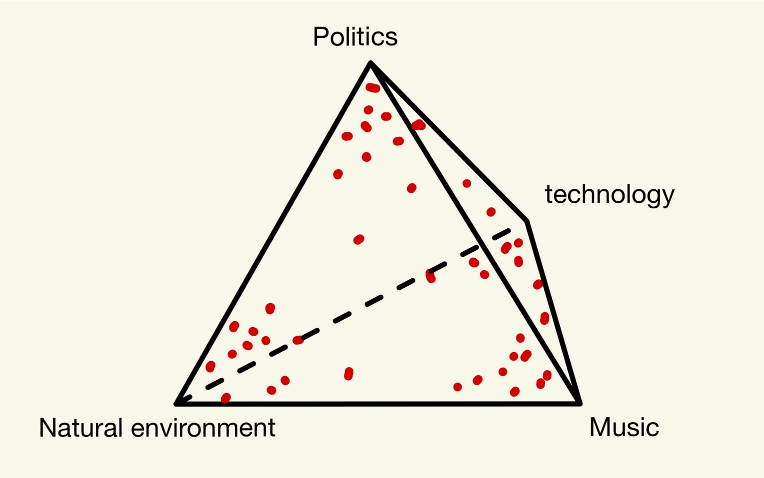
- Give the topic name created by LDA (ex: economy, technology, humanities...)
- Make the conclusion

Example:Aftre LDA



RED:Music,PURPLE:Politics,GREEN:Natural environment,BLUE:technology

Example:Aftre LDA



Data Analytic Demo

Natural Language Processing (1)

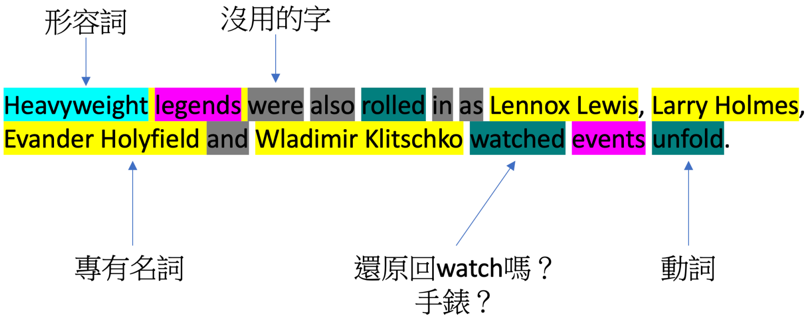


圖 1: 英文句子拆分

Natural Language Processing (2)

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.

Natural Language Processing (3)

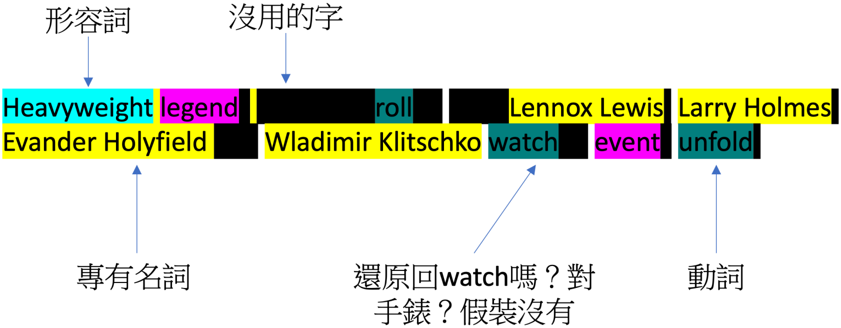


圖 2: Processing Result

LDA: Number of Topics Selection

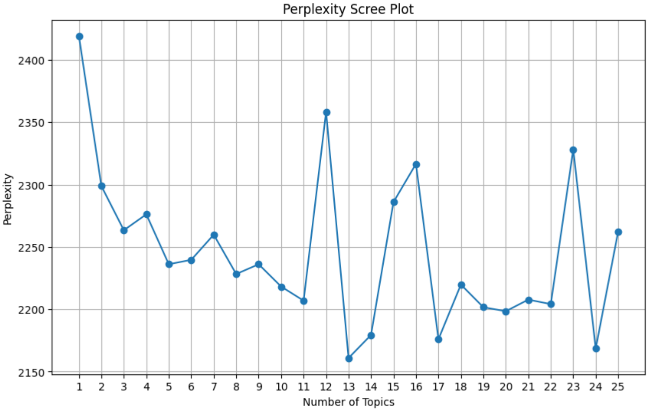


圖 3: Perplexity Scree Plot

LDA: Will cross validation work?

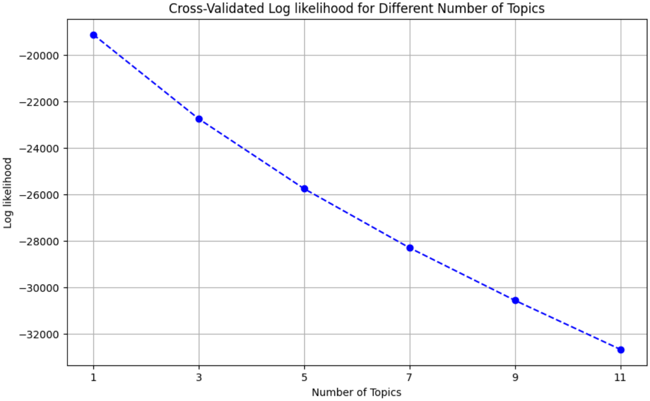


图 4: CV log-likelihood

LDA: T-SNE visualization

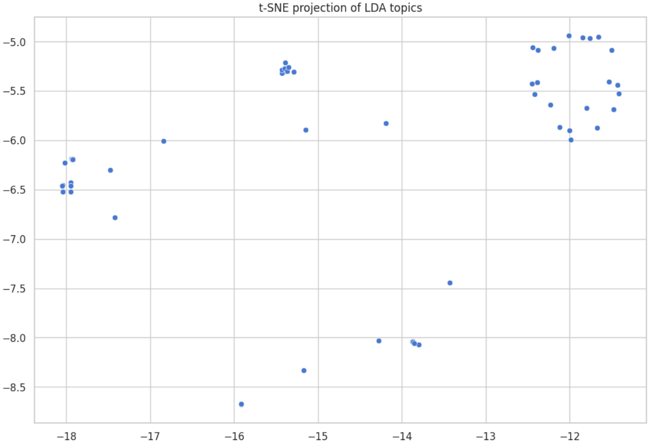


圖 5: T-SNE plot

LDA: Topic Explanation

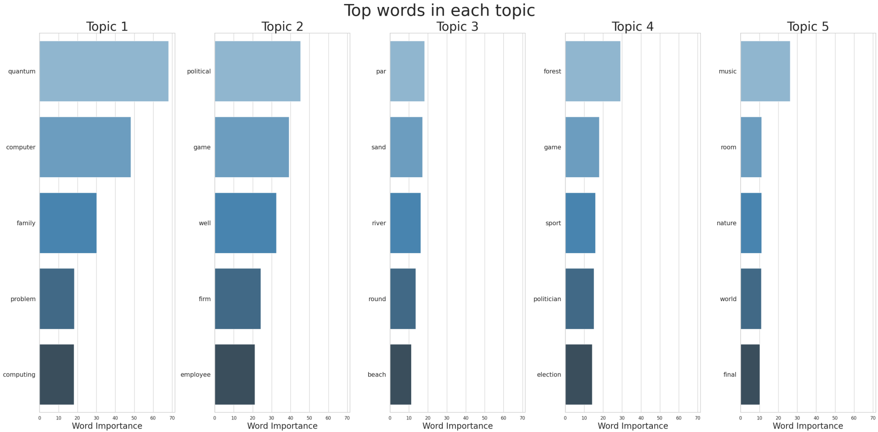


圖 6: Topic

Topic 1: Tech

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. ...

Take Home Challenge

Take Home Challenge

Prove :

$$\int \sum_{\mathbf{z}} q(\mathbf{z}, \theta) ln \frac{q(\mathbf{z}, \theta)}{p(\mathbf{z}, \theta | \mathbf{w}, \alpha, \beta)} d\theta \geq 0$$

(Hint:KL divergnece ≥ 0)

Data Analytic: Job Topic Modeling

GitHub Link

1. Topic modeling using job_description (Data sources: 104, 1111, meetjob, etc., sampled around 80,000 job postings)
2. Compare the effectiveness of LDA and the topic model in the csv file (combined with BerTopic and GPT3.5-turbo for topic classification)
3. Since job_description is in Chinese text and word extraction is complex, NER extracted skill_tag can be used for topic modeling

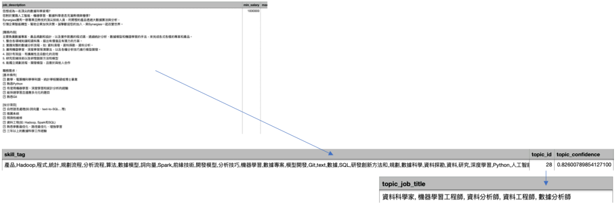


圖 7: Data Table

Reference

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

The End