Topic Model and Visualization of Chinese Ancient Poems

Yanyi Lu (University of Nottingham, Ningbo, China)

Introduction

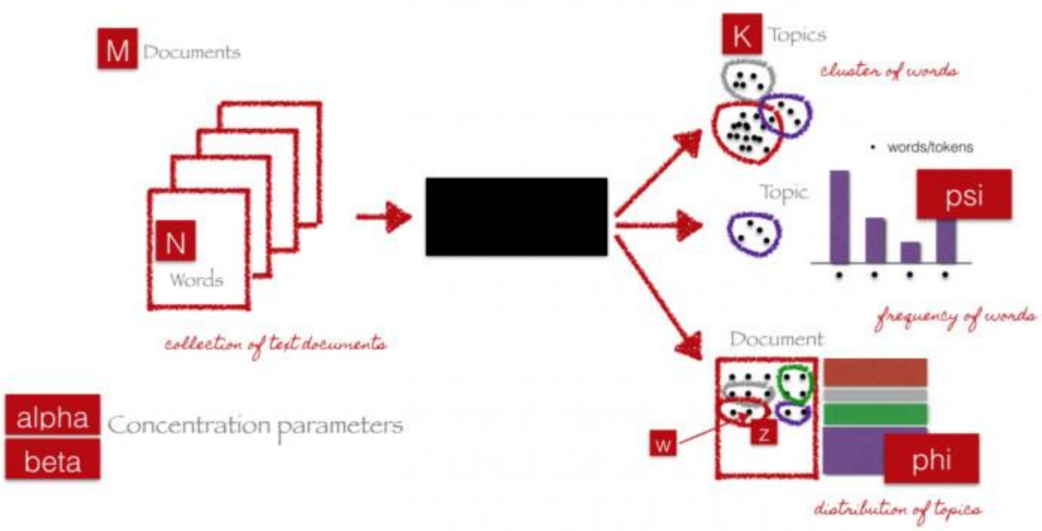
In the age of information, large amount of data is generated every second, which means a tool helping to organize and offer insights to understand large collections of unstructured text bodies, is needed. Developed as a text-mining tool, a method named topic model has been used to detect instructive structures in data such as genetic information, images, and networks.

In this project, the texts used are Chinese ancient poems in Tang dynasty, over 40000 poems. Generally, the gist and emotion of literary works interpreted by readers are subjective, to some extent. Just as a famous saying goes, *there are a thousand Hamlets in a thousand people's eyes.* However, is it possible to explore out the deadly right answer for Hamlets using some quantitative methods, for example topic models? This is one of the interesting points to explore in this project.

This report focuses on two major objectives. One of the objectives is using the topic model (LDA) to find and define the topics of Chinese ancient poems. Another objective is exploring some methods of visualization to make the topics or information extracted from great texts more readable and valuable. The whole report consists of theory, procedure, results and analysis, discussion, conclusion, reference and website of Github.

Theory

In machine learning and natural language processing, topic modeling is a type of statistical model for discovering the abstract ‘topics’ that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovering hidden semantic structures in a text body. Given that a document is about a specific topic, one would expect special words to appear in the document more or less frequently. Just as an example from wiki, "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.



One of the method for realizing topic model is Latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. Based on variational methods and an EM algorithm for empirical Bayes parameter estimation, efficient approximate inference techniques are presented. (Blei, Ng & Jordan, 2003).

Procedure

Preparation of texts

The texts chosen to do the topic modeling are Chinese ancient poems from Tang Dynasty, which are downloaded from Github (https://github.com/todototry/AncientChinesePoemsDB). In the poem file (all-TANG-poems from zhengzhou uni), every single poem has already been disposed to the text without title or author but only body of the poem, and stored in the txt file individually. The cleanness of raw texts facilitates the later operations, to some extent.

Preprocess of texts

1. **Read** all poem texts and store them in a list. During this procedure, the blank texts will be deleted automatically.

2. **Tokenization** by 'jieba', a special module for Chinese tokenization.

3. Because of the specificity of poems, the text of **stopword** is created manually through three-time cleaning.

   a) **First cleaning.** Create a stopwords.txt only with punctuation mark, then sort the frequency of every vocabulary by the function of dictionary. Observe the vocabulary with the frequency over 200, and select the totally useless words to add into the stopwords text. The criteria of selection are that these words reflect neither the language style of ancient poem nor the topic and content of poems.

   b) **Second cleaning.** Sort the frequency of vocabulary by the function of dictionary again using new stopword text. Observing the vocabulary with the frequency over 200, some of the linguistic styles peculiar to poetry will be found.

   c) **Third cleaning.** Select these half-useless words adding into the stopwords and now the whole cleaning process is finished finally. The criteria defined for non-stopword are retaining the vocabularies which can imply concreate image. Therefore, mostly useful vocabularies are noun and double-character words.

Visualization

Three-type wordcloud and interactive map of topic model (see details in next part 'Result and Analysis').

Results and Analysis:

Linguistic features of poems

From half-useful words selected after first cleaning, some linguistic features about Chinese ancient poem in Tang Dynasty are extracted.

1. The most popular vocabulary used in poem is 'moon'(月), which is mentioned 4739 times totally among 42153 poems. The proportion is about 11.24%, which means one of ten poems involves the moon on average, if considering no repeated words in most four-line poems. It is true in fact that most poets are willing to express their feelings and emotions implicitly through the moon, and moon exactly can be attached on various tags, including homesick, tranquility, loneliness and so on.

2. As for some highly ranking words, the functions of them in poems are modal particle, syntactic expletive and auxiliary word, which are rarely or almost no longer used in modern Chinese, but used in extremely high frequency in ancient poems. To some extent, these words with unique features can help to distinct the language of old dynasty and contemporary. Part of these words are still used in written Chinese by some writers, which makes the style of articles more antique. However, an increasing number of these words are disappearing gradually.

3. Interesting phenomena of 'the most' which is ignored or hard to notice without text mining.

    a) Most popular verb: '归' (return, back home)

    b) Most popular place: '洞庭' (one of the famous lakes in China)

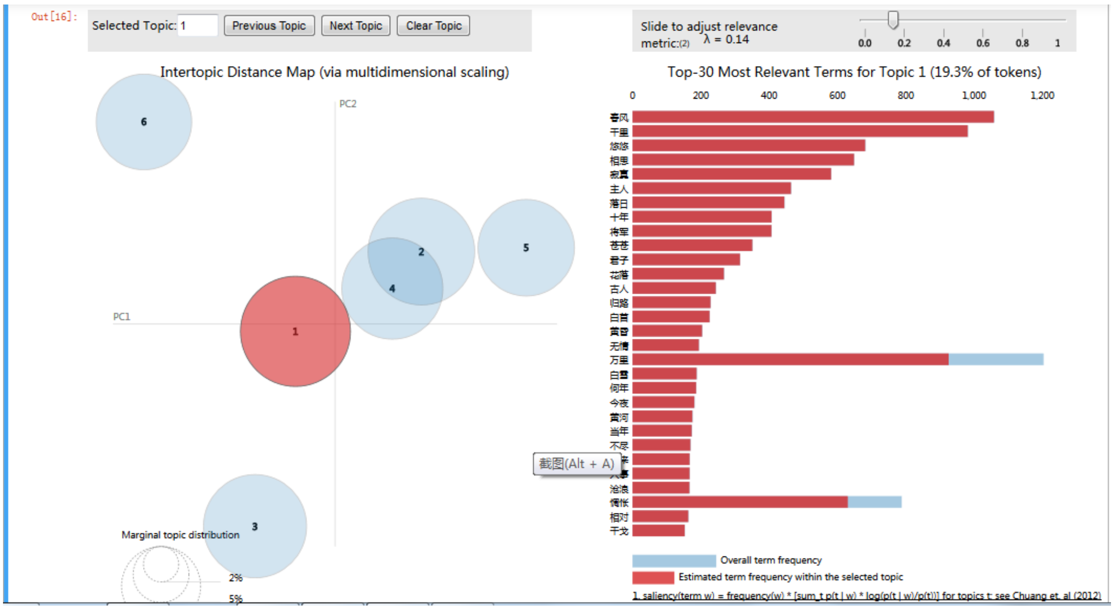
    c) Most popular number: '一' (one)
    second popular number: 万里(thousands of miles)

    for c), this interesting phenomenon exactly verifies one of the ancient Chinese syntaxes. If the amount of object needs to be described in poems, it is often described as number ‘one’ rather than a specific number, three or four. Similarly, if it is a vast number, it will be called as ‘thousands’ directly rather than two thousand or five thousand. One of the possible explanations about the usage of these two inaccurate number is to make the ancient poems more concise and present the veiled beauty of poems.

Analysis of topics

Meaning of each topic

With the mouse hovering over the bubbles on the left, the right panel will display the words related to the topic accordingly, and the meaning of the topic can be summarized by glancing over these words.


Out[16]:
Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.14
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
6
2
5
4
1
3
Marginal topic distribution
2%
5%
Top-30 Most Relevant Terms for Topic 1 (19.3% of tokens)
0
200
400
600
800
1,000
1,200
春风
千里
悠悠
相思
寂寞
主人
落日
十年
将军
苍苍
君子
花落
古人
归路
白首
黄昏
无情
万里
白雪
何年
今夜
黄河
当年
不尽
沧浪
惆怅
相对
干戈
截图(Alt + A)
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

(example: interactive visualization)

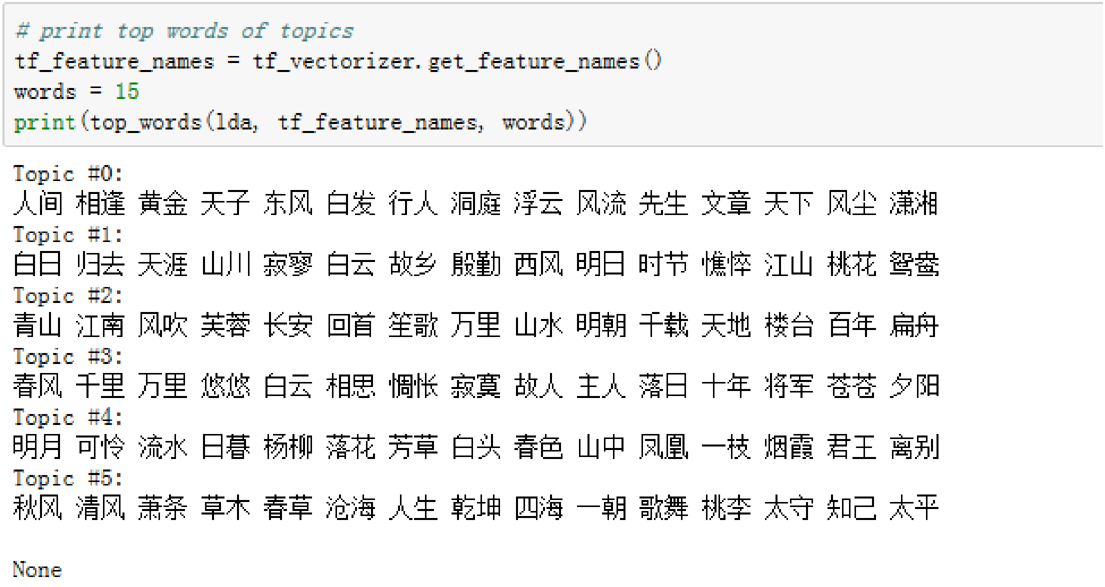
However, among these words listed at right, which one is more relevant to the topic and what is the criteria used for finding the relevant words? According to the algorithm proposed by the authors of pyLDAvis, Sievert and Sirley (2014), the relevance between a topic and a word is decided by a parameter, lambda (λ).

$$\textbf{Relevance (term w | topic t)} = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w)$$

If λ is close to 1, the word appearing more frequently in the topic is more relevant to the topic; If λ is closer to 0, more special and exclusive of the word, more relevant to the topic. Therefore, the value of λ can be adjusted practically by users for actual aims.

In this project, due to the expectations for topic finding are extensive classification and, to some extent no-overlapping, the number determined ultimately for topics is six after some attempts and experiments.

From observation of two extreme situations, the differences between λ= 0 and λ= 1 is not very determining in final topic words. However, it is still more likely to choose λ= 0, which will make the topic classification of poems more specific and distinguishing. As for the key words chosen to define the topics of poem, they shall not be selected simply from top words in the ranking list of each topic, but defined manually through adjusting λ to find the optimal solution aiming to each specific area.

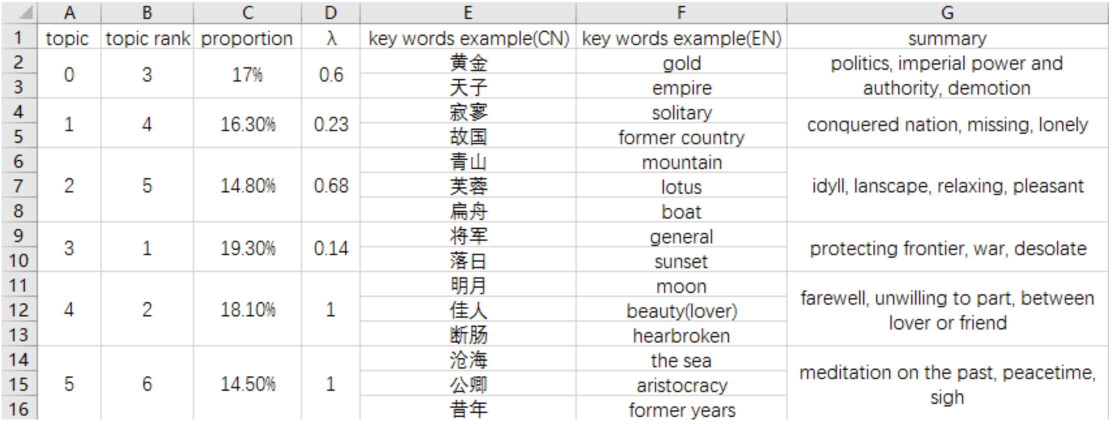


```
# print top words of topics
tf_feature_names = tf_vectorizer.get_feature_names()
words = 15
print(top_words(lda, tf_feature_names, words))
```

```
Topic #0:
人间 相逢 黄金 天子 东风 白发 行人 洞庭 浮云 风流 先生 文章 天下 风尘 潇湘
Topic #1:
白日 归去 天涯 山川 寂寥 白云 故乡 殷勤 西风 明日 时节 憔悴 江山 桃花 鸳鸯
Topic #2:
青山 江南 风吹 芙蓉 长安 回首 笙歌 万里 山水 明朝 千载 天地 楼台 百年 扁舟
Topic #3:
春风 千里 万里 悠悠 白云 相思 惆怅 寂寞 故人 主人 落日 十年 将军 苍苍 夕阳
Topic #4:
明月 可怜 流水 日暮 杨柳 落花 芳草 白头 春色 山中 凤凰 一枝 烟霞 君王 离别
Topic #5:
秋风 清风 萧条 草木 春草 沧海 人生 乾坤 四海 一朝 歌舞 桃李 太守 知己 太平

None
```

(topics selected if all default λ = 1, not very satisfied)



	A	B	C	D	E	F	G
1	topic	topic rank	proportion	λ	key words example(CN)	key words example(EN)	summary
2	0	3	17%	0.6	黄金	gold	politics, imperial power and authority, demotion
3					天子	empire	
4	1	4	16.30%	0.23	寂寥	solitary	conquered nation, missing, lonely
5					故国	former country	
6	2	5	14.80%	0.68	青山	mountain	idyll, lanscape, relaxing, pleasant
7					芙蓉	lotus	
8					扁舟	boat	
9	3	1	19.30%	0.14	将军	general	protecting frontier, war, desolate
10					落日	sunset	
11	4	2	18.10%	1	明月	moon	farewell, unwilling to part, between lover or friend
12					佳人	beauty(lover)	
13					断肠	hearbroken	
14	5	6	14.50%	1	沧海	the sea	meditation on the past, peacetime, sigh
15					公卿	aristocracy	
16					昔年	former years	

(topics selected manually through adjusting λ to find optimal results, more satisfied)

Meaning of frequency

Frequency of topics: The size of the bubble represents the number of poems in this topic, which also stands for the frequency of this topic. At the situation of six topics, distribution is basically on average.

Frequency of words: the proportions of same word in different topics are different, which means one word can represent various meanings or imply diverse emotions in different literacy environment; which also means the roles played by these words (also called contribution degrees) on different topics are very diverse, high or low sometimes.

Relevance between topics

In pyLDAvis, Intertopic Distance Map via multidimensional scaling is used for visualizing the relevance between each topic. The main component is extracted to make the dimensions and the topics are distributed to these two dimensions. Therefore, the distance between topics in the map presents the level of proximity or relevance between topics.

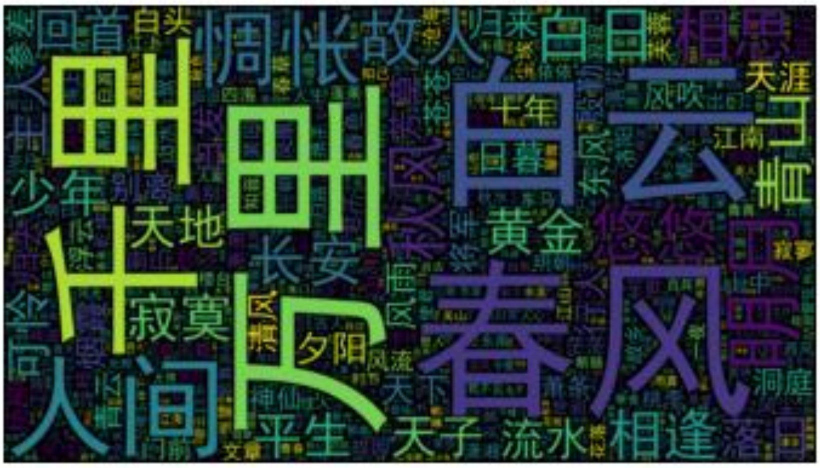
Topic 1,3,4 lie nearly by each other and topic 1,4 are overlapping in some areas, the probable reason for which is that all of these three are negative emotion in domination.

The aggregation degree of topic 2 ranks only second to the other three topics. However, the emotion is positive in topic 2, which is totally opposite to other three topics. One of the explanations is possibly that poets repose their implicit emotion on realistic objects but one visual object sometimes can present various meanings or diverse emotions in different environment. Therefore, choosing partly overlapped objects to express opposite emotion in poems maybe indirectly lead to the phenomenon of aggregation in distance map.

The distribution of Topic 0 and 5 is far away from the main distribution areas, which is probably because that the theme of these two topics are special, so the keywords are unique and not easy for overlapping.

Optimized wordcloud

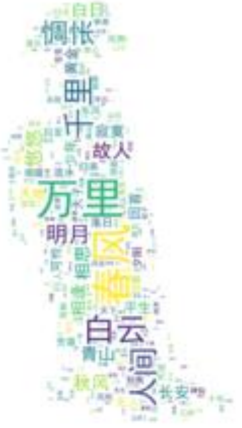
Type 1: original wordcloud



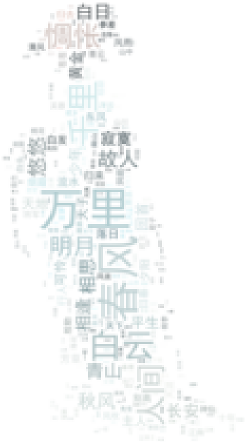
Type 2: wordcloud with customized shape. Because of the texts used are about Chinese ancient poems, the mask chosen, a poet, is related to this theme too.

Type 3: the difference between type 2 and type is the color of words in wordcloud. In the case of third type, the color of words keeps the same with the mask. Therefore, the function of mask is not only providing shape to the wordcloud but also color. Moreover, it is better to choose strongly colorful pictures as mask rather than close colors.

(mask)



(type 2)



(type 3)

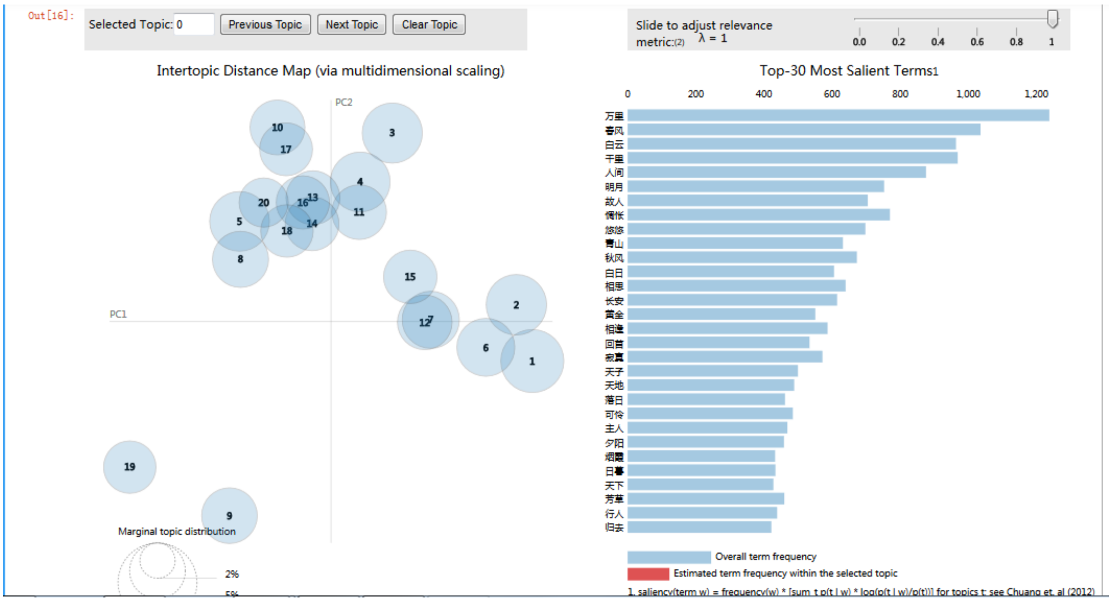
Discussion

Perplexity

How to decide the optimal number of topics automatically?

In information theory, perplexity is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the sample. In natural language processing, perplexity is a way of evaluating language models. A language model is a probability distribution over entire sentences or texts.

LDA perplexity can be roughly interpreted as the degree of uncertain for the topic belonging of an article. In most cases, more topics means smaller perplexity, but the overfitting of topics is hard to avoid. As for this project, the number of texts used for topic model is over 40000, which means the optimal number of topics is possibly huge. However, whether the vast number of topics is what the researchers want, is still a problem.


Out [16]:
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
10
17
3
4
20
16
13
5
14
11
18
8
15
2
12
7
6
1
19
9
Marginal topic distribution
2%
5%
Top-30 Most Salient Terms1
0
200
400
600
800
1,000
1,200
万里
春风
白云
千里
人间
明月
故人
惆怅
悠悠
青山
秋风
白日
相思
长安
黄金
相逢
回首
寂寞
天子
天地
落日
可怜
主人
夕阳
烟霞
日暮
天下
芳草
行人
归去
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

(case of 20 topics)

$$\textbf{Perplexity = exp^\{- (}\sum\textbf{log(p(w))) / (N) \}}$$

P (W) is the probability of each word appearing in the test set.

$$P(W) = \sum z \, P(z \mid d) * P(W \mid z).$$

N = all words appearing in the test set, or the total length of the test set with duplication
z = training topic
d = documents in test set

Hierarchical Dirichlet Process

Sometimes, overfitting is hard to avoid when the substantial number of texts is used to generate topics. In fact, overfitting is necessary in complex topics, which means current topics need structure of layer or branch for more detailed classification. The relationship of these topics shall be hierarchical rather than totally paratactic.

One of methods for automatically determining the number of topics is called 'hierarchical Dirichlet process' (HDP). The Hierarchical Dirichlet Process (HDPs) is a stochastic process that can be used to define a nonparametric distribution on a mixture of mixtures (or admixture) model. That is, each grouping of data is a draw from a mixture model, and the mixture components are shared among the diverse groups. Using a hierarchy of Dirichlet processes allows the number of mixture components to be inferred from the data. HDPs are most commonly used in topic modeling, where the top mixture corresponds to the global set of topics shared among the entire corpus (all documents) and the secondary mixture corresponds to the topic mixture for a given document (Teh, et al., 2005).

Future applications

In addition, based on the topics and relevant keywords extracted from great texts, more interesting and useful applications can be developed. For example,

1. Combined with sentiment analysis to extract the distribution of emotion in topics
2. Taking use of the representative words of different topics to make an automatic poem-generator
3. Based on the taste of users to search automatically and give recommendations of poems in segmented topics
4. For literature and linguistic researchers, compare the topics of two poems in qualitative way easily.

Conclusion

At first part, this report explains the theory of topic model and LDA. Then, this report explains the procedure of whole experiment especially the three-time cleaning and manual creation of stopwords. After these preprocesses, LDA topic model is generated and the results are presented through an interactive module named pyLDAvis, and three-type wordcloud as well. Through analyzing the frequency of words, some linguistic features of poems in ancient China can be found or verified. Moreover, with the interactive topic map, three typical questions about LDA topic model are explained. What is the meaning of each topic, what is implied through the feature of frequency, and what is the relationship between topics? In the final part, the discussions are about optimal number and hierarchical relationship of topics, and future applications of topic model.

Reference

Blei, D., Ng, A., & Jordan, M., 2003. Latent Dirichlet Allocation. [Online] Available at: http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf [Accessed 11 August 2017].

Teh, Y. et al., 2005. Hierarchical Dirichlet Processes. [Online] Available at: https://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf [Accessed 12 August 2017].

Sievert, C., & Shirley, K., 2014. LDAvis: A method for visualizing and interpreting topics. [Online] Available at: https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf [Accessed 13 August 2017].

Github

ID: Yanyi1996

https://github.com/Yanyi1996/topic-model-and-visualization-on-Chinese-ancient-poems-