Coursework 2: Understanding Data

Pornphapat Innwon
31378218
pi1e19@soton.ac.uk

ABSTRACT

This paper describes an approach for understanding the data by using data mining technique. Data is transformed into usable format by using pre-processing method. LDA, NMF and dimension reduction technique are used to presents the visualisation for analysis. The methods for classification in this paper are k-means and hierarchical clustering. The result shows the relationship between data using knowledge from data analysis.

1. INTRODUCTION

The raw data usually collects from the primary source such as sensors and history logs, however, most of these data contain information that is not useful or is difficult to understand. Therefore, data mining is used to process raw data such that people can discover the significant trend or pattern of data, understand how each data relate and then use this knowledge in order to evaluate and solve the problems. One example is fraud card detection.

This paper will describe a method to analyze the data set using the data mining tool. It will present how a data set is extracted and pre-processed using varied tools, shows the feature extraction and then applies appropriate techniques to explore and understand data set.

2. DATA SET

The data set that is used for this paper is a collection of 24 textbooks about Antiquity. These books have been scanned using OCR or Optical Character Recognition to produce text web documents. Each book has been separated into HTML pages and needs to be extracted. For this, the Pandas library is an appropriate tool to convert raw data into a transformable form. Also in this part, The library called BeautifulSoup is used to extract text and remove all HTML tags from the files. These HTML tags are not useful for analyzing the data.

3. PRE-PROCESSING

All of the data are texts that are from scanned textbooks. To analyze this type of data, all text must separate into words or tokens and should be processed before analysis. Each word can be used to identify the relationship between books. Note that, all advanced processes have been done by NLTK or Natural Language Toolkit that is a suitable coding library with these text data. The processes are explained below:

1. `Lower case` Many tools are usually lettered sensitive. Thus, all text is lower case.
2. `Punctuation Removal` Punctuation does not contain useful information for this analysis and must be removed.
3. `Number Removal` Same reason as punctuation. This also includes text number such as "one" and "iii".
4. `Word Tokenization` Texts need to be separated into a list of words. This process must be done for the next process.
5. `Stemming`. This is used to reduce words into the original form. For example, the word such as "ran" or "running" can be reduced to "run". This can make dense data which is great for analysis. Note that, SnowballStemmer from NLTK library is used for this task.
6. `Noise Removal` Some words such as 2 or fewer letters words and stop words have no value and should be removed. Some words such as "hath" and "lib" are also not useful for understanding data.

Note that, the data came from scanning books using OCR. Because of that, there are inevitably some errors such as spelling error and text with no spaces. These problems can reduce performance in analyzing the data.

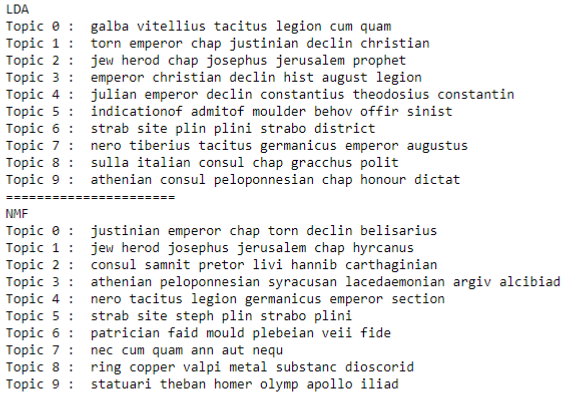
To convert these data into evaluation form, all words should be converted to numeric form. tf(term frequency) and tf-idf(term frequency-inverse document frequency) are fitting feature extraction for text analysis. tf is word count from each document while tf-idf also give importance to uncommon words. Such words can greatly improve text analysis performance. Most of the task use tf-idf to produce results, however, tf is also used in some cases of text analysis that tf-idf is not suitable with.

Note that, these tools are from sklearn library and parameters are adjusted to further improve performance for this context. "max-df" is set to 0.9 to ignore terms that appear more than 90% of all documents and "min-df" is set to 0.1 to ignore terms that appear less than 10% of all documents. These adjustments can also remove noises that occur from OCR. From the result, there are about 33000 features.

4. DATA ANALYSIS

Topic modelling is the method that can present the hidden structure from a set of text documents. There are varied algorithms for this. Two of the most popular one that will be used are LDA and NMF from sklearn library.[1]

LDA or Latent Dirichlet Allocation is a probabilistic model and NMF or Non-Negative Matrix Factorization is a model that comes from linear algebra. LDA use tf features as it is a fitting tool which based on term count while NMF uses tf-idf features.[2]



```
LDA
Topic 0 :  galba vitellius tacitus legion cum quam
Topic 1 :  torn emperor chap justinian declin christian
Topic 2 :  jew herod chap josephus jerusalem prophet
Topic 3 :  emperor christian declin hist august legion
Topic 4 :  julian emperor declin constantius theodosius constantin
Topic 5 :  indicationof admitof moulder behov offir sinist
Topic 6 :  strab site plin plini strabo district
Topic 7 :  nero tiberius tacitus germanicus emperor augustus
Topic 8 :  sulla italian consul chap gracchus polit
Topic 9 :  athenian consul peloponnesian chap honour dictat
======================
NMF
Topic 0 :  justinian emperor chap torn declin belisarius
Topic 1 :  jew herod josephus jerusalem chap hyrcanus
Topic 2 :  consul samnit pretor livi hannib carthaginian
Topic 3 :  athenian peloponnesian syracusan lacedaemonian argiv alcibiad
Topic 4 :  nero tacitus legion germanicus emperor section
Topic 5 :  strab site steph plin strabo plini
Topic 6 :  patrician faid mould plebeian veii fide
Topic 7 :  nec cum quam ann aut nequ
Topic 8 :  ring copper valpi metal substanc dioscorid
Topic 9 :  statuari theban homer olymp apollo iliad
```

Figure 1. LDA and NMF for topic modeling

From figure 2, it is shown that there are some readable topics from each algorithm. In overall, most of the topics are about ancient European history as there are famous names like Justinian and Josephus. For example, topic 6 is about society class. It can be seen that NMF is better than LDA for this task. In NMF, terms in each topic are more nearly related, unlike LDA. For instance, there are terms about late Roman empire in topic 0, Jew history in topic 1 and Greek history in topic 3.

For data visualisation, SVD or singular value decomposition is a tool to build illustrate model using tf-idf as the parameter. In this context, SVD is great for finding similarities between data. Note that, the method also comes from sklearn library.

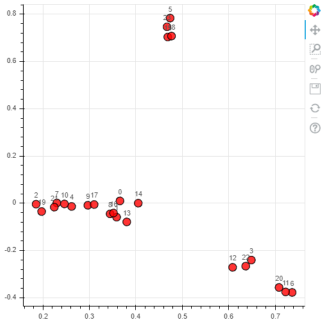


Figure 2. Data visualisation

From figure 2, there is a very clear separation between three groups: a small top group, a large left-bottom group and a medium right-bottom group. For example, the top group contains document 5, 15, 18 and 23.

In addition, LDA can apply with multidimensional scaling to visualise the topics inside the data. It can detect what topic dominate in all documents, terms are related to the topic and the relationship between topics. The result is close to the topics in figure 1 as several dominating topics are about a period in European history.

5. CLASSFICATION

As the task is unsupervised learning, k-means and hierarchical clustering from sklearn library are appropriate clustering tools. For k-means, the silhouette score is used to find the most suitable number of clustering. Note that, this score use cosine similarity which is suitable for this context. In addition, this measure also uses in hierarchical clustering. With tf-idf features, the result is shown below.[3]

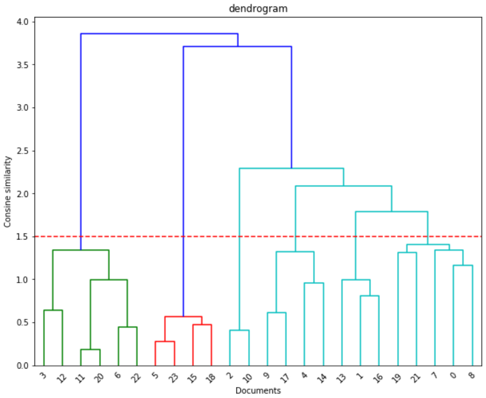


Figure 3. Hierarchical Clustering

From the result, hierarchical clustering produces six groups while k-means produces eight. The notable group from blue documents are document 2 and 10 as it also has very low cosine similarity distance. When comparing to figure 2 with figure 3, the members of top group are exactly the same as members of red group and the members of bottom-right group are also exactly the same as members of green group.

On the performance wise, the hierarchical clustering is better than k-means. This may be because the noises and outlier sensitive property of k-means. For example, there are some documents are not grouped with any other documents despite the similarity distance.

By looking into topic group terms from k-means, it presents some key terms that made some documents group together. According to figure 3, red documents is related to Jew history and green documents is about Roman history. Even though blue documents are sort of ambiguous because of varied key terms, the noted document 2 and 10 are related to Greek history. In assumption, this can means that this group is not directly related to history but can be some topics like government, resources and social hierarchy during ancient time. The reason is because there are related topic terms about these things.

5. CONCLUSION

The result shows the visible distinction between documents. The pre-processing approach can extract useful information. The topic modelling technique give insights about document groups from the clustering algorithm and can use to determine the difference between each group.

6. REFERENCES

[1]Complete Guide to Topic Modeling - NLP-FOR-HACKERS: 2020. https://nlpforhackers.io/topic-modeling/.

[2]Topic Modeling with LDA and NMF on the ABC News Headlines dataset: 2020. https://medium.com/ml2vec/topic-modeling-is-an-unsupervised-learning-approach-to-clustering-documents-to-discover-topics-fdfbf30e27df.

[3]Document Clustering with Python: 2020. http://brandonrose.org/clustering#Multidimensional-scaling.