Name: Kimanh Pham (38936)
Phuong Lieu (38636)

Assignment 3: Text Mining

Q1a:

After tokenizing the text, visually, we can determine that terms that occur only once are 80. The most common term is “the” with 34 times.

Q1b:

The most frequent words in the document are nouns, for example “cash, acceptance, transaction, data, system, management, secure, containers, constituted, terminal” besides the numbers and other non-descriptive words such as “the, and, to, a, are, with, or, each, by, of, in”. Those terms are mixed in terms of frequency; however, they partly can describe the content of the document which is possibly about finance management system or logistics management.

Q1c:

“Text data is sparse” means the small amounts of text presented but spread all over the document.

Usually, the frequent terms are non-descriptive while less frequent terms are more descriptive, which are so sparse all over the document. Therefore, there is challenging to understand fully and precisely the content based on the small present in small amounts of these descriptive terms and explain over a whole content of the document. So, it makes it really hard to find a similarity between two documents based on words counts.

Q2:

Using bigrams we can see that the most frequent texts express more meaning than just using unigrams. These words “cash acceptance”, “acceptance terminal”, “the cash”, “secure container”, “terminal 112” explain more about what is the contain of the document. It could be related to a contract or terms in logistics or transportation. Moreover, it helps us eliminate non-meaning words, such as articles or conjunction words. However, results in biagrams are less frequent than one in unigrams. Therefore, we can miss some other key words which also represent the similarity between documents. To get the gist or measuring the similarity of two documents, looking at both bigrams and unigrams could give us the most precise result than just using one.

Q3a:

These key words are well descriptive and related to the text of the documents. Basically, the method can clarify what are the most relevant terms and also frequently occur in the documents of analyzed corpus regardless the most frequent term such as “the”. In detail, let’s take 10 out of top 20 key words shown in the table, they are very similar to the topic of the text. In other words, these key words can represent the topic of the document and useful for information retrieval.

Hence, this method works quite well in terms of keyword extraction method and raw word level. However, TF-IDF suffers from the problem of data sparsity. This method depends on whether you want to put emphasis on the TF or rather on the IDF part it might be different for the ranking behavior in your use case and apply adjustments to it until you are happy with the result. In this case, the weight is more emphasized on IF, which focus on local of the text.

Q3b:

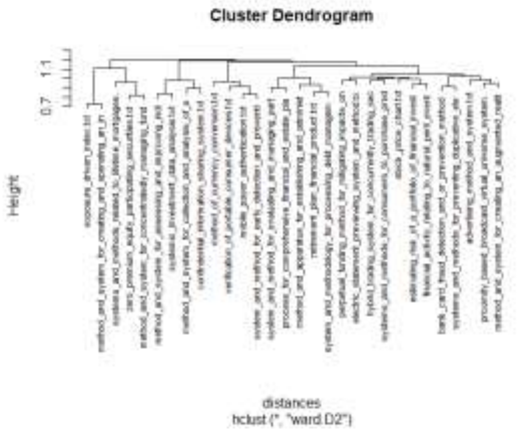
There are several alternative text preprocessing techniques. We think that choosing the stemming technique which is used to find the root/stem of a word is possible to improve the effectiveness of information retrieval. Because it matches similar words with the same roots that leads to reduce index size of the text. In overall, it shows more keywords and help to understand the topic easier.

Besides, Stopwords removal technique is used to eliminate meaningless words in text mining (e.g. the, a, and…) which also help to reduce index size. Because stop words usually account 20-30% of total word counts. However, comparing to TF-IDF, this technique is not very useful.

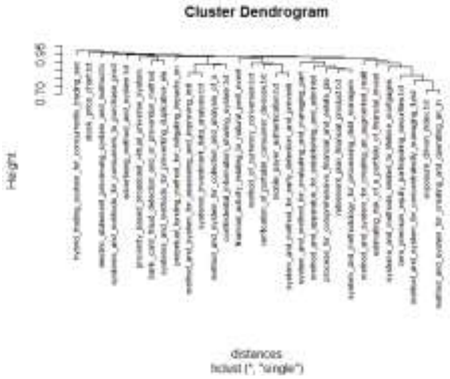
Q4:

We will measure similarity between 30 first documents given using hierarchical clustering in three different methods. They can be visualized as a dendrogram.

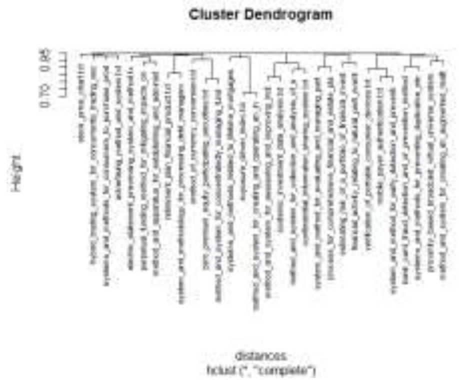
- With Ward method: used to find compact clusters



- Single linkage: adopt a friends of friends, the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors)



- Complete linkage: find similar clusters, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (furthest neighbors)



There are several main differences in global structure between the above dendograms using three different methods, including:

- The single linkage plot appears to be more hierarchical, more flat. The resulting clusters tend to represent long chains. This method seems to cover topics in one stem, thus, it is more difficult to interpret the result.
- The complete method seems to give nearly the same result as the single method. The plot shows a clearer hierarchy
- The ward method appears to be the most efficient method, in which its dendograms is clear, cover all the similar topics in different stems by different distances. However, it tends to create clusters of small size.

Q5:

If we run k-means at 20, the clusters will be partitioned into at 20 sections with the smallest cluster included only one document and the largest included three relevant documents. However, it is hard to understand and have an overview about general group topic if there are many specific clusters like that. Therefore, we would like to divide them into only 6 categories, which based on not only the content of the documents but also the relative similarity between them.

- [11],[26]: Trade index

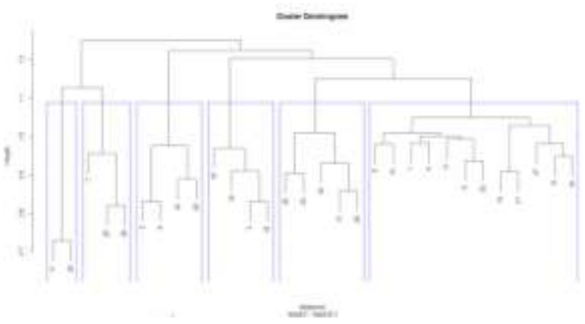
- [7],[25],[28]: Funds and Securities management system

- [2],[6],[19],[24]: Analysis in customers, shareholders and investment data

- [16],[18],[3],[15]: System on mobile device

- [20],[23][30],[13],[29]: Financial Managing

- [9],[12],[1],[4],[17],[5],[22],[10],[21],[27],[8],[14]: Others concerns about financial system (risks, bank card’s frauds, advertising and so on)



Q6:

When we model all documents into 10 topics, we could name the topics based on its top key words as below:

1. Risk management
2. Stock index
3. Payment transaction on device
4. Investment assessment
5. Currency exchange
6. Data analysis
7. Company assets management
8. Payment System
9. User transaction
10. User system

We also can modify the number of topics from 10 to 6 and have less models with more general meanings

1. Assets management
2. Risk management
3. Stock index
4. User payment system
5. Device transaction
6. Payment system

When comparing these two modeling with partitioning clusters method, we can see they are somehow the same. For example, the six partitioning clusters could be grouped into the ten topics we generated.

1. Funds and Securities management system =>Risk management
2. Trade Index => Stock index
3. System on mobile device => Payment transaction on device
4. Analysis in customers, shareholders and investment data => Data analysis
5. Financial Managing => Company assets management
6. Others concerns about financial system (risks, bank card's frauds, advertising and so on) => User system

Moreover, we can extract some new insights. For instance, there are topics about Investment assessment, Payment system, User transaction, Currency exchange that were not shown in the result of 30 documents in partitioning cluster method.

Q7a:

Feature engineering play an important role in discovering knowledge by extracting unstructured data. It is a process of transforming raw data into basic structure/features that represent better the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. Mainly there are modeling for prediction and modeling for explanations that are usually easier to interpret.

Let's take health care data which mainly focuses on diabetes. Doctors, physicians, insurer can predict complicated cases of diabetes based on patterns extracted from feature engineering such as number of prescriptions of an individual regardless examining the healthy status of each prescription. Moreover, demographic information such as patient age can be combined with medical history (number of prescriptions) in order to evaluate a complicated case of diabetes. For example, from extraction, a case that age>36 and medication>6 is defined as a 100% complicated case. Therefore, we can create a pattern to predict case of diabetes in future by having similar information of a patient.

However, we just only can discover the pattern that we provide in the data. It means this feature engineering method just answer our questions. If we ask an inappropriate question or give it an unclear indicator, it will generate less meaningful results. We usually do not know yet what pattern we are looking for from the unknown and big data. Therefore, we need to use the human expertise and specific knowledge in order to ensure that the data contains relevant indicators for the prediction task.

Q7b:

Exploratory analysis method analyzes big data to summarize their main characteristics in accurate predictive models without necessity of extracting causal structure of the data. In some cases, we may not be capable of observing the causal model. If we assume that we have observed all relevant variables, it will cause complication of a case. We still need exploratory analysis to tell us outcomes beyond the causality in order to evaluate the possibility of causation existing. However, social sciences emphasize on theories that rely on causality without serious consideration of predictive power. Predictive models are accurate due to large amount of analyzed data used to test such models and give reliable error limit. Sample estimates become reasonable proxies, which eliminates inappropriate bias on the problem.

In text mining, exploratory analysis helps to explain unknown information in prior, finding currently non-predictive answers. The language in text is ambiguous comparing to data. Using text in exploratory analysis forms hypotheses of the phenomena or topics. Similar to data mining, it is also combined with human expertise to draw a high probability of result.