COMP6237 Data Mining Coursework 2: Understanding Data

Yan li 30844789
University of Southampton
ly1u19@soton.ac.uk

1 INTRODUCTION

In this report, I will explain how to group documents using different machine learning and data mining methods. I plan to found the potential connection between 24 documents. This report includes the following parts:

- Data Extraction
- Text Data Cleaning
- Tokenizing and Stemming
- Feature Extraction Calculating Cosine Similarity
- Dimensionality Reduction
- K-means clustering
- Hierarchical Clustering

2 DATA EXTRACTION

Following the instruction of data mining courses, I download a Zip file containing the HTML pages with the OCR results, there are 24 folders in total, each of them containing a passage. The first task is to extract data from HTML files. Beautifulsoup is a perfect package to do web scraping for separating HTML TAG which I used to obtain the original text content and save it in separate 24 txt files.

3 TEXT DATA CLEANING

I use regular expressions to match all text data, it is better to delete punctuation marks as these are not helpful for text classification. Then I remove words with a length of less than three and delete stop words as these words have no practical meaning. The orc used to extract the image for text data produces wrong results, need to correct the wrong words to get the correct results.

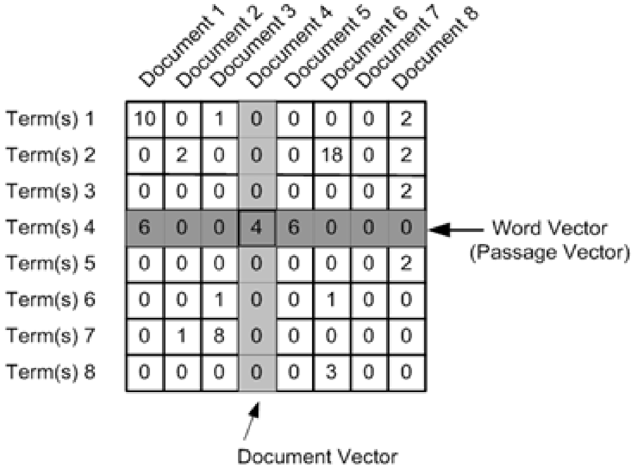
4 TOKENIZING AND STEMMING

Tokenizing is a process of converting continuous data into singular strings to perform text classification. first, tokenizing data and convert it to strings.[1] I chose Stemming instead of Lemmatization because this method is simpler has a similar result. After these two steps, I cleaning the data and correct the wrong words again to get clean data

5 FEATURE EXTRACTION AND CALCULATING COSINE SIMILARITY

TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and data mining. TF is the term frequency (Term Frequency), and IDF is the inverse document frequency index (Inverse Document Frequency).The following figure show the similarity among these texts[2]

I used the scikit-learn package to calculate the features of each article and get 2656 Features,then calculate the cosine similarity.



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Figure 1: TF-IDF

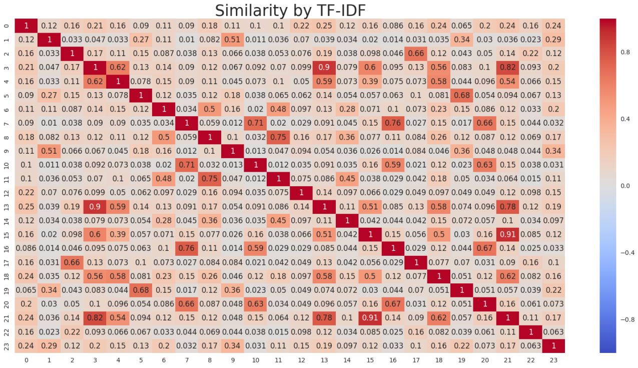


Figure 2: similarity

6 DIMENSIONALITY REDUCTION

In the previous step, 2656 Features were obtained, but we need to extract principal components from so many eigenvalues. In this course, I learned principal component analysis (PCA) and multidimensional scaling (MDS) for dimension reduction.Therefore,I choosed these two methods and compared the cluster results obtained by the two methods in the next part. The following visualization is the results obtained by dimensionality reduction to two dimensions using PCA.

7 K-MEANS CLUSTERING

The k-means clustering algorithm (k-means clustering algorithm) is an iterative solution clustering analysis algorithm. The first step

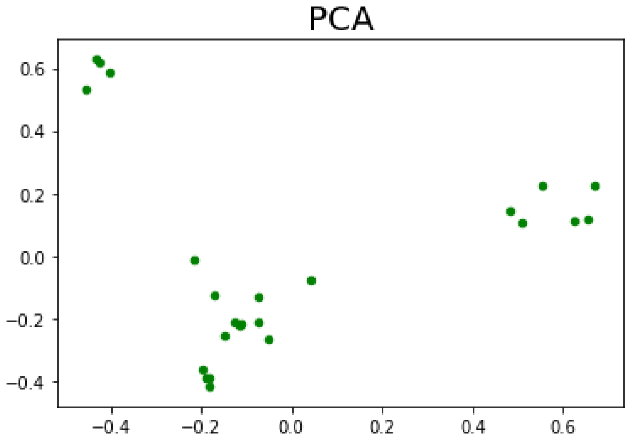


Figure 3: principal component analysis (PCA)

is to pre-divide the data into K groups, then randomly select K objects as the initial clustering center, and then calculate The distance between each object and each seed cluster center, each object is assigned to the nearest cluster center. The cluster centers and the objects assigned to them represent a cluster. Each time a sample is assigned, the clustering center of the cluster will be recalculated based on the existing objects in the cluster. This process will continue to repeat until a termination condition is met. The termination condition may be that no (or minimum number) objects are reassigned to different clusters, no (or minimum number) cluster centers change again, and the sum of squared errors is locally minimum.[3]

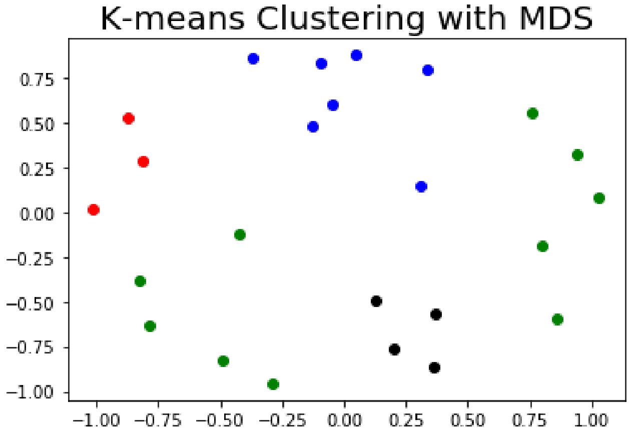


Figure 4: K-means with MDS

By using the features extracted in the previous step and clustering applying the k means to cluster, the following visualization results were obtained. It can be seen that the PCA method can group the text into three categories, and the cluster results obtained by MDS are not very ideal.

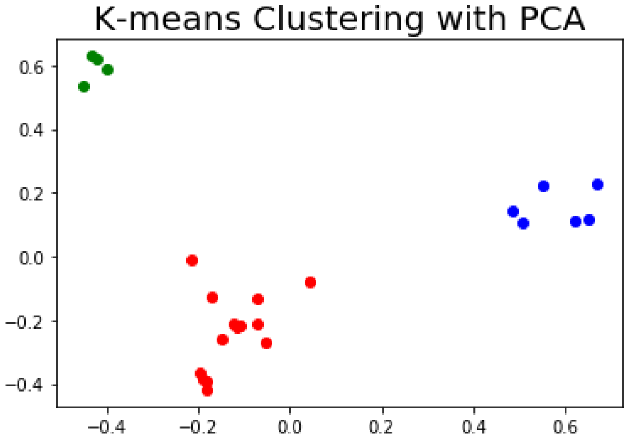


Figure 5: K-means with PCA

8 HIERARCHICAL CLUSTERING

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. [4] The following figure shows the result of clustering, we can see the association between these texts

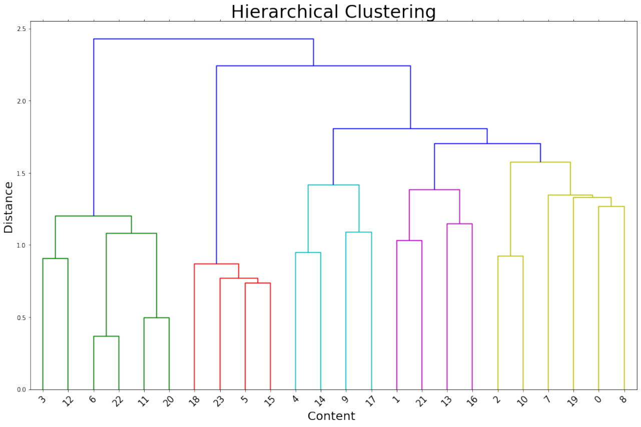


Figure 6: ward$_{c}lusters$

REFERENCES

[1] Nizar Habash, Owen Rambow, and Ryan Roth. Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, volume 41, page 62, 2009.

[2] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ, 2003.

[3] Dan Pelleg and Andrew Moore. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–281, 1999.

[4] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.