## Abstract

This paper provides an overview of a standard descriptive data mining processing pipeline including data collection, feature extraction, clustering and data-analytic visualization. This pipeline was applied to a set of 24 text documents about Antiquity. By mapping all the documents to a multidimensional space, K-Means and Hierarchical Clustering were used to explore the relationships in between documents. These documents were clustered into different number of groups among which clustering them into 7 groups were the most accurate according the title of each book. And then these documents were visualized on a 2-dimensional space through Multidimensional Scaling to present their relationships.

Keyword: data mining, feature extraction, K-Means, Hierarchical Clustering, Multidimensional Scaling.

## Introduction

The 24 documents

## Data Collection

The 24 text documents are 24 books of which the content of each page is stored in HTML page. By analyzing the structure of the HTML source code, all the text was found inside “Span” tag with class attributes as “ocr\_cinfo”. BeautifulSoup, a Python library, was used to pull all the text out of all the HTML files. And data were store as JSON format:

{

“book\_name” : book name,

“text”: text

}

## Feature Extraction

Vector Space Model(VSM) was used in this experiment to represent all the documents as vectors in N-dimensional space. Term Frequency (TF) and Term Frequency- Inverse Document Frequency(TF-IDF) were the two main ways to convert each documents to vectors.

The first step to model documents was to build up a term indexing dictionary by select all distinct terms present in every documents, using term as the key and index as the value denoted as :

E(t) = {T0: 0, … , Ti: i, …, Tn:n}.

There are many stop-words (e.g. the, a, is and in etc.) that present in virtually every documents. But the aim of VSM is to extract important features that distinguish one document from another, thus these words were ignored.

### TF

TF is the simplest way of representing each document into N-dimensional space. It simply counts how many times each term in E(t) present in a certain document di. The function is defined as follows:

where fr(x, t) is defined is:

Therefore, documents are represented as vectors:

Every single term maps to the corresponding dimension, of which the value is the times of the term occurs in the document. And the dimensionality of vectors is the number of vocabularies in the dictionary.

The set of documents were then represented as matrix:

Note: This matrix tends to be very sparse since one term may occur in one document but may not occur in any other documents.

#### Limitation of TF

TF scales up the terms that occurs many times in a document and scales down terms that might be more informative. Frequency of a term in a document might not to accurately reflect the significance of the term. TF-IDF is introduce to solve this problem.

### TF-IDF

### 

TF-IDf measures how important each term is to a document. TF is the frequency count of a term in a document. IDF stands for inverse document frequency which assigns high values to rare terms and low values to common words. Tf-idf(t) is defined as:

where tf(t) is the same as before:

where idf(t) is:

idf weights vector is then calculated as:

In order to easliy calculate Mtf-idf, a diagonal matrix Midf was created:

And then *Mtf-idf* can be calculated as:

Finally, we applied L2-normalization process to *Mtf-idf. (Note, the normalization is applied on row level not matrix level.)*

## Clustering

The 24 text documents had been mapped into a multi-dimensional space and were ready for further processing.

Grouping these documents into different groups based to their textual similarity would be helpful to better understand the documents. Two most common and practical clustering techniques, K-Means Clustering and Hierarchical Clustering were used in this experiment.

Before I applied clustering techniques, I took a simple look at of each document. Roughly these documents can be grouped into 7 groups according to their content.

### K-Means

### Hierarchical Clustering

## Multidimensional