# KEYPHRASE EXTRACTION

# CLUSTERING TO FIND EXEMPLAR TERMS FOR KEYPHRASE EXTRACTION

A
PROJECT REPORT
submitted in partial fulfillment of
the requirements of the subject

**Statistical Methods in Artificial Intelligence**

Submitted by:

| | |
|---|---|
| **Sonakshi Sharma** | **2018201090** |
| **Manik Langer** | **2018201092** |
| **Aman Raj** | **2018201085** |
| **Varun Bhatt** | **2018201086** |

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY

HYDERABAD

# ABSTRACT

Extraction of keyphrases from individual documents is a research area in which one try to extract a small set of keyphrases that describe the content of a single document. The advantages with this form ofextraction is that it retains most of the semantic context from the document.Keywords and keyphrasesprovide an essential way to present the topic of a given document and can help  readers access core information in text. Keyword and keyphrase extraction techniques are typically used in information retrieval tasks.These are very useful in many natural language processing (NLP) applications such as document summarization, classification and clustering. But it is an expensive and time-consuming job for users to tag keyphrases of a document. These needs motivate methods for automatic keyphrase extraction.

This project presents an unsupervised method for keyphrase extraction.

# TABLE OF CONTENTS

# 1. INTRODUCTION

With the information explosion we face nowadays, and large storage space available at reduced costs, scientific information is expanding and the number of scientific articles published is increasing. Scientific articles are recorded in text and text is unstructured data within digital forms, which may contain the underlying information.With such an overwhelming amount of literature(text),there is much interest in techniques to explore underlying information and uncover new knowledge.

In natural language processing a common theme has for many areasbeen the amount of semantic context that can be used. Individual document keyphrase extraction tries to maintain the important semantic context from the document while also identifying those terms or phrases that are describing for the document.

Keyphrases, as a brief summary of a document, provide a solution to help organize, manage and retrieve documents, and are widely used in digital libraries and information retrieval.

From the observation of human-assigned keyphrases, we conclude that good keyphrases of a document should satisfy the following properties:

1. **Understandable :** The keyphrases are understandable to people. This indicates the extracted keyphrases should be grammatical.For example, “machine learning” is a grammatical phrase, but “machine learned” is not.

2. **Relevant :** The keyphrases are semantically relevant with the document theme. For example, for a document about “machine learning”, we want the keyphrases all about this theme.

3. **Good coverage :** The keyphrases should cover the whole document well.
Suppose we have a document describing “Beijing” from various aspects of “location”,“atmosphere” and “culture”, the extracted keyphrases should cover all the three aspects,instead of just a partial subset of them.

Therefore selecting a keyphrase must conserve the above properties.

# 2. PROBLEM STATEMENT

Keyphrases in articles of journals and books are usually assigned by **authors**. However,most articles on the web usually do not have human-assigned keyphrases.

With the scientific information  expanding the number of scientific articles published is  also increasing. Scientific articles are recorded in text. And text is unstructured data within digital forms, which may contain underlying information.

This is particularly the case of **medical domain**.Medical practitioners and clinicians have clinical questions which need to be answered todiagnose patients. To make best decisions, practitioners synthesise all of the important information about the patient, relevant research, and experience with previous patients to determine the best course of action .However, it is still extremely difficult for physicians to obtain the most useful evidence ina large literature.This is motivated by the desire of many important clinical practices which are quicker, more convenient, moreconsistent and more efficient than current practice. Hidden evidence in text can be terms including keywords and keyphrases.Since manually searching terms is time-consuming and expensive,automated term extraction techniques can save time and economy. Therefore, automated extraction and automated knowledge acquisition are necessary and important.

# 3. LITERATURE REVIEW

| ARTICLE | DESCRIPTION | FINDINGS |
|---|---|---|
| • A straight forward method for keyphrase extraction | • This technique select keyphrases according to frequency criteria. | • This is a inefficient method and its poor performance drives people to explore other methods. |
| • Paper on The supervised approach (Turney, 1999) | • Regards keyphrase extraction as a classification task.In this approach, a model is trained to determine whether a candidate term of the document is a keyphrase, based on statistical and linguistic features. | • For the supervised keyphrase extraction approach, a document set with human-assigned keyphrases is required as training set. However,human labelling is time-consuming. |
| • Mihalcea and Tarau, 2004 | • As an example of an unsupervised keyphrase extraction approach, the graph-based ranking (Mihalcea and Tarau, 2004) regards keyphrase extraction as a ranking task, where a document is represented by a term graph based on term relatedness,and then a graph-based ranking algorithm is used to assign importance scores to each term. | • Existing methods usually use term cooccurrences within a specified window size in the given document as an approximation of term relatedness . |

# 4. ALGORITHM OVERVIEW

The method in this project is mainly inspired by the nature of appropriate keyphrases, namely understandable, semantically relevant with the document and high coverage of the whole document.

Under the bag-of-words assumption, each term in the document, except for function words, is used to describe an aspect of the theme. Based on these aspects, terms are grouped into different clusters.

The terms in the same cluster are more relevant with each other than with the ones in other clusters.
Based on above description the clustering-based method for keyphrase extraction has the following steps:

**1. Candidate term selection :** We first filter out the stop words and select candidate terms for keyphrase extraction.

**2. Calculating term relatedness :** We use some measures to calculate the semantic relatedness of candidate terms.

**3. Term clustering :** Based on term relatedness, we group candidate terms into clusters and find the exemplar terms of each cluster.

**4. From exemplar terms to keyphrases :** Finally, we use these exemplar terms to extract keyphrases from the document.

## 4.1. Candidate term selection

A brute-force method might consider all words and/or phrases in a document as candidate keyphrases. However, given computational costs and the fact that not all words and phrases in a document are equally likely to convey its content, heuristics are typically used to identify a smaller subset of better candidates. Common heuristics rules are used for this purpose that include removing stop words and punctuation; filtering for words with certain parts of speech .
This step proceeds as follows:

- Firstly the stop words are removed.
- Secondly the remaining text is tokenized for English or segmented into words for Chinese and other languages without word-separators.

- Thenconsider the remaining single terms as candidates for calculating semantic relatedness and clustering.

In methods like (Turney, 1999; Elbeltagy and Rafea, 2009), candidate keyphrases were first found using n-gram. Instead, in this method, we just find the single-word terms as the candidate terms at the beginning. After identifying the exemplar terms within the candidate terms, we extract multi-word keyphrases using the exemplars.

## 4.2. Calculating term relatedness

An intuitive method for measuring term relatedness is based on term cooccurrence relations within the given document. The cooccurrence relation expresses the cohesion relationships between terms.This method assumes that more important candidates are related to a greater number of other candidates, and that more of those related candidates are also considered important.

In the paper **“Clustering to Find Exemplar Terms for Keyphrase Extraction by Zhiyuan Liu, Peng Li, Yabin Zheng, Maosong Sun”,**that is being implemented in this project uses the technique of **Cooccurrence-based Term relatedness**.But for getting different results we have additionally calculated the term relatedness using **word to vec** technique .

### 4.2.1 Word to Vec

- It is a method for representing text as machine learning does not accept text and must be converted to numbers.
- It maps all the words of a language into a vector space of a given dimension.
- W2V objective function causes the words that occur in similar contexts to have similar embeddings.
- There are two algorithms used for this purpose:

  1) Continuous Bag of words(CBOW):
    - Predicts the target word from the context
    - Use one hot vector for every word and selects a window size for iterating
    - In a window it predicts the center word from the surrounding two words using a simple neural network.
    - The window is then slided and same process is repeated until the weights are obtained.

  2) Skip Gram
    - Choose athe window size and give center word as input trying to predict the context words.

- The weights are updated in similar way as that in CBOW.

To obtain the word vector weight matrix is multiplied by the one hot vector.

## 4.3. Term clustering

Unsupervised machine learning methods attempt to discover the underlying structure of a dataset without the assistance of already-labeled examples i.e “training data”**.**In this paper, we use three widely used clustering algorithms, hierarchical clustering and spectral clustering , to cluster the candidate terms of a given document based on the semantic relatedness between them.

### 4.3.1. Hierarchical clustering

Hierarchical clustering groups data over a variety of scales by creating a cluster tree. The tree is a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The hierarchical clustering follows this procedure:
1. Find the distance or similarity between every pair of data points in the dataset
2. Group the data points into a binary and hierarchical cluster tree
3.Determine where to cut the hierarchical tree into clusters. In hierarchical clustering, we have to specify the cluster number “m” in advance.

### 4.3.2. Spectral clustering

Spectral clustering makes use of the spectrum of the similarity matrix of the data to perform dimensionality reduction for clustering into fewer dimensions, which is simple to implement and often outperforms traditional clustering methods such as k-means.Since the cooccurrence-based term relatedness is usually sparse, the traditional eigen value decomposition in spectral clustering will sometimes get run-time error. In this project, we use the singular value decomposition (SVD) technique for spectral clustering instead.

## 4.4. From exemplar terms to keyphrases

After term clustering, we select the exemplar terms of each clusters as seed terms.Most manually assigned keyphrases turn out to be noun groups.Therefore we use a regular expression to fetch the keyphrase .The pattern can be represented using regular expression as follows:

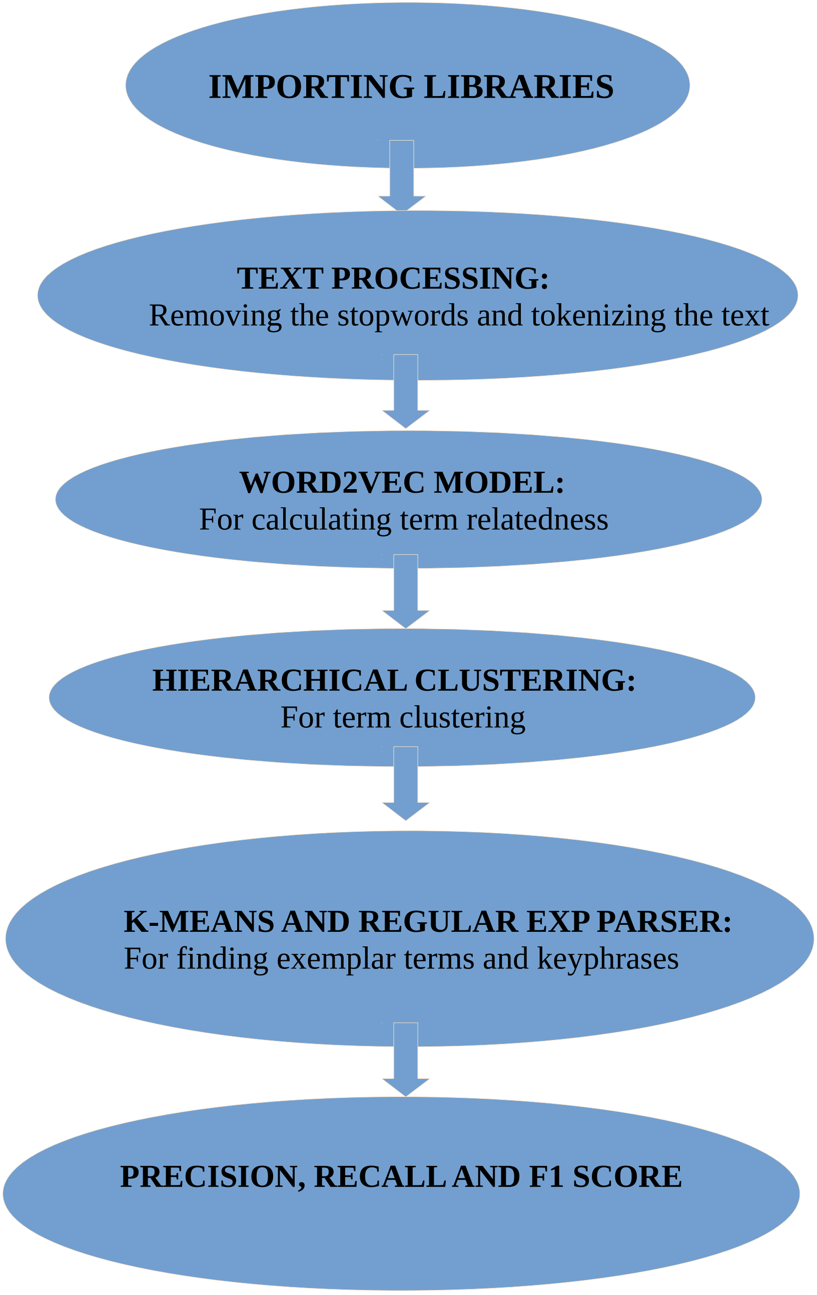
**(JJ)*(NN | NNS | NNP)+**

where JJ indicates adjectives and various forms of nouns are represented using NN, NNS and NNP. From these noun groups, we select the ones that contain one or more exemplar terms to be the keyphrases of the document. In this process, we may find single-word keyphrases. In practice, only a small fraction of keyphrases are single-word.

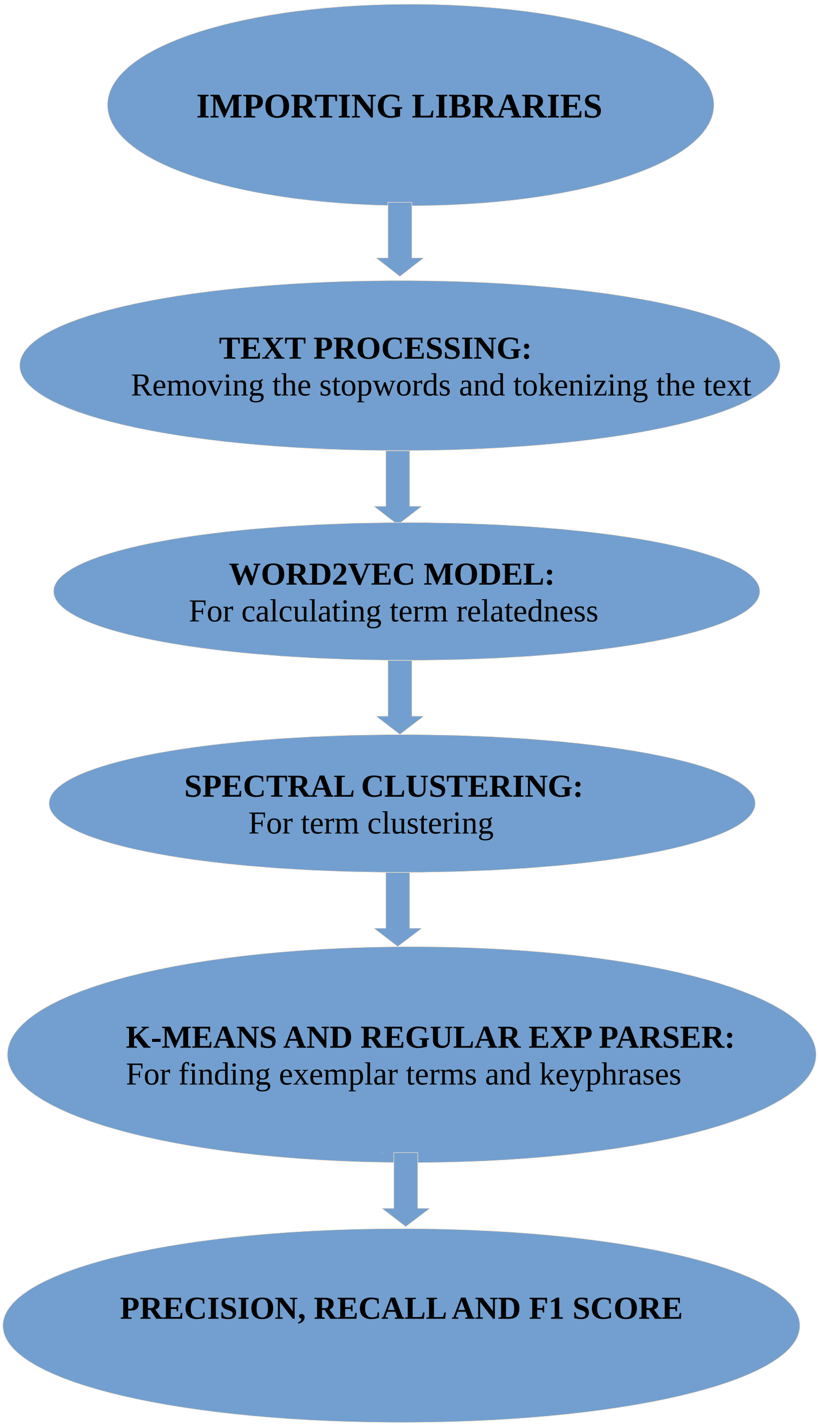
## 5. IMPLEMENTATION

The various models that we implemented are:

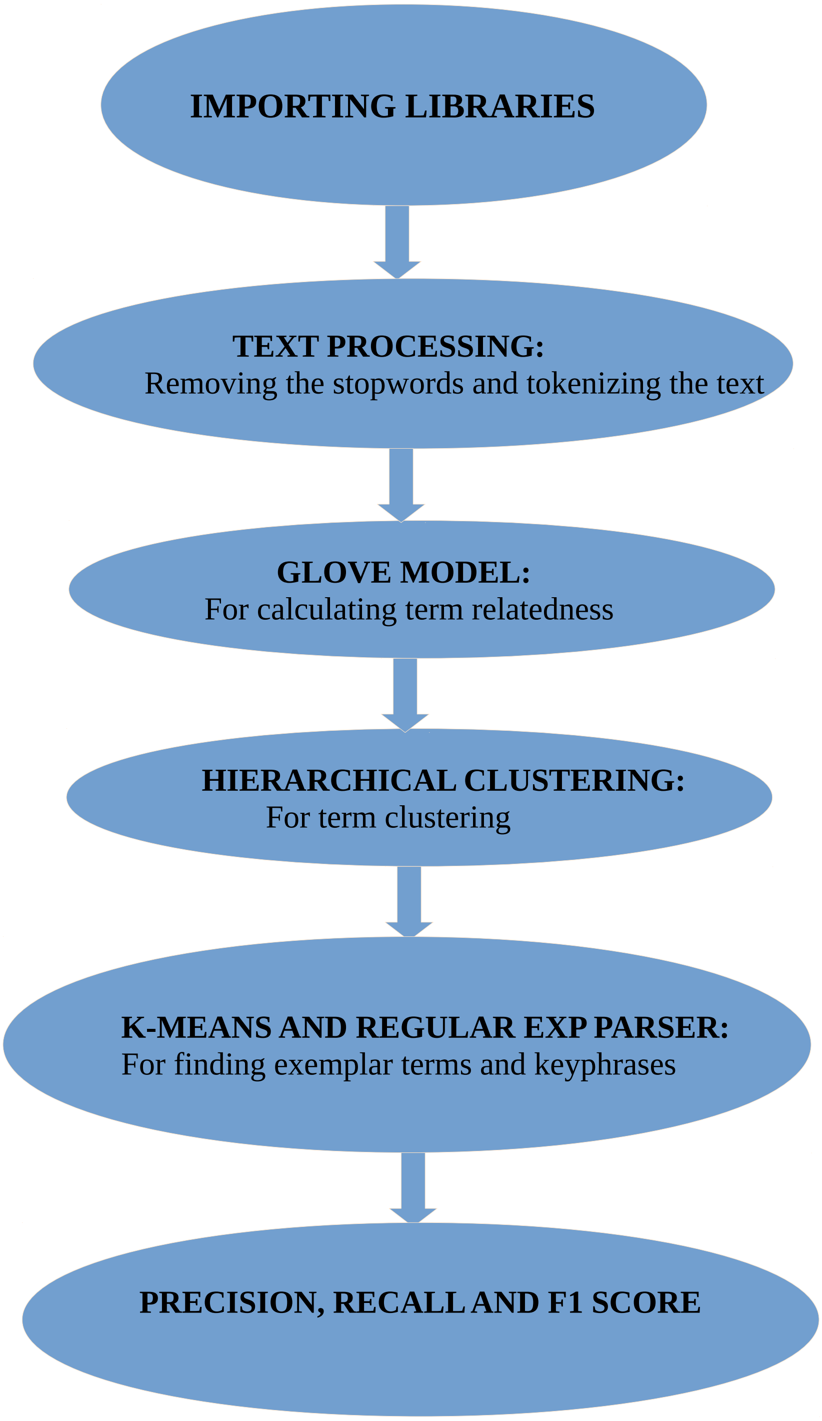
### 1.Word to vec and Hierarchical Clustering


IMPORTING LIBRARIES
TEXT PROCESSING:
Removing the stopwords and tokenizing the text
WORD2VEC MODEL:
For calculating term relatedness
HIERARCHICAL CLUSTERING:
For term clustering
K-MEANS AND REGULAR EXP PARSER:
For finding exemplar terms and keyphrases
PRECISION, RECALL AND F1 SCORE

## 2.Word to vec and Spectral Clustering


IMPORTING LIBRARIES
TEXT PROCESSING:
Removing the stopwords and tokenizing the text
WORD2VEC MODEL:
For calculating term relatedness
SPECTRAL CLUSTERING:
For term clustering
K-MEANS AND REGULAR EXP PARSER:
For finding exemplar terms and keyphrases
PRECISION, RECALL AND F1 SCORE

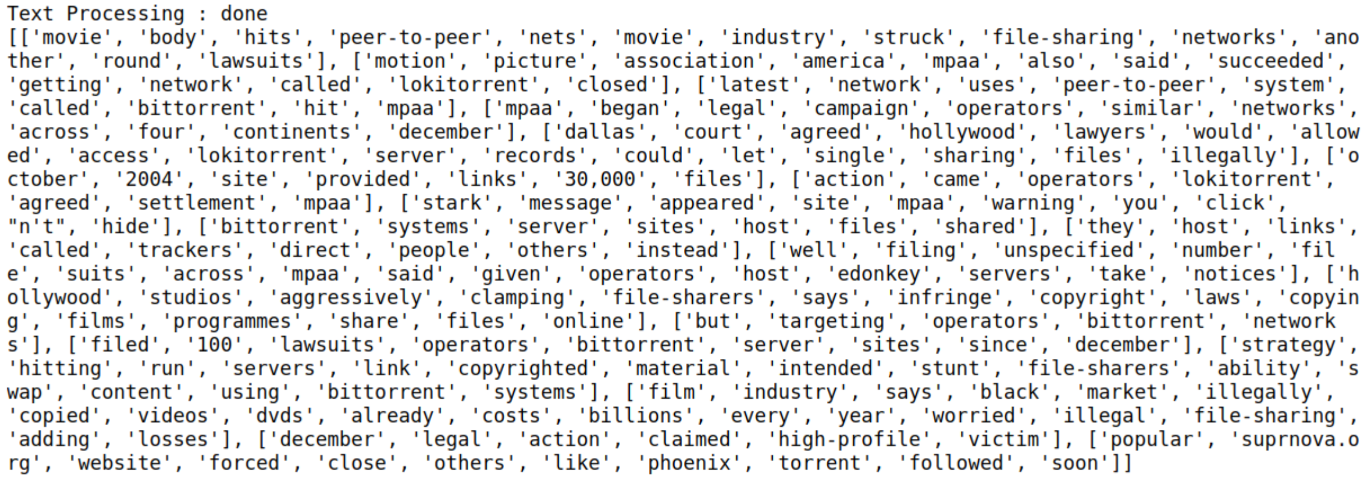
## 3. Cooccurence and Hierarchical Clustering


IMPORTING LIBRARIES
TEXT PROCESSING:
Removing the stopwords and tokenizing the text
GLOVE MODEL:
For calculating term relatedness
HIERARCHICAL CLUSTERING:
For term clustering
K-MEANS AND REGULAR EXP PARSER:
For finding exemplar terms and keyphrases
PRECISION, RECALL AND F1 SCORE

# 6. RESULTS AND OBSERVATIONS

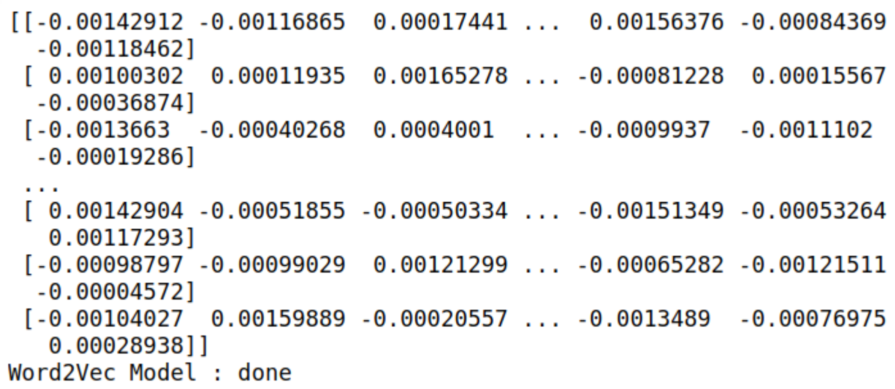
Input : A text file

Step 1: After removing stop words and tokenizing



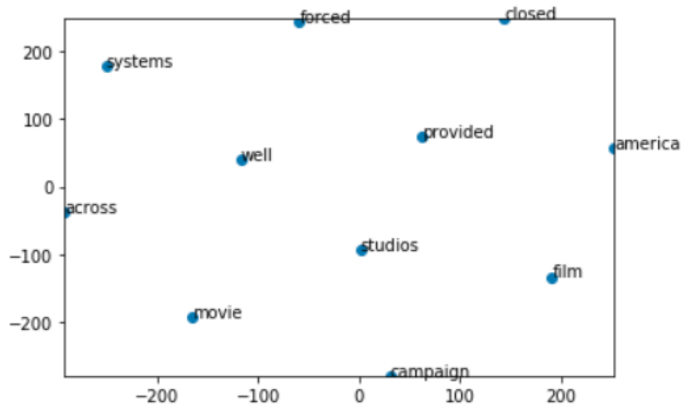
```
Text Processing : done
[['movie', 'body', 'hits', 'peer-to-peer', 'nets', 'movie', 'industry', 'struck', 'file-sharing', 'networks', 'ano
ther', 'round', 'lawsuits'], ['motion', 'picture', 'association', 'america', 'mpaa', 'also', 'said', 'succeeded',
'getting', 'network', 'called', 'lokitorrent', 'closed'], ['latest', 'network', 'uses', 'peer-to-peer', 'system',
'called', 'bittorrent', 'hit', 'mpaa'], ['mpaa', 'began', 'legal', 'campaign', 'operators', 'similar', 'networks',
'across', 'four', 'continents', 'december'], ['dallas', 'court', 'agreed', 'hollywood', 'lawyers', 'would', 'allow
ed', 'access', 'lokitorrent', 'server', 'records', 'could', 'let', 'single', 'sharing', 'files', 'illegally'], ['o
ctober', '2004', 'site', 'provided', 'links', '30,000', 'files'], ['action', 'came', 'operators', 'lokitorrent',
'agreed', 'settlement', 'mpaa'], ['stark', 'message', 'appeared', 'site', 'mpaa', 'warning', 'you', 'click',
"n't", 'hide'], ['bittorrent', 'systems', 'server', 'sites', 'host', 'files', 'shared'], ['they', 'host', 'links',
'called', 'trackers', 'direct', 'people', 'others', 'instead'], ['well', 'filing', 'unspecified', 'number', 'fil
e', 'suits', 'across', 'mpaa', 'said', 'given', 'operators', 'host', 'edonkey', 'servers', 'take', 'notices'], ['h
ollywood', 'studios', 'aggressively', 'clamping', 'file-sharers', 'says', 'infringe', 'copyright', 'laws', 'copyin
g', 'films', 'programmes', 'share', 'files', 'online'], ['but', 'targeting', 'operators', 'bittorrent', 'network
s'], ['filed', '100', 'lawsuits', 'operators', 'bittorrent', 'server', 'sites', 'since', 'december'], ['strategy',
'hitting', 'run', 'servers', 'link', 'copyrighted', 'material', 'intended', 'stunt', 'file-sharers', 'ability', 's
wap', 'content', 'using', 'bittorrent', 'systems'], ['film', 'industry', 'says', 'black', 'market', 'illegally',
'copied', 'videos', 'dvds', 'already', 'costs', 'billions', 'every', 'year', 'worried', 'illegal', 'file-sharing',
'adding', 'losses'], ['december', 'legal', 'action', 'claimed', 'high-profile', 'victim'], ['popular', 'suprnova.o
rg', 'website', 'forced', 'close', 'others', 'like', 'phoenix', 'torrent', 'followed', 'soon']]
```

Step 2 : Word2vec matrix for term relatedness



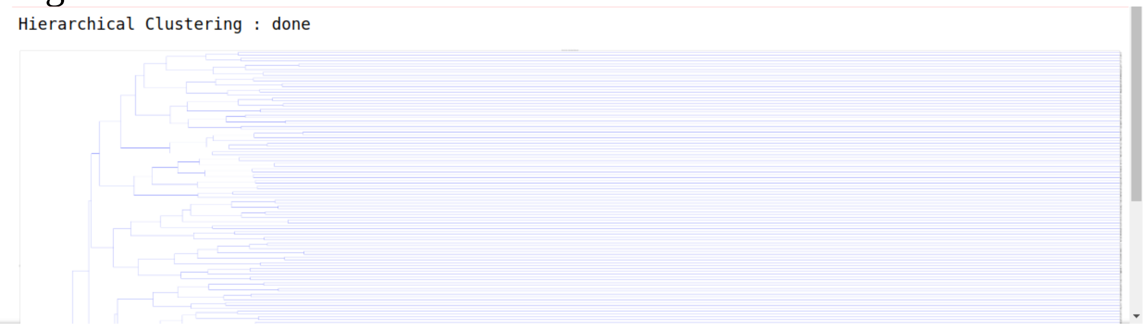
```
[[-0.00142912 -0.00116865  0.00017441 ...  0.00156376 -0.00084369
  -0.00118462]
 [ 0.00100302  0.00011935  0.00165278 ... -0.00081228  0.00015567
  -0.00036874]
 [-0.0013663  -0.00040268  0.0004001  ... -0.0009937  -0.0011102
  -0.00019286]
 ...
 [ 0.00142904 -0.00051855 -0.00050334 ... -0.00151349 -0.00053264
   0.00117293]
 [-0.00098797 -0.00099029  0.00121299 ... -0.00065282 -0.00121511
  -0.00004572]
 [-0.00104027  0.00159889 -0.00020557 ... -0.0013489  -0.00076975
   0.00028938]]
Word2Vec Model : done
```

Displaying the relatedness among the words



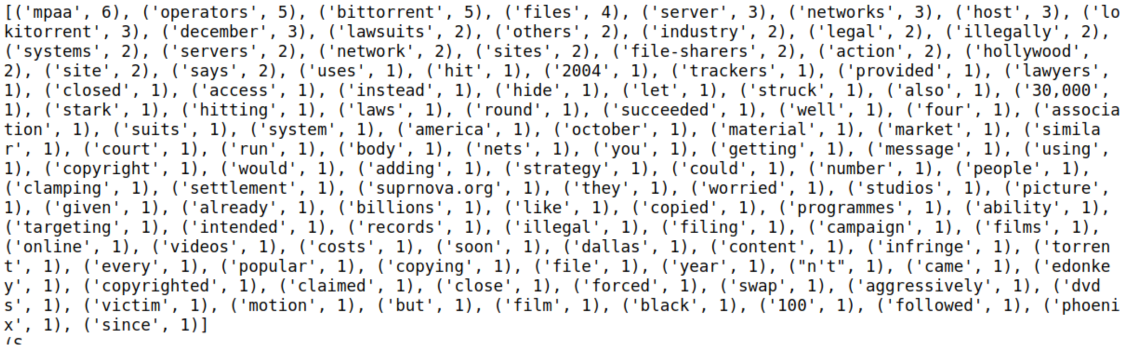
## Step 3 : Hierarchical Clustering

## Dendogram obtained:



```
Hierarchical Clustering : done
```

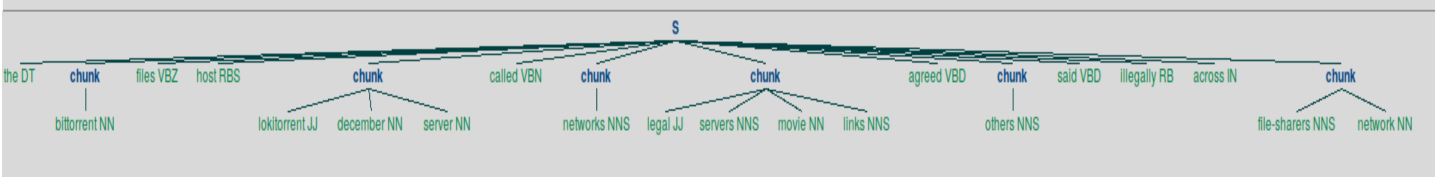
## Step 4 :Finding exemplar terms from clusters using K means with term frequencies

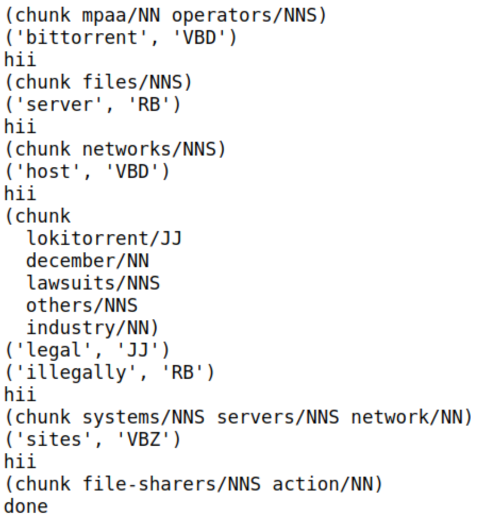


```
[('mpaa', 6), ('operators', 5), ('bittorrent', 5), ('files', 4), ('server', 3), ('networks', 3), ('host', 3), ('lo
kitorrent', 3), ('december', 3), ('lawsuits', 2), ('others', 2), ('industry', 2), ('legal', 2), ('illegally', 2),
('systems', 2), ('servers', 2), ('network', 2), ('sites', 2), ('file-sharers', 2), ('action', 2), ('hollywood',
2), ('site', 2), ('says', 2), ('uses', 1), ('hit', 1), ('2004', 1), ('trackers', 1), ('provided', 1), ('lawyers',
1), ('closed', 1), ('access', 1), ('instead', 1), ('hide', 1), ('let', 1), ('struck', 1), ('also', 1), ('30,000',
1), ('stark', 1), ('hitting', 1), ('laws', 1), ('round', 1), ('succeeded', 1), ('well', 1), ('four', 1), ('associa
tion', 1), ('suits', 1), ('system', 1), ('america', 1), ('october', 1), ('material', 1), ('market', 1), ('simila
r', 1), ('court', 1), ('run', 1), ('body', 1), ('nets', 1), ('you', 1), ('getting', 1), ('message', 1), ('using',
1), ('copyright', 1), ('would', 1), ('adding', 1), ('strategy', 1), ('could', 1), ('number', 1), ('people', 1),
('clamping', 1), ('settlement', 1), ('suprnova.org', 1), ('they', 1), ('worried', 1), ('studios', 1), ('picture',
1), ('given', 1), ('already', 1), ('billions', 1), ('like', 1), ('copied', 1), ('programmes', 1), ('ability', 1),
('targeting', 1), ('intended', 1), ('records', 1), ('illegal', 1), ('filing', 1), ('campaign', 1), ('films', 1),
('online', 1), ('videos', 1), ('costs', 1), ('soon', 1), ('dallas', 1), ('content', 1), ('infringe', 1), ('torren
t', 1), ('every', 1), ('popular', 1), ('copying', 1), ('file', 1), ('year', 1), ("n't", 1), ('came', 1), ('edonke
y', 1), ('copyrighted', 1), ('claimed', 1), ('close', 1), ('forced', 1), ('swap', 1), ('aggressively', 1), ('dvd
s', 1), ('victim', 1), ('motion', 1), ('but', 1), ('film', 1), ('black', 1), ('100', 1), ('followed', 1), ('phoeni
x', 1), ('since', 1)]
(s
```

## Step 5 : Select top 20 most frequent terms

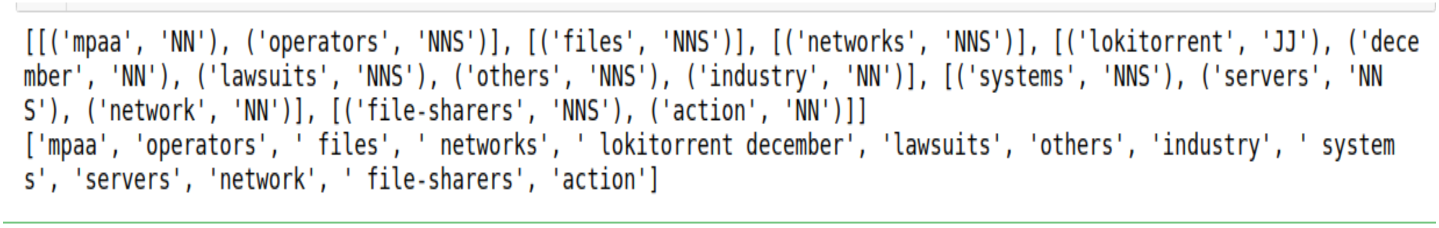
## Step 6 :Chunking





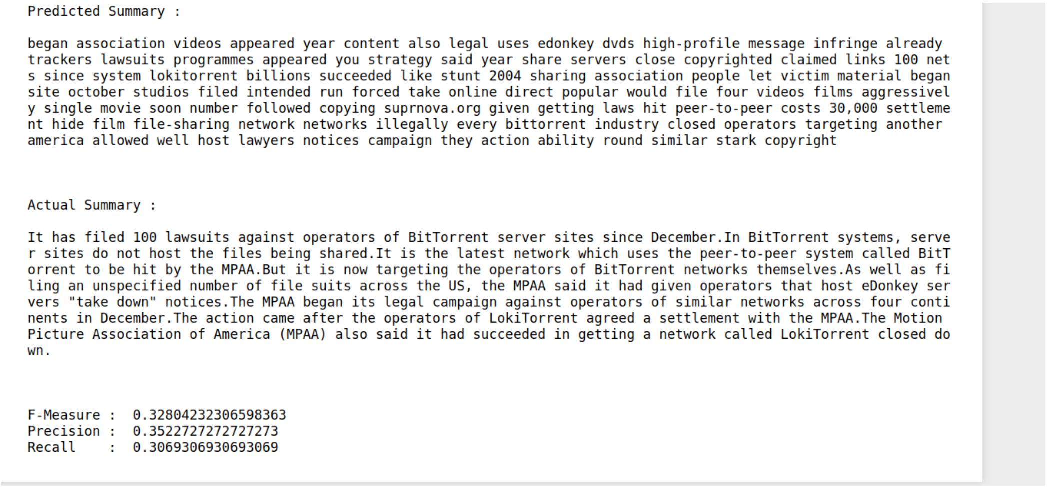
```
(chunk mpaa/NN operators/NNS)
('bittorrent', 'VBD')
hii
(chunk files/NNS)
('server', 'RB')
hii
(chunk networks/NNS)
('host', 'VBD')
hii
(chunk
  lokitorrent/JJ
  december/NN
  lawsuits/NNS
  others/NNS
  industry/NN)
('legal', 'JJ')
('illegally', 'RB')
hii
(chunk systems/NNS servers/NNS network/NN)
('sites', 'VBZ')
hii
(chunk file-sharers/NNS action/NN)
done
```

## Step 7 : Finally getting the keyphrases from these chunks



```
[[('mpaa', 'NN'), ('operators', 'NNS')], [('files', 'NNS')], [('networks', 'NNS')], [('lokitorrent', 'JJ'), ('dece
mber', 'NN'), ('lawsuits', 'NNS'), ('others', 'NNS'), ('industry', 'NN')], [('systems', 'NNS'), ('servers', 'NN
S'), ('network', 'NN')], [('file-sharers', 'NNS'), ('action', 'NN')]]
['mpaa', 'operators', ' files', ' networks', ' lokitorrent december', 'lawsuits', 'others', 'industry', ' system
s', 'servers', 'network', ' file-sharers', 'action']
```

## Step 8 : Precision ,recall and f1 score



```
Predicted Summary :

began association videos appeared year content also legal uses edonkey dvds high-profile message infringe already
trackers lawsuits programmes appeared you strategy said year share servers close copyrighted claimed links 100 net
s since system lokitorrent billions succeeded like stunt 2004 sharing association people let victim material began
site october studios filed intended run forced take online direct popular would file four videos films aggressivel
y single movie soon number followed copying suprnova.org given getting laws hit peer-to-peer costs 30,000 settleme
nt hide film file-sharing network networks illegally every bittorrent industry closed operators targeting another
america allowed well host lawyers notices campaign they action ability round similar stark copyright



Actual Summary :

It has filed 100 lawsuits against operators of BitTorrent server sites since December.In BitTorrent systems, serve
r sites do not host the files being shared.It is the latest network which uses the peer-to-peer system called BitT
orrent to be hit by the MPAA.But it is now targeting the operators of BitTorrent networks themselves.As well as fi
ling an unspecified number of file suits across the US, the MPAA said it had given operators that host eDonkey ser
vers "take down" notices.The MPAA began its legal campaign against operators of similar networks across four conti
nents in December.The action came after the operators of LokiTorrent agreed a settlement with the MPAA.The Motion
Picture Association of America (MPAA) also said it had succeeded in getting a network called LokiTorrent closed do
wn.



F-Measure :  0.32804232306598363
Precision :  0.3522727272727273
Recall    :  0.3069306930693069
```

## Observation:

1)Word2Vec with hierarchical clustering

| **Parameters** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|
| | | | |
| m=(1/4)n | 0.13 | 0.32 | 0.19 |
| m=(1/2)n | 0.22 | 0.26 | 0.24 |
| m=(2/3)n | 0.30 | 0.27 | 0.28 |
| m=(4/5)n | 0.36 | 0.26 | 0.30 |

Where m=number of clusters
n=total number of candidate terms

2) Word2Vec with Spectral clustering

| Parameters | Precision | Recall | F1 score |
|---|---|---|---|
| | | | |
| m=(1/4)n | 0.18 | 0.42 | 0.25 |
| m=(1/2)n | 0.31 | 0.36 | 0.34 |
| m=(2/3)n | 0.40 | 0.35 | 0.38 |
| m=(4/5)n | 0.46 | 0.33 | 0.39 |

3)Cooccurence matrix for relatedness with Hierarchical clustering

a)Set value of w,window size=10

| Parameters | Precision | Recall | F1 score |
|---|---|---|---|
| | | | |
| m=(1/4)n | 0.125 | 0.28 | 0.17 |
| m=(1/2)n | 0.23 | 0.27 | 0.25 |
| m=(2/3)n | 0.31 | 0.27 | 0.29 |
| m=(4/5)n | 0.39 | 0.28 | 0.33 |

b)w=20

| Parameters | Precision | Recall | F1 score |
|---|---|---|---|
| | | | |
| m=(1/4)n | 0.11 | 0.26 | 0.15 |
| m=(1/2)n | 0.25 | 0.28 | 0.26 |
| m=(2/3)n | 0.35 | 0.30 | 0.32 |
| m=(4/5)n | 0.39 | 0.28 | 0.33 |

c)w=40

| Parameters | Precision | Recall | F1 score |
|---|---|---|---|
| | | | |
| m=(1/4)n | 0.13 | 0.31 | 0.19 |
| m=(1/2)n | 0.21 | 0.25 | 0.23 |
| m=(2/3)n | 0.31 | 0.27 | 0.29 |
| m=(4/5)n | 0.38 | 0.27 | 0.32 |

## Analysis and Discussion

From the above experiment results, we can see the clustering-based method is both robust and effective for keyphrase extraction as an unsupervised method.

We calculated on clustering methods namely hierarchical and spectral and on different cluster values and observed different behaviour.

As the cluster number increases more exemplar terms are identified from these clusters and more keyphrases will be extracted from the document based on exemplar terms.

If we set the cluster number to m=n all terms will be selected as exemplar terms. Thus it is important for this method to appropriately specify the cluster number.

In the experiment we also noticed that frequent word list is important for keyphrase extraction.

## Conclusion and future work

In this project, we propose an unsupervised clustering-based keyphrase extraction algorithm.
This method groups candidate terms into clusters and identify the exemplar terms.
Then keyphrases are extracted from the document based on the exemplar terms.
The clustering based on term semantic relatedness guarantees the extracted keyphrases have a good coverage of the document.

# 7. REFERENCES

[1] Zhiyuan Liu, Peng Li, Yabin Zheng, Maosong Sun,”**Clustering to Find Exemplar Terms for Keyphrase Extraction”**

[2] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing