CS6109-Compiler design Project

Team No-15

Project Title

Genetic Semantic Graph Approach for Multi-document Summarization

Team Members:

1. RAGHURAJ SV 2019103563

2. DHAMODHARAN S 2019103515

3. VIGNESH G 2018103078

Genetic Semantic Graph Approach for Multi-document Summarization

ABSTRACT:

Obtaining training data for multi-document summarization (MDS) is time consuming and resource intensive, so recent neural models can only be trained for limited domains. In this paper, we propose an unsupervised method for multi-document summarization, in which we convert the original documents to a sentences, removing stop words, forming Semantic similarity matrix then apply page rank algorithm to obtain rank for multiple clusters of sentences, and finally taking top ranked sentences to generate the final summary. Human evaluation shows our system produces consistent and complete summaries compared to human written ones. The aim of automatic multi-document abstractive summarization is to create a compressed version of the source text and preserves the salient information. Existing Graph based summarization methods treat sentence as bag of words, rely on content similarity measure and did not consider semantic relationships between sentences. These methods may fail in determining redundant sentences that are semantically equivalent. This paper introduces a genetic semantic Graph based approach for multi-document summarization. Semantic graph from the document set is constructed in such a way that the graph nodes represent the predicate argument structures (PAS) and the edges of graph correspond to semantic similarity weight determined from PAS-to-PAS semantic similarity, and PAS-to-document set relationship. The PAS-to-document set relationship is represented by different features, weighted and optimized by genetic algorithm. The salient graph nodes (PASs) are ranked based on modified graph based ranking algorithm.

DELIVERABLES:

Input: Multiple Text documents for summary generation

Output: Single summary Text Summarized document.

Introduction:

Nowadays, users of internet are flooded with the huge amount of textual data due to the explosive growth of information overload over the internet. The current age of information overload demands need of text summarization, which is a significant and timely tool for user to speedily comprehend the massive volume of information. Multi-document text summarization aims to select the most significant information from the given source documents and produce a concise summary that can satisfy user's needs .The idea of document summarization is a bit different from key-Phrase extraction or topic modelling . In this case, the end result is still in the form of some document, but with a few sentences based on the length we might want the summary to be. This is similar to an abstract or an executive summary in a research paper. The main objective of automated document summarization is to perform this summarization without involving human input, except for running computer programs. Mathematical and statistical models help in building and automating the task of summarizing documents by observing their content and context. Text summarization approaches can be broadly separated into two groups: extractive summarization and abstractive summarization. Extractive summarization extracts the most important representative sentences from the source documents and group them to produce a summary. However, abstractive summarization requires natural language processing techniques such as semantic representation, natural language generation, and compression techniques. Abstractive summarization aims to interpret and examine the source text and creates a concise summary that usually contain compressed sentences or may contain some novel sentences not present in the original source text .

Problem statement:

For recovering data, people generally use the web, for example, Google, Bing, Yahoo etc. Since the amount of material on the web is evolving quickly, for clients it isn't simple to discover pertinent and fitting data according to the prerequisites. When a client transmits a query on an Internet search engine for information or data then the reaction in most of the occasions is a great many documents and the client needs to confront the repetitive assignment of finding the fitting data from this ocean of responses. This issue is known as "Data Overloading".

Objective:

The essential objective of various multi-document summarization techniques is to create summaries which provide extensive inclusion, less redundancy in the information and extensive consistency between sentences. In other words, the important content is removed from each data source and at that point is re-structured to generate summaries for multiple documents.so we propose a multidocument summarizer for making people get the crisp, important data from the enormous amount of data, a time saving in this busy modern world.

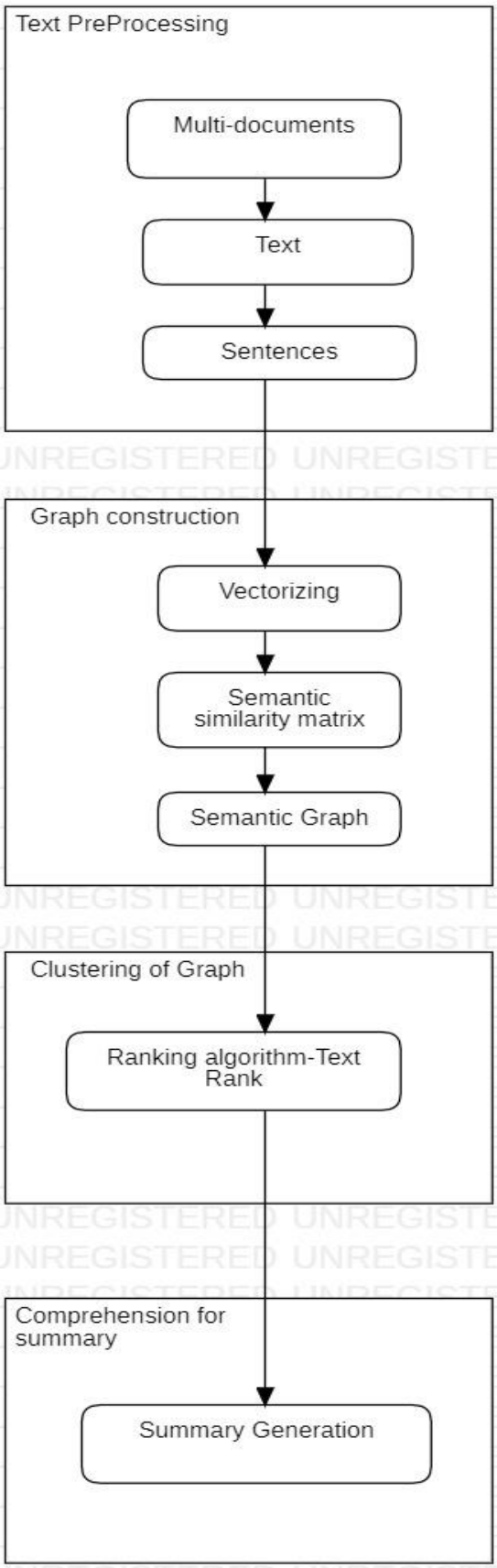
Literature survey:

N0.	Title	Author, Year	Methodology	Advantages	Limitations
1.	Genetic Semantic Graph Approach for Multi-document Abstractive Summarization	Atif Khan, Naomie Salim, Yogan Jaya Kumar(2015)	semantic role labeling, semantic graph, semantic similarity measure, genetic algorithm, abstractive summarization	. They leverage language semantics to create representations and use natural language generation (NLG) techniques where the machine uses knowledge bases and semantic representations to generate text on its own and create summaries just like a human would write them.	They require a lot of data and compute.
2.	Extraction based summarization using a shortest path algorithm	Jonas Sj¨obergh, Kenji Arak(2006)	shortest path from the first sentence to the last sentence in a graph representing the original text. Traditional sentence weights.	This content can be words, phrases, or even sentences. The end result from this approach is a short executive summary of a couple of lines extracted from the original document.	No new content is generated in this technique, hence the name extraction-based.

Method used in this project:

In our project genetic semantic extraction based summarization method is since abstraction based summarization involves larger dataset pre-processed and cannot be used for all kind of data sets.

Block Diagram:



Module:1 Text Pre-Processing

Step-1: multi-documents

Step-2: Text

Step-3: Sentences

Sentences:

We concatenate documents and apply minimal text processing, mainly sentence split, as we want to keep documents raw for subsequent processing. We remove unwanted stop words, whitespaces, newlines in our document .The output here is a list of sentences, which are fed to the sentence graph construction step.

Pseudo code:

```
import re
DOCUMENT = re.sub(r'\n|\r', ' ', DOCUMENT)
DOCUMENT = DOCUMENT.strip()

print(DOCUMENT)


sentences=nltk.sent_tokenize(DOCUMENT)
print(sentences)
len(sentences)
```

Input: Mutliple document after recognition of tokens, Removing ambiguious sentences.

Output: List of sentences from all the documents combined.

Module-2: Graph construction:

Step-1: vectorizing

Step-2: Semantic similarity matrix

Step-3: Semantic matrix construction

Vectorizing: TF factor(term frequency):

The weight of a term that occurs in a document is simply proportional to the term frequency.

Formula : tf(t,d) = count of t in d / number of words in d

IDF factor(inverse document frequency):

While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an *inverse document frequency* factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

Pseudo code:

```
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd

tv = TfidfVectorizer(min_df=0., max_df=1., use_idf=True)
dt_matrix = tv.fit_transform(norm_sentences)
dt_matrix = dt_matrix.toarray()

vocab = tv.get_feature_names()
td_matrix = dt_matrix.T
print(td_matrix.shape)
pd.DataFrame(np.round(td_matrix, 2), index=vocab).head(15)
```

Semantic similarity matrix:

- The goal at this step is to identify pairwise sentence connection which represents the discourse structure of documents D.
- Decide on the number of sentences, k, that we want in the final summary
- Build a document-term feature matrix using weights like TF-IDF or Bag of Words.
- Compute a document similarity matrix by multiplying the matrix by its transpose.
- Use these documents (sentences in our case) as the vertices and the similarities between each pair of documents as the weight or score.

Pseudo code:

```
similarity_matrix = np.matmul(dt_matrix, dt_matrix.T)
print(similarity_matrix.shape)
np.round(similarity_matrix, 3)
```

Input: List of sentences from the previous step.

Output: After vectorization and semantic similarity matrix only semantically connected sentences are obtained, sent into graph construction constructor, Semantic graph is output.

Module-3: Clustering from graph:

Steps: Ranking algorithm – Text Rank

Most graph clustering approaches try to identify communities of nodes in a graph based on the edges linking them.

The TextRank summarization algorithm internally uses the popular PageRank algorithm, which is used by Google for ranking websites and pages. This is used by the Google search engine when providing relevant web pages based on search queries. To understand TextRank better, we need to understand some of the concepts surrounding PageRank. The core algorithm in PageRank is a graph-based scoring or ranking algorithm, where pages are scored or ranked based on their importance.

Websites and pages contain further links embedded in them which link to more pages having more links and this continues across the Internet. This can be represented as a graph-based model where vertices indicate the web pages and edges indicate links among them. This can be used to form a voting or recommendation system such so when one vertex links to another one in the graph it is basically casting a vote.

Vertex importance is decided not only on the number of votes or edges but also the importance of the vertices that are connected to it and their importance.

Pseudo code:

```
scores = networkx.pagerank(similarity_graph)
ranked_sentences = sorted(((score, index) for index, score
in scores.items()), reverse=True)
ranked_sentences[:20]
```

Input: semantic Graph is fed as input for optimization phase

output: Semantic graph is clustered based on Text Rank algorithm.

Module-4: Comprehension for summary:

Steps: Summary Generation

Text ranking algorithm result is comprehensive based on the no. of lines output we need and the final output a summarized document is obtained.

Pseudo code:

```
num_sentences=7;
top_sentence_indices = [ranked_sentences[index][1]
                        for index in range(num_sentences)]
top_sentence_indices.sort()
```

Input: sentence clusters is fed as input.

Output: optimize the clusters into single text document

Dataset description:

Data set used in our project are text files such as news articles or a any kind of documents for which summarization is required for us.

Text files which are related to each other can be used as a data set in our project so that the overall information from documents is summarized to give a summary.

If text files which doesn't have related information are used, our project tries to find common happenings in between them and provide a summarization.

Sample text files used in this implementation is shown below.

Result implementation:

This project is implemented in python on google research colaboratory.

Module-1: Text Pre-processing.

Step-1: Downloading requirements:

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
```

output:

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

Importing requirements NLTK- a Natural Language Processing based library. Where we download, punkt for tokenizing our document, stop words such as is, a, are, that, this etc..

Step-2: concatenating documents:

```
DOCUMENT=""
for i in range(1,11):
  file_name=f"{i}.txt"
  DOCUMENT+=(open(file_name).read())
print(DOCUMENT)
```

output:

```
A visible improvement in Delhi's air quality was recorded on Sunday although it was
in the ''very poor'' category while the city's Environment Minister Gopal Rai said
his government will submit a lockdown proposal to the Supreme Court on Monday to
reduce pollution further.
The national capital recorded a 24-hour average air quality index (AQI) of 330 on
Sunday as against 437 the previous day as emissions from farm fires in Haryana and
Punjab dropped significantly. The AQI was and 471 on Friday, the worst this season
so far.

The air quality index of neighbouring Ghaziabad, Gurgaon, Noida, Faridabad, Greater
Noida was recorded at 331, 287, 321, 298 and 310, respectively.

An AQI between zero and 50 is considered ''good'', 51 and 100 ''satisfactory'', 101
and 200 ''moderate'', 201 and 300 ''poor'', 301 and 400 ''very poor'', and 401 and
500 ''severe''.

The India Meteorological Department said visibility levels ranged from 1,500 to
2,200 metres at the Indira Gandhi International Airport and from 1,000 to 1,500
metres at the Safdarjung Airport.

Delhi Environment Minister Gopal Rai said the city government will on Monday submit
to the Supreme Court a proposal on clamping a lockdown and its modalities.

The Supreme Court had on Saturday termed the rise in pollution levels an "emergency
situation" and suggested clamping a lockdown in the national capital.

The Delhi government has already announced the closure of physical classes in
schools, colleges and other educational institutions, except those where exams are
being conducted, for a week from Monday.President Xi Jinping said on Saturday that
China can no longer rely on its previous economic development model of depending on
global exports and must build self-controlled, safe and reliable domestic
production and supply system to ensure industrial and national security.
The just-concluded plenary session of the ruling Communist Party of China (CPC),
headed by Xi, adopted his proposals to make the 14th Five-Year Plan (2021-2025).
While the 14th five-year plan envisages a massive overhaul of the country's
domestic market to boost consumption in order to reduce China's reliance on
shrinking exports markets, the Vision 2035 visualises a long-term plan, reflecting
the development vision of Xi. Politically, Xi's Vision 2035 plan has sparked
speculation that he could continue in power for the next 15 years.
Xi, 67, has emerged as the CPC's most powerful leader after its founder Mao Zedong,
holding the posts of CPC General Secretary, head of the military besides the
presidency with prospects of a life-long tenure.
```

A constitutional amendment in 2018 removed the two 5year term limit for the president, which would enable Xi to continue in power for life. His second term as the president is due to end in 2022.Vitamin D nutrient is in the news more than ever for its greater implications during the time of Coronavirus pandemic. Health experts have raised concerns over the growing cases of vitamin D deficiency in general population as people are staying at home and are not able to obtain the 'sunshine vitamin' from natural sunlight. In fact, a new study discovered over 80% COVID-19 patients suffering from vitamin D deficiency. Vitamin D is known to aid many bodily functions and its insufficiency may lead to weak bones, heart-related ailments, low immunity and even respiratory problems.
The study that was published in 'The Journal of Clinical Endocrinology & Metabolism', found 80 percent of 216 COVID-19 patients admitted in a hospital in Spain to be vitamin D-deficient. The researchers also noticed that men had lower vitamin D levels than women.
"Vitamin D-deficient COVID-19 patients had a greater prevalence of hypertension and cardiovascular diseases, raised serum ferritin and troponin levels, as well as a longer length of hospital stay. We did not find any relationship between vitamin D concentrations or vitamin deficiency and the severity of the disease," wrote co-author Jose L. Hernandez, Ph.D., of the University of Cantabria in Santander, Spain.
Another recent study, published in 'Plos One' journal, had claimed that vitamin D sufficiency may lessen the oxygen requirement in COVID-19 patients and fasten the treatment process.
Vitamin D is also known as 'sunshine vitamin'.
Apart from exposing yourself to sunlight regularly to obtain the vitamin naturally, a diet rich in foods with high vitamin D content may also help.

Documents are named 1.txt, 2.txt, 3.txt and using for loop we have concatenated it into a single document.

Sample documents are uploaded:

The files used are:

1.txt

A visible improvement in Delhi's air quality was recorded on Sunday although it was in the ''very poor'' category while the city's Environment Minister Gopal Rai said his government will submit a lockdown proposal to the Supreme Court on Monday to reduce pollution further.

The national capital recorded a 24-hour average air quality index (AQI) of 330 on Sunday as against 437 the previous day as emissions from farm fires in Haryana and Punjab dropped significantly. The AQI was and 471 on Friday, the worst this season so far.

The air quality index of neighbouring Ghaziabad, Gurgaon, Noida, Faridabad, Greater Noida was recorded at 331, 287, 321, 298 and 310, respectively.

An AQI between zero and 50 is considered ''good'', 51 and 100 ''satisfactory'', 101 and 200 ''moderate'', 201 and 300 ''poor'', 301 and 400 ''very poor'', and 401 and 500 ''severe''.

The India Meteorological Department said visibility levels ranged from 1,500 to 2,200 metres at the Indira Gandhi International Airport and from 1,000 to 1,500 metres at the Safdarjung Airport.

Delhi Environment Minister Gopal Rai said the city government will on Monday submit to the Supreme Court a proposal on clamping a lockdown and its modalities.

The Supreme Court had on Saturday termed the rise in pollution levels an "emergency situation" and suggested clamping a lockdown in the national capital.

The Delhi government has already announced the closure of physical classes in schools, colleges and other educational institutions, except those where exams are being conducted, for a week from Monday.

2.txt:

President Xi Jinping said on Saturday that China can no longer rely on its previous economic development model of depending on global exports and must build self-controlled, safe and reliable domestic production and supply system to ensure industrial and national security.

The just-concluded plenary session of the ruling Communist Party of China (CPC), headed by Xi, adopted his proposals to make the 14th Five-Year Plan (2021-2025).

While the 14th five-year plan envisages a massive overhaul of the country's domestic market to boost consumption in order to reduce China's reliance on shrinking exports markets, the Vision 2035 visualises a long-term plan, reflecting the development vision of Xi. Politically, Xi's Vision 2035 plan has sparked speculation that he could continue in power for the next 15 years.

Xi, 67, has emerged as the CPC's most powerful leader after its founder Mao Zedong, holding the posts of CPC General Secretary, head of the military besides the presidency with prospects of a life-long tenure.

A constitutional amendment in 2018 removed the two 5year term limit for the president, which would enable Xi to continue in power for life. His second term as the president is due to end in 2022.

3.txt:

Vitamin D nutrient is in the news more than ever for its greater implications during the time of Coronavirus pandemic. Health experts have raised concerns over the growing cases of vitamin D deficiency in general population as people are staying at home and are not able to obtain the 'sunshine vitamin' from natural sunlight. In fact, a new study discovered over 80% COVID-19 patients suffering from vitamin D deficiency. Vitamin D is known to aid many bodily functions and its insufficiency may lead to weak bones, heart-related ailments, low immunity and even respiratory problems.

The study that was published in 'The Journal of Clinical Endocrinology & Metabolism', found 80 percent of 216 COVID-19 patients admitted in a hospital in Spain to be vitamin D-deficient. The researchers also noticed that men had lower vitamin D levels than women.

"Vitamin D-deficient COVID-19 patients had a greater prevalence of hypertension and cardiovascular diseases, raised serum ferritin and troponin levels, as well as a longer length of hospital stay. We did not find any relationship between vitamin D concentrations or vitamin deficiency and the severity of the disease," wrote co-author Jose L. Hernandez, Ph.D., of the University of Cantabria in Santander, Spain.

Another recent study, published in 'Plos One' journal, had claimed that vitamin D sufficiency may lessen the oxygen requirement in COVID-19 patients and fasten the treatment process.

Vitamin D is also known as 'sunshine vitamin'.

Apart from exposing yourself to sunlight regularly to obtain the vitamin naturally, a diet rich in foods with high vitamin D content may also help.

Step-3: Tokenizing into sentences:

```
import re
DOCUMENT = re.sub(r'\n|\r', ' ', DOCUMENT)
DOCUMENT = DOCUMENT.strip()

print(DOCUMENT)
```

Output:

```
A visible improvement in Delhi's air quality was recorded on Sunday although it was
in the ''very poor'' category while the city's Environment Minister Gopal Rai said
his government will submit a lockdown proposal to the Supreme Court on Monday to
reduce pollution further. The national capital recorded a 24-hour average air
quality index (AQI) of 330 on Sunday as against 437 the previous day as emissions
from farm fires in Haryana and Punjab dropped significantly. The AQI was and 471 on
Friday, the worst this season so far.  The air quality index of neighbouring
Ghaziabad, Gurgaon, Noida, Faridabad, Greater Noida was recorded at 331, 287, 321,
298 and 310, respectively.  An AQI between zero and 50 is considered ''good'', 51
and 100 ''satisfactory'', 101 and 200 ''moderate'', 201 and 300 ''poor'', 301 and
400 ''very poor'', and 401 and 500 ''severe''.  The India Meteorological Department
said visibility levels ranged from 1,500 to 2,200 metres at the Indira Gandhi
International Airport and from 1,000 to 1,500 metres at the Safdarjung Airport.
Delhi Environment Minister Gopal Rai said the city government will on Monday submit
to the Supreme Court a proposal on clamping a lockdown and its modalities.  The
Supreme Court had on Saturday termed the rise in pollution levels an "emergency
situation" and suggested clamping a lockdown in the national capital.  The Delhi
government has already announced the closure of physical classes in schools,
colleges and other educational institutions, except those where exams are being
conducted, for a week from Monday.President Xi Jinping said on Saturday that China
can no longer rely on its previous economic development model of depending on
global exports and must build self-controlled, safe and reliable domestic
production and supply system to ensure industrial and national security. The just-
concluded plenary session of the ruling Communist Party of China (CPC), headed by
Xi, adopted his proposals to make the 14th Five-Year Plan (2021-2025). While the
14th five-year plan envisages a massive overhaul of the country's domestic market
to boost consumption in order to reduce China's reliance on shrinking exports
markets, the Vision 2035 visualises a long-term plan, reflecting the development
vision of Xi. Politically, Xi's Vision 2035 plan has sparked speculation that he
could continue in power for the next 15 years. Xi, 67, has emerged as the CPC's
most powerful leader after its founder Mao Zedong, holding the posts of CPC General
Secretary, head of the military besides the presidency with prospects of a life-
long tenure. A constitutional amendment in 2018 removed the two 5year term limit
for the president, which would enable Xi to continue in power for life. His second
term as the president is due to end in 2022.Vitamin D nutrient is in the news more
than ever for its greater implications during the time of Coronavirus pandemic.
Health experts have raised concerns over the growing cases of vitamin D deficiency
in general population as people are staying at home and are not able to obtain the
'sunshine vitamin' from natural sunlight. In fact, a new study discovered over 80%
```

COVID-19 patients suffering from vitamin D deficiency. Vitamin D is known to aid many bodily functions and its insufficiency may lead to weak bones, heart-related ailments, low immunity and even respiratory problems. The study that was published in 'The Journal of Clinical Endocrinology & Metabolism', found 80 percent of 216 COVID-19 patients admitted in a hospital in Spain to be vitamin D-deficient. The researchers also noticed that men had lower vitamin D levels than women. "Vitamin D-deficient COVID-19 patients had a greater prevalence of hypertension and cardiovascular diseases, raised serum ferritin and troponin levels, as well as a longer length of hospital stay. We did not find any relationship between vitamin D concentrations or vitamin deficiency and the severity of the disease," wrote co-author Jose L. Hernandez, Ph.D., of the University of Cantabria in Santander, Spain. Another recent study, published in 'Plos One' journal, had claimed that vitamin D sufficiency may lessen the oxygen requirement in COVID-19 patients and fasten the treatment process. Vitamin D is also known as 'sunshine vitamin'. Apart from exposing yourself to sunlight regularly to obtain the vitamin naturally, a diet rich in foods with high vitamin D content may also help.

```
sentences=nltk.sent_tokenize(DOCUMENT)
print(sentences)
len(sentences)
```

Output:

["A visible improvement in Delhi's air quality was recorded on Sunday although it was in the ''very poor'' category while the city's Environment Minister Gopal Rai said his government will submit a lockdown proposal to the Supreme Court on Monday to reduce pollution further.", 'The national capital recorded a 24-hour average air quality index (AQI) of 330 on Sunday as against 437 the previous day as emissions from farm fires in Haryana and Punjab dropped significantly.', 'The AQI was and 471 on Friday, the worst this season so far.', 'The air quality index of neighbouring Ghaziabad, Gurgaon, Noida, Faridabad, Greater Noida was recorded at 331, 287, 321, 298 and 310, respectively.', "An AQI between zero and 50 is considered ''good'', 51 and 100 ''satisfactory'', 101 and 200 ''moderate'', 201 and 300 ''poor'', 301 and 400 ''very poor'', and 401 and 500 ''severe''.", 'The India Meteorological Department said visibility levels ranged from 1,500 to 2,200 metres at the Indira Gandhi International Airport and from 1,000 to 1,500 metres at the Safdarjung Airport.', 'Delhi Environment Minister Gopal Rai said the city government will on Monday submit to the Supreme Court a proposal on clamping a lockdown and its modalities.', 'The Supreme Court had on Saturday termed the rise in pollution levels an "emergency situation" and suggested clamping a lockdown in the national capital.', 'The Delhi government has already announced the closure of physical classes in schools, colleges and other educational institutions, except those where exams are being conducted, for a week from Monday.President Xi Jinping said on Saturday that China can no longer rely on its previous economic development model of depending on global exports and must build self-controlled, safe and reliable domestic production and supply system to ensure industrial and national security.', 'The just-concluded plenary session of the ruling Communist Party of China (CPC), headed by Xi, adopted his proposals to make the 14th Five-Year Plan (2021-2025).', "While the 14th five-year plan envisages a massive overhaul of the country's domestic market to boost consumption in order to reduce China's reliance on shrinking exports markets, the Vision 2035 visualises a long-term plan, reflecting the development vision of Xi.", "Politically, Xi's Vision 2035 plan has sparked speculation that he could continue in power for the next 15 years.", "Xi, 67, has emerged as the CPC's most powerful leader after its founder Mao Zedong, holding the posts of CPC General Secretary, head of the military besides the presidency with prospects of a life-long tenure.", 'A constitutional amendment in 2018 removed the two 5year term limit for the president, which would enable Xi to continue in power for life.', 'His second term as the president is due to end in 2022.Vitamin D nutrient is in the news more than ever for its greater implications during the time of Coronavirus pandemic.', "Health experts have raised concerns over the growing cases of vitamin D deficiency in general population as people are staying at home and are not able to obtain the 'sunshine vitamin' from natural sunlight.", 'In fact, a new study discovered over 80% COVID-19 patients suffering from vitamin D deficiency.', 'Vitamin D is known to aid many bodily functions and its insufficiency may lead to weak bones, heart-related ailments, low immunity and even respiratory problems.', "The study that was published in 'The Journal of Clinical Endocrinology & Metabolism', found 80 percent of 216 COVID-19 patients admitted in a hospital in Spain to be vitamin D-deficient.", 'The researchers also noticed that men had lower vitamin D levels than women.', '"Vitamin D-deficient COVID-19 patients had a greater prevalence of hypertension and cardiovascular diseases, raised serum ferritin and troponin levels, as well as a longer length of hospital stay.', 'We did not find any relationship between vitamin D concentrations or vitamin deficiency and the severity of the disease," wrote co-author Jose L. Hernandez, Ph.D., of the University of Cantabria in Santander, Spain.', "Another recent study, published in 'Plos One' journal, had claimed that vitamin D sufficiency may lessen the oxygen requirement in COVID-19 patients and fasten the treatment process.",

"Vitamin D is also known as 'sunshine vitamin'.", 'Apart from exposing yourself to sunlight regularly to obtain the vitamin naturally, a diet rich in foods with high vitamin D content may also help.']
25

Regex library meant for regular expression processing re is imported. Newline are removed, whitespaces are removed.

nltk.sent_tokenize is used to tokenize the document into sentences. In last line output , indicates that the documents are sentenced.

Step-4: Removing stop words from the documents and making it into a numpy are to make it easy for vectorization purpose:

```
import numpy as np

stop_words = nltk.corpus.stopwords.words('english')

def normalize_document(doc):
    #remove special characters\whitespaces
    doc = re.sub(r'[^a-zA-Z\s]', '', doc, re.I|re.A)
    #make into lowercase
    doc = doc.lower()
    doc = doc.strip()
    # tokenize document
    tokens = nltk.word_tokenize(doc)
    # filter stopwords out of document
    filtered_tokens = [token for token in tokens if token not in st
op_words]
    # re-create document from filtered tokens
    doc = ' '.join(filtered_tokens)
    return doc
#making into numpy array for further processing
normalize_corpus = np.vectorize(normalize_document)

norm_sentences = normalize_corpus(sentences)
norm_sentences[:4]
```

Output:

```
array(['visible improvement delhis air quality recorded sunday although poor
category citys environment minister gopal rai said government submit lockdown
proposal supreme court monday reduce pollution',
       'national capital recorded hour average air quality index aqi sunday
previous day emissions farm fires haryana punjab dropped significantly',
       'aqi friday worst season far',
       'air quality index neighbouring ghaziabad gurgaon noida faridabad greater
noida recorded respectively'],
      dtype='<U359')
```

Normalize document function is created to remove special characters, make into lowercase, tokenize based on word to remove the stop words. Then join it again to a sentence.

Numpy array is used to make an array of sentences for further processing.

Module-2 Graph construction:

Step-1 vectorization:

We will be vectorizing our normalized sentences using the TF-IDF feature engineering scheme. We keep things simple and don't filter out any words based on document frequency. But feel free to try that out and maybe even leverage n-grams as features.

```
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd

tv = TfidfVectorizer(min_df=0., max_df=1., use_idf=True)
dt_matrix = tv.fit_transform(norm_sentences)
dt_matrix = dt_matrix.toarray()

vocab = tv.get_feature_names()
td_matrix = dt_matrix.T
print(td_matrix.shape)
pd.DataFrame(np.round(td_matrix, 2), index=vocab)
```

Output:

index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
able	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0	0.0
admitted	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.29	0.0
adopted	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aid	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.24	0.0	0.0
ailments	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.24	0.0	0.0
air	0.18	0.2	0.0	0.23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
airport	0.0	0.0	0.0	0.0	0.0	0.47	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
already	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
also	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.32
although	0.23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
amendment	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.28	0.0	0.0	0.0	0.0	0.0	0.0
announced	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
another	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
apart	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aqi	0.0	0.2	0.37	0.0	0.26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Step-2 : Semantic Similarity matrix

- Compute a document similarity matrix by multiplying the matrix by its transpose.

Numpy python module has matrix multiplication function to which tdidf vectorized matrix and its transpose is passed as arguments.

```
similarity_matrix = np.matmul(dt_matrix, dt_matrix.T)
print(similarity_matrix.shape)
np.round(similarity_matrix, 3)
```

Output:

```
(25, 25)
array([[1.   , 0.152, 0.   , 0.125, 0.113, 0.029, 0.557, 0.183, 0.043,
        0.   , 0.035, 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.152, 1.   , 0.074, 0.192, 0.051, 0.   , 0.   , 0.105, 0.056,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.   , 0.074, 1.   , 0.   , 0.096, 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.125, 0.192, 0.   , 1.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.052, 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.047, 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.113, 0.051, 0.096, 0.   , 1.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.029, 0.   , 0.   , 0.   , 0.   , 1.   , 0.037, 0.039, 0.021,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.051, 0.033, 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.557, 0.   , 0.   , 0.   , 0.   , 0.037, 1.   , 0.233, 0.09 ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.183, 0.105, 0.   , 0.   , 0.   , 0.039, 0.233, 1.   , 0.068,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.065, 0.041, 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.043, 0.056, 0.   , 0.   , 0.   , 0.021, 0.09 , 0.068, 1.   ,
        0.053, 0.088, 0.   , 0.017, 0.021, 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.032, 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.053,
        1.   , 0.174, 0.054, 0.077, 0.036, 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.035, 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.088,
        0.174, 1.   , 0.178, 0.022, 0.026, 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.054, 0.178, 1.   , 0.   , 0.14 , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.017,
        0.077, 0.022, 0.   , 1.   , 0.031, 0.   , 0.044, 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.021,
        0.036, 0.026, 0.14 , 0.031, 1.   , 0.125, 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ],
       [0.   , 0.   , 0.   , 0.052, 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.   , 0.125, 1.   , 0.034, 0.027, 0.017,
        0.02 , 0.028, 0.064, 0.036, 0.017, 0.078, 0.038],
       [0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
        0.   , 0.   , 0.   , 0.044, 0.   , 0.034, 1.   , 0.109, 0.03 ,
        0.035, 0.048, 0.079, 0.103, 0.03 , 0.243, 0.169],
       [0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   , 0.   ,
```

```
 0.  , 0.  , 0.  , 0.  , 0.  , 0.027, 0.109, 1.  , 0.023,
 0.225, 0.038, 0.132, 0.113, 0.193, 0.107, 0.052],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.017, 0.03 , 0.023, 1.  ,
 0.018, 0.024, 0.015, 0.031, 0.054, 0.175, 0.075],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.02 , 0.035, 0.225, 0.018,
 1.  , 0.028, 0.214, 0.094, 0.257, 0.08 , 0.038],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.051, 0.  , 0.065, 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.028, 0.048, 0.038, 0.024,
 0.028, 1.  , 0.079, 0.05 , 0.024, 0.251, 0.121],
[0.  , 0.  , 0.  , 0.047, 0.  , 0.033, 0.  , 0.041, 0.032,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.064, 0.079, 0.132, 0.015,
 0.214, 0.079, 1.  , 0.032, 0.085, 0.07 , 0.034],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.036, 0.103, 0.113, 0.031,
 0.094, 0.05 , 0.032, 1.  , 0.031, 0.141, 0.068],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.017, 0.03 , 0.193, 0.054,
 0.257, 0.024, 0.085, 0.031, 1.  , 0.069, 0.076],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.078, 0.243, 0.107, 0.175,
 0.08 , 0.251, 0.07 , 0.141, 0.069, 1.  , 0.246],
[0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.  ,
 0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.038, 0.169, 0.052, 0.075,
 0.038, 0.121, 0.034, 0.068, 0.076, 0.246, 1.  ]])
```

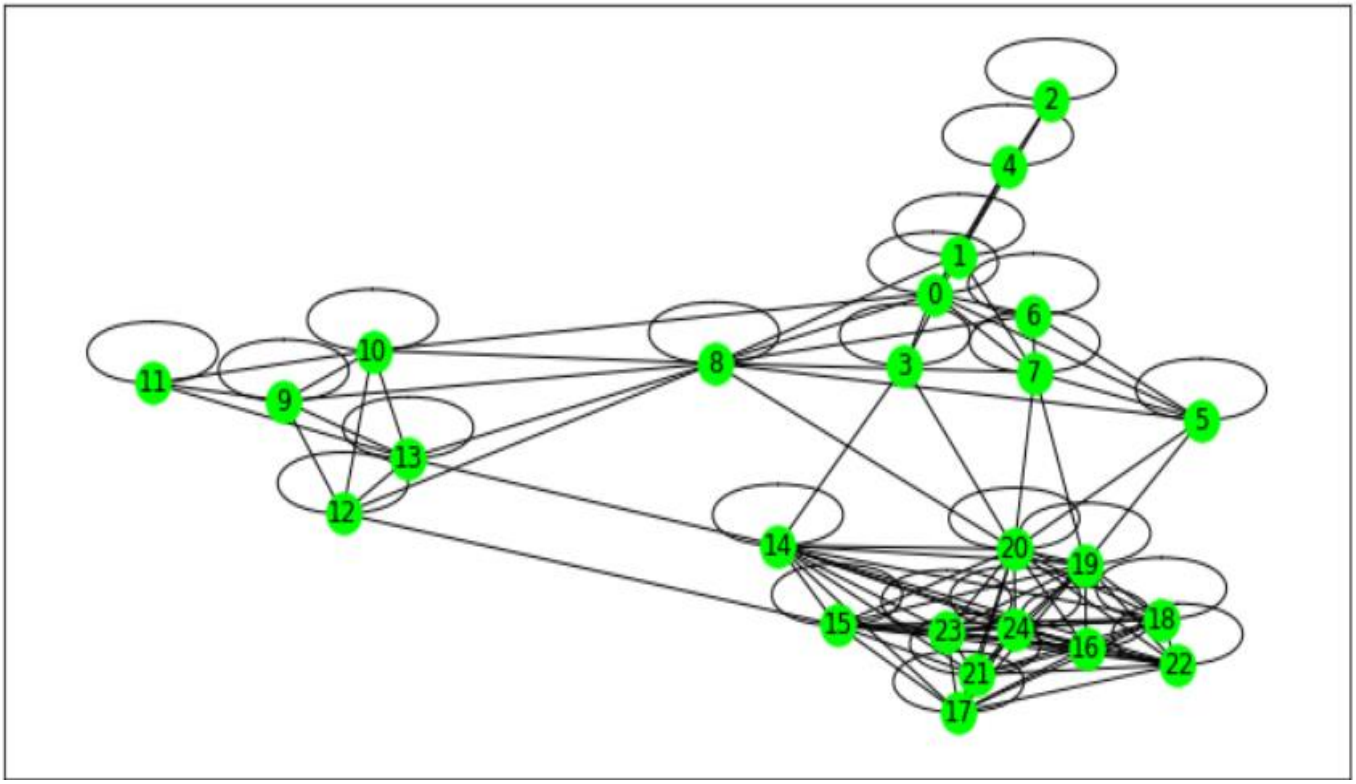
Step-3 semantic similarity Graph:

```
import networkx

similarity_graph = networkx.from_numpy_array(similarity_matrix)
similarity_graph

import matplotlib.pyplot as plt
%matplotlib inline

plt.figure(figsize=(12,12))
networkx.draw_networkx(similarity_graph, node_color='lime')
```



Module-3:Ranking algorithms:

```
scores = networkx.pagerank(similarity_graph)
ranked_sentences = sorted(((score, index) for index, score
                                                in scores.items()),
                          reverse=True)
ranked_sentences[:20]
```

Output:

```
[(0.050945782511250085, 23),
 (0.050900845134889634, 0),
 (0.043580471810056434, 6),
 (0.04269699685279467, 16),
 (0.04260678150553241, 20),
 (0.042527541930082534, 18),
 (0.04162906733430659, 15),
 (0.04114402722540945, 24),
 (0.04112474692403288, 10),
 (0.04103534753917604, 1),
 (0.040558319467001336, 7),
 (0.03984092587850705, 19),
 (0.03964281911991289, 22),
 (0.039084043438304156, 9),
 (0.038540659135477384, 11),
 (0.03829541229344317, 8),
 (0.03812770097996277, 13),
 (0.03759480882574738, 21),
 (0.03728044524253565, 14),
 (0.03631889674420968, 2)]
```

Ranking of the sentence is shown in percentage along with its sentence number starting from highest lowest.

Module-4:

Summary generation:

```
num_sentences=7;
top_sentence_indices = [ranked_sentences[index][1]
                        for index in range(num_sentences)]
top_sentence_indices.sort()

print('\n'.join(np.array(sentences)[top_sentence_indices]))
```

Output:

```
A visible improvement in Delhi's air quality was recorded on Sunday although it was
in the ''very poor'' category while the city's Environment Minister Gopal Rai said
his government will submit a lockdown proposal to the Supreme Court on Monday to
reduce pollution further.
Delhi Environment Minister Gopal Rai said the city government will on Monday submit
to the Supreme Court a proposal on clamping a lockdown and its modalities.
Health experts have raised concerns over the growing cases of vitamin D deficiency
in general population as people are staying at home and are not able to obtain the
'sunshine vitamin' from natural sunlight.
In fact, a new study discovered over 80% COVID-19 patients suffering from vitamin D
deficiency.
The study that was published in 'The Journal of Clinical Endocrinology &
Metabolism', found 80 percent of 216 COVID-19 patients admitted in a hospital in
Spain to be vitamin D-deficient.
"Vitamin D-deficient COVID-19 patients had a greater prevalence of hypertension and
cardiovascular diseases, raised serum ferritin and troponin levels, as well as a
longer length of hospital stay.
Vitamin D is also known as 'sunshine vitamin'.
```

Conclusion:

The proposed approach assumes semantic structure of sentence - predicate argument structure as graph node, and establish semantic relationships between PASs using Jiang semantic similarity measure. The

semantic similarity measures assists in detecting redundancy by capturing semantically equivalent predicate argument structures. The proposed Graph based approach incorporates PAS-to-PAS semantic similarity and PAS-to-document set relationship into the graph-based ranking algorithm, and experimental results demonstrate that modified ranking algorithm improves summarization results. The approach is promising enough to be applicable to any domain and does not require any intervention of human experts.

References:

1.Genetic Semantic Graph Approach for Multi- document Abstractive Summarization(October 2015)Conference: Fifth International Conference on Digital Information Processing and Communications (ICDIPC), 2015At: Sierre, Switzerland

Authors: Atif khan, Naomie salim, Yogan jaya kumar.

2. Extraction based summarization using a shortest path algorithm(February 2006)

Conference: Proceedings of the 12th Annual Natural Language Processing Conference NLP2006

Authors:Jonas Sjobergh, Kenji Araki .