In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Corpus Creation

In [2]:
```
df = pd.read_csv("Dataset/Articles.csv")
data = df["content"].values

print("Number of Articles : ", len(data))
```

```
Number of Articles :  1303
```

Pre-Processing

In [3]:
```
import re
from nltk.tokenize import sent_tokenize

def pre_processing(text):

    # text to sentence
    tokenized = sent_tokenize(text)

    # Remove Punctuation
    # Lower Case
    # Strip White Spaces
    pattern   = re.compile(r'[^a-zA-Z0-9\s]')
    tokenized = [pattern.sub('', sent).strip().lower() for sent in tokenized]

    return tokenized

corpus = []
for doc in data:
    corpus.extend(pre_processing(doc))

print("Number of Sentences in Corpus : ", len(corpus))
```

```
Number of Sentences in Corpus :  31603
```

Pre-Processing on Input Text

In [4]:
```
input_text = """
Millions go missing at China bank.

Two senior officials at one of China's top commercial banks have reportedly disappeared after funds
worth up to $120m (£64m) went missing. The pair both worked at Bank of China in the northern city
of Harbin, the South China Morning Post reported.

The latest scandal at Bank of China will do nothing to reassure foreign investors that China's
big four banks are ready for international listings. Government policy sees the bank listings as vital
economic reforms. Bank of China is one of two frontrunners in the race to list overseas. The other is
China Construction Bank. Both are expected to list abroad during 2005. They shared a $45bn state
bailout in 2003, to help clean up their balance sheets in preparation for a foreign stock market debut.

However, a report in the China-published Economic Observer said on Monday that the two banks may
have scrapped plans to list in New York because of the cost of meeting regulatory requirements
imposed since the Enron scandal. Bank of China is the country's biggest foreign exchange dealer,
while China Construction Bank is the largest deposit holder.

China's banking sector is burdened with at least $190bn of bad debt according to official data,
though most observers believe the true figure is far higher. Officially, one in five loans is not
being repaid. Attempts to strengthen internal controls and tighten lending policies have uncovered
a succession of scandals involving embezzlement by bank officials and loans-for-favours. The most
high-profile case involved the ex-president of Bank of China, Wang Xuebing, jailed for 12 years in
2003. Another high-profile fraud, China Life, is facing shareholder lawsuits and a probe by the
US Securities and Exchange Commission following its 2004 New York listing over its failure to
disclose accounting irregularities at its parent company.
"""

input_text = input_text.replace("\n" , " ")
sentences = sent_tokenize(input_text)
input_tok = pre_processing(input_text)
```

ROUGE Score

In [5]:
```
from rouge_scorer import rouge_scorer

expected = """
The other is China Construction Bank. The latest scandal at Bank of China will do nothing to reassure
foreign investors that China's big four banks are ready for international listings.
Bank of China is the country's biggest foreign exchange dealer, while China Construction Bank is the
largest deposit holder. Bank of China is one of two frontrunners in the race to list overseas.
Although he committed the offences whilst running Bank of China in New York, Mr. Wang was head of
China Construction Bank when the scandal broke. Earlier this month, a China Construction Bank branch
manager was jailed for life in a separate case. The pair both worked at Bank of China in the
northern city of Harbin, the South China Morning Post reported. The most high-profile case involved
the ex-president of Bank of China, Wang Xuebing, jailed for 12 years in 2003.
Two senior officials at one of China's top commercial banks have reportedly disappeared after funds
worth up to $120m (£64m) went missing. China's banks used to act as cash offices for state
enterprises and did not require checks on credit worthiness.
"""

expected = expected.replace("\n", " ").strip()

def rouge_metrics(summary):

    scorer = rouge_scorer.RougeScorer(['rouge1'], use_stemmer=True)
    scores = scorer.score(summary, expected)

    print("Rouge Score : ", scores, end="\n\n")
```

Summarize Function

In [6]:
```
from sklearn.metrics.pairwise import cosine_similarity
import networkx as nx

def summarize(input_vec):
    # Cosine Similarity
    similarity_matrix = cosine_similarity(input_vec, input_vec)

    # Matrix to Graph
    G = nx.from_numpy_array(similarity_matrix)

    # PageRank Algorithm
    pagerank_scores = nx.pagerank(G)

    # Sort sentences based on PageRank Scores
    sorted_sentences = sorted(pagerank_scores, key=pagerank_scores.get, reverse=True)

    # Select top 10
    top_k = 10
    summary = [sentences[i] for i in sorted_sentences[:top_k]]

    rouge_metrics(" ".join(summary))
    print(" ".join(summary))
```

Vectorization

Bag of Words

In [7]:
```
from sklearn.feature_extraction.text import CountVectorizer

bag_of_words = CountVectorizer()

corpus_bow = bag_of_words.fit_transform(corpus)
input_bow  = bag_of_words.transform(input_tok)
```

In [8]:
```
summarize(input_bow)
```

```
Rouge Score :  {'rouge1': Score(precision=0.7905759162303665, recall=0.6894977168949772, fmeasure=0.7365853658536586)}

Bank of China is one of two frontrunners in the race to list overseas. The pair both worked at Bank of China in the northern city of Harbin, the South China Morning Post reported. The most hi
gh-profile case involved the ex-president of Bank of China, Wang Xuebing, jailed for 12 years in 2003. Although he committed the offences whilst running Bank of China in New York, Mr. Wang wa
s head of China Construction Bank when the scandal broke. Bank of China is the country's biggest foreign exchange dealer, while China Construction Bank is the largest deposit holder. However,
 a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements imposed
 since the Enron scandal. The latest scandal at Bank of China will do nothing to reassure foreign investors that China's big four banks are ready for international listings. The other is China
 Construction Bank. China's leaders see reforming the top four banks as vital to distribute capital to profitable companies and protect the health of China's economic boom. The introduction of
 market reforms has been accompanied by attempts to modernize the banking sector, but links between banks and local government remain strong.
```

TF - IDF

In [9]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

tf_idf = TfidfVectorizer()

corpus_idf = tf_idf.fit_transform(corpus)
input_idf = tf_idf.transform(input_tok)
```

In [10]:
```
summarize(input_idf)
```

```
Rouge Score :  {'rouge1': Score(precision=0.8481675392670157, recall=0.9, fmeasure=0.8733153638814016)}

Bank of China is one of two frontrunners in the race to list overseas. The other is China Construction Bank. The latest scandal at Bank of China will do nothing to reassure foreign investors
 that China's big four banks are ready for international listings. Bank of China is the country's biggest foreign exchange dealer, while China Construction Bank is the largest deposit holder.
 Although he committed the offences whilst running Bank of China in New York, Mr. Wang was head of China Construction Bank when the scandal broke.  Millions go missing at China bank. The pair
 both worked at Bank of China in the northern city of Harbin, the South China Morning Post reported. The most high-profile case involved the ex-president of Bank of China, Wang Xuebing, jailed
 for 12 years in 2003. Earlier this month, a China Construction Bank branch manager was jailed for life in a separate case. China's leaders see reforming the top four banks as vital to distrib
ute capital to profitable companies and protect the health of China's economic boom.
```

Continuous Bag of Words

In [11]:
```
from gensim.models import Word2Vec
from nltk.tokenize import word_tokenize

g_model = Word2Vec(sentences=[word_tokenize(sent) for sent in corpus], vector_size=200, window=5, workers=5, epochs=500)
```

In [12]:
```
def get_embeddings(sent_1):
    vec = np.array([g_model.wv[word] if word in g_model.wv else np.zeros((200)) for word in sent_1])
    vec = vec.sum(axis=0)
    return vec

input_cbow = np.array([get_embeddings(sent) for sent in [word_tokenize(sent) for sent in input_tok]])
```

In [13]:
```
summarize(input_cbow)
```

```
Rouge Score :  {'rouge1': Score(precision=0.7382198952879581, recall=0.6077586206896551, fmeasure=0.6666666666666666)}

Bank of China is one of two frontrunners in the race to list overseas. The latest scandal at Bank of China will do nothing to reassure foreign investors that China's big four banks are ready
 for international listings. China's leaders see reforming the top four banks as vital to distribute capital to profitable companies and protect the health of China's economic boom. Bank of Ch
ina is the country's biggest foreign exchange dealer, while China Construction Bank is the largest deposit holder. The other is China Construction Bank. However, a report in the China-publish
ed Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements imposed since the Enron scandal. The
 introduction of market reforms has been accompanied by attempts to modernize the banking sector, but links between banks and local government remain strong. The most high-profile case involve
d the ex-president of Bank of China, Wang Xuebing, jailed for 12 years in 2003. Although he committed the offences whilst running Bank of China in New York, Mr. Wang was head of China Constru
ction Bank when the scandal broke. Last year, China's premier, Wen Jiabao, targeted bank lending practices in a series of speeches, and regulators ordered all big loans to be scrutinized, in
 an attempt to cool down irresponsible lending.
```

Skip gram

In [14]:
```
from gensim.models import Word2Vec
from nltk.tokenize import word_tokenize

g_model = Word2Vec(sentences=[word_tokenize(sent) for sent in corpus], vector_size=200, window=5, workers=5, epochs=500, sg=1)
```

In [15]:
```
def get_embeddings(sent_1):
    vec = np.array([g_model.wv[word] if word in g_model.wv else np.zeros((200)) for word in sent_1])
    vec = vec.sum(axis=0)
    return vec

input_sg = np.array([get_embeddings(sent) for sent in [word_tokenize(sent) for sent in input_tok]])
```

In [16]:
```
summarize(input_sg)
```

```
Rouge Score :  {'rouge1': Score(precision=0.7853403141361257, recall=0.5555555555555556, fmeasure=0.6507592190889372)}

The latest scandal at Bank of China will do nothing to reassure foreign investors that China's big four banks are ready for international listings. Although he committed the offences whilst r
unning Bank of China in New York, Mr. Wang was head of China Construction Bank when the scandal broke. Last year, China's premier, Wen Jiabao, targeted bank lending practices in a series of s
peeches, and regulators ordered all big loans to be scrutinized, in an attempt to cool down irresponsible lending. Two senior officials at one of China's top commercial banks have reportedly
 disappeared after funds worth up to $120m (£64m) went missing. However, a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in
 New York because of the cost of meeting regulatory requirements imposed since the Enron scandal. Bank of China is the country's biggest foreign exchange dealer, while China Construction Bank
 is the largest deposit holder. Another high-profile financial firm, China Life, is facing shareholder lawsuits and a probe by the US Securities and Exchange Commission following its 2004 New
 York listing over its failure to disclose accounting irregularities at its parent company. China's banking sector is burdened with at least $190bn of bad debt according to official data, thou
gh most observers believe the true figure is far higher. The introduction of market reforms has been accompanied by attempts to modernize the banking sector, but links between banks and local
 government remain strong. Earlier this month, a China Construction Bank branch manager was jailed for life in a separate case.
```

Word2Vec - PreTrained Embeddings

In [17]:
```
import gensim.downloader as api

model = api.load("glove-wiki-gigaword-200")

def get_embeddings(sent_1):
    vec = np.array([model[word] if word in model else np.zeros((200)) for word in sent_1])
    vec = vec.sum(axis=0)
    return vec

input_wv = np.array([get_embeddings(sent) for sent in [word_tokenize(sent) for sent in input_tok]])
```

In [18]:
```
summarize(input_wv)
```

```
Rouge Score :  {'rouge1': Score(precision=0.6544502617801047, recall=0.4578754578754586, fmeasure=0.5387931034482758)}

However, a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements
 imposed since the Enron scandal. The latest scandal at Bank of China will do nothing to reassure foreign investors that China's big four banks are ready for international listings. The introd
uction of market reforms has been accompanied by attempts to modernize the banking sector, but links between banks and local government remain strong. Bank of China is one of two frontrunners
 in the race to list overseas. China's banking sector is burdened with at least $190bn of bad debt according to official data, though most observers believe the true figure is far higher. Last
 year, China's premier, Wen Jiabao, targeted bank lending practices in a series of speeches, and regulators ordered all big loans to be scrutinized, in an attempt to cool down irresponsible le
nding. Although he committed the offences whilst running Bank of China in New York, Mr. Wang was head of China Construction Bank when the scandal broke. Another high-profile financial firm, C
hina Life, is facing shareholder lawsuits and a probe by the US Securities and Exchange Commission following its 2004 New York listing over its failure to disclose accounting irregularities a
t its parent company. They shared a $45bn state bailout in 2003, to help clean up their balance sheets in preparation for a foreign stock market debut. China's leaders see reforming the top f
our banks as vital to distribute capital to profitable companies and protect the health of China's economic boom.
```

GloVe

In [19]:
```
from gensim.scripts.glove2word2vec import glove2word2vec
from gensim.models import KeyedVectors

model = KeyedVectors.load_word2vec_format("GloVe/glove.6B.50d.word2vec", binary=False)

def get_embeddings(sent_1):
    vec = np.array([model[word] if word in model else np.zeros((50)) for word in sent_1])
    vec = vec.sum(axis=0)
    return vec

input_glove = np.array([get_embeddings(sent) for sent in [word_tokenize(sent) for sent in input_tok]])
```

In [20]:
```
summarize(input_glove)
```

```
Rouge Score :  {'rouge1': Score(precision=0.6858638743455497, recall=0.4962121212121212, fmeasure=0.5758241758241759)}

However, a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements
 imposed since the Enron scandal. The introduction of market reforms has been accompanied by attempts to modernize the banking sector, but links between banks and local government remain stron
g. Last year, China's premier, Wen Jiabao, targeted bank lending practices in a series of speeches, and regulators ordered all big loans to be scrutinized, in an attempt to cool down irrespon
sible lending. They shared a $45bn state bailout in 2003, to help clean up their balance sheets in preparation for a foreign stock market debut. The latest scandal at Bank of China will do no
thing to reassure foreign investors that China's big four banks are ready for international listings. Bank of China is one of two frontrunners in the race to list overseas. China's banking se
ctor is burdened with at least $190bn of bad debt according to official data, though most observers believe the true figure is far higher. Another high-profile financial firm, China Life, is
 facing shareholder lawsuits and a probe by the US Securities and Exchange Commission following its 2004 New York listing over its failure to disclose accounting irregularities at its parent c
ompany. Although he committed the offences whilst running Bank of China in New York, Mr. Wang was head of China Construction Bank when the scandal broke. Earlier this month, a China Construct
ion Bank branch manager was jailed for life in a separate case.
```

FastText

In [21]:
```
from gensim.models import FastText
from nltk.tokenize import word_tokenize

f_model = FastText(sentences=[word_tokenize(sent) for sent in corpus], vector_size=200, window=5, workers=5, epochs=500)
```

In [22]:
```
def get_embeddings(sent_1):
    vec = np.array([f_model.wv[word] if word in f_model.wv else np.zeros((200)) for word in sent_1])
    vec = vec.sum(axis=0)
    return vec

input_ft = np.array([get_embeddings(sent) for sent in [word_tokenize(sent) for sent in input_tok]])
```

In [23]:
```
summarize(input_ft)
```

```
Rouge Score :  {'rouge1': Score(precision=0.7120418848167539, recall=0.5551020408163265, fmeasure=0.6238532110091743)}

Attempts to strengthen internal controls and tighten lending policies have uncovered a succession of scandals involving embezzlement by bank officials and loans-for-favours. China's leaders s
ee reforming the top four banks as vital to distribute capital to profitable companies and protect the health of China's economic boom. The introduction of market reforms has been accompanied
 by attempts to modernize the banking sector, but links between banks and local government remain strong. However, a report in the China-published Economic Observer said on Monday that the two
 banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements imposed since the Enron scandal. Bank of China is one of two frontrunners in the race
 to list overseas. The latest scandal at Bank of China will do nothing to reassure foreign investors that China's big four banks are ready for international listings. Last year, China's premie
r, Wen Jiabao, targeted bank lending practices in a series of speeches, and regulators ordered all big loans to be scrutinized, in an attempt to cool down irresponsible lending. The most high
-profile case involved the ex-president of Bank of China, Wang Xuebing, jailed for 12 years in 2003. The pair both worked at Bank of China in the northern city of Harbin, the South China Morn
ing Post reported. Bank of China is the country's biggest foreign exchange dealer, while China Construction Bank is the largest deposit holder.
```