In [1]:

```
#!/usr/bin/env python
# coding: utf-8
import nltk
from nltk.corpus import stopwords
from nltk.cluster.util import cosine_distance
import numpy as np
import networkx as nx

def read_article(file_name):
    file = open(file_name, "r")
    filedata = file.readlines()
    article = filedata[0].split(". ")
    sentences = []

    for sentence in article:
        print(sentence)
        sentences.append(sentence.replace("[^a-zA-Z]", " ").split(" "))
    sentences.pop()

    return sentences

def sentence_similarity(sent1, sent2, stopwords=None):
    if stopwords is None:
        stopwords = []

    sent1 = [w.lower() for w in sent1]
    sent2 = [w.lower() for w in sent2]

    all_words = list(set(sent1 + sent2))

    vector1 = [0] * len(all_words)
    vector2 = [0] * len(all_words)

    # build the vector for the first sentence
    for w in sent1:
        if w in stopwords:
            continue
        vector1[all_words.index(w)] += 1

    # build the vector for the second sentence
    for w in sent2:
        if w in stopwords:
            continue
        vector2[all_words.index(w)] += 1

    return 1 - cosine_distance(vector1, vector2)

def build_similarity_matrix(sentences, stop_words):
    # Create an empty similarity matrix
    similarity_matrix = np.zeros((len(sentences), len(sentences)))

    for idx1 in range(len(sentences)):
        for idx2 in range(len(sentences)):
            if idx1 == idx2: #ignore if both are same sentences
                continue
            similarity_matrix[idx1][idx2] = sentence_similarity(sentences[idx1], sente

    return similarity_matrix


def generate_summary(file_name, top_n):
#     nltk.download("stopwords")
    stop_words = stopwords.words('english')
    summarize_text = []

    # Step 1 - Read text anc split it
    sentences =  read_article(file_name)
```

```
    # Step 2 - Generate Similary Martix across sentences
    sentence_similarity_martix = build_similarity_matrix(sentences, stop_words)

    # Step 3 - Rank sentences in similarity martix
    sentence_similarity_graph = nx.from_numpy_array(sentence_similarity_martix)
    scores = nx.pagerank(sentence_similarity_graph)

    # Step 4 - Sort the rank and pick top sentences
    ranked_sentence = sorted(((scores[i],s) for i,s in enumerate(sentences)), reverse=
    print("Indexes of top ranked_sentence order are ", ranked_sentence)

    for i in range(top_n):
        summarize_text.append(" ".join(ranked_sentence[i][1]))

    # Step 5 - Offcourse, output the summarize text
    print("Summarize Text: \n", ". ".join(summarize_text))
```

In [2]:
```
sample = " The September 11 attacks, commonly known as 9/11, were four coordinated Isl
sample
```

Out[2]:
```
" The September 11 attacks, commonly known as 9/11, were four coordinated Islamist s
uicide terrorist attacks carried out by Al-Qaeda against the United States on Septem
ber 11, 2001. That morning, 19 terrorists hijacked four commercial airliners schedul
ed to travel from the East Coast to California. The hijackers crashed the first two
planes into the Twin Towers of the World Trade Center in New York City, two of the w
orld's five tallest buildings at the time, and aimed the next two flights toward tar
gets in or near Washington, D.C., in an attack on the nation's capital. The third te
am succeeded in striking the Pentagon, the headquarters of the U.S. Department of De
fense in Arlington County, Virginia, while the fourth plane crashed in rural Pennsyl
vania during a passenger revolt. The September 11 attacks killed 2,977 people, makin
g them the deadliest terrorist attack in history, and instigated the multi-decade gl
obal war on terror, fought in Afghanistan, Iraq, and elsewhere."
```

In [3]:
```
generate_summary("sample1.txt", 1)
```

```
The September 11 attacks, commonly known as 9/11, were four coordinated Islamist sui
cide terrorist attacks carried out by Al-Qaeda against the United States on Septembe
r 11, 2001
That morning, 19 terrorists hijacked four commercial airliners scheduled to travel f
rom the East Coast to California
The hijackers crashed the first two planes into the Twin Towers of the World Trade C
enter in New York City, two of the world's five tallest buildings at the time, and a
imed the next two flights toward targets in or near Washington, D.C., in an attack o
n the nation's capital
The third team succeeded in striking the Pentagon, the headquarters of the U.S
Department of Defense in Arlington County, Virginia, while the fourth plane crashed
in rural Pennsylvania during a passenger revolt
The September 11 attacks killed 2,977 people, making them the deadliest terrorist at
tack in history, and instigated the multi-decade global war on terror, fought in Afg
hanistan, Iraq, and elsewhere.
Indexes of top ranked_sentence order are  [(0.2409638268462948, ['The', 'hijackers',
'crashed', 'the', 'first', 'two', 'planes', 'into', 'the', 'Twin', 'Towers', 'of',
'the', 'World', 'Trade', 'Center', 'in', 'New', 'York', 'City,', 'two', 'of', 'the',
"world's", 'five', 'tallest', 'buildings', 'at', 'the', 'time,', 'and', 'aimed', 'th
e', 'next', 'two', 'flights', 'toward', 'targets', 'in', 'or', 'near', 'Washingto
n,', 'D.C.,', 'in', 'an', 'attack', 'on', 'the', "nation's", 'capital']), (0.2409638
268462948, ['The', 'September', '11', 'attacks,', 'commonly', 'known', 'as', '9/1
1,', 'were', 'four', 'coordinated', 'Islamist', 'suicide', 'terrorist', 'attacks',
'carried', 'out', 'by', 'Al-Qaeda', 'against', 'the', 'United', 'States', 'on', 'Sep
tember', '11,', '2001']), (0.2409638268462948, ['That', 'morning,', '19', 'terrorist
s', 'hijacked', 'four', 'commercial', 'airliners', 'scheduled', 'to', 'travel', 'fro
m', 'the', 'East', 'Coast', 'to', 'California']), (0.2409638268462948, ['Departmen
t', 'of', 'Defense', 'in', 'Arlington', 'County,', 'Virginia,', 'while', 'the', 'fou
rth', 'plane', 'crashed', 'in', 'rural', 'Pennsylvania', 'during', 'a', 'passenger',
'revolt']), (0.03614469261482073, ['The', 'third', 'team', 'succeeded', 'in', 'strik
ing', 'the', 'Pentagon,', 'the', 'headquarters', 'of', 'the', 'U.S']), (0.03614469261482073, ['The', 'third', 'team', 'succeeded', 'in', 'strik
ing', 'the', 'Pentagon,', 'the', 'headquarters', 'of', 'the', 'U.S'])]
Summarize Text:
 The hijackers crashed the first two planes into the Twin Towers of the World Trade
Center in New York City, two of the world's five tallest buildings at the time, and
aimed the next two flights toward targets in or near Washington, D.C., in an attack
on the nation's capital
```

In [4]:
```
sent1 = "The September 11 attacks, commonly known as 9/11, were four coordinated Islar
sent2 = "The September 11 attacks, commonly known as 9/11, were four coordinated Islar
stop_words = stopwords.words('english')

import nltk
# nltk.download('stopwords')
# stop_words = stopwords.words('english')
sentence_similarity(sent1, sent2, stop_words)
```

Out[4]: 1.0000000000000002

In [5]:
```
s = read_article("sample1.txt")
```

```
The September 11 attacks, commonly known as 9/11, were four coordinated Islamist sui
cide terrorist attacks carried out by Al-Qaeda against the United States on Septembe
r 11, 2001
That morning, 19 terrorists hijacked four commercial airliners scheduled to travel f
rom the East Coast to California
The hijackers crashed the first two planes into the Twin Towers of the World Trade C
enter in New York City, two of the world's five tallest buildings at the time, and a
imed the next two flights toward targets in or near Washington, D.C., in an attack o
n the nation's capital
The third team succeeded in striking the Pentagon, the headquarters of the U.S
Department of Defense in Arlington County, Virginia, while the fourth plane crashed
in rural Pennsylvania during a passenger revolt
The September 11 attacks killed 2,977 people, making them the deadliest terrorist at
tack in history, and instigated the multi-decade global war on terror, fought in Afg
hanistan, Iraq, and elsewhere.
```

In [6]:
```
len(s)
```

Out[6]:
```
5
```

In [7]:
```
sm = build_similarity_matrix(s, stop_words)
sm
```

Out[7]:
```
array([[0.        , 0.06299408, 0.        , 0.        , 0.        ],
       [0.06299408, 0.        , 0.        , 0.        , 0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.0474579 ],
       [0.        , 0.        , 0.        , 0.        , 0.        ],
       [0.        , 0.        , 0.0474579 , 0.        , 0.        ]])
```

In [8]:
```
sentence_similarity_graph = nx.from_numpy_array(sm)
scores = nx.pagerank(sentence_similarity_graph)
scores
```

Out[8]:
```
{0: 0.2409638268462948,
 1: 0.2409638268462948,
 2: 0.2409638268462948,
 3: 0.03614469261482073,
 4: 0.2409638268462948}
```

In [9]:
```
print(sentence_similarity_graph)
```

```
Graph with 5 nodes and 2 edges
```

In [11]:
```
generate_summary("sample1.txt", 2)
```

```
The September 11 attacks, commonly known as 9/11, were four coordinated Islamist sui
cide terrorist attacks carried out by Al-Qaeda against the United States on Septembe
r 11, 2001
That morning, 19 terrorists hijacked four commercial airliners scheduled to travel f
rom the East Coast to California
The hijackers crashed the first two planes into the Twin Towers of the World Trade C
enter in New York City, two of the world's five tallest buildings at the time, and a
imed the next two flights toward targets in or near Washington, D.C., in an attack o
n the nation's capital
The third team succeeded in striking the Pentagon, the headquarters of the U.S
Department of Defense in Arlington County, Virginia, while the fourth plane crashed
in rural Pennsylvania during a passenger revolt
The September 11 attacks killed 2,977 people, making them the deadliest terrorist at
tack in history, and instigated the multi-decade global war on terror, fought in Afg
hanistan, Iraq, and elsewhere.
Indexes of top ranked_sentence order are  [(0.2409638268462948, ['The', 'hijackers',
'crashed', 'the', 'first', 'two', 'planes', 'into', 'the', 'Twin', 'Towers', 'of',
'the', 'World', 'Trade', 'Center', 'in', 'New', 'York', 'City,', 'two', 'of', 'the',
"world's", 'five', 'tallest', 'buildings', 'at', 'the', 'time,', 'and', 'aimed', 'th
e', 'next', 'two', 'flights', 'toward', 'targets', 'in', 'or', 'near', 'Washingto
n,', 'D.C.,', 'in', 'an', 'attack', 'on', 'the', "nation's", 'capital']), (0.2409638
268462948, ['The', 'September', '11', 'attacks,', 'commonly', 'known', 'as', '9/1
1,', 'were', 'four', 'coordinated', 'Islamist', 'suicide', 'terrorist', 'attacks',
'carried', 'out', 'by', 'Al-Qaeda', 'against', 'the', 'United', 'States', 'on', 'Sep
tember', '11,', '2001']), (0.2409638268462948, ['That', 'morning,', '19', 'terrorist
s', 'hijacked', 'four', 'commercial', 'airliners', 'scheduled', 'to', 'travel', 'fro
m', 'the', 'East', 'Coast', 'to', 'California']), (0.2409638268462948, ['Departmen
t', 'of', 'Defense', 'in', 'Arlington', 'County,', 'Virginia,', 'while', 'the', 'fou
rth', 'plane', 'crashed', 'in', 'rural', 'Pennsylvania', 'during', 'a', 'passenger',
'revolt']), (0.03614469261482073, ['The', 'third', 'team', 'succeeded', 'in', 'strik
ing', 'the', 'Pentagon,', 'the', 'headquarters', 'of', 'the', 'U.S'])]
Summarize Text:
 The hijackers crashed the first two planes into the Twin Towers of the World Trade
Center in New York City, two of the world's five tallest buildings at the time, and
aimed the next two flights toward targets in or near Washington, D.C., in an attack
on the nation's capital. The September 11 attacks, commonly known as 9/11, were four
coordinated Islamist suicide terrorist attacks carried out by Al-Qaeda against the U
nited States on September 11, 2001
```