```
import numpy as np
import pandas as pd
import nltk
from nltk.tokenize import sent_tokenize
nltk.download('punkt') # one time execution
import re
```

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

```
# Upload the CSV file
from google.colab import files
uploaded = files.upload()
```

Choose files tennis_articles_v4.csv

- **tennis_articles_v4.csv**(text/csv) - 13798 bytes, last modified: 14/12/2022 - 100% done

```
Saving tennis_articles_v4.csv to tennis_articles_v4.csv
```

```
# Read the CSV file
import io
df = pd.read_csv(io.StringIO(uploaded['tennis_articles_v4.csv'].decode("utf-8")))
```

```
df.head()
```

	article_id	article_text	source
0	1	Maria Sharapova has basically no friends as te...	https://www.tennisworldusa.org/tennis/news/Mar...
1	2	BASEL, Switzerland (AP), Roger Federer advance...	http://www.tennis.com/pro-game/2018/10/copil-s...
2	3	Roger Federer has revealed that organisers of ...	https://scroll.in/field/899938/tennis-roger-fe...
3	4	Kei Nishikori will try to end his	http://www.tennis.com/pro-

```
# split the the text in the articles into sentences
sentences = []
for s in df['article_text']:
  sentences.append(sent_tokenize(s))
```

```
# flatten the list
sentences = [y for x in sentences for y in x]
```

```
# remove punctuations, numbers and special characters
clean_sentences = pd.Series(sentences).str.replace("[^a-zA-Z]", " ")

# make alphabets lowercase
clean_sentences = [s.lower() for s in clean_sentences]
```

```
<ipython-input-7-57e05bf8eb2b>:2: FutureWarning: The default value of regex will change from True to False in a future ve
  clean_sentences = pd.Series(sentences).str.replace("[^a-zA-Z]", " ")
```

```
nltk.download('stopwords')# one time execution
```

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

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
```

```
# function to remove stopwords
def remove_stopwords(sen):
  sen_new = " ".join([i for i in sen if i not in stop_words])
  return sen_new
```

```
# remove stopwords from the sentences
clean_sentences = [remove_stopwords(r.split()) for r in clean_sentences]
```

```
# download pretrained GloVe word embeddings
! wget http://nlp.stanford.edu/data/glove.6B.zip
```

```
    --2023-04-12 17:19:40--  http://nlp.stanford.edu/data/glove.6B.zip
    Resolving nlp.stanford.edu (nlp.stanford.edu)... 171.64.67.140
    Connecting to nlp.stanford.edu (nlp.stanford.edu)|171.64.67.140|:80... connected.
    HTTP request sent, awaiting response... 302 Found
    Location: https://nlp.stanford.edu/data/glove.6B.zip [following]
    --2023-04-12 17:19:40--  https://nlp.stanford.edu/data/glove.6B.zip
    Connecting to nlp.stanford.edu (nlp.stanford.edu)|171.64.67.140|:443... connected.
    HTTP request sent, awaiting response... 301 Moved Permanently
    Location: https://downloads.cs.stanford.edu/nlp/data/glove.6B.zip [following]
    --2023-04-12 17:19:40--  https://downloads.cs.stanford.edu/nlp/data/glove.6B.zip
    Resolving downloads.cs.stanford.edu (downloads.cs.stanford.edu)... 171.64.64.22
    Connecting to downloads.cs.stanford.edu (downloads.cs.stanford.edu)|171.64.64.22|:443... connected.
    HTTP request sent, awaiting response... 200 OK
    Length: 862182613 (822M) [application/zip]
    Saving to: ‘glove.6B.zip’

    glove.6B.zip        100%[===================>] 822.24M  5.01MB/s    in 2m 39s

    2023-04-12 17:22:19 (5.18 MB/s) - ‘glove.6B.zip’ saved [862182613/862182613]
```

```
! unzip glove*.zip
```

```
    Archive:  glove.6B.zip
      inflating: glove.6B.50d.txt
      inflating: glove.6B.100d.txt
      inflating: glove.6B.200d.txt
      inflating: glove.6B.300d.txt
```

```
# Extract word vectors
word_embeddings = {}
f = open('glove.6B.100d.txt', encoding='utf-8')
for line in f:
    values = line.split()
    word = values[0]
    coefs = np.asarray(values[1:], dtype='float32')
    word_embeddings[word] = coefs
f.close()
```

```
sentence_vectors = []
for i in clean_sentences:
  if len(i) != 0:
    v = sum([word_embeddings.get(w, np.zeros((100,))) for w in i.split()])/(len(i.split())+0.001)
  else:
    v = np.zeros((100,))
  sentence_vectors.append(v)
```

```
len(sentence_vectors)
```

```
    119
```

The next step is to find similarities among the sentences. We will use cosine similarity to find similarity between a pair of sentences. Let's create an empty similarity matrix for this task and populate it with cosine similarities of the sentences.

```
# similarity matrix
sim_mat = np.zeros([len(sentences), len(sentences)])
```

```
from sklearn.metrics.pairwise import cosine_similarity
```

```
for i in range(len(sentences)):
  for j in range(len(sentences)):
    if i != j:
      sim_mat[i][j] = cosine_similarity(sentence_vectors[i].reshape(1,100), sentence_vectors[j].reshape(1,100))[0,0]
```

```
import networkx as nx

nx_graph = nx.from_numpy_array(sim_mat)
scores = nx.pagerank(nx_graph)
```

```
ranked_sentences = sorted(((scores[i],s) for i,s in enumerate(sentences)), reverse=True)
```

```
# Specify number of sentences to form the summary
sn = 10

# Generate summary
```

```
for i in range(sn):
  print(ranked_sentences[i][1])
```

```
When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether
Major players feel that a big event in late November combined with one in January before the Australian Open will mean tc
Speaking at the Swiss Indoors tournament where he will play in Sundays final against Romanian qualifier Marius Copil, the
"I felt like the best weeks that I had to get to know players when I was playing were the Fed Cup weeks or the Olympic we
Currently in ninth place, Nishikori with a win could move to within 125 points of the cut for the eight-man event in Lonc
He used his first break point to close out the first set before going up 3-0 in the second and wrapping up the win on his
The Spaniard broke Anderson twice in the second but didn't get another chance on the South African's serve in the final s
"We also had the impression that at this stage it might be better to play matches than to train.
The competition is set to feature 18 countries in the November 18-24 finals in Madrid next year, and will replace the cla
Federer said earlier this month in Shanghai in that his chances of playing the Davis Cup were all but non-existent.
```

✓ 0s completed at 10:53 PM