IT 458

Assignment 2

Name: Niraj Nandish

Roll no.: 191IT234

Import the corpus and required libraries

```
import string
import nltk
import re

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

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
urls = [
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/india-gate.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/raj-ghat.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/chhatarpur-mandir.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/bahai-lotus-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/connaught-place.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/chandni-chowk.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/akshardham.html",
```

"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/garden-of-five-senses.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/delhi/dilli-haat.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/cholamandal-artistsa-village.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/fort-st-george.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/krishnamacharya-yoga-mandiram.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/chettinad-cuisine.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/kalakshetra.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/filter-coffee.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/tanjore-paintings.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/chennai/kapaleeswarar-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/experiences/spiritual/buddhism.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/experiences/spiritual/jainism.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/experiences/spiritual/sikhism.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/experiences/spiritual/christianity.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/experiences/spiritual/islam.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/experiences/spiritual/hinduism.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/ranthambore-national-park.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/mandawa-fort.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/amber-fort.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/govind-devji-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/jantar-mantar.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/abhaneri.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/kanak-vrindavan.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/jaipur-wax-museum.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/samode-palace.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/jhunjhunu.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/jal-mahal.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/jaipur/hawa-mahal.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/elephanta-caves.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/chhatrapati-shivaji-maharaj-terminus.h
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/global-vipasana-pagoda.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/deolali.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/bhandardara.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/ganeshpuri-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/mahalaxmi-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/mumbai/haji-ali-mosque.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/sarnath.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/kaimoor-wildlife-sanctuary.html",

"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/ganga-ghats.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/dashashwamedh-ghats.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/vishwanath-khand.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/banaras-hindu-university.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/varanasi/ramnagar-fort.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/ahar.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/city-palace.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/dungarpur.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/bagore-ki-haveli.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/charbhuja.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/fateh-sagar-lake.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/ekling-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/udaipur/haldighati.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/haridwar/har-ki-pauri.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/haridwar/patanjali-yogapeeth.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/haridwar/daksh-mahadev-temple-and-sati-kund.h
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/haridwar/mansa-devi-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/haridwar/kumbh-mela.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/haridwar/chandi-devi-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/dakshineswar-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/vidyasagar-setu.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/academy-of-fine-arts.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/writers-building.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/victoria-memorial.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/alipore-zoological-garden.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/star-theatre.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/armenian-church.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/sovabazar-rajbari.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/race-course.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/bankura.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kolkata/belur-math.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/vidhana-soudha.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/bannerghatta-national-park.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/belur.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/bangalore-palace.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/halebeedu--halebidu-.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/br-hills.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/cubbon-park.html",

```
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/hesaraghatta-lake.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/bull-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/bengaluru/kolar-gold-fields.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/chowmahalla-palace.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/falaknuma-palace.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/hussain-sagar-lake.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/nagarjuna-sagar-dam.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/charminar.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/salar-jung-museum.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/birla-mandir.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/ananda-buddha-vihara.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/chilkur-balaji-temple.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kochi/fort-kochi.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/hyderabad/purani-haveli.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kochi/athirapally-waterfalls.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kochi/bishop-s-house.html",
"https://www.incredibleindia.org/content/incredible-india-v2/en/destinations/kochi/chinese-fishing-nets.html"
]

docs_list = []
```

```
import urllib.request
from bs4 import BeautifulSoup

for url in urls:
    html = urllib.request.urlopen(url)
    htmlParse = BeautifulSoup(html, 'html.parser')
    para = htmlParse.find("div", {"class": "read-more-hide"}).p.get_text()
    docs_list.append(para)
```

```
import copy
docs = copy.deepcopy(docs_list)
docs_list[1]
```

```
'Built in the memory of Mohandas Karamchand Gandhi, who is revered as the Father of the Nation, Raj Ghat is a marble
```

```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import word_tokenize

stemmer = PorterStemmer()

def text_lowercase(text):
    return text.lower()

def remove_non_alpha(text):
    res = ""
    for elem in text:
        if elem.isalnum() or elem == " ":
          res += elem
    return res

def remove_num(text):
    res = re.sub(r'\d+', '', text)
    return res

def remove_punc(text):
    trans = str.maketrans('', '', string.punctuation)
    return text.translate(trans)

def remove_whitespace(text):
    return  " ".join(text.split())

def remove_stopwords(text):
    stop_words = set(stopwords.words("english"))
    word_tokens = word_tokenize(text)
    filtered_text = [word for word in word_tokens if word not in stop_words]
    return filtered_text

def stem_words(text):
    word_tokens = word_tokenize(text)
```

```
    stems = [stemmer.stem(word) for word in word_tokens]
    return stems

def vocab_size():
    uniqueWords = set()
    for doc in docs:
        for word in doc.split():
            uniqueWords.add(word)
    return len(uniqueWords)
```

```
print("Initial Vocabulary size: " + str(vocab_size()))
```

```
Initial Vocabulary size: 5085
```

```
for i in range(len(docs)):
    docs[i] = text_lowercase(docs[i])

print("After text lowercase prepocessing: " + str(vocab_size()))
```

```
After text lowercase prepocessing: 4800
```

```
for i in range(len(docs)):
    docs[i] = remove_non_alpha(docs[i])

print("After removing non-alphanumeric characters: " + str(vocab_size()))
```

```
After removing non-alphanumeric characters: 4016
```

```
for i in range(len(docs)):
    docs[i] = remove_num(docs[i])

print("After removing numbers: " + str(vocab_size()))
```

```
After removing numbers: 3842
```

```
for i in range(len(docs)):
    docs[i] = remove_punc(docs[i])

print("After removing punctutaion marks: " + str(vocab_size()))
```

```
 After removing punctutaion marks: 3842
```

```
for i in range(len(docs)):
    docs[i] = remove_whitespace(docs[i])

print("After removing extra whitespace: " + str(vocab_size()))
```

```
 After removing extra whitespace: 3842
```

```
for i in range(len(docs)):
    docs[i] = remove_stopwords(docs[i])

vocab = set()
for doc in docs:
    for word in doc:
        vocab.add(word)

print("After removing stop words: " + str(len(vocab)))
```

```
 After removing stop words: 3735
```

```
# Dictionary to store index for each word
total_documents = len(docs)
vocab = set(vocab)
index_dict = {}

i = 0
for word in vocab:
```

```
    index_dict[word] = i
```

```
# Create a count dictionary
def count_dict(sentences):
    word_count = {}
    for word in vocab:
        word_count[word] = 0
        for sent in sentences:
            if word in sent:
                word_count[word] += 1
    return word_count

word_count = count_dict(docs)
```

```
# Term Frequency
def termfreq(document, word):
    total_length = len(document)
    occurance = len([token for token in document if token == word])
    return occurance/total_length
```

```
import numpy as np

def inverse_doc_freq(word):
    try:
        word_occurance = word_count[word] + 1
    except:
        word_occurance = 1
    return np.log(total_documents/word_occurance)
```

```
def tf_idf(sentence):
    tf_idf_vec = np.zeros((len(vocab),))
    for word in sentence:
        tf = termfreq(sentence,word)
        idf = inverse_doc_freq(word)
        value = tf*idf
```

```
        tf_idf_vec[index_dict[word]] = value
    return tf_idf_vec
```

```
query1 = "India Gate in Delhi"
query2 = "Temple in India"
query3 = "Caves in Maharashtra"
```

```
def doc_preprocess(query):
    query = text_lowercase(query)
    query = remove_non_alpha(query)
    query = remove_num(query)
    query = remove_punc(query)
    query = remove_whitespace(query)
    query = remove_stopwords(query)

    return query
```

Query 1 - India Gate in Delhi

```
from sklearn.metrics.pairwise import cosine_similarity

query1 = doc_preprocess(query1)
res = []
for doc in docs_list:
    doc_pre = doc_preprocess(doc)
    doc_vec = tf_idf(doc_pre)
    query_vec = tf_idf(query1)

    cos_sim = cosine_similarity(doc_vec.reshape(1,-1), query_vec.reshape(1,-1))
    res.append([doc,cos_sim])
```

```
res = sorted(res, key=lambda x:x[1], reverse = True)
```

```
import pandas as pd
res = pd.DataFrame(res)
```

```
for i in range(5):
    print("doc {}:".format(i+1) + res[0][i])
```

```
doc 1:A stunning archway standing as a tribute to the brave soldiers who sacrificed their lives for the country, Indi
doc 2:The heartbeat of Delhi, Connaught Place is a heritage neighbourhood that has been modeled after the Georgian st
doc 3:Preserving the antiquity of Old Delhi, Chandni Chowk is a bustling neighbourhood that is one of the oldest in t
doc 4:Rising from the bank of Lake Pichola, the imposing City Palace is a grand icon of Udaipur's and Rajasthan's roy

With a facade that is 244 m long and 30.4 m high, the palace complex has 11 smaller palaces. Construction of the pala

It has a maze of courtyards, pavilions, terraces, corridors, rooms and hanging gardens. Encircled by fortifications,
doc 5:India boasts some of the finest mosques and dargahs in the world, chief of which is the Jama Masjid, in Delhi.

Another popular site, echoing with spiritual fervour is the Hazratbal Shrine, located on the banks of the Dal Lake in

Ajmer Sharif, in Rajasthan, is believed to be the site where no prayer goes unanswered. The resting place of Gharib N

The majestic complex of Fatehpur Sikri, in Agra, houses the tomb of Hazrat Sheikh Salim Chishti, a great saint. Every

One of the more popular attractions of Mumbai, the Haji Ali complex houses the tomb of Muslim saint, Pir Haji Ali Sha

One of the most prominent spiritual sites in Delhi, Hazrat Nizamuddin Dargah is the shrine of Muslim sufi saint, Niza
```

Query 2 - Temple in India

```
from sklearn.metrics.pairwise import cosine_similarity

query2 = doc_preprocess(query2)
res2 = []
for doc in docs_list:
```

```
    doc_pre = doc_preprocess(doc)
    doc_vec = tf_idf(doc_pre)
    query_vec = tf_idf(query2)

    cos_sim = cosine_similarity(doc_vec.reshape(1,-1), query_vec.reshape(1,-1))
    res2.append([doc,cos_sim])
```

```
res2 = sorted(res2, key=lambda x:x[1], reverse = True)
```

```
import pandas as pd
res2 = pd.DataFrame(res2)
```

```
for i in range(5):
    print("doc {}:".format(i+1) + res2[0][i])
```

```
doc 1:The Chilkur Balaji Temple, located on the outskirts of Hyderabad, on the banks of the Osman Sagar Lake, is a pc

doc 2:Dedicated to Goddess Katyayani, an incarnation of Goddess Durga, Chhatarpur Temple is said to be the second-lar
doc 3:One of the most prominent shrines in the city, Dakshineswar Kali Temple is dedicated to Goddess Kali. Located c
doc 4:The Mahalaxmi Temple is an ancient shrine dedicated to Goddess Laxmi, Goddess Mahakali and Goddess Mahasaraswat

doc 5:Situated 22 km away from Udaipur, the Ekling Temple is one of the most famous pilgrimage sites in the state of
```

Query 3 - Caves in Maharashtra

```
from sklearn.metrics.pairwise import cosine_similarity

query3 = doc_preprocess(query3)
res3 = []
for doc in docs_list:
    doc_pre = doc_preprocess(doc)
    doc_vec = tf_idf(doc_pre)
    query_vec = tf_idf(query3)
```

```
    cos_sim = cosine_similarity(doc_vec.reshape(1,-1), query_vec.reshape(1,-1))
    res3.append([doc,cos_sim])
```

```
res3 = sorted(res3, key=lambda x:x[1], reverse = True)
```

```
import pandas as pd
res3 = pd.DataFrame(res3)
```

```
for i in range(5):
  print("doc {}:".format(i+1) + res3[0][i])
```

```
doc 1:One of the best things about Mumbai is the sea and the access it provides for quick day-time trips to nearby be

doc 2:Located about 16 km from Nashik and 262 km from Mumbai, the hill station of Deolali is a popular getaway. Also

doc 3:The word 'Sikh' translates into a disciple and the religion of Sikhism is based on the path of discipleship. Th

The most prominent of Sikh spiritual sites is Golden Temple or Sri Harmandir Sahib, in Amritsar, Punjab. The temple i

Baba Atal Rai tower is situated to the south of the Golden Temple. At a height of 40 m, it is a nine-storey tower and

Takht Sri Harmandir Sahib, also known as Patna Sahib, in Bihar, marks the birth site of Guru Gobind Singh, the last c

A quiet and serene spot amid the bustling marketplace, Gurudwara Bangla Sahib is possibly the most popular attraction

Nanded, in Maharashtra, is the perfect example of spiritual and philosophical antiquity blending seamlessly with the

doc 4:A stunning archway standing as a tribute to the brave soldiers who sacrificed their lives for the country, Indi
doc 5:Built in the memory of Mohandas Karamchand Gandhi, who is revered as the Father of the Nation, Raj Ghat is a ma
```

Variant 1

```
def termfreq(document, word):
```

```
    N = len(document)
    occurance = len([token for token in document if token == word])
    return 1 + np.log(occurance/N)
```

```
import numpy as np

def inverse_doc_freq(word):
    try:
        word_occurance = word_count[word] + 1
    except:
        word_occurance = 1
    return np.log(1 + np.log(total_documents/word_occurance))
```

Variant 2

```
def termfreq(document, word, K = 0.4):
    N = len(document)
    occurance = len([token for token in document if token == word])
    return K + (1-K) * occurance/N
```

```
import numpy as np

def inverse_doc_freq(word):
    try:
        word_occurance = word_count[word] + 1
    except:
        word_occurance = 1
    return np.log((total_documents- word_occurance)/word_occurance)
```

Colab paid products - Cancel contracts here