**Kulvir Singh**

**19BCE2074**

**Experiment 4 (01.03.2021)**

*Question 1*

***Problem Statement :***

Write a python program to find the important words from the text using TF-IDF. Use minimum of 5 documents with the real text source from a web page of some relevance.

***Procedure :***

1. We need to count the imp terms ( i.e. words that occur the most in the doc)

2. We will use TextBlob for breaking up text to words and get a word count. We will count the freq of words in a document, normalized by dividing with the number of words in the doc.

3. n\_containing will return the number of occurrences of word in different documents.

4. idf(word, bloblist) computes "inverse document frequency" which measures how common a word is among all documents in bloblist. The more common a word is, the lower its idf. We take the ratio of the total number of documents to the number of documents containing word, then take the log of that. Add 1 to the divisor to prevent division by zero.

5. tfidf(word, blob, bloblist) computes the TF-IDF score. It's the product of tf and idf.

6. For each document, we store the TF-IDF scores in a dictionary scores mapping word => score using a dict comprehension. We then sort the words by their scores and output the top 3 words.

***Code :***

import nltk

nltk.download('punkt')

import math

from textblob import TextBlob as tb

print('\n\*\*\*OUTPUT\*\*\*\n')

def tf(word, blob):

 return blob.words.count(word) / len(blob.words)

def n\_containing(word, bloblist):

 return sum(1 for blob in bloblist if word in blob)

def idf(word, bloblist):

 return math.log(len(bloblist) / (1 + n\_containing(word, bloblist)))

def tfidf(word, blob, bloblist):

 return tf(word, blob) \* idf(word, bloblist)

document1 = tb("""Python is a 2000 made-for-TV horror movie directed by Richard Clabaugh. The film features several cult favorite actors, including William Zabka of The Karate Kid fame, Wil Wheaton, Casper Van Dien, Jenny McCarthy, Keith Coogan, Robert Englund (best known for his role as Freddy Krueger in the A Nightmare on Elm Street series of films), Dana Barron, David Bowe, and Sean Whalen. The film concerns a genetically engineered snake, a python, that escapes and unleashes itself on a small town. It includes the classic final girl scenario evident in films like Friday the 13th. It was filmed in Los Angeles, California and Malibu, California. Python was followed by two sequels: Python II (2002) and Boa vs. Python (2004), both also made-for-TV films.""")

document2 = tb("""Python, from the Greek word (πύθων/πύθωνας), is a genus of nonvenomous pythons[2] found in Africa and Asia. Currently, 7 species are recognised.[2] A member of this genus, P. reticulatus, is among the longest snakes known.""")

document3 = tb("""The Colt Python is a .357 Magnum caliber revolver formerly manufactured by Colt's Manufacturing Company of Hartford, Connecticut. It is sometimes referred to as a "Combat Magnum".[1] It was first introduced in 1955, the same year as Smith & Wesson's M29 .44 Magnum. The now discontinued Colt Python targeted the premium revolver market segment. Some firearm collectors and writers such as Jeff Cooper, Ian V. Hogg, Chuck Hawks, Leroy Thompson, Renee Smeets and Martin Dougherty have described the Python as the finest production revolver ever made.""")

document4 = tb("""Web mining is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by mining the web. It makes utilization of automated apparatuses to reveal and extricate data from servers and web2 reports, and it permits organizations to get to both and unstructured information from browser activities, server logs, website and link structure, page content and different sources.""")

document5 =tb("""In computing, a hyperlink, or simply a link, is a reference to data that the user can follow by clicking or tapping.[1] A hyperlink points to a whole document or to a specific element within a document. Hypertext is text with hyperlinks. The text that is linked from is called anchor text. A software system that is used for viewing and creating hypertext is a hypertext system, and to create a hyperlink is to hyperlink (or simply to link). A user following hyperlinks is said to navigate or browse the hypertext.""")

bloblist = [document1, document2, document3, document4, document5]

for i, blob in enumerate(bloblist):

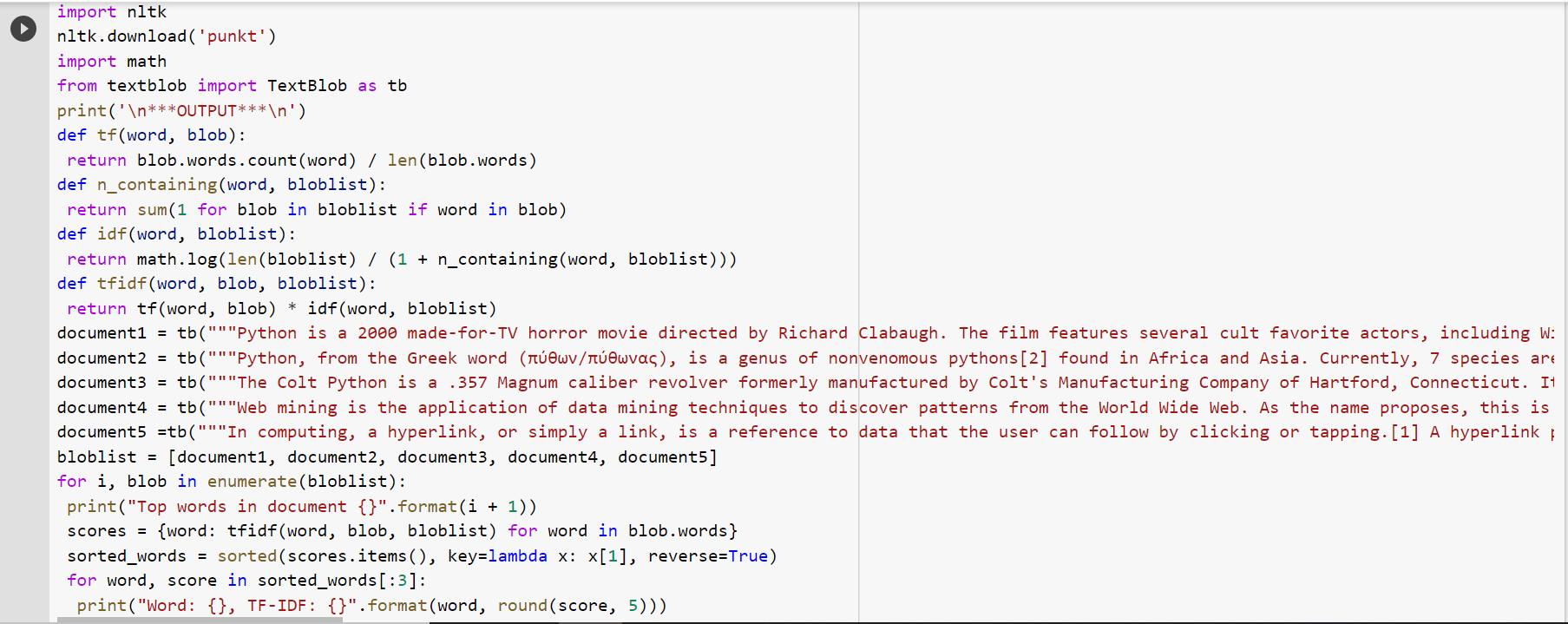
 print("Top words in document {}".format(i + 1))

 scores = {word: tfidf(word, blob, bloblist) for word in blob.words}

 sorted\_words = sorted(scores.items(), key=lambda x: x[1], reverse=True)

 for word, score in sorted\_words[:3]:

  print("Word: {}, TF-IDF: {}".format(word, round(score, 5)))

***Code Screenshot:*** 

***Output Screenshots :***