**Web Mining (CSE3024)**

**Lab Assignment 5**

Name: **Kritika Mishra**

Registration Number: **16BCI0041**

Slot: L15+L16

Date: 1st September 2018

Question: **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.**

**Code:**

# -\*- coding: utf-8 -\*-

"""

Created on Sat Sep 1 10:24:52 2018

@author: Kritika Mishra

"""

import math

from textblob import TextBlob as tb

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.words)

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 &amp; 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(""" Python's large standard library, commonly cited as one of its greatest strengths,[91] provides tools suited to many tasks. For Internet-facing applications, many standard formats and protocols such as MIME and HTTP are supported. It includes modules for creating graphical user interfaces, connecting to relational databases, generating pseudorandom numbers, arithmetic with arbitrary precision decimals,[92] manipulating regular expressions, and unit testing. Some parts of the standard library are covered by specifications (for example, the Web Server Gateway Interface (WSGI) implementation wsgiref follows PEP 333[93]), but most modules are not. They are specified by their code, internal documentation, and test suites (if supplied). However, because most of the standard library is cross-platform Python code, only a few modules need altering or rewriting for variant implementations.""")

document5=tb(""" Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by metaprogramming[42] and metaobjects (magic methods)).[43] Many other paradigms are supported via extensions, including design by contract[44][45] and logic programming.[46] Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution. Python's design offers some support for functional programming in the Lisp tradition. It has filter(), map(), and reduce() functions; list comprehensions, dictionaries, and sets; and generator expressions.[47] The standard library has two modules (itertools and functools) that implement functional tools borrowed from Haskell and Standard ML.""")

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[:5]:

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

**Output:**

