Information Retrieval – Assignment 11 Screenshot

# Code:

# IR11A.py CS5154/6054 cheng 2022  
# TfidfVectorizer is used to generate vocabulary  
# a random term is the query and the top 5 cosine similarity  
# in Tfidf are considered as the initial pseudo relevant for  
# probabilistic pseudo relevance feedback (IIR 11.3.4)  
# then pt, ut, ct are computed for terms and documents are ranked  
# with sum of ct for t in docs   
# Usage: python IR11A.py  
  
import re  
import numpy as np  
import random  
import math  
from sklearn.feature\_extraction.text import TfidfVectorizer  
from sklearn.metrics.pairwise import cosine\_similarity  
pseudorelevant = 5  
  
f = open("bible.txt", "r")  
docs = f.readlines()  
f.close()  
  
# Step 1  
tfidf = TfidfVectorizer(binary=True, max\_df=0.04, min\_df=8)  
dt = tfidf.fit\_transform(docs)  
docsets = [set(d.indices) for d in dt]  
  
N = len(docs)  
terms = list(tfidf.vocabulary\_)  
T = len(terms)  
  
query = random.choice(terms)  
print("Query = ", query, tfidf.vocabulary\_.get(query)) # printing query  
q = tfidf.transform([query])  
sim = cosine\_similarity(q[0], dt)  
  
# Step 2  
relevantset = set()  
for d in np.argsort(sim[0])[::-1][0:pseudorelevant]:  
 print(sim[0][d], docs[d])  
 relevantset.add(d)  
  
print(relevantset)  
  
dfs = np.zeros(T, dtype=int)  
for d in range(N):  
 for t in docsets[d]:  
 dfs[t] = dfs[t] + 1  
  
# Repeating Steps 2 and 3  
  
counter = 1  
while counter < 3:  
 ct = np.zeros(T)  
 for t in range(T):  
 Vt = 0  
 for d in relevantset:  
 if t in docsets[d]:  
 Vt = Vt + 1  
 pt = (Vt + 0.5)/(pseudorelevant + 1.0)  
 ut = (dfs[t] - Vt + 0.5)/(N - pseudorelevant + 1.0)  
 ct[t] = math.log(pt/(1 - pt) \* (1 - ut)/ut)  
 counter += 1  
  
  
rsv = np.zeros(N)  
for d in range(N):  
 for t in docsets[d]:  
 rsv[d] = rsv[d] + ct[t]  
  
relevantset.clear()  
for d in np.argsort(rsv)[::-1][0:pseudorelevant]:  
 print(rsv[d], docs[d])  
 relevantset.add(d)  
  
print(relevantset)

# Results

# A computer screen capture Description automatically generated with medium confidence

A picture containing background pattern

Description automatically generated