Felix Khmenlnitsky 207241217
Rotem Bahalker 208032748

Lab3 In this lab we will work on TFIDF using NLTK and SKLEARN libaries.

To do TFIDF using sklearn, we will use TfidfVectorizer,this class combines all the things we need to process tfidf. these thing are : 1)Tokenization 2)calculation of term frequencies(the TF part) 3)calcuate the inverse document frequencies(IDF) using the function TfidfVectorizer.fit_transform wwe can combine the fit and transform(our vocabulary(from traing sets) and the feature matrix)

The function works using the TFIDF equations ,It first tokenize then count the term frequencies and lasly compute the IDF In sklearn the default way to normalize our answers is to use L2 Normalization.

Cosine sim Cosine similarity is the a metric that measures how similer two vectors are in multi dimensional space. The close the angle to 0,the similar they are.

We can use from Sklearn libary , the class sklearn.metrics.pairwise and from there to take cosine_similarity that gets two TFIDF matrixies and compute them.

In [1]:

```python
#Building TFIDF
#We will use the preprocessed data.
import pickle
import pandas as pd
import numpy as np
with open('preprocessed_data.pkl', 'rb') as f:
    data = pickle.load(f)


from sklearn.feature_extraction.text import TfidfVectorizer
documents = [item['text'] for item in data] #Beacuse data is a list of dicts, we will take the text from each o
vector = TfidfVectorizer()
tfidf_matrix  = vector.fit_transform(documents)

terms = vector.get_feature_names_out()
sum_tfidf_scores = np.array(tfidf_matrix.sum(axis=0)).flatten()

words_and_scores = [(terms[i], sum_tfidf_scores[i]) for i in range(len(terms))]

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

# Print all words and their total TF-IDF scores across all documents
print("\nAll Words and Their Combined TF-IDF Scores Across All Documents:")
for word, score in words_and_scores:
    print(f"{word}: {score}")
```

```
All Words and Their Combined TF-IDF Scores Across All Documents:
space: 43.38092822992172
would: 37.80955320581766
subject: 36.66598210331874
lines: 36.39002777059808
organization: 35.64226891637338
window: 32.42440120025617
writes: 32.04250395053071
nntppostinghost: 28.9524865607723
like: 28.498528564880942
article: 27.76030329139654
one: 27.67121281518911
university: 25.868895350782726
get: 25.581628164541936
use: 22.905415643779175
dont: 22.480555536712792
know: 22.190060988368867
im: 21.34841632083912
could: 20.789123180628348
program: 20.056237944343422
thanks: 19.85341117943643
distribution: 19.53928087499174
server: 19.44240038739587
also: 19.139526137743513
...
xsoftware: 0.005765449518573399
xtos: 0.005765449518573399
xtosinfoxsoftuucp: 0.005765449518573399
xtreme: 0.005765449518573399
```

In [18]:

```python
#Lets build cosinesim using Our query.
from sklearn.metrics.pairwise import cosine_similarity
query = "would the organization use space?"
query_vector = vector.transform([query]) # making a vector to add.

tfidf_matrix_query = np.vstack([tfidf_matrix.toarray(),query_vector.toarray()])

#cosine.
cosine_sim = cosine_similarity(tfidf_matrix_query[-1:], tfidf_matrix_query[:-1])
```

```
simlaries = cosine_sim.flatten()
sorted = np.argsort(simlaries[::-1]) #cal with the last vector -> our query vector.

N = 5  # You can change N to any number
most_similar_docs = sorted[:5]

print("Top {} most relevant documents to the query:".format(N))
for idx in most_similar_docs:
    print(f"Document {idx + 1}: (Similarity Score: {simlaries[idx]:.4f}) : {documents[idx]} ")
```

```
Top 5 most relevant documents to the query:
Document 1366: (Similarity Score: 0.0272) : mccallmksoldsegticom fred j mccall subject hst servicing mission sch
eduled days organization texas instruments inc lines aprlambdamsfcnasagov bdaylambdamsfcnasagov brian day writes
rdouglasstsciedu rob douglas writes try land shuttle big huge telescope back could problems shuttle isnt designe
d land much weight payload hst really much heavier spacelab cant speak sheer mass part problem hst wasnt built e
ver brought back built kinds jolt forces support cradle additional weight would required insisting perfect safet
y people dont balls live real world mary shafer nasa ames dryden fredmccalldsegticom dont speak others dont spea
k
Document 1349: (Similarity Score: 0.0968) : pefquadsuchicagoedu enrico palazzo subject gamma ray bursters replyt
o pefmidwayuchicagoedu organization university chicago lines graydon saundrsgqucdnqueensuca things detected spac
e anyone looked possible problems detectors mechanism cosmic rays whatever could cause dector think seeing one t
hings graydon would explain widely separated detectors ulysses pvo ginga et al would see burst time fact fore ba
tse widely separated interplanetary network sure way locate random burst one detector one locate burst except sa
y somewhere field view two detectors one use time burst seen detector narrow location thin annulus sky three det
ectors one gets intersecting annuli giving two possible locations one locations impossible say earth blocked par
t sky voila error box batse detectors location determination within degrees would someone gsfc like david like c
omment current state location determination inde pendent sightings detectors helps drive uncertainty touch somet
hing didnt mean though believe reference somewhere absorptionlike features seen fraction grbs actually caused de
tector would mean nasty god though would nai crystal act like gauss neutron starbut getting far afield peter pet
erfoddjobuchicagoedu
Document 1530: (Similarity Score: 0.0065) : pjaquescamborneschoolofminesacuk paul jaques subject polygon raster
converter required organization internet lines nntppostinghost enterpoopmitedu xpertexpolcsmitedu anybody tell k
now obtain source code polygon filling algorithm graphics orientated mailing lists may able help thanks paul pau
l jaques systems engineer camborne school mines rosemanowes herniss penryn cornwall email pjaquescsmacuk tel sti
thians fax
Document 1880: (Similarity Score: 0.0082) : moorehalleyestmcom richard moore subject x interactive performance i
nreplyto afieldencbnewsbcbattcoms message mon apr gmt organization company center minnesota usa lines image disp
lay frames per second seems lower limit interactive operations bringing image viewing less second seems good num
ber course measure response time based applications planning run
Document 751: (Similarity Score: 0.0049) : edwindlimslercnasagov tom nguyen subject resolve different font forma
ts organization nasa lewis research center lines distribution world nntppostinghost limslercnasagov newssoftware
vaxvms vnews hi tried run softpc pc emulation software program installed silicon graphics workstation human desi
gn system hds x terminal everything went fine except fonts could converted one type format hds uses different fo
nt format sgi worksation following questions resolve different font formats different machines program convert o
ne type font format another similar problemsexperiences found solution please let know help greatly appreciated
thank advance help information tom nguen edwindlimslercnasagov
```

In [12]:

```
#Let's try another query.
query = "How does subject matter affect the lines of communication?"
query_vector = vector.transform([query]) # making a vector to add.

tfidf_matrix_query = np.vstack([tfidf_matrix.toarray(),query_vector.toarray()])

#cosine.
cosine_sim = cosine_similarity(tfidf_matrix_query[-1:], tfidf_matrix_query[:-1])

simlaries = cosine_sim.flatten()
sorted = np.argsort(simlaries[::-1]) #cal with the last vector -> our query vector.

N = 5  # You can change N to any number
most_similar_docs = sorted[:5]

print("Top {} most relevant documents to the query:".format(N))
for idx in most_similar_docs:
    print(f"Document {idx + 1}: (Similarity Score: {simlaries[idx]:.4f}) : {documents[idx]} ")
```

```
Top 5 most relevant documents to the query:
Document 1641: (Similarity Score: 0.0059) : prbaccessdigexnet pat subject interesting dcx cost anecdote organiza
tion express access online communications usa lines nntppostinghost accessdigexnet thats assuming could get buil
t course would probably sport cruise missile racks sidewinder missile tubes bomb points extra drop tanks full ec
m suite terrain following radar stealth materials might fly technology demonstrator require actual flight pat
Document 1086: (Similarity Score: 0.0046) : manlicsuhedu man lung li subject malloc problem organization compute
r science dept univ houston main campus lines distribution world nntppostinghost rodincsuhedu solve problem mess
age perform malloc shows xtcreatemanagedwidget call application written xr running decstation using athena widge
ts soon added codes remote procedure call program refused work also program working using xlib calls rpc executa
ble code dont idea much memory decstation help appreciated thanks man l li manlicsuhedu dept computer science un
iv houston houstontx
Document 741: (Similarity Score: 0.0021) : dealynaryagsfcnasagov brian dealy csc subject xwindows always opaque
organization nasagoddard space flight center lines distribution comp nntppostinghost naryagsfcnasagov keywords x
window parentchild relation originator dealynaryagsfcnasagov article hessswt hessswtinformatikunihamburgde hauke
hess writes hi wonder possible parent window paint area childs could possible implement rubberband across multip
le xwindows select objects displayed one window hauke specify rootwindow creating gc may use xlib draw multiple
windows application something similar rubber banding cursmove xcreatefontcursor dispdata xccrosshair geomvalsfor
eground blckpixl greydark geomvalsplanemask allplanes geomvalslinewidth geomvalsfunction gxxor geomvalssubwindow
mode includeinferiors evntmask gcforeground gcplanemask gclinewidth gcfunction gcsubwindowmode geomgcon xcreateg
c dispdata rootiden evntmask geomvals later move rubber band bands using following logic void tselectmovebands i
nt deltxloc int deltyloc stuff deleted xdrawrectangle dispdata rootiden geomgcon selepntrrootx selepntrrooty sel
epntrxlnth selepntrylnth undraw old one selepntrpapax deltxloc selepntrpapay deltyloc selepntrrootx deltxloc sel
epntrrooty deltyloc xdrawrectangle dispdata rootiden geomgcon selepntrrootx selepntrrooty selepntrxlnth selepntr
ylnth draw new one stuff deleted hope helps brian dealy knowing dealykonggsfcnasagov thats importantits knowing
uunetdftsrvkongdealy bdylan brian dealy knowing dealykonggsfcnasagov thats importantits knowing uunetdftsrvkongd
ealy bdylan
Document 723: (Similarity Score: 0.0041) : chatterjhajilcsmitedu shash chatterjee subject help sunview olwmxview
xr articleid hajiaa organization internet lines nntppostinghost enterpoopmitedu xpertexpolcsmitedufinlcsmitedu h
i compiled xr distribution sunsunos also compiled public domain xview olwm distribution old rdparty application
binaries sunview programs get work xview olwm tried using openwindows version svenv program work news accessthat
s mailing directly also email alias questions get compwindowsx compwindowsopenlook please respond fwrbvfinafmil
thanks shash shash chatterjee email fwrbvfinafmil ec software phone lockheed fort worth company fax po box mz ft
worth tx
Document 1071: (Similarity Score: 0.0314) : henryzootorontoedu henry spencer subject moonbase race organization
u toronto zoology lines article hgfbwwshakalacom danteshakalacom charlie prael writes doug actually memory serve
s atlas outgrowth old titan icbm nope youre confusing separate programs atlas firstgeneration us icbm titan seco
ndgeneration one titan ii titan launchers based thirdgeneration heavy icbm essentially nothing common three prog
rams yes three programs despite similarity names titan titan ii completely different missiles didnt even use fue
ls never mind launch facilities theres probably quite old pads albeit need serious reconditioning still able buy
turf pad bunkers including prep facility midwest farmland prices strikes pretty damned cheap sorry titan silos c
ant handle titan launchers large srbs b cant handle sort launcher without massive violations normal rangesafety
rules nobody cares things event nuclear war peacetime matter c scrapped years ago
```

In [13]:

```
#Last one.
query = "Would the organization change if the subject of the lines is altered?"
query_vector = vector.transform([query]) # making a vector to add.

tfidf_matrix_query = np.vstack([tfidf_matrix.toarray(),query_vector.toarray()])

#cosine.
cosine_sim = cosine_similarity(tfidf_matrix_query[-1:], tfidf_matrix_query[:-1])

simlaries = cosine_sim.flatten()
sorted = np.argsort(simlaries[::-1]) #cal with the last vector -> our query vector.

N = 5  # You can change N to any number
most_similar_docs = sorted[:3]

print("Top {} most relevant documents to the query:".format(N))
for idx in most_similar_docs:
    print(f"Document {idx + 1}: (Similarity Score: {simlaries[idx]:.4f}) : {documents[idx]} ")
```

```
Top 5 most relevant documents to the query:
Document 1071: (Similarity Score: 0.0050) : henryzootorontoedu henry spencer subject moonbase race organization
u toronto zoology lines article hgfbwwshakalacom danteshakalacom charlie prael writes doug actually memory serve
s atlas outgrowth old titan icbm nope youre confusing separate programs atlas firstgeneration us icbm titan seco
ndgeneration one titan ii titan launchers based thirdgeneration heavy icbm essentially nothing common three prog
rams yes three programs despite similarity names titan titan ii completely different missiles didnt even use fue
ls never mind launch facilities theres probably quite old pads albeit need serious reconditioning still able buy
turf pad bunkers including prep facility midwest farmland prices strikes pretty damned cheap sorry titan silos c
ant handle titan launchers large srbs b cant handle sort launcher without massive violations normal rangesafety
rules nobody cares things event nuclear war peacetime matter c scrapped years ago
Document 1086: (Similarity Score: 0.0078) : manlicsuhedu man lung li subject malloc problem organization compute
r science dept univ houston main campus lines distribution world nntppostinghost rodincsuhedu solve problem mess
age perform malloc shows xtcreatemanagedwidget call application written xr running decstation using athena widge
ts soon added codes remote procedure call program refused work also program working using xlib calls rpc executa
ble code dont idea much memory decstation help appreciated thanks man l li manlicsuhedu dept computer science un
iv houston houstontx
Document 1908: (Similarity Score: 0.0132) : buennekemontyrandorg richard buenneke subject white house outlines o
ptions station russian cooperation xadded forwarded space digest organization via international space university
originalsender isuvacationvenaricscmuedu distribution sci lines blindcarboncopy spacenewsaustenrandorg ctiausten
randorg subject white house outlines options station russian cooperation date tue apr pdt richard buenneke buenn
ekeaustenrandorg gibbons outlines space station redesign guidance nasa headquarters washington dc april release
dr john h gibbons director office science technology policy outlined membersdesignate advisory committee redesig
n space station april three budget options guidance committee deliberations redesign space station low option bi
llion midrange option billion high option billion considered committee option would cover total expenditures spa
ce station fiscal year would include funds development operations utilization shuttle integration facilities res
earch operations support transition cost also must include adequate program reserves insure program implementati
on within available funds next years billion reserved within nasa budget presidents new technology investment re
sult station options billion must accompanied offsetting reductions rest nasa budget example space station optio
n billion would require billion offsets nasa budget next years gibbons presented information organizational sess
ion advisory committee generally membersdesignate focused upon administrative topics used session get acquainted
also received legal ethics briefing orientation process station redesign team following develop options advisory
committee consider gibbons also announced united states international partners europeans japanese canadians deci
ded consultation give full consideration use russian assets course space station redesign process end russians a
sked participate redesign effort asneeded consulting basis redesign team make use expertise assessing capabiliti
es mir possible use mir russian capabilities systems us international partners hope benefit expertise russian pa
rticipants assessing russian systems technology overall goal redesign effort develop options reducing station co
sts preserving key research exploration capabilitiaes careful integration russian assets could key factor achiev
ing goal gibbons reiterated president clinton committed redesigned space station making every effort preserve sc
ience technology jobs space station program represents however also committed space station well managed one con
sume national resources used invest future industry nation nasa administrator daniel goldin said russian partici
pation accomplished eastwest space science center university maryland leadership roald sagdeev end blindcarbonco
py
```