

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

To do TFIDF using sklearn, we will use TfidfVectorizer, this class combines all the things we need to process tfidf. these things are :

1)Tokenization 2)calculation of term frequencies(the TF part) 3)calculate the inverse document frequencies(IDF) using the function TfidfVectorizer.fit_transform we can combine the fit and transform(our vocabulary(from training sets) and the feature matrix)

The function works using the TFIDF equations ,It first tokenizes then counts the term frequencies and lastly computes 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 similar two vectors are in multi dimensional space. The closer the angle to 0, the more similar they are.

We can use from Sklearn library , the class sklearn.metrics.pairwise and from there to take cosine_similarity that gets two TFIDF matrixes and computes them.

```
In [1]: #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] #Because data is a list of dicts, we will take the text from each dict
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: 22.48055536712792
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]: #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 scheduled days organization texas instruments inc lines aprlambdamsfncnasagov bdaylambdamsfncnasagov brian day writes rdouglasstsciedu rob douglas writes try land shuttle big huge telescope back could problems shuttle isnt designed land much weight payload hst really much heavier spacelab cant speak sheer mass part problem hst wasnt built ever brought back built kinds jolt forces support cradle additional weight would required insisting perfect safety people dont balls live real world mary shafer nasa ames dryden fredmccalldsegticom dont speak others dont speak

Document 1349: (Similarity Score: 0.0968) : pefquadsuchicagoedu enrico palazzo subject gamma ray bursters reply to pefmidwayuchicagoedu organization university chicago lines graydon saundrsgqucdnqueensuca things detected space anyone looked possible problems detectors mechanism cosmic rays whatever could cause detector think seeing one things graydon would explain widely separated detectors ulysses pvo ginga et al would see burst time fact fore batse widely separated interplanetary network sure way locate random burst one detector one locate burst except say somewhere field view two detectors one use time burst seen detector narrow location thin annulus sky three detectors one gets intersecting annuli giving two possible locations one locations impossible say earth blocked part sky voila error box batse detectors location determination within degrees would someone gsfc like david like comment current state location determination independent sightings detectors helps drive uncertainty touch something didnt mean though believe reference somewhere absorptionlike features seen fraction grbs actually caused detector would mean nasty god though would nai crystal act like gauss neutron starbut getting far afield peter peterfoddjobuchicagoedu

Document 1530: (Similarity Score: 0.0065) : pjaquescamborneschoolofminesacuk paul jaques subject polygon raster converter required organization internet lines nntppostinghost enterpoopmitedu xpertexpolcsmitedu anybody tell know obtain source code polygon filling algorithm graphics orientated mailing lists may able help thanks paul paul jaques systems engineer camborne school mines rosemanowes herniss penryn cornwall email pjaquescsmacuk tel stithians fax

Document 1880: (Similarity Score: 0.0082) : moorehalleyestmcom richard moore subject x interactive performance interactive nreplyto afieldencbnewsbcbattcoms message mon apr gmt organization company center minnesota usa lines image display frames per second seems lower limit interactive operations bringing image viewing less second seems good number course measure response time based applications planning run

Document 751: (Similarity Score: 0.0049) : edwindlimslerncnasagov tom nguyen subject resolve different font formats organization nasa lewis research center lines distribution world nntppostinghost limslerncnasagov newssoftware vaxvms vnews hi tried run softpc pc emulation software program installed silicon graphics workstation human design system hds x terminal everything went fine except fonts could converted one type format hds uses different font format sgi workstation following questions resolve different font formats different machines program convert one type font format another similar problemsexperiences found solution please let know help greatly appreciated thank advance help information tom nguyen edwindlimslerncnasagov

```

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 organization express access online communications usa lines nntppostinghost accessdigexnet thats assuming could get built course would probably sport cruise missile racks sidewinder missile tubes bomb points extra drop tanks full ECM suite terrain following radar stealth materials might fly technology demonstrator require actual flight pat

Document 1086: (Similarity Score: 0.0046) : manlicsuhehu man lung li subject malloc problem organization computer science dept univ houston main campus lines distribution world nntppostinghost rodincsuhehu solve problem message perform malloc shows xcreatemanagedwidget call application written xr running decstation using athena widgets soon added codes remote procedure call program refused work also program working using xlib calls rpc executable code dont idea much memory decstation help appreciated thanks man l li manlicsuhehu dept computer science univ 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 xwindow parentchild relation originator dealynaryagsfcnasagov article hessswt hessswtinformatikunihamburgde haukehess writes hi wonder possible parent window paint area childs could possible implement rubberband across multiple 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 blkpixl greydark geomvalsplanemask allplanes geomvalslinewidth geomvalsfunction gx xor geomvalssubwindow mode includeinferiors evtmask gcforeground gclplanemask gcllinewidth gcfunction gcsubwindowmode geomgcon xcreategc dispdata rootiden evtmask geomvals later move rubber band bands using following logic void tselectmovebands int deltxloc int deltyloc stuff deleted xdrawrectangle dispdata rootiden geomgcon selepntrootx selepntrooty selepntxl nth selepntryl nth undraw old one selepntropapax deltxloc selepntropapay deltyloc selepntrootx deltxloc selepntrooty deltyloc xdrawrectangle dispdata rootiden geomgcon selepntrootx selepntrooty selepntxl nth selepntryl nth draw new one stuff deleted hope helps brian dealy knowing dealykonggsfcnasagov thats importantits knowing uunetdftsrvkongdealy bdylan brian dealy knowing dealykonggsfcnasagov thats importantits knowing uunetdftsrvkongdealy bdylan

Document 723: (Similarity Score: 0.0041) : chatterjhajilcsmitedu shash chatterjee subject help sunview olwmxview xr articleid hajiaa organization internet lines nntppostinghost enterpoompitedu xpertexpolcsmitedufinlcsmitedu hi 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 compwindowsex compwindowsoopenlook 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 hgfbwshakalacom danteshakalacom charlie prael writes doug actually memory serve s atlas outgrowth old titan icbm nope youre confusing separate programs atlas firstgeneration us icbm titan secondgeneration one titan ii titan launchers based thirdgeneration heavy icbm essentially nothing common three programs yes three programs despite similarity names titan titan ii completely different missiles didnt even use fuels 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 cant 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])

similarities = cosine_sim.flatten()
sorted = np.argsort(similarities[::-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: {similarities[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 hgfbwshakalacom danteshakalacom charlie prael writes doug actually memory serve s atlas outgrowth old titan icbm nope youre confusing separate programs atlas firstgeneration us icbm titan secondgeneration one titan ii titan launchers based thirdgeneration heavy icbm essentially nothing common three programs yes three programs despite similarity names titan titan ii completely different missiles didnt even use fuels 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 cant 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) : manlicsuhehu man lung li subject malloc problem organization compute r science dept univ houston main campus lines distribution world nntppostinghost rodincsuhehu solve problem message perform malloc shows xtcreatemanagedwidget call application written xr running decstation using athena widgets soon added codes remote procedure call program refused work also program working using xlib calls rpc executable code dont idea much memory decstation help appreciated thanks man l li manlicsuhehu dept computer science univ houston houstontx

Document 1908: (Similarity Score: 0.0132) : buennekemontyrandorg richard buenneke subject white house outlines options station russian cooperation xadded forwarded space digest organization via international space university originalsender isuvacationvenaricscmuedu distribution sci lines blindcarboncopy spacenewsastenrandorg ctiaustenrandorg subject white house outlines options station russian cooperation date tue apr pdt richard buenneke buennekeastenrandorg 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 redesign space station april three budget options guidance committee deliberations redesign space station low option billion midrange option billion high option billion considered committee option would cover total expenditures space station fiscal year would include funds development operations utilization shuttle integration facilities research operations support transition cost also must include adequate program reserves insure program implementation within available funds next years billion reserved within nasa budget presidents new technology investment result station options billion must accompanied offsetting reductions rest nasa budget example space station option billion would require billion offsets nasa budget next years gibbons presented information organizational session 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 decided consultation give full consideration use russian assets course space station redesign process end russians asked participate redesign effort asneeded consulting basis redesign team make use expertise assessing capabilities mir possible use mir russian capabilities systems us international partners hope benefit expertise russian participants assessing russian systems technology overall goal redesign effort develop options reducing station costs preserving key research exploration capabilities careful integration russian assets could key factor achieving goal gibbons reiterated president clinton committed redesigned space station making every effort preserve science technology jobs space station program represents however also committed space station well managed one consume national resources used invest future industry nation nasa administrator daniel goldin said russian participation accomplished eastwest space science center university maryland leadership roald sagdeev end blindcarboncopy