Report of Phase Four

Steps to run the program:

You can run this program with or without using weights.

1. To run the program without using weights, you can give:

   py simple.py <query terms> Example: py simple.py diet.

2. To run the program using weights, you can give:

   py simple.py<query_terms with weights> Example: py simple.py Wt 0.1 diet.

Please note that for this program to run you need to have 'files' folder (containing all the html documents) in your project folder and you need to have beautifulsoup package installed.

Detailed summary of the updates made to the code:

In this phase, since we are not taking the input and output directory as the input from the user we are using getpwd() function available in the os package to get the current working directory and attach \\files\\ to the cwd so that we can access files folder which contains all the html documents. The idea is to construct the inverted index on the fly instead of using the tdm and postings file created from phase three. Post that, we take the query terms from the command line arguments. If the query contains Wt then we assign the corresponding weight to that corresponding word in the query_dict. Else, we assign a default weight to each word. Post that we make all the query words to lowercase and check if any of the query words are stopwords. If that is the case, we simply remove it from the query_dict.

We have used "Doc-at-a-time" approach to find the document similarity scores. We have defined a postings_list dictionary which we use to store the postings of all the query words. We take each query word and check if it exists in term document matrix. If it does then we create a record for it in postings_list and using the position and number of postings information we have available in the term document matrix, we retrieve the postings of each individual query term. Using this postings_list, we create a list named docs_occur which we use to store all the documents in which any of the query terms occur. Post that, we iterate through each of the document available in docs_occur list and calculate the document query similarity score for all the documents. We iterate through all the terms in each relevant document and for all the query terms available in that document we update dp with the product of its tfidf score and its weight. We update dl with sum of square of all the tfidf scores of all terms in that document. We compute document similarity score using dp over the product of square root of dl and square root of sum of square of all the weights.

After having the document query similarity for all the relevant documents, we sort them based on their score in descending order and take the top ten documents and display them.

Displaying top 10 tfidf terms of all the retrieved documents:

After displaying all the relevant documents, we take each document name and get its tfidf dictionary from global_tfidf_list and we sort the tfidf dictionary using sorted function based on their tfidf scores in the descending order and post that we display the top ten terms i.e., the terms which have the top tfidf scores.

If there are no relevant documents i.e., there are no documents which contain the query term or terms then we display a message saying “There are no files with these key words in the given corpus.”

Data Structures:

We have mainly used dictionaries to maintain all the tfidf scores of all the terms in a document where word acts as key and tfidf score acts as value. We have used a dictionary to maintain the term document matrix and we have maintained a list for all the postings of the corpus. We have not created any temporary files along the process.

Complexity:

The complexity of my code to retrieve relevant documents is O (n * m) where n is the number of documents with any of the query words present and m is average number of words present in each of those documents. And the complexity of the code to retrieve the documents which has any of the query terms is O (n * m) where n is the number of documents in the corpus and m is the number of query terms.

The following is the output for all the sample queries:

Output for **py simple.py diet**:

018.html 0.19

263.html 0.05

009.html 0.05

252.html 0.04

050.html 0.04

152.html 0.02

353.html 0.02

Top ten terms in 018.html are :

deceased : 0.28603877677367767

weight : 0.25226248955475644

information : 0.21320071635561041

please : 0.21320071635561041

diet : 0.19069251784911845

loss : 0.16514456476895406

indicate : 0.1348399724926484

address : 0.1348399724926484

email : 0.1348399724926484

fax : 0.1348399724926484

Top ten terms in 263.html are :

yr : 0.21579765395152142

pgs : 0.21579765395152142

fn : 0.19802950859533525

cardiovascular : 0.17149858514250918

heart : 0.16419739435729666

system : 0.1565560727712877

blood : 0.1309842079988962

cardiac : 0.12126781251816672

study : 0.12126781251816672

studies : 0.12126781251816672

Top ten terms in 009.html are :

turkey : 0.17822655773580193

turkeys : 0.16087993330796924

birds : 0.11881770515720128

bird : 0.11375929179890457

animals : 0.0970142500145335

meat : 0.09074852129730329

animal : 0.09074852129730329

people : 0.08401680504168084

food : 0.08401680504168084

united : 0.07669649888473729

Top ten terms in 252.html are :

yr : 0.23012097910583817

pgs : 0.23012097910583817

fn : 0.2000615668784767

children : 0.13591507055488938

afn : 0.11371470653683449

education : 0.11097419040461784

teaching : 0.10816426114554825

learning : 0.10816426114554825

child : 0.10527936095153853

handicapped : 0.10527936095153853

Top ten terms in 050.html are :

cup : 0.16552117772047392

fat : 0.14334554477024927

fruit : 0.13848495294356314

calories : 0.13848495294356314

juice : 0.1334474397584392

fibre : 0.12275379077928711

lemon : 0.1170411471961308

tofu : 0.11103498152964102

protein : 0.10468478451804296

soy : 0.10468478451804296

Top ten terms in 152.html are :

hawai : 0.13467682667066172

school : 0.10432022136310715

university : 0.10432022136310715

experience : 0.09835404792097975

president : 0.08865514073213371

owner : 0.08865514073213371

japanese : 0.08865514073213371

instructor : 0.08517710406460051

business : 0.08517710406460051

chef : 0.08155086839163112

Top ten terms in 353.html are :

az : 0.18816842583146234

hogy : 0.14120241876946066

es : 0.11429308034995797

nem : 0.10488252361925077

magyar : 0.08891848558108607

egy : 0.08081741215815424

february : 0.07866189271443808

szerint : 0.07416314367913046

uj : 0.07180815509291504

nemzet : 0.0693732685933489

Output for **py simple.py international affairs**:

133.html 0.14

161.html 0.12

138.html 0.11

117.html 0.11

219.html 0.10

205.html 0.10

247.html 0.09

229.html 0.09

125.html 0.09

143.html 0.09

Top ten terms in 133.html are :

law : 0.35355339059327456

canada : 0.1732050807568881

library : 0.152752523165195

rights : 0.14719601443879776

international : 0.14719601443879776

canadian : 0.14719601443879776

human : 0.13540064007726632

internet : 0.12909944487358085

constitutional : 0.12909944487358085

property : 0.12247448713915918

Top ten terms in 161.html are :

rights : 0.28421021140160296

women : 0.27846799904898606

human : 0.24448686826915356

international : 0.17052612684096177

amnesty : 0.16572160707065603

conference : 0.16077357421162855

world : 0.1503899078867471

governments : 0.13330640546608272

action : 0.12710267051871402

un : 0.11368408456064119

Top ten terms in 138.html are :

florence : 0.1747408113322076

art : 0.1747408113322076

students : 0.1747408113322076

isu : 0.1747408113322076

international : 0.15132998169159548

university : 0.1235604126430431

studies : 0.1235604126430431

available : 0.1235604126430431

arts : 0.1235604126430431

design : 0.1235604126430431

Top ten terms in 117.html are :

livermore : 0.2601329908572362

east : 0.24525573579398652

airport : 0.21239769762143676

hour : 0.21239769762143676

pcmdi : 0.1938916835823705

continue : 0.17342199390482413

minutes : 0.17342199390482413

directions : 0.15018785229652776

laboratory : 0.15018785229652776

address : 0.15018785229652776

Top ten terms in 219.html are :

public : 0.20851441405707505

larry : 0.19504737440137374

relations : 0.18057877962865404

irvine : 0.14744195615489733

affairs : 0.14744195615489733

nelson : 0.12768847961381247

communications : 0.12768847961381247

serves : 0.12768847961381247

uci : 0.12768847961381247

foundation : 0.12768847961381247

Top ten terms in 205.html are :

launch : 0.2515773027133144

space : 0.21683236581372337

international : 0.1344737247202767

site : 0.1344737247202767

vehicle : 0.127572976668769

facilities : 0.12027695586485303

facility : 0.11250879009260266

technical : 0.10416289926882244

station : 0.10416289926882244

support : 0.10416289926882244

Top ten terms in 247.html are :

yr : 0.2248000829380206

pgs : 0.2248000829380206

fn : 0.2234738171864762

business : 0.13355114986922023

international : 0.13355114986922023

marketing : 0.11945177983233543

multinational : 0.1117369085932381

japanese : 0.1117369085932381

management : 0.10628300005118843

foreign : 0.10344827586206816

Top ten terms in 229.html are :

gary : 0.19781414201873646

orange : 0.1744556751977297

county : 0.1744556751977297

public : 0.1474419561548974

relations : 0.1474419561548974

agency : 0.1474419561548974

media : 0.1474419561548974

press : 0.1474419561548974

irvine : 0.13187609467915765

clients : 0.13187609467915765

Top ten terms in 125.html are :

program : 0.17000510022951165

uppsala : 0.15205718425394124

university : 0.15205718425394124

swedish : 0.15205718425394124

sweden : 0.13168538439184427

students : 0.13168538439184427

available : 0.13168538439184427

fall : 0.13168538439184427

international : 0.13168538439184427

cost : 0.13168538439184427

Top ten terms in 143.html are :

leicester : 0.1833396994056423

university : 0.15877683720748897

isu : 0.15877683720748897

available : 0.15877683720748897

credit : 0.15877683720748897

hall : 0.15877683720748897

cost : 0.15877683720748897

england : 0.1296407447104329

studies : 0.1296407447104329

full : 0.1296407447104329

Output for **py simple.py Zimbabwe**:

There are no files with these key words in the given corpus.

Output for **py simple.py computer network**:

140.html 0.13

060.html 0.13

156.html 0.12

223.html 0.11

164.html 0.10

135.html 0.09

502.html 0.09

128.html 0.09

145.html 0.08

064.html 0.08

Top ten terms in 140.html are :

home : 0.32732683535398865

sheet : 0.32732683535398865

learn : 0.26726124191242445

nheri : 0.26726124191242445

official : 0.18898223650461365

web : 0.18898223650461365

site : 0.18898223650461365

view : 0.18898223650461365

title : 0.18898223650461365

ib : 0.18898223650461365

Top ten terms in 060.html are :

computer : 0.18291322825490808

motion : 0.1724522454236911

aesop : 0.1724522454236911

voice : 0.14934802547658516

surgical : 0.13633547078730324

medical : 0.13633547078730324

control : 0.13633547078730324

company : 0.13633547078730324

world : 0.12194215216993871

surgeons : 0.12194215216993871

Top ten terms in 156.html are :

nuclear : 0.16427218513662048

network : 0.1469295091111656

abolition : 0.13883533620176539

weapons : 0.13603029278483134

fax : 0.11780568923589896

tel : 0.107541388994909

email : 0.10011558470434938

statement : 0.09209294356482031

germany : 0.09209294356482031

apc : 0.08780717642256178

Top ten terms in 223.html are :

internet : 0.24309705084550082

access : 0.18644692689054815

service : 0.12273439277698847

affordable : 0.1179193903426055

web : 0.1179193903426055

communications : 0.10764518342983646

commission : 0.10764518342983646

architecture : 0.10764518342983646

cost : 0.10212118763546978

decentralized : 0.10212118763546978

Top ten terms in 164.html are :

server : 0.23658011170369236

ibm : 0.18325416653445817

software : 0.18325416653445817

servers : 0.17277368511627234

secure : 0.16161498378886383

os : 0.16161498378886383

internet : 0.1496264004161452

connection : 0.13658959117703853

information : 0.13658959117703853

family : 0.12216944435630545

Top ten terms in 135.html are :

bible : 0.26261286571944503

learn : 0.1856953381770518

home : 0.1856953381770518

activities : 0.1856953381770518

time : 0.1856953381770518

answer : 0.1856953381770518

waldo : 0.1856953381770518

official : 0.13130643285972252

web : 0.13130643285972252

site : 0.13130643285972252

Top ten terms in 502.html are :

send : 0.17960530202677485

comments : 0.17960530202677485

please : 0.17960530202677485

messages : 0.17960530202677485

thank : 0.17960530202677485

Top ten terms in 145.html are :

home : 0.22782254977690725

education : 0.18601633295108144

time : 0.18601633295108144

special : 0.18601633295108144

children : 0.1764705882352944

dr : 0.1764705882352944

duvall : 0.1764705882352944

classrooms : 0.15563243006262323

academic : 0.15563243006262323

school : 0.14408763192842247

Top ten terms in 064.html are :

mission : 0.20023571015840727

ve : 0.1610695384807916

shuttle : 0.14569287935358993

thank : 0.1373605639486893

aaron : 0.11895773785772187

jsc : 0.10859306069076759

incredible : 0.10859306069076759

information : 0.10859306069076759

world : 0.10859306069076759

report : 0.09712858623572664

Output for **py simple.py hydrotherapy**:

273.html 0.06

Top ten terms in 273.html are :

yr : 0.23784948889522697

pgs : 0.23784948889522697

fn : 0.22342069135256795

nutrition : 0.19132604955399282

food : 0.14130365221672006

nutritional : 0.1289919962949374

obesity : 0.10792244704280571

habits : 0.09991677068886828

effects : 0.09991677068886828

paper : 0.09121111529894027

Output for **py simple.py identity theft**:

379.html 0.05

301.html 0.04

245.html 0.04

380.html 0.04

328.html 0.04

332.html 0.03

298.html 0.03

397.html 0.02

027.html 0.02

235.html 0.02

Top ten terms in 379.html are :

criminal : 0.17054233423793125

crime : 0.16019662853723265

person : 0.15197586589727494

law : 0.14328422047021397

dui : 0.14026737269443793

illinois : 0.12748806952014846

crimes : 0.12748806952014846

property : 0.12408777488865559

misdemeanors : 0.1169910761148242

driving : 0.10131724393151663

Top ten terms in 301.html are :

yr : 0.25185876165203164

pgs : 0.25185876165203164

fn : 0.24624038734173628

black : 0.18069104496721178

analysis : 0.11427905097845878

ellison : 0.10580184237878919

examines : 0.10129755194401611

invisible : 0.09658342616078149

baldwin : 0.09658342616078149

american : 0.09658342616078149

Top ten terms in 245.html are :

yr : 0.21899171497017295

pgs : 0.21899171497017295

fn : 0.21072487598054526

economic : 0.16322678706085503

rates : 0.12643492558832717

paper : 0.11150512337989017

analysis : 0.11150512337989017

economy : 0.11150512337989017

study : 0.10323368445272058

examines : 0.10323368445272058

Top ten terms in 380.html are :

crime : 0.17539312294644552

person : 0.16033046361894146

criminal : 0.13922910991765308

crimes : 0.12819864484887225

fraud : 0.12654439948110605

law : 0.10467375519867203

illinois : 0.10467375519867203

computer : 0.10467375519867203

white : 0.10056729103085396

collar : 0.09844984776133968

Top ten terms in 328.html are :

yr : 0.23577842852172132

pgs : 0.23577842852172132

fn : 0.22740490955443934

philosophy : 0.13453455879926152

machiavelli : 0.11651034560709175

analysis : 0.10169856723618355

kant : 0.09846921430972355

descartes : 0.09513029883089813

hume : 0.09513029883089813

examines : 0.09513029883089813

Top ten terms in 332.html are :

yr : 0.21750475796862181

pgs : 0.21750475796862181

fn : 0.21750475796862181

african : 0.16155847834440418

nigeria : 0.15112411847332444

south : 0.15112411847332444

africa : 0.1399137464430121

apartheid : 0.1399137464430121

policy : 0.12772319171982657

political : 0.1142390955955082

Top ten terms in 298.html are :

yr : 0.23908130208913056

pgs : 0.23908130208913056

fn : 0.23468258446469278

analysis : 0.09911787167740733

examines : 0.09606735753462627

american : 0.08965548828342397

comparison : 0.08452800498106686

character : 0.08452800498106686

novel : 0.08274826871296237

review : 0.08092940333343618

Top ten terms in 397.html are :

health : 0.19057389107692282

care : 0.18794515977959364

patient : 0.1445528359737699

illinois : 0.13382992102364089

medicare : 0.12418891082596399

medical : 0.1201160301396835

law : 0.10927167294163039

abuse : 0.10697095900127143

person : 0.10461966178420169

consent : 0.09975093361076309

Top ten terms in 027.html are :

court : 0.1670015407569001

government : 0.16044989042404456

internet : 0.12471474558430241

speech : 0.12363494391310227

district : 0.12144654188931557

cda : 0.1134552055594804

id : 0.11226716019737303

act : 0.09688082545915179

children : 0.09117686495269127

app : 0.09117686495269127

Top ten terms in 235.html are :

yr : 0.21898852673774175

pgs : 0.21898852673774175

fn : 0.21353567448257665

political : 0.13151272746117196

policy : 0.12221745750194116

israel : 0.1189577378577208

analysis : 0.10859306069076666

iran : 0.1085930606907666

iraq : 0.1085930606907666

war : 0.1085930606907666

Output for **py simple.py Wt 0.3 dog 0.4 cat 0.3 bird**:

009.html 0.06

001.html 0.03

022.html 0.02

251.html 0.02

119.html 0.02

390.html 0.01

118.html 0.01

309.html 0.01

307.html 0.01

303.html 0.01

Top ten terms in 009.html are :

turkey : 0.17822655773580193

turkeys : 0.16087993330796924

birds : 0.11881770515720128

bird : 0.11375929179890457

animals : 0.0970142500145335

meat : 0.09074852129730329

animal : 0.09074852129730329

people : 0.08401680504168084

food : 0.08401680504168084

united : 0.07669649888473729

Top ten terms in 001.html are :

blancornelas : 0.17421664013428995

mexican : 0.15987215339548677

journalists : 0.14410681164656797

zeta : 0.13845334619821412

tijuana : 0.12639003479139002

press : 0.11990411504661506

government : 0.11990411504661506

political : 0.11304668378884462

félix : 0.1057454898622071

newspapers : 0.09790129997472591

Top ten terms in 022.html are :

http : 0.24287602608109216

www : 0.2052677068139892

html : 0.18359701840863063

kids : 0.17037082270290352

com : 0.15899968200095335

edu : 0.14990633779917167

information : 0.1402245390376251

children : 0.13338014159643297

provides : 0.13338014159643297

offers : 0.12239801227242042

Top ten terms in 251.html are :

yr : 0.22775688540033762

pgs : 0.22775688540033762

fn : 0.2213399070815271

education : 0.17395219456573827

school : 0.13149549909567565

black : 0.12589735538214894

educational : 0.11387844270016881

cultural : 0.10043134635865515

studies : 0.10043134635865515

children : 0.10043134635865515

Top ten terms in 119.html are :

energy : 0.23846480795382632

free : 0.14423743716582932

revolution : 0.11776937428767735

power : 0.11776937428767735

manning : 0.11172583840857912

air : 0.0985329278164296

clean : 0.0985329278164296

space : 0.09122376506188846

inventors : 0.0832755231749133

shared : 0.07448389227238608

Top ten terms in 390.html are :

property : 0.18716086928760983

tenant : 0.1732772946773724

landlord : 0.15006253908964662

real : 0.1225255500920297

estate : 0.1136988985564246

lease : 0.10997004672173234

illinois : 0.10412669513882493

home : 0.10004169272643107

law : 0.09578262852211519

title : 0.09578262852211519

Top ten terms in 118.html are :

poetry : 0.1742130722779687

getsi : 0.13666397770281732

books : 0.10437883709948627

read : 0.10437883709948627

manon : 0.10437883709948627

horse : 0.09663602537758927

book : 0.09252195028233338

world : 0.08821621827824555

reading : 0.08821621827824555

school : 0.08821621827824555

Top ten terms in 309.html are :

yr : 0.2528010792437211

pgs : 0.2528010792437211

fn : 0.2520693813871055

literature : 0.12155864986286773

examines : 0.12002954937252404

novel : 0.08377851779740171

kafka : 0.08154402130394031

analysis : 0.07924654425591939

characters : 0.07688044057231533

death : 0.07688044057231533

Top ten terms in 307.html are :

yr : 0.25312209792939344

pgs : 0.25312209792939344

fn : 0.2503095428796865

analysis : 0.12621094857350182

poetry : 0.11899282346174633

poem : 0.10808059922670497

chaucer : 0.09593508129463857

poems : 0.09407208683836006

comparison : 0.09217144471710854

examines : 0.09023077590464373

Top ten terms in 303.html are :

yr : 0.25000000000000144

pgs : 0.25000000000000144

fn : 0.24394257259323343

examines : 0.09986693274212073

comparison : 0.09474209111998394

analysis : 0.09297105474972679

dickens : 0.09297105474972679

novel : 0.08932370012631258

theme : 0.08552093186602991

characters : 0.08552093186602991

Output for **py simple.py Wt 0.9 baltimore**:

141.html 0.04

076.html 0.03

352.html 0.02

437.html 0.01

435.html 0.01

433.html 0.01

Top ten terms in 141.html are :

children : 0.23950235212423368

power : 0.1802908209386238

super : 0.17490778777865085

robot : 0.16361125403656945

robots : 0.151474587536049

tour : 0.13827674747047486

special : 0.13827674747047486

program : 0.13118084083398812

positive : 0.11569062720769316

hospitals : 0.10710870802417553

Top ten terms in 076.html are :

surgeon : 0.16375648758367892

robotic : 0.14976008767869695

procedures : 0.14976008767869695

field : 0.14812331626406067

arm : 0.14138687583786083

instrument : 0.12490854772583855

laparoscopic : 0.11890939480786995

operative : 0.11890939480786995

visual : 0.11684124756739729

tracking : 0.11040460140637993

Top ten terms in 352.html are :

az : 0.20117847640746267

es : 0.15946723621777115

hogy : 0.14536782710625273

nem : 0.13777811155231762

magyar : 0.13382193189304817

volt : 0.11035239204385006

meg : 0.09462639551305867

ha : 0.08876742736194708

egy : 0.08249337908658667

volna : 0.08029315913582889

Top ten terms in 437.html are :

edu : 0.33044960236560145

com : 0.31392690921338906

net : 0.24500794050408764

ca : 0.15789802199912062

ac : 0.11165076209152086

gov : 0.10421536410973682

uk : 0.10214121907486905

au : 0.09620702447031951

org : 0.09508830010164415

se : 0.08928466335356122

Top ten terms in 435.html are :

com : 0.3245274795261376

edu : 0.30390767187898005

net : 0.2681024817702892

ca : 0.15080483014555776

gov : 0.10488772554392105

max : 0.10237005312851276

ac : 0.10215744635046269

uk : 0.09510340687308722

org : 0.09487451747535323

uu : 0.09208347774676369

Top ten terms in 433.html are :

com : 0.3269717076845477

edu : 0.2988166097691778

net : 0.28241964892231475

ca : 0.15667550204736377

max : 0.12007851601677076

ac : 0.09940919195451028

uu : 0.09852079887412167

ms : 0.09816319043864519

org : 0.09708246183523304

gov : 0.0963552421437809
