“basic vector space retrieval system.”

The files look something like this and can ONLY keep the DOCNO and the TITLE text and TEXT text, merge them into one corpus/DOC for analysis + merge ea of those corpuses into one grand corpus for analysis:

<DOC>

<DOCNO>

6

</DOCNO>

<TITLE>

popular programming languages

for Data Science

internal .

</TITLE>

<AUTHOR>

wasserman,b.

</AUTHOR>

<BIBLIO>

j. ae. scs. 24, 1957, 924.

</BIBLIO>

<TEXT>

"Python and Java are very popular programming

languages for Data Science",

</TEXT>

</DOC>

Path to arrive at **freq\_word\_Corpus**.

[NB: dashed lines = different blocks of code & . . . = skipped lines in same block of code.]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

def getFiles(**dir\_path**):

**files** = os.listdir(dir\_path)

-----

for **f** in files:

with open(dir\_path+'/'+os.path.basename(**f**), 'r') as ipfile:

i = 0

for line in **ipfile**:

line = line.strip()

.

.

.

elif i == 5:

strtemp += line

strtemp += " "

review.append(line)

i += 1

while line != '</TITLE>':

for **line** in ipfile:

**line** = line.strip()

if line == '</TITLE>':

review.append(line)

i += 1

else:

**strtemp += line**

**strtemp += " "**

review.append(line)

i += 1

break

**titles**.append(strtemp)

strtemp = ""

elif line == '<TEXT>':

review.append(line)

i += 1

while line != '</TEXT>':

for line in ipfile:

**line** = line.strip()

if line == '</TEXT>':

review.append(line)

i += 1

else:

**strtemp += line**

**strtemp += " "**

review.append(line)

i += 1

break

**texts**.append(strtemp)

-----

def getPerDocCorp(titles, texts):

strtemp = ""

corpustemp = ""

for i in range(len(titles)):

**strtemp += titles[i]**

**strtemp += texts[i**]

**corpustemp** += strtemp

perDocCorp.append(strtemp)

strtemp = ""

**corpus**.append(corpustemp)

return(perDocCorp, corpus)

-----

tokens = str(corpus)

tokens = tokens.lower()

.

.

.

**tokens** = [word for word in tokens if word not in stoplist]

-----

**fdist** = nltk.FreqDist(tokens)

-----

freq\_word = []

# two values or columns in fdist\_a

j = 0

for k, v in **fdist**.items():

**freq\_word.**append([k, v])

j += 1

-----

**wordlist** = pd.DataFrame(freq\_word)

-----

**wordlistSorted** = wordlist.sort\_values(by=[1, 0],

ascending=[False, True])

-----

**freq\_word\_Corpus**.append(wordlistSorted)

DF | IDF 🡪 IDF Dict

idf = getIDF(word)

def getDocVecLen(file\_names, freq\_word\_Corpus):

for docid in (range(len(file\_names))):

sumSquares = 0

for word in freq\_word\_Corpus[0][0]:

weight = getWeight(word, docid)

weight\_sq = weight\*\*2

sumSquares += weight\_sq

docVecLen[docid] = math.sqrt(sumSquares)

return(docVecLen)

<script type="text/javascript"

src="https://events.k-state.edu/widget/combo?schools=k-state&venues=advanced\_manufacturing\_institute%2Cberney\_family\_welcome\_center%2Ccollege\_of\_business\_133%2Cdurland\_hall%2Cengineering\_hall\_349%2Cenglishcounseling\_services%2Cfiedler\_auditorium\_engineering\_complex%2Cfiedler\_hall%2Cfrank\_myers\_field%2Ck-state\_alumni\_center%2Cksu\_foundation\_center%2Cleadership\_studies\_building%2Cmanhattank-state\_innovation\_center%2Crathbone\_hall\_64%2Ctointon\_family\_baseball\_stadium\_at\_frank\_meyer\_field&types=133293%2C133294%2C133296%2C138209%2C133299%2C133300%2C133301%2C133302%2C133303%2C133304%2C133310%2C133311%2C133315%2C133316%2C133313%2C131607%2C131608%2C131609%2C131611%2C131612%2C131615%2C131616%2C141252%2C131618&groups=beta\_theta\_pi\_gamma\_epsilon\_chapter\_of%2Cbusiness\_education\_club%2Cbusiness\_leaders\_of\_america%2Ccliftonstrengths%2Cadvertising\_club\_ksu%2Cengineering\_student\_council%2Cgraduate\_student\_council%2Cmarketing\_club\_ksu%2Cschool\_of\_leadership\_studies\_ambassadors%2Cstrengths\_advocates%2Cwildcats\_forever&departments=biosecurity\_research\_institute%2Ccenter\_for\_advancement\_of\_entrepreneurship%2Ccollege\_of\_business\_administration%2Ccollege\_of\_engineering%2Ccomputer\_science\_department\_of%2Cdeveloping\_scholars\_program%2Ceconomics\_department\_of%2Ceducational\_supportive\_services%2Cengineering\_research\_and\_graduate\_programs%2Cexcite\_exploring\_science\_technology\_and\_engineering%2Cfirst\_scholars%2Cgeographic\_information\_systems\_spatial\_analysis\_laboratory%2Cgiscience\_virtual\_collaboratory%2Cglobal\_campus%2Cgraduate\_school%2Cgreat\_plains\_idea%2Cgrow\_project\_girls\_researching\_our\_world%2Chomepage\_calendar%2Chomepage\_featured%2Cindustrial\_and\_manufacturing\_systems\_engineering\_imse\_department\_of%2Ck-state\_alumni\_association%2Ck-state\_advancement\_of\_women\_in\_science\_and\_engineering%2Ck-state\_family\_center%2Ck-state\_first%2Ck-state\_olathe%2Ck-state\_research\_and\_extension%2Ckansas\_state\_polytechnic%2Cks\_lsamp%2Cksre\_riley\_county%2Cksu\_research\_foundation%2Cmanagement\_department\_of%2Cmarketing\_department\_of%2Cmathematics\_department\_of%2Cmcnair\_scholars\_program%2Cmulticultural\_engineering\_program%2Cnew\_student\_services%2Cnon-traditional\_and\_veteran\_student\_services%2Coffice\_of\_first-generation\_students%2Coffice\_of\_graduate\_studies\_coe%2Coffice\_of\_research\_development%2Coffice\_of\_the\_president%2Coffice\_of\_the\_provost%2Coffice\_of\_the\_vice\_provost\_for\_undergraduate\_studies%2Coffice\_of\_undergraduate\_research\_creative\_inquiry%2Cparents\_and\_family\_association%2Cphilosophy\_department\_of%2Cphysics\_department\_of%2Cpowercat\_financial%2Csocial\_media%2Cstaley\_school\_of\_leadership\_studies%2Cstatistics\_department\_of%2Cstay\_strong\_stay\_healthy%2Cstudent\_access\_center%2Cstudent\_financial\_assistance%2Cwomen\_in\_engineering\_and\_science\_program%2Cwildcat\_wellness\_coalition%2Cweb\_services%2Cvice\_president\_for\_research%2Cteaching\_and\_learning\_center&days=31&num=50&all\_instances=1&template=modern"></script><div id="lclst\_widget\_footer"><a rel="nofollow" style="margin-left:auto;margin-right:auto;display:block;width:81px;margin-top:10px;" title="Widget powered by Localist Event Calendar Software" href="https://www.localist.com?utm\_source=widget&utm\_campaign=widget\_footer&utm\_medium=branded%20link"><img src="//d3e1o4bcbhmj8g.cloudfront.net/assets/platforms/default/about/widget\_footer.png" alt="Localist Online Calendar Software" style="vertical-align: middle;" width="81" height="23"></a></div>