FINAL PROJECT REPORT

SEARCH RETRIEVAL SYSTEM

NAWAAZ SHARIF
SYED NOOR RAZI ALI
MOHAMMED RASHIDUDDIN

CSC 575: INTELLIGENT INFORMATION RETRIEVAL

PROFESSOR BAMSHAD MOBASHER

Table of Contents:

Introduction:

Many clinical trials require patients for the studies that are to be conducted for the trials in the medical research, but due to the high rate of insufficient patients recruitment, there's a possibility that the study or trial may need to be terminated permanently.

The TREC clinical trials track has concluded to perform the trials in a way such that a patient who has similar disease or who has the similar symptoms related to a particular study, all their medical history and records are needed to be updated on the clinicaltrials.gov website so that the different doctor/researcher can refer to the previous trial that has already been performed. The data set is broken down into eligible and non-relevant queries to distinguish between patients who have sufficient information from people who have insufficient information about the disease or trial being conducted.

The structured data is helpful for clinical trials as the description of the patient that is mentioned in the data set is only limited to his or her disease topics. The TREC clinical trials provides a platform for doctors or researchers a platform, for evaluating patients matching symptoms for clinical trial requirements, for example if a researcher comes all the way to the middle of the research and then something bad happens to the patient and then after some time a person researcher gets a similar patient from the point where the previous research was held he can continue from there instead of starting the whole research from the beginning which can be more effective and save time in the field of medical research.

About Dataset:

In the dataset of 2021 Clinical Trials Track, we have data of several different patient's cases that are created by people who have received medical training.

The information of patients is recorded in an .xml file, where we have:

1. Task: which is "2021 TREC Clinical Trials"
2. Number: patient case number recorded.
3. Description: in the form of corpus, where we have detailed description of patient, like how old he/she is, their gender, which department dealt with their case, and what disease or symptoms they had, and the detailed medical report of what all things have been identified by the practitioner.

Functions and Implementation:

Since we have five zip files we are combining into a single zip:

1. ClinicalTrials.2021-04-27.part1.zip
2. ClinicalTrials.2021-04-27.part2.zip
3. ClinicalTrials.2021-04-27.part3.zip
4. ClinicalTrials.2021-04-27.part4.zip
5. ClinicalTrials.2021-04-27.part5.zip

This will give us Final_combined_ClinicalTrials.zip.

Each of the zip file contains 500+ .xml files. Looping through each xml and extracting only useful tags and converting it into a Data Frame.

Important Columns in the clinical trials dataset:

['org_study_id', 'nct_id', 'brief_title', 'sponsor_agency', 'brief_summary' ,'detailed_description' , 'condition' , 'eligibility_condition' , 'eligibility_gender' , 'eligibility_minage' , 'eligibility_maxage']

```
df_clinical_trial.head()
```

	org_study_id	nct_id	brief_title	sponsor_agency	brief_summary	detailed_description	condition	eligibility_condition	eligibility_gender	eligibility
0	NCRR-M01RR01070-0506	D000102	Congenital Adrenal Hyperplasia: Calcium Channe...	National Center for Research Resources (NCRR)	This study will test the ability of extended r...	This protocol is designed to assess both acute...	Congenital Adrenal Hyperplasia	Inclusion Criteria: - diagnosed ...	All	
1	NCRR-M01RR00400-0587	D000104	Does Lead Burden Alter Neuropsychological Deve...	National Center for Research Resources (NCRR)	Inner city children are at an increased risk f...		Lead Poisoning	Inclusion Criteria: - Pregnant m...	Female	
2	2002LS032	D000105	Vaccination With Tetanus and KLH to Assess Imm...	Masonic Cancer Center, University of Minnesota	The purpose of this study is to learn how the ...	Patients will receive each vaccine once only c...	Cancer	Inclusion Criteria: - Patients m...	All	
3	NCRR-M01RR03186-9943	D000106	41.8 Degree Centigrade Whole Body Hyperthermia...	National Center for Research Resources (NCRR)	Recently a non-toxic system for whole body hyp...		Rheumatic Diseases	Inclusion Criteria: - Patients a...	All	

1. Corpus

Converting of pd.DataFrame to a dictionary where ‘nct_id’ is our index, ‘brief_summary', 'detailed_description' , 'condition' , 'eligibility_condition' and ‘brief_title’ and these columns are combined to form our document data for each document ID.

2. Queries

Fetching the data from the URL 'https://www.trec-cds.org/topics2021.xml' and converting it into pd.DataFrame for queries.

3. Pre-Processing of Corpus:

First, we have removed punctuations, then we converted our string to lower case and then split it into a list, and used a nltk package to fetch the stop words and performing stemming. This removed the stop words and do the stemming for the whole corpus dataset.

```
preprocessing_documents()
```

```
{'D000102': ['protocol',
  'design',
  'assess',
  'acut',
  'chronic',
  'effect',
  'calcium',
  'channel',
  'antagonist',
  'nifedipin',
  'hypothalamicpituitaryadren',
  'axi',
```

Counting the occurrences of each word in the document with document id and frequency, we get the output as below.

Output:

```
Word    Total Freq      Document Freq   Posting(Doc-ID, Count)
aa              18              11              [('D000451', 1), ('D000640', 3), ('D000655', 2), ('D000715', 2), ('D000717', 2),
('D000722', 2), ('D000724', 2), ('D000781', 1), ('D000863', 1), ('D000881', 1), ('D001013', 1)]
aactg           2               2               [('D001103', 1), ('D001117', 1)]
aactgsponsor    2               1               [('D001103', 2)]
aand            1               1               [('D000478', 1)]
abacavir        13              10              [('D000864', 2), ('D000885', 1), ('D000903', 1), ('D000912', 1), ('D000919', 1),
('D000922', 1), ('D000940', 1), ('D001086', 2), ('D001118', 2), ('D001119', 1)]
abandon         3               2               [('D000126', 1), ('D000586', 2)]
abat            1               1               [('D000604', 1)]
abbott          5               5               [('D000679', 1), ('D000704', 1), ('D000718', 1), ('D000735', 1), ('D001009', 1)]
abbrevi         2               2               [('D000468', 1), ('D001097', 1)]
abc             48              9               [('D000864', 10), ('D000872', 1), ('D000885', 11), ('D000912', 4), ('D000919',
4), ('D000922', 2), ('D000940', 2), ('D001086', 11), ('D001119', 3)]
abdomen         2               2               [('D000571', 1), ('D001118', 1)]
abdomin         9               7               [('D000410', 1), ('D000502', 1), ('D000619', 2), ('D000668', 2), ('D000742', 1),
('D000927', 1), ('D001054', 1)]
abductor        1               1               [('D000394', 1)]
aberr           2               2               [('D000938', 1), ('D001101', 1)]
```

4. Calculating IDF

Writing a function to compute the IDF values of the terms present in the documents, we are creating a idf_dict which holds the IDF values for all the words.

```
def compute_tfidf(term, doc_id):
    tf = sorted_dict[term][doc_id]
    idf = math.log2(N / len(sorted_dict[term]))
    idf_dict[term]=math.log2(N / len(sorted_dict[term]))
    return tf * idf
```

Output:

```
idf_dict
```

```
{'aa': 6.50635266602479,
 'aactg': 8.965784284662087,
 'aactgsponsor': 9.965784284662087,
 'aand': 9.965784284662087,
 'abacavir': 6.643856189774724,
 'abandon': 8.965784284662087,
 'abat': 9.965784284662087,
 'abbott': 7.643856189774724,
 'abbrevi': 8.965784284662087,
 'abc': 6.795859283219775,
 'abdomen': 8.965784284662087,
 'abdomin': 7.158429362604483,
 'abductor': 9.965784284662087,
 'aberr': 8.965784284662087,
 'abg': 8.965784284662087,
 'abi': 9.965784284662087,
 'abil': 3.0350469470992008,
 'abl': 3.3658714424749596,
 'ablat': 7.380821783940931,
```

5. Synonyms for Query Expansion and Vector Representation of Query.

Our Original Query

```
q=queries_topics()['Queries'].loc[1]
q
```

```
'48 M with a h/o HTN hyperlipidemia, bicuspid aortic valve, and tobacco abuse who presented to his cardiologist on [**2148-10-1
**] with progressive SOB and LE edema. TTE revealed severe aortic stenosis with worsening LV function. EF was 25%. RV pressure
was 41 and had biatrial enlargement. Noted to have 2+ aortic insufficiency with mild MR. He was sent home from cardiology clini
c with Lasix and BB (which he did not tolerate), continued to have worsening SOB and LE edema and finally presented here for ev
aluation.\nDuring this admission repeat echo confirmed critical aortic stenosis showing left ventricular hypertrophy with cavit
y dilation and severe global hypokinesis, severe aortic valve stenosis with underlying bicuspid aortic valve, dilated ascending
aorta, mild pulmonary artery systolic hypertension. The patient underwent a preop workup for valvular replacement with preop ch
est CT scan and carotid US (showing moderate heterogeneous plaque with bilateral 1-39% ICA stenosis). He also underwent a cardi
ac cath with right heart cath to evaluate his pulm art pressures which showed no angiographically apparent flow-limiting corona
ry artery disease.'
```

We have used this word2vec model and used to find different words/synonyms for our query expansion.

After synonym expansion

```
word2vec_model = gensim.models.KeyedVectors.load_word2vec_format(r'GoogleNews-vectors-negative300.bin', binary=True)
```

```
word2vec_expand(q, num_words=5, threshold=0.5)
```

```
"48 m with a h/o htn hyperlipidemia, bicuspid aortic valve, and tobacco abuse who presented to his cardiologist on [**2148-10-1
**] with progressive sob and le edema. tte revealed severe aortic stenosis with worsening lv function. ef was 25%. rv pressure
was 41 and had biatrial enlargement. noted to have 2+ aortic insufficiency with mild mr. he was sent home from cardiology clini
c with lasix and bb (which he did not tolerate), continued to have worsening sob and le edema and finally presented here for ev
aluation. during this admission repeat echo confirmed critical aortic stenosis showing left ventricular hypertrophy with cavity
dilation and severe global hypokinesis, severe aortic valve stenosis with underlying bicuspid aortic valve, dilated ascending a
orta, mild pulmonary artery systolic hypertension. the patient underwent a preop workup for valvular replacement with preop che
st ct scan and carotid us (showing moderate heterogeneous plaque with bilateral 1-39% ica stenosis). he also underwent a cardia
c cath with right heart cath to evaluate his pulm art pressures which showed no angiographically apparent flow-limiting coronar
y artery disease. sure_theyâ Zach_Sivan TW_Incl meters -----f----- wtih withthe wth wih fontanelle decayed_tooth ileocecal_va
lve overbites bicuspids aortic_valve aortic_arch mitral_valve thoracic_aortic thoracic_aorta Tobacco cigarette cigarettes curin
g_huts smokeless_tobacco sexual_abuse Abuse abusers abuses abused whom whose presenting Presenting present Presented submitted
His her him he himself Cardiologist cardiac_surgeon gastroenterologist interventional_cardiologist cardiologists onthe On upon
wtih withthe wth wih n_CHRIST_EPISCOPAL inflammatory_neuropathy liberal multiple_sclerosis_SPMS progessive sobbing sobs weeping
weep cried à l' les du qui 谷 rit #x##E# C'est_un ami reveals reveal revealing confirmed divulged Severe severest serious mild d
ebilitating aortic_valve aortic_arch mitral_valve thoracic_aortic thoracic_aorta stenoses mitral_regurgitation left_ventricular
_dysfunction carotid_stenosis stenotic wtih withthe wth wih deteriorating rapidly_deteriorating worsened deterioration worsen d
pa_cb dpa_sm dpa_sk dpa_cp dpa_dc ¬ nes ty ser ..... is had Was wasn'ta became raleigh vw duluth louisville atv Pressure presu
re pressue pressures presssure is had Was wasn'ta became have has was Had been noting emphasized explained acknowledged pointed
've had haven_t already been aortic_valve aortic_arch mitral_valve thoracic_aortic thoracic_aorta deficiency insufficient inade
quacy untimeliness wtih withthe wth wih milder Mild mildest severe moderate_atopic_dermatitis He him his she himself is had Was
wasn'ta became sending send sends forwarded mailed house Superfast_WiFi homes fromthe interventional_cardiology pediatric_cardi
ology radiation_oncology gastroenterology radiology clinics Clinic dental_clinic outpatient_clinic clinc wtih withthe wth wih p
ropecia prozac plavix levitra zoloft aaa bb_+ He him his she himself didn_`_t not does didn_t do do did anymore necessarily any
thing continues continuing continue continued_unabated began 've had haven_t already been deteriorating rapidly_deteriorating w
orsened deterioration worsen sobbing sobs weeping weep cried à l' les du qui ulceration edema_swelling pericardial_effusions as
cites cerebral_ischemia FINALLY belatedly eventually again fi_nally presenting Presenting present Presented submitted Archived_
```

After expanding the query, I have used the same pre-processing steps for queries as we have used for corpus main dataset by removing punctuations, stop words and stemming of each query.

```
expandedquery=word2vec_expand(q, num_words=5, threshold=0.5)
q=userquery(expandedquery)
q
```

```
{'ho': 10.01262453886506,
 'aortic': 5.312184820723967,
 'valv': 5.842699537422747,
 'tobacco': 10.01262453886506,
 'present': 2.4052942251154485,
 'progress': 4.764697025421474,
 'le': 8.427662038143904,
 'edema': 6.690696443977697,
 'reveal': 4.727222320002811,
 'sever': 3.1797345247003177,
 'stenosi': 6.205269616807455,
 'worsen': 10.01262453886506,
 'function': 3.727222320002811,
 'ef': 10.01262453886506,
 'pressur': 4.368768349090335,
 'enlarg': 6.312184820723967,
 'note': 4.842699537422747,
 'insuffici': 5.690696443977697,
```

We get the doclength, queries length, which helps us in getting the inverted_index values.

To calculate the document length, we are taking the sum of the squares of document weights and multiplying them with their respective IDF values of their term. To calculate the cosine similarity, we are taking the square root of the tf_idf, and similarly to calculate dice and jaccard similarity, we take the sum of the tf_idf values.

To calculate the queries length, we take the sum of square of their respective weights.

To store the documents with their respective similarity score we create a hashmap.

To get the similar documents for the query we create a new hash map with zero scores and using the respective similarity we find the score for each document which is then run iteratively to find the term weights for the whole query and document ID.

```
doclength('Cosine')
```

```
{88: 52.34522400686566,
 391: 68.8333182632571,
 827: 36.479099045668626,
 860: 51.48244626564064,
 1031: 123.85800401823795,
 594: 72.43232056054718,
 187: 88.29420531833908,
 451: 56.43759276363038,
 1027: 59.42673819874818,
 75: 101.07253823659555,
 676: 64.32699841183141,
 247: 62.275927178992696,
 415: 76.11012066446108,
```

Queries Length

```
querieslength(q,"Cosine")
```

```
85.96852949851773
```

This gives us the output”

Cosine similarity values for our documents:

```
documents_contents('COSINE',q).iloc[:10]
```

	Document ID	Cosine	brief_title
0	D000411	0.108254	Spine Patient Outcomes Research Trial (SPORT):...
1	D000424	0.104150	Tidal Lavage in Knee Osteoarthritis
2	D000409	0.102447	Spine Patient Outcomes Research Trial (SPORT):...
3	D000408	0.097654	Low Back Pain Patient Education Evaluation
4	D000406	0.091465	Effects of Strength Training on Knee Osteoarth...
5	D000369	0.090714	Maintenance Therapies in Bipolar Disorders
6	D000410	0.075182	Spine Patient Outcomes Research Trial (SPORT) ...
7	D000615	0.072945	Girls Health Enrichment Multi-Site Studies (GEMS)
8	D000404	0.072791	Effects of Comprehensive Care for Knee OA
9	D000672	0.072135	An Efficacy Study of 2',3'-Dideoxyinosine (ddI...

Jaccard similarity values for our documents:

```
documents_contents('Jaccard',q).iloc[:10]
```

	Document ID	Jaccard	brief_title
0	D000206	0.000015	Clinical Rescue Protocol - 2
1	D000411	0.000011	Spine Patient Outcomes Research Trial (SPORT):...
2	D000409	0.000011	Spine Patient Outcomes Research Trial (SPORT):...
3	D000426	0.000011	Treatment of Calcium Deficiency in Young Women
4	D000369	0.000010	Maintenance Therapies in Bipolar Disorders
5	D000408	0.000010	Low Back Pain Patient Education Evaluation
6	D000116	0.000010	Randomized Trial for Retinitis Pigmentosa
7	D000797	0.000009	Women's Interagency HIV Study (WIHS)
8	D000702	0.000009	A Multicenter Placebo-Controlled Double-Blind ...
9	D000484	0.000009	Treatment of Hypertension

Dice similarity values for our documents:

```
documents_contents('DICE',q).iloc[:10]
```

	Document ID	Dice	brief_title
0	D000206	0.000031	Clinical Rescue Protocol - 2
1	D000411	0.000022	Spine Patient Outcomes Research Trial (SPORT):...
2	D000409	0.000021	Spine Patient Outcomes Research Trial (SPORT):...
3	D000426	0.000021	Treatment of Calcium Deficiency in Young Women
4	D000369	0.000020	Maintenance Therapies in Bipolar Disorders
5	D000408	0.000020	Low Back Pain Patient Education Evaluation
6	D000116	0.000019	Randomized Trial for Retinitis Pigmentosa
7	D000797	0.000018	Women's Interagency HIV Study (WIHS)
8	D000702	0.000018	A Multicenter Placebo-Controlled Double-Blind ...
9	D000484	0.000018	Treatment of Hypertension

From our original corpus data, we have selected the title information and the similarity to show with respective to the document id.

<u>Query Expansion and Relevance Feedback:</u>

Sometimes when a user enters a query, the terms may be not in his dictionary, so this is where rocchhio relevance feedback comes which ask the user about relevant and non-relevant documents and gives the new query.

This new query may have few added terms and may have few terms disregarded when compared to the original query.

By default, we have used alpha =0.5 and beta=0.25 to calculate the modified query.

Output

	abil	access	accid	affect	alloc	andor	anesthesia	annual	anoth	anserin	wide	wilson	womac	work	worker	world	wors	write	xray
Q	0.0000	0.000	0.000	0.00	0.0000	0.0000	0.00	0.00	5.489063	0.000	0.000	0.0000	0.00	0.00	0.000	0.000	0.00	0.0000	6.105734
Q1	0.1875	0.375	0.125	0.25	0.1875	0.3125	0.25	0.25	5.614063	0.125	0.375	0.0625	0.25	0.75	0.125	0.375	0.25	0.0625	6.355734

```
Increased columns: Index(['abil', 'access', 'accid', 'affect', 'alloc', 'andor', 'anesthesia',
       'annual', 'anoth', 'anserin',
       ...
       'wide', 'wilson', 'womac', 'work', 'worker', 'world', 'wors', 'write',
       'xray', 'year'],
      dtype='object', length=425)
Decreased columns: Index(['addit', 'analyz', 'appar', 'cardiovascular', 'clinic', 'continu',
       'diseas', 'emphas', 'evalu', 'forth', 'heart', 'howev', 'hypertens',
       'import', 'key', 'physician', 'point', 'prior', 'return', 'right',
       'serious', 'show', 'vascular'],
      dtype='object')
```

	addit	analyz	appar	cardiovascular	clinic	continu	diseas	emphas	evalu	forth	...	import	key	physician	point
Q	4.058428	6.312185	4.58636	6.20527	3.218209	4.427662	3.001397	6.312185	4.553193	6.8427	...	4.081887	9.012625	7.20527	5.427662
Q1	3.808428	6.249685	4.46136	5.95527	2.593209	4.240162	2.626397	6.249685	4.428193	6.7802	...	3.956887	8.950125	7.14277	5.365162

Similarly, we have also found the cosine similarity of the old and new query with the documents.

Output:

```
sim_documents
```

	D000451	D000640	D000655	D000715	D000717	D000722	D000724	D000781	D000863	D000881	...	D000804	D000195	D000201	D000248	D000267
Q	0.030363	0.042787	0.062930	0.034433	0.053253	0.030803	0.037903	0.052951	0.043168	0.046173	...	0.015526	0.000000	0.000000	0.012360	0.012360
Q1	0.049905	0.062739	0.077363	0.052517	0.070795	0.042710	0.058867	0.063837	0.057259	0.062395	...	0.028502	0.004361	0.008367	0.021112	0.017589

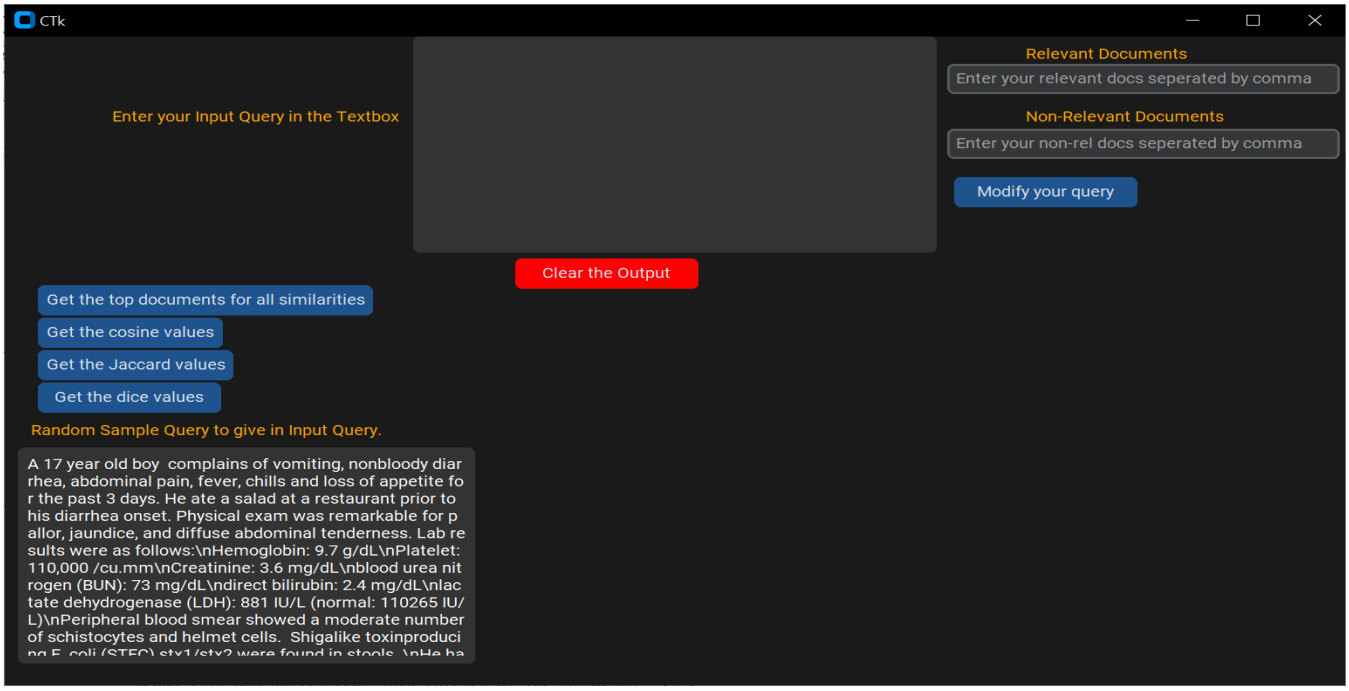
<u>GUI APPLICATION:</u>

We have created an interactive implementation of our search retrieval system using Tkinter module, where we enter the query and it gives us the combined results of the cosine similarity, jaccard similarity and dice similarity.

We also ask the user about the relevant and non-relevant documents and shows the suggested vocabulary terms which can be helpful in retrieving the desired sets of documents.

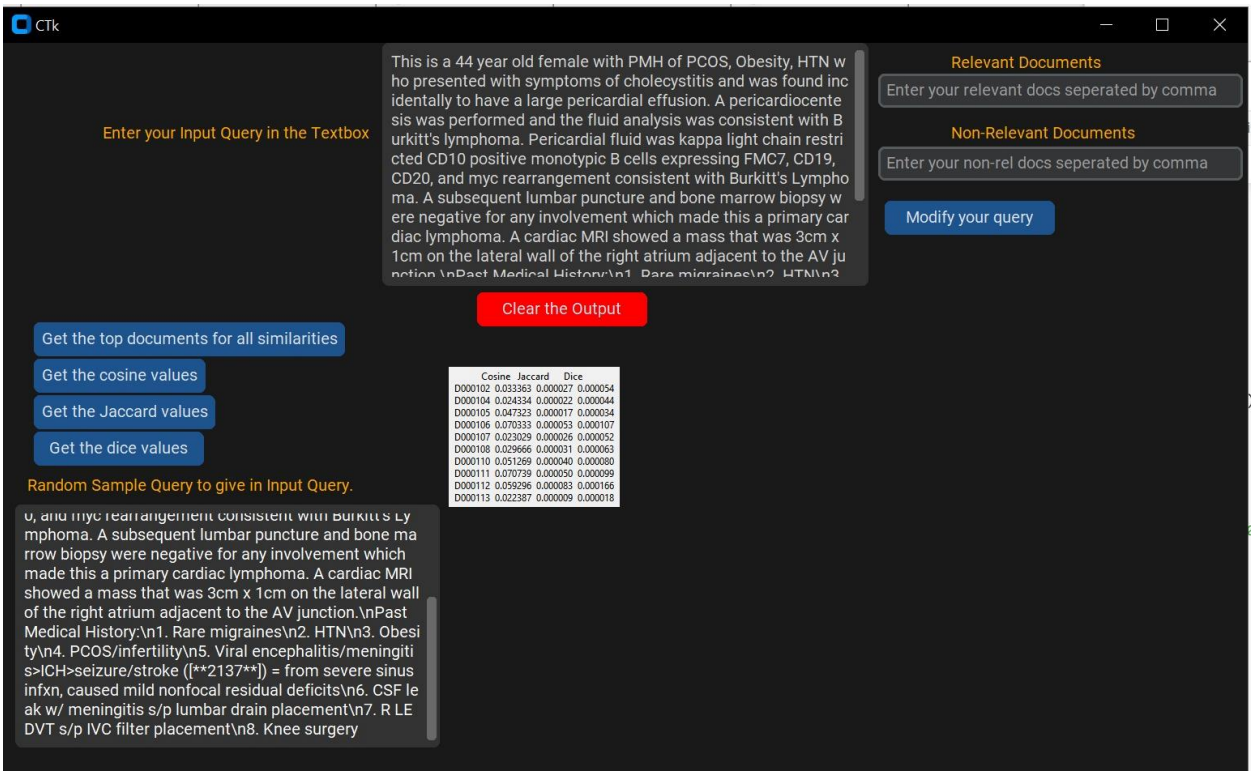
Since the queries are large, we have also created a random generator of query which is given to the user to select and see the result.

Output:

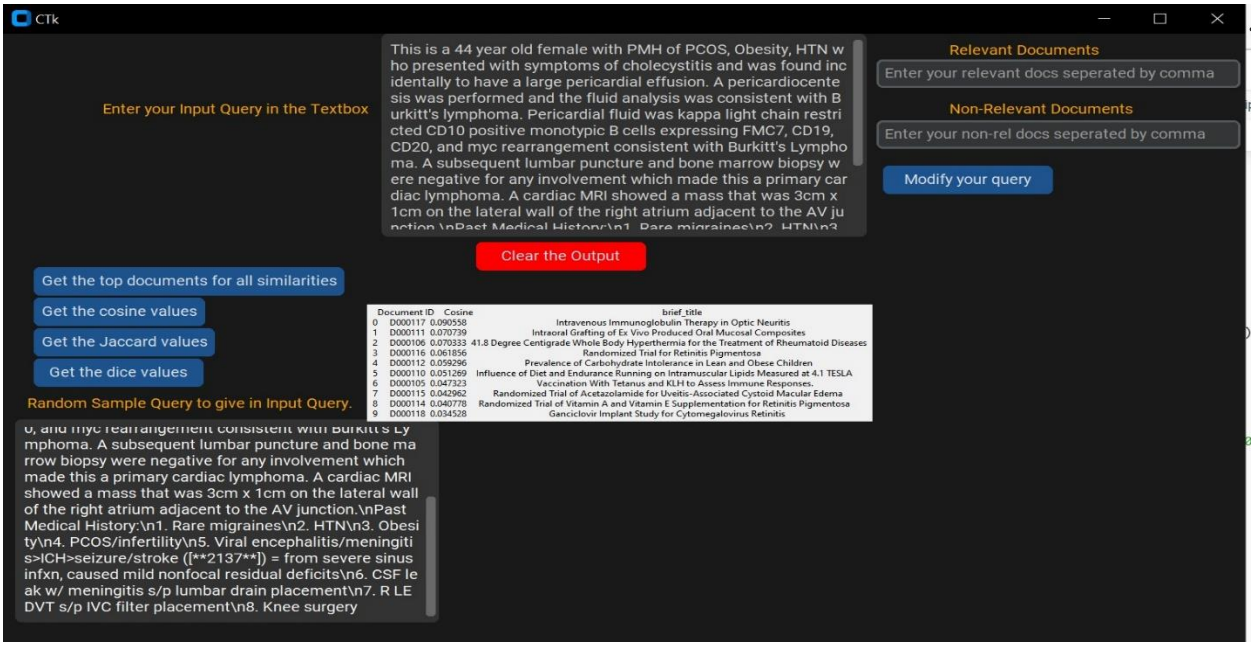


Entering the query:

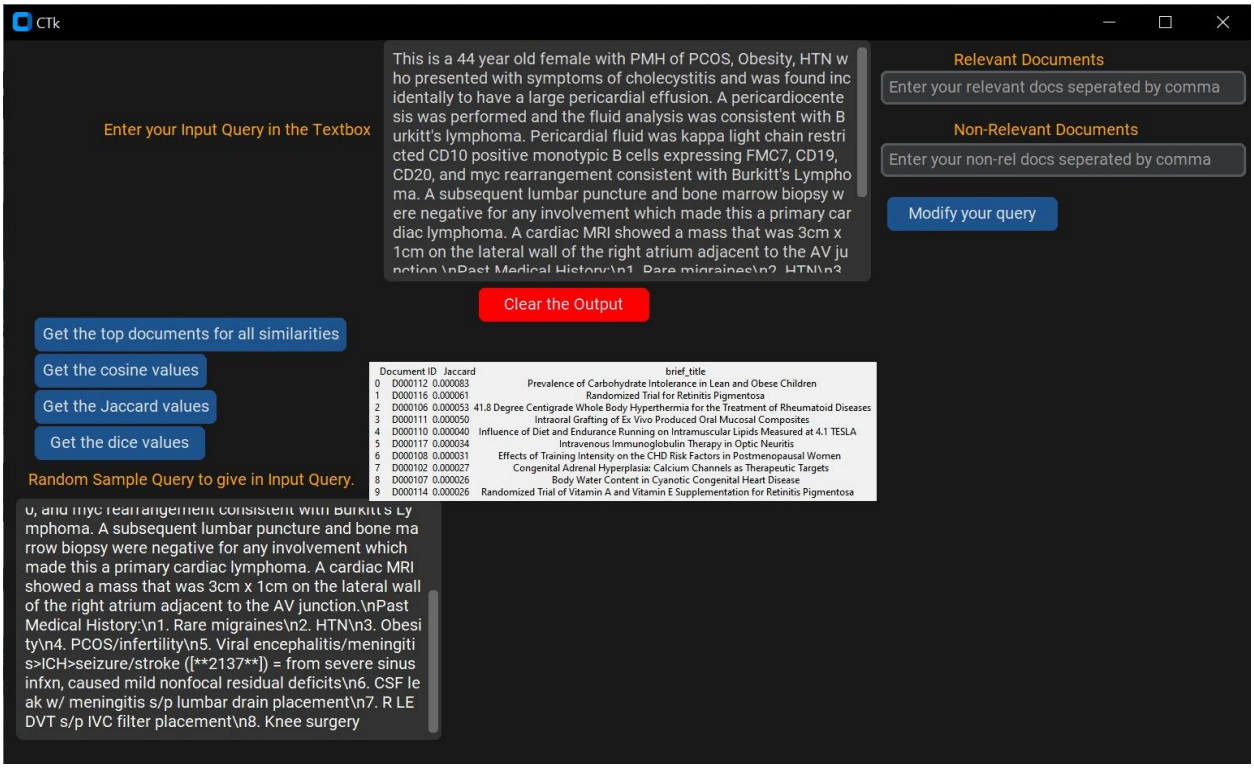
Getting all the similarity scores, with respect to the top 10 similar documents.



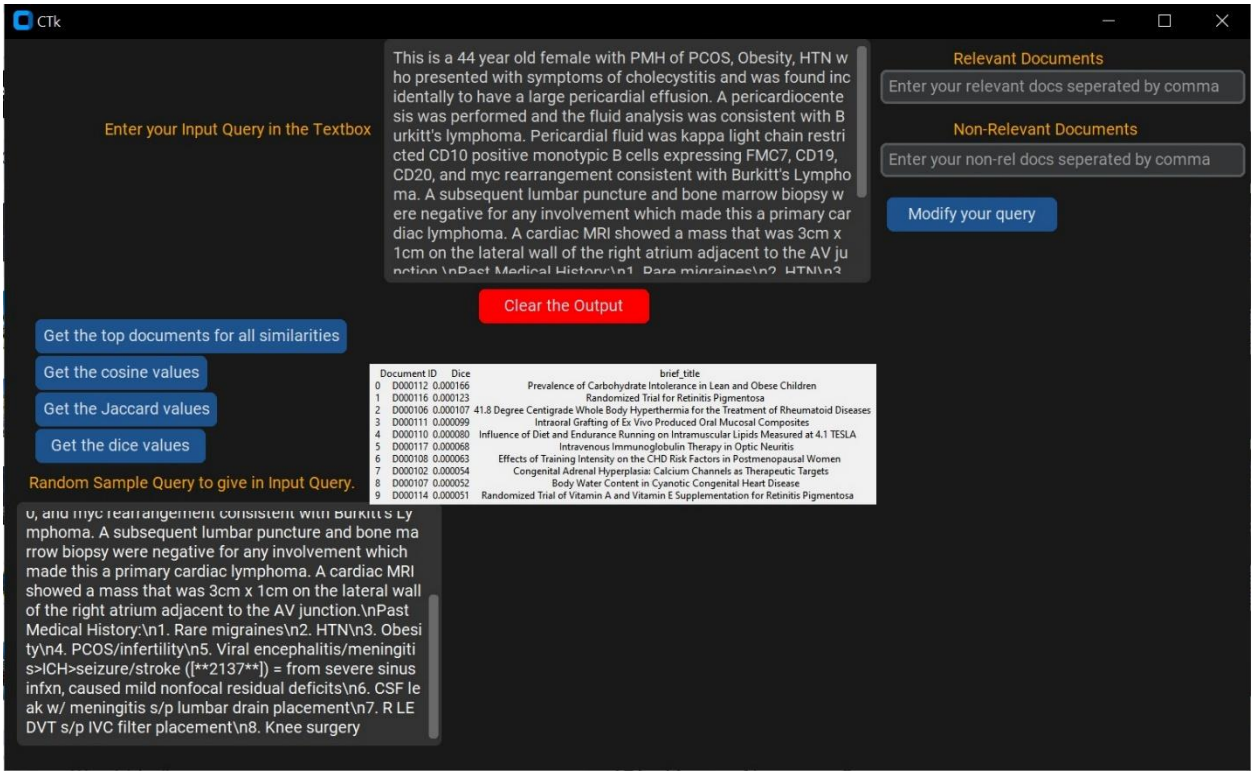
Displaying the cosine similarity scores:



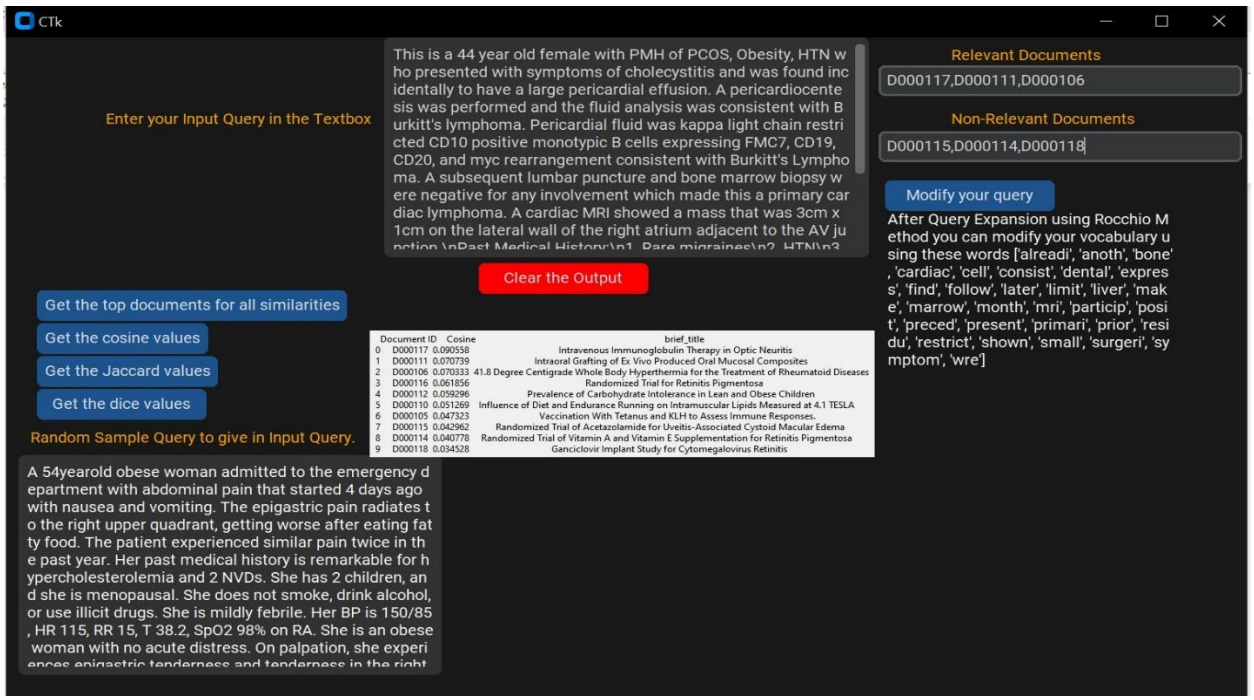
Displaying the jaccard similarity scores:



Displaying the dice similarity scores:



Giving the relevant and non-relevant documents it gives us the words that we can use in our vocabulary.



Evaluation:

The dataset used to evaluate our search/retrieval system is Medline dataset from the University of Glasgow.

The dataset can be obtained on http://ir.dcs.gla.ac.uk/resources/test_collections/medl/

We have performed same pre-processing steps, term-weighting and inverted index as we have performed above and we got an precision of 0.6 and recall of 0.25.

From the above precision and recall values, we can say that our search/retrieval system works fairly well.

Appendices:

Here is the Jupyter Notebook and the HTML file for the search engine retrieval.

CSC_575_Final_Project.html

CSC_575_Final_Project.ipynb

YouTube Link:
https://youtu.be/Kz1P5lzgqic