FINAL PROJECT REPORT

CSC 575 : Intelligent Information Retrieval

Search Retrieval System

AYESHA ALI , SWATHI BABU

CONTENTS

ABOUT THE PROJECT:

The project is a SEARCH RETRIEVAL SYSTEM that runs on the local directory. The system has a crawler, indexer, and query processing components. The system recommends the top 10 documents based on the given query using the inverted index and cosine ranking. It also uses Rocchio relevance feedback to improve the system by receiving more relevant documents for the given query.

DESCRIPTION OF THE SYSTEM:

1. **Crawler**

   We have a local based crawler that crawls through, local directory containing documents. This function implements some of the os module functions like the walk function. Generally this crawler will go through the local directory, each file as document is preprocessed to remove stop words, punctuations and converting to lowercase. This preprocessing is done by a separate function **preprocess_words(line)**. We have imported nltk modules such as PorterStemmer to stem the tokens and remove stopwords to preprocess our tokens. After preprocessing, the system creates an inverted index for each document.

2. **Inverted Index**

   Primarily, our inverted index is a dictionary, where key is token and its value is a nested dictionary containing the docId as key and count of the token in that document as value. After the preprocessing step, the inverted index construction is done by **create_inverted_index(doc, docid, inverted_index):** .
   The docid is the name of the file identified when the crawler crawls through the directory.

   Simultaneously , the doc lengths dictionary is also constructed using the function **doc_lengths(inverted_index, docLengths, NDoc)** . This function utilizes the **tf_idf(dict, N) ,** such the for each token in inverted index ,we computed the IDF weight for each token , and incremented the length of D by $(I * C)^2$. The document length in Doc Lengths dictionary is set to the square root of the current document length. (referred from Implementation Notes).

   The entire process of creating Inverted Index and Doc Lengths is done for one time initially by calling the function **first_time(filepath = '/Users/swathib/Downloads/med/MED.ALL').** (filepath given initially for test purposes.). It saves the inverted index, doc_lengths dictionaries as json files in our local storage which are retrieved each time the system starts.

3. **Query Processing**

The main process of the system starts now. The **main()** function is called where the inverted index file (inverted_index.json) and the doc lengths file (doc_lengths.json) that were previously saved are retrieved to start the system. Then the total number of documents **NDoc** is calculated. This happens everytime the system starts. Then the system asks the user for a query using the input() function in python and is stored in the **query** variable. Once the user enters a query, the query undergoes the same preprocessing that the documents did using the **query_process(query)** function i.e., the leading spaces are removed, punctuation and stop words are removed from the query, and it is converted to lowercase and the tokens in the query are stemmed (using Porter stemmer algorithm). This function returns the preprocessed tokens of the query **p_query** and the vector form of the query required for further steps in the retrieval system in the form of a python dictionary **QtermDict** where key is the token and the value is the count of the token in the query (raw weight of the token for now).

4. Retrieval Scores

Using the query terms, query vector (in dictionary form), total number of documents and the inverted index, the scores of each document for the given query is calculated using the function **retrieval_scores(p_query, QtermDict, NDoc, inverted_index).** This function goes through each token in the query and calculates the idf value for that term using the **tf_idf(inverted_index[token], NDoc)** function. The idf value is used to calculate the query length **Qlength** by incrementing the score by the square of the idf value. The scores of the documents are stored in a dictionary **rankDict** with key as document id and value as the scores. Then for each document that contains the term, the score of that document is incremented if the document already had a previous parsed token (the docid is in the rankDict dictionary) by **count of the token in that document * idf * count of the token in the query *idf**. If the document is not in the rankDict, it is initialized with the same value rather than being incremented.The function then applies square root on the query length. The function returns the calculated query length and the dictionary with the scores for the documents retrieved for the given query. Then the control flow comes back to the main function where the **cosine_Ranking(rankDict, Qlength,docLengths)** function is called to calculate the cosine ranking for the retrieved documents using the previously calculated scores, query length and the doc lengths that was retrieved from local storage. This cosine ranking function returns the dictionary **rankedDocs** sorted by the cosine ranking with key as the retrieved document id and the value as cosine ranking. The main function then calls the **printDocs(rankedDocs)** function to print the results in a user-friendly way. The printDocs function goes through the top 10 retrieved documents and prints the document id and using the function **print_title(docid)** prints the title of the document retrieved. The function print_title is especially for the dataset used in the evaluation of this project but it can very easily be customized to work on other documents. Followed by this, relevance feedback is applied based on user input.

5. Rocchio Algorithm

The system asks the user whether they want to provide feedback on the relevance of the documents. On confirmation, the users would be prompted again to enter the list of the relevant document IDs. Once the list is given, it is passed as argument to the function **feedback_user(QtermDict, relevant, 10,rankedDocs,inverted_index,docLengths,NDoc)** . This function returns the ranked Documents .

The **feedback_user(QtermDict, relevant, 10,rankedDocs,inverted_index,docLengths,NDoc)** implements all the process in **Step 4** but after the function of Rocchio Algorithm.

We implemented the standard IR approach Rocchio Method.

The **rocchio_process(rankedDocs, rel, QtermDict,k,inverted_index):** We create a dataframe of the relevant and the non relevant documents from the top 10 ranked Documents and counts from the inverted index (not a full document -term matrix). It's only the dataframe document term matrix of 10 documents.

We also considered a scenario where a query token may not exist in the inverted index, in that case the query token is added to the dataframe with count 0 for each of the documents.

We create numpy arrays for each of the Q vector, Relevant and Non Relevant documents
And call the **modify_query_rocchio(Q, R, NR, alpha, beta).** This function implements the following formula:

Rocchio formula : Q1= Q+(alpha*mean_R)-(beta*mean_NR)
We considered alpha= 0.5 and beta =0.25

We get the new query with modified term weights from the existing query.

Using the modified query, we perform the same processes in **Step 4 : Retrieval Scores** to get retrieval scores and cosine ranked top 10 documents.

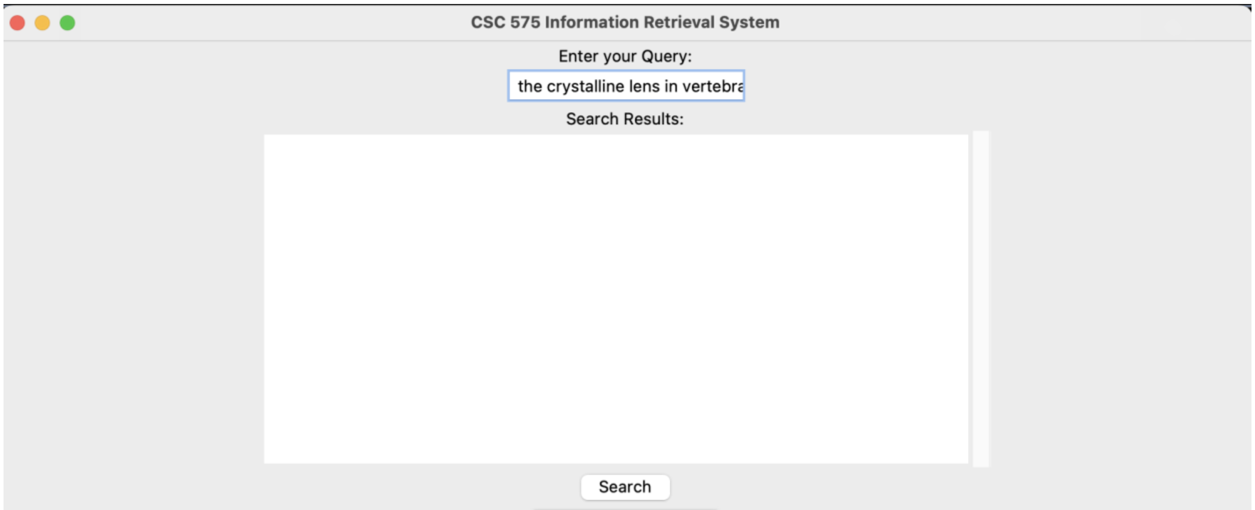
Upon getting the result of 10 documents, we observed that we got better results in terms of more relevant documents than the previous result without relevance feedback . We have cross checked from our benchmark test collections that contain the relevance documents for each query.

The user also has the option to skip providing feedback on the relevance of the documents, otherwise the user can enter the relevant documents and get the ranked documents . If the user would like to ask another query, they would choose the ‘Click for new Search’ button which would let them repeat this process in a new window.

TEST RUNS ON MULTIPLE QUERY INPUTS AND THE INTERMEDIATE OUTPUTS:

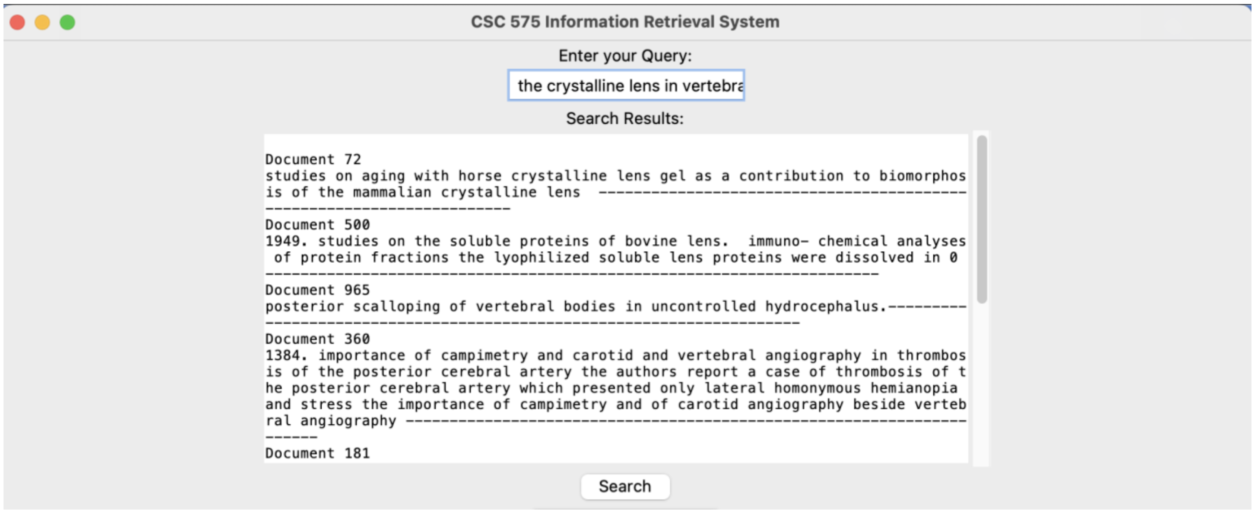
TEST 1 → Running a Query with Relevance feedback

Asking the user for the query



the crystalline lens in vertebra

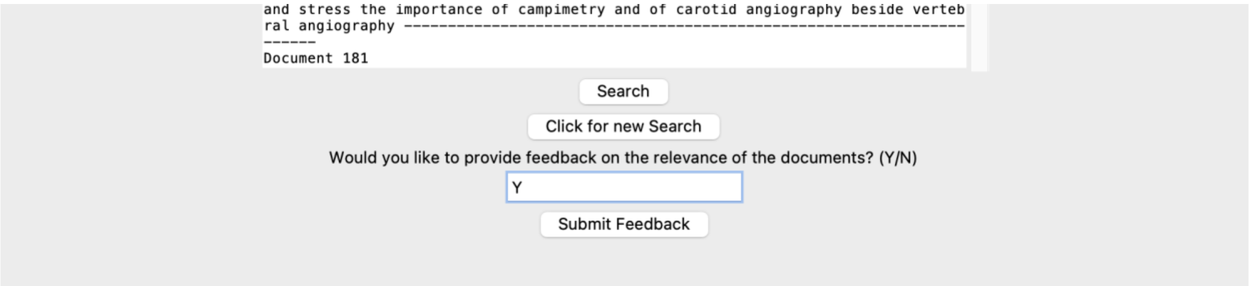
The user then gets the top 10 retrieved documents for the above query which can be seen by scrolling



the crystalline lens in vertebra

```
Document 72
studies on aging with horse crystalline lens gel as a contribution to biomorphos
is of the mammalian crystalline lens  ------------------------------------------
----------------------------
Document 500
1949. studies on the soluble proteins of bovine lens.  immuno- chemical analyses
 of protein fractions the lyophilized soluble lens proteins were dissolved in 0
----------------------------------------------------------------------
Document 965
posterior scalloping of vertebral bodies in uncontrolled hydrocephalus.---------
-------------------------------------------------------------
Document 360
1384. importance of campimetry and carotid and vertebral angiography in thrombos
is of the posterior cerebral artery the authors report a case of thrombosis of t
he posterior cerebral artery which presented only lateral homonymous hemianopia
and stress the importance of campimetry and of carotid angiography beside verteb
ral angiography ----------------------------------------------------------------
------
Document 181
```

Following this, the user is asked if they would like to give feedback

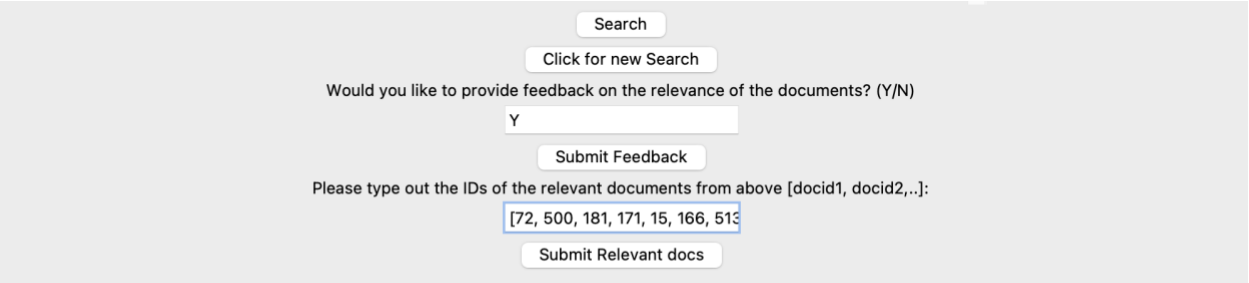


```
and stress the importance of campimetry and of carotid angiography beside verteb
ral angiography ----------------------------------------------------------------
------
Document 181
```

Would you like to provide feedback on the relevance of the documents? (Y/N)

Y

If the user says yes, then the user is asked to give a list of relevant documents



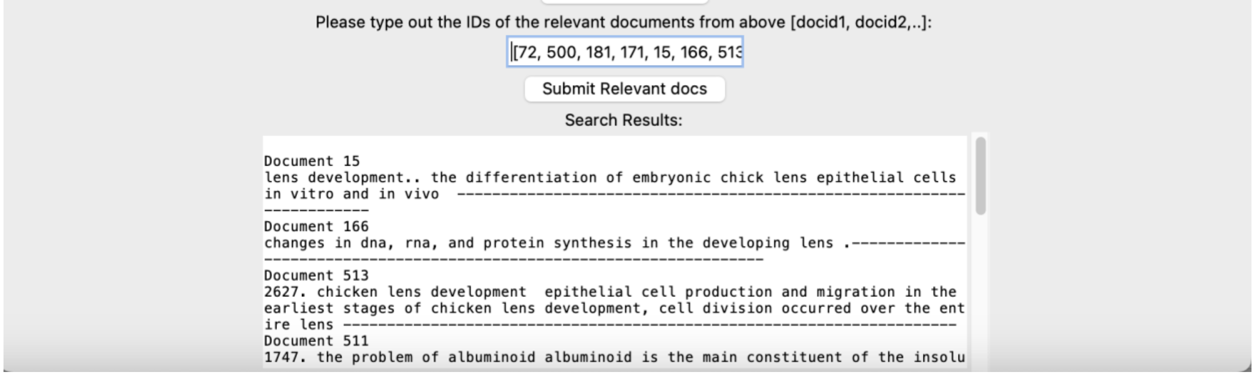
Would you like to provide feedback on the relevance of the documents? (Y/N)

Y

Please type out the IDs of the relevant documents from above [docid1, docid2,..]:

[72, 500, 181, 171, 15, 166, 513

Then the list of the top 10 retrieved documents for the Rocchio altered query is printed



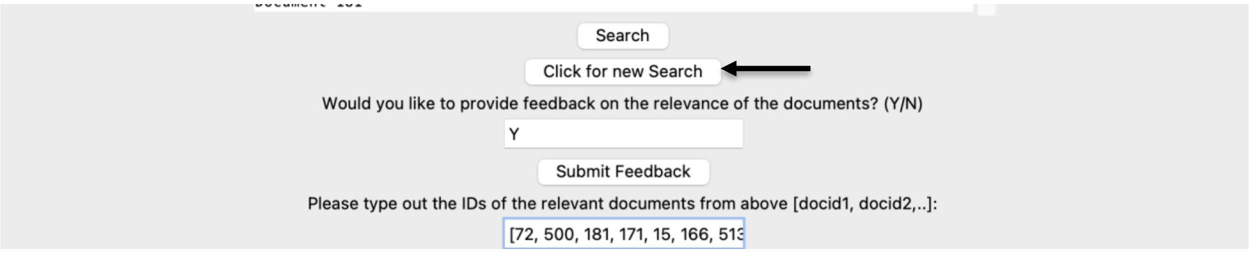
Please type out the IDs of the relevant documents from above [docid1, docid2,..]:

[72, 500, 181, 171, 15, 166, 513

Search Results:

```
Document 15
lens development.. the differentiation of embryonic chick lens epithelial cells
in vitro and in vivo  ----------------------------------------------------------
------------
Document 166
changes in dna, rna, and protein synthesis in the developing lens .-------------
-------------------------------------------------------
Document 513
2627. chicken lens development  epithelial cell production and migration in the
earliest stages of chicken lens development, cell division occurred over the ent
ire lens -----------------------------------------------------------------------
Document 511
1747. the problem of albuminoid albuminoid is the main constituent of the insolu
```

Then the system can be closed by shutting down the window or they can click on new search to continue



Would you like to provide feedback on the relevance of the documents? (Y/N)

Y

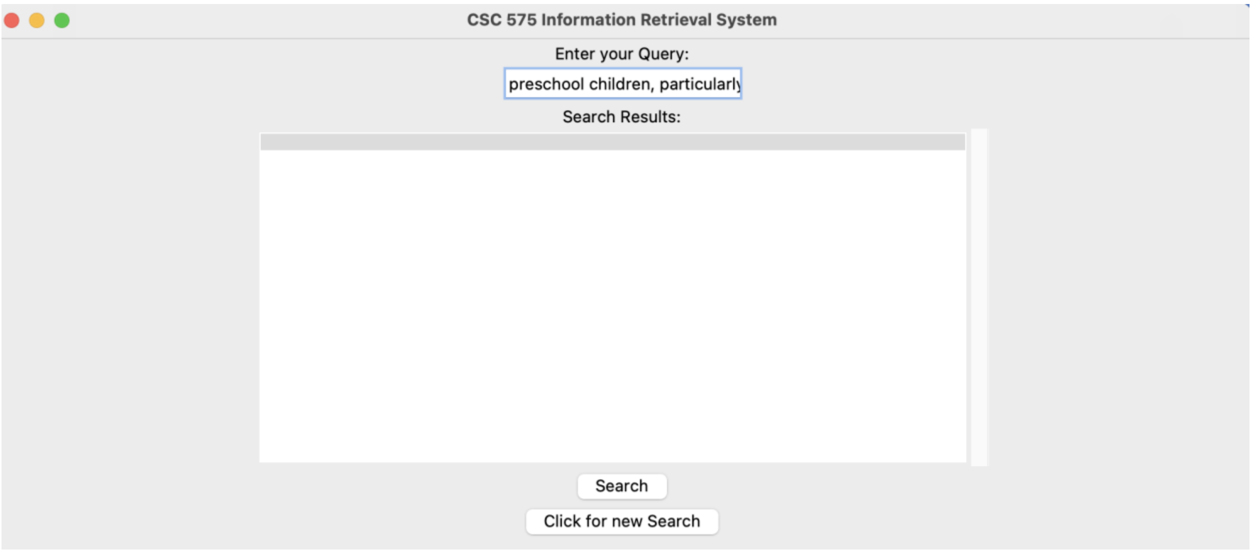
Please type out the IDs of the relevant documents from above [docid1, docid2,..]:

[72, 500, 181, 171, 15, 166, 513

Since the user clicked on it, the system asks for the next query in a new window.

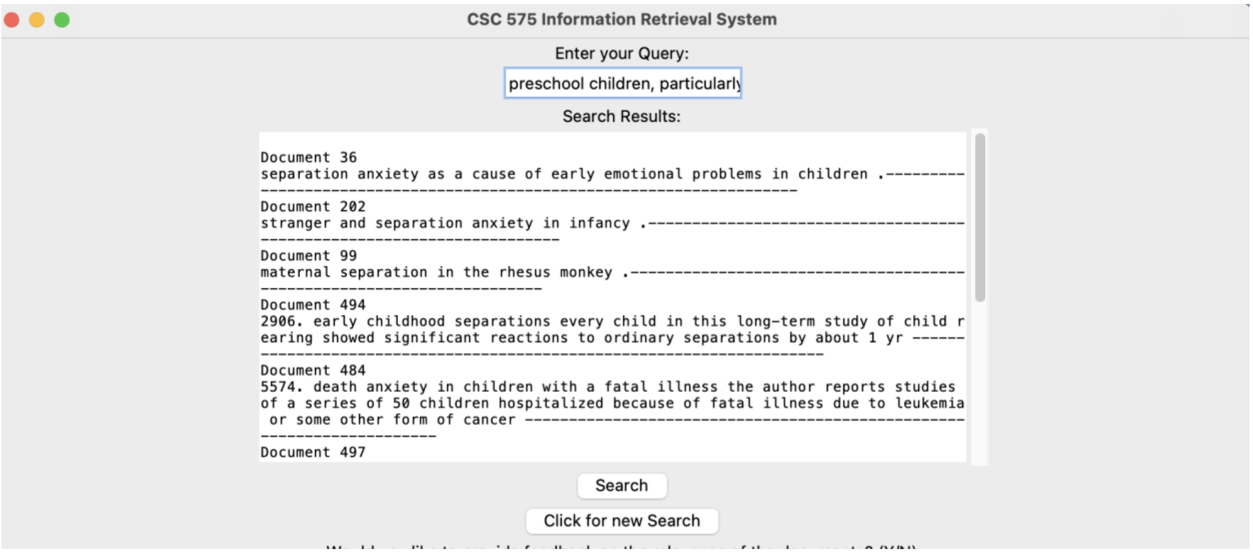
TEST 2 → Running another query without Relevance feedback

The system asks the user for the query



preschool children, particularly

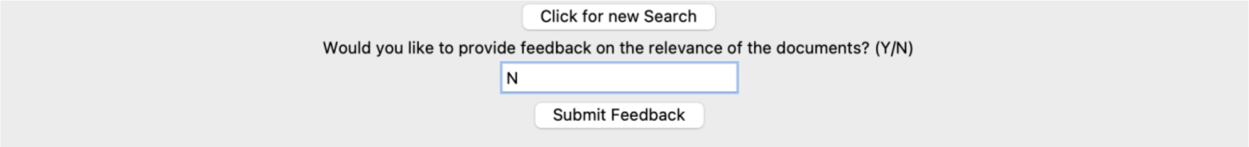
The system prints the top 10 documents



preschool children, particularly

```
Document 36
separation anxiety as a cause of early emotional problems in children .---------
-------------------------------------------------------------
Document 202
stranger and separation anxiety in infancy .------------------------------------
----------------------------------
Document 99
maternal separation in the rhesus monkey .--------------------------------------
--------------------------------
Document 494
2906. early childhood separations every child in this long-term study of child r
earing showed significant reactions to ordinary separations by about 1 yr ------
----------------------------------------------------------------
Document 484
5574. death anxiety in children with a fatal illness the author reports studies 
of a series of 50 children hospitalized because of fatal illness due to leukemia
 or some other form of cancer --------------------------------------------------
--------------------
Document 497
```

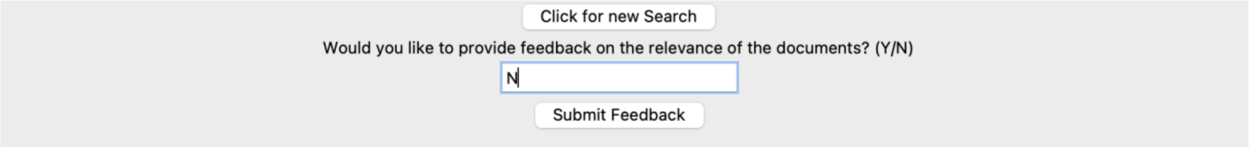
The system asks if the user would like to give feedback.



Would you like to provide feedback on the relevance of the documents? (Y/N)

N

If the user says no, the system stops. Then the system can be closed by shutting down the window or they can click on new search to continue



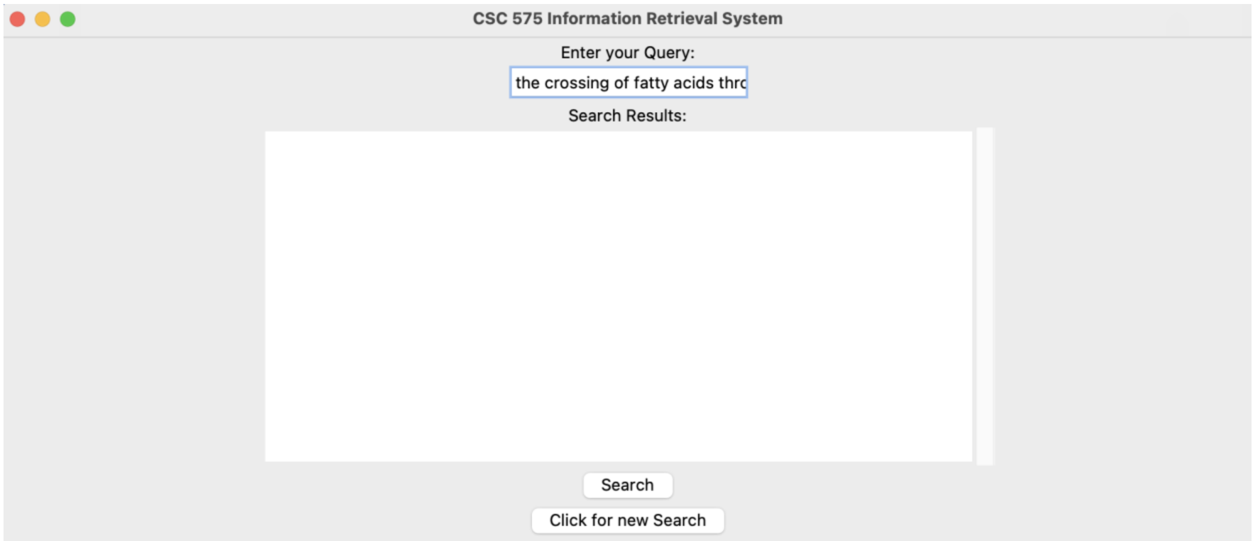
Would you like to provide feedback on the relevance of the documents? (Y/N)

N

The system proceeds as usual since we the user clicked on new search. We can run two more queries to test the efficiency of the system

TEST 3 → Testing another query with Relevance feedback

Asking the user for the query

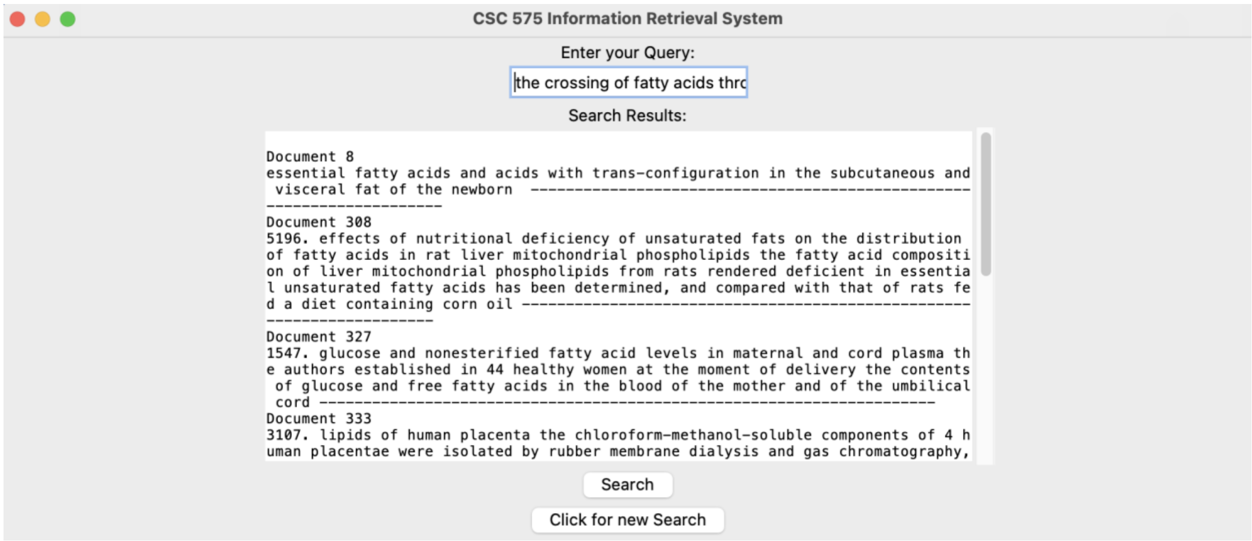


Enter your Query:

the crossing of fatty acids thro

Search Results:

The user then gets the top 10 retrieved documents for the above query



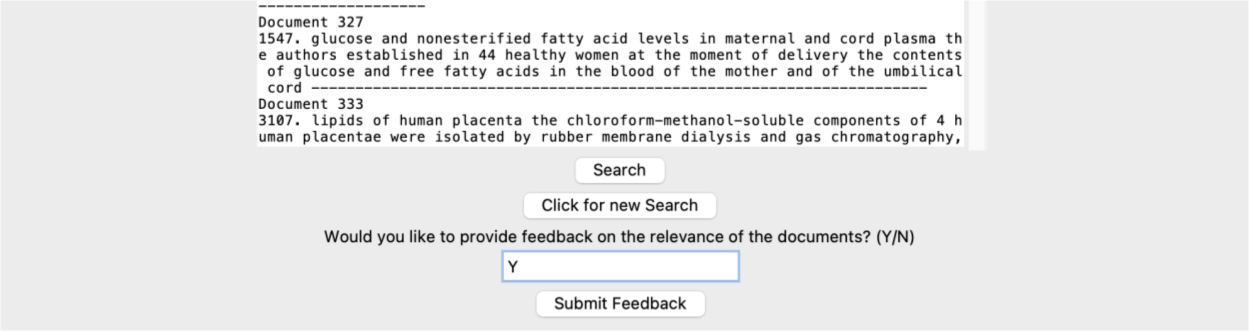
Enter your Query:

the crossing of fatty acids thro

Search Results:

```
Document 8
essential fatty acids and acids with trans-configuration in the subcutaneous and
 visceral fat of the newborn  --------------------------------------------------
--------------------
Document 308
5196. effects of nutritional deficiency of unsaturated fats on the distribution
of fatty acids in rat liver mitochondrial phospholipids the fatty acid compositi
on of liver mitochondrial phospholipids from rats rendered deficient in essentia
l unsaturated fatty acids has been determined, and compared with that of rats fe
d a diet containing corn oil ---------------------------------------------------
-------------------
Document 327
1547. glucose and nonesterified fatty acid levels in maternal and cord plasma th
e authors established in 44 healthy women at the moment of delivery the contents
 of glucose and free fatty acids in the blood of the mother and of the umbilical
 cord ----------------------------------------------------------------------
Document 333
3107. lipids of human placenta the chloroform-methanol-soluble components of 4 h
uman placentae were isolated by rubber membrane dialysis and gas chromatography,
```

Following this, the user is asked if they would like to give feedback

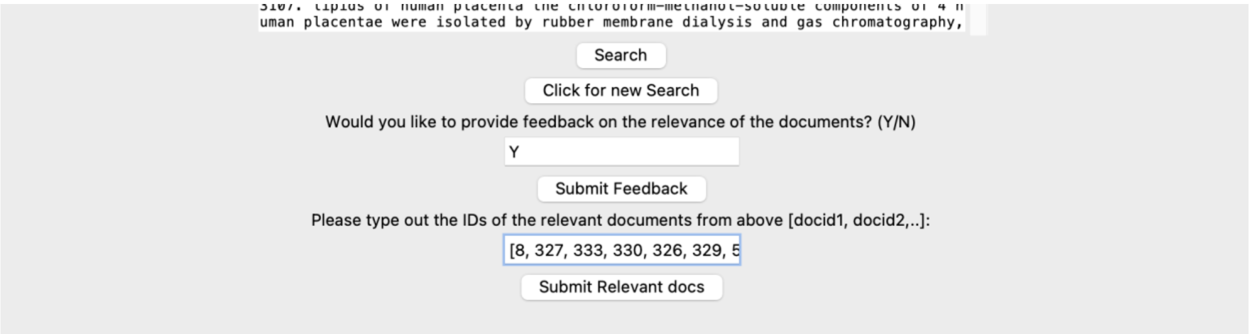


```
--------------------
Document 327
1547. glucose and nonesterified fatty acid levels in maternal and cord plasma th
e authors established in 44 healthy women at the moment of delivery the contents
 of glucose and free fatty acids in the blood of the mother and of the umbilical
 cord ---------------------------------------------------------------------------
Document 333
3107. lipids of human placenta the chloroform-methanol-soluble components of 4 h
uman placentae were isolated by rubber membrane dialysis and gas chromatography,
```

Would you like to provide feedback on the relevance of the documents? (Y/N)

Y

If the user says yes, then the user is asked to give a list of relevant documents



```
3107. lipids of human placenta the chloroform-methanol-soluble components of 4 h
uman placentae were isolated by rubber membrane dialysis and gas chromatography,
```

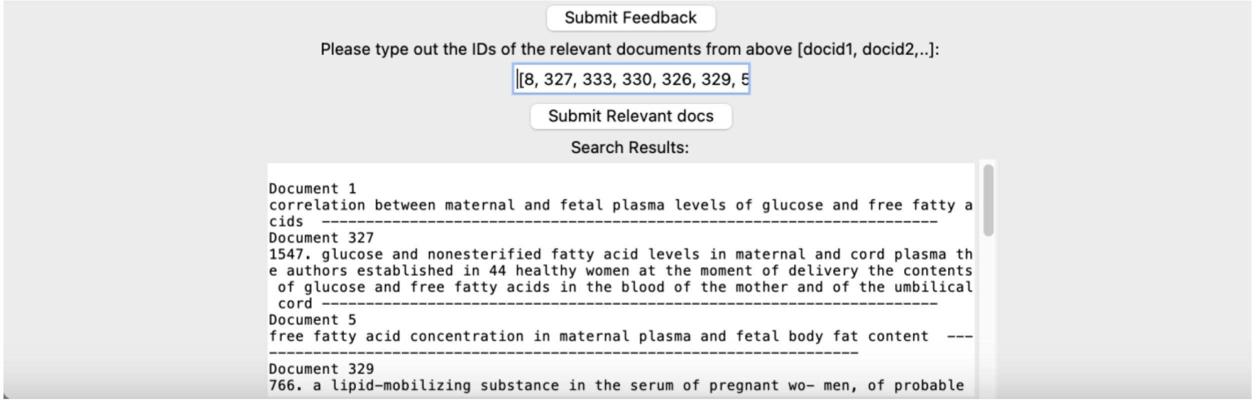
Would you like to provide feedback on the relevance of the documents? (Y/N)

Y

Please type out the IDs of the relevant documents from above [docid1, docid2,..]:

[8, 327, 333, 330, 326, 329, 5

Then the list of the top 10 retrieved documents for the Rocchio altered query is printed



Please type out the IDs of the relevant documents from above [docid1, docid2,..]:

[8, 327, 333, 330, 326, 329, 5

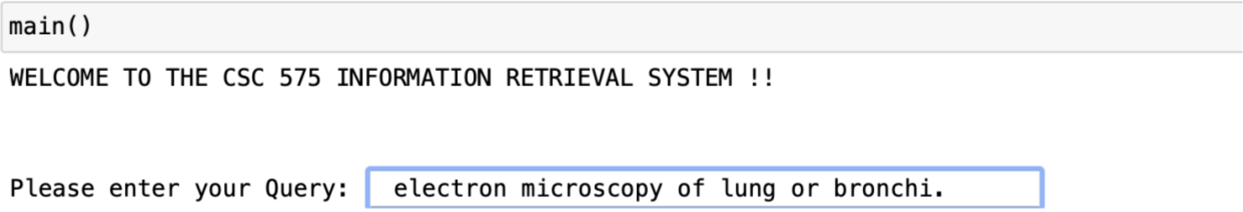
Search Results:

```
Document 1
correlation between maternal and fetal plasma levels of glucose and free fatty a
cids  ---------------------------------------------------------------------------
Document 327
1547. glucose and nonesterified fatty acid levels in maternal and cord plasma th
e authors established in 44 healthy women at the moment of delivery the contents
 of glucose and free fatty acids in the blood of the mother and of the umbilical
 cord ---------------------------------------------------------------------------
Document 5
free fatty acid concentration in maternal plasma and fetal body fat content  ---
----------------------------------------------------------------------------
Document 329
766. a lipid-mobilizing substance in the serum of pregnant wo- men, of probable
```

Then the user can continue or stop the system.

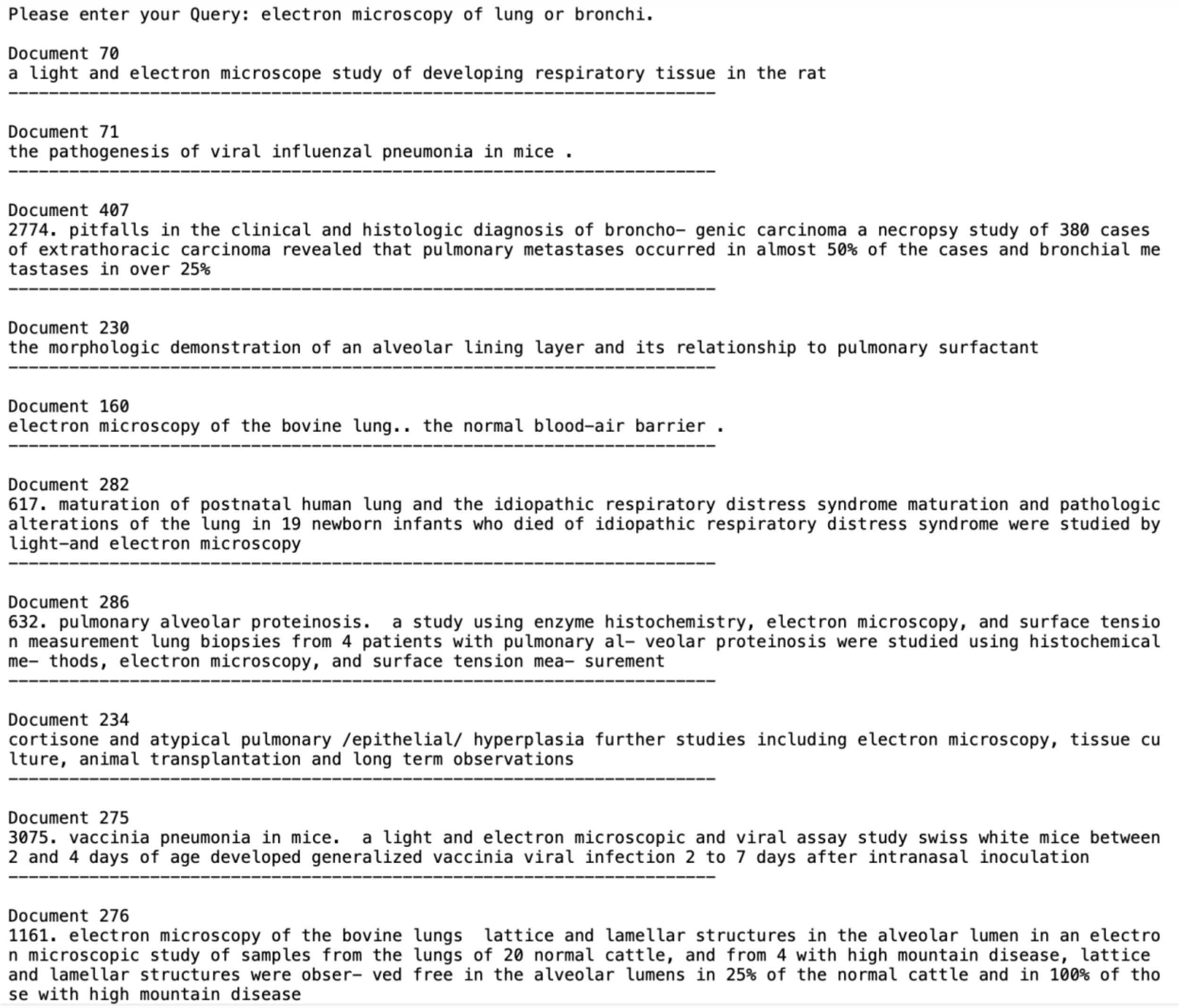
TEST 4 → Testing another query with Relevance feedback with all the intermediate outputs printed (using Jupyter Notebook to see the entire output instead of our GUI)

Entering the query



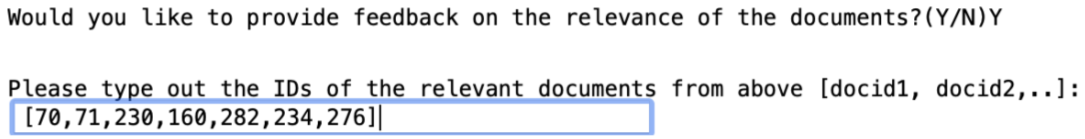
```
main()

WELCOME TO THE CSC 575 INFORMATION RETRIEVAL SYSTEM !!


Please enter your Query:  electron microscopy of lung or bronchi.
```

The user then gets the top 10 retrieved documents for the above query



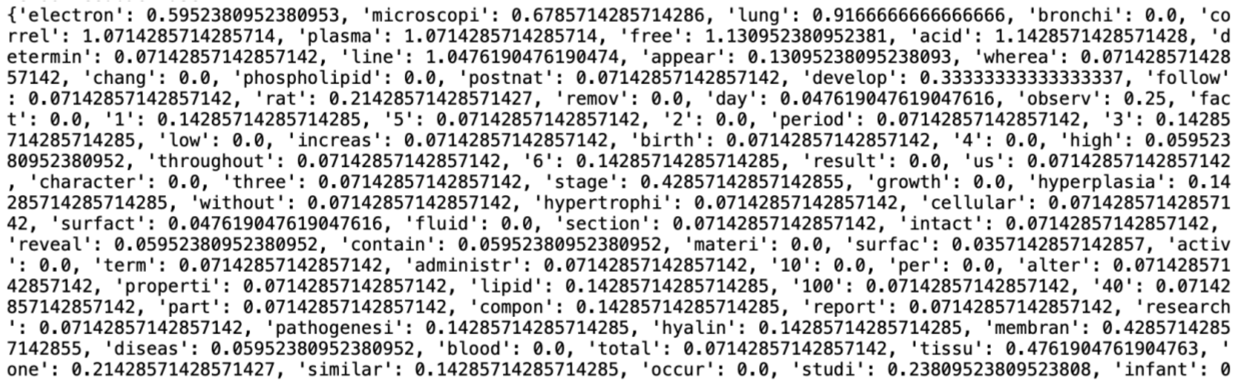
```
Please enter your Query: electron microscopy of lung or bronchi.

Document 70
a light and electron microscope study of developing respiratory tissue in the rat
----------------------------------------------------------------------

Document 71
the pathogenesis of viral influenzal pneumonia in mice .
----------------------------------------------------------------------

Document 407
2774. pitfalls in the clinical and histologic diagnosis of broncho- genic carcinoma a necropsy study of 380 cases
of extrathoracic carcinoma revealed that pulmonary metastases occurred in almost 50% of the cases and bronchial me
tastases in over 25%
----------------------------------------------------------------------

Document 230
the morphologic demonstration of an alveolar lining layer and its relationship to pulmonary surfactant
----------------------------------------------------------------------

Document 160
electron microscopy of the bovine lung.. the normal blood-air barrier .
----------------------------------------------------------------------

Document 282
617. maturation of postnatal human lung and the idiopathic respiratory distress syndrome maturation and pathologic
alterations of the lung in 19 newborn infants who died of idiopathic respiratory distress syndrome were studied by
light-and electron microscopy
----------------------------------------------------------------------

Document 286
632. pulmonary alveolar proteinosis.  a study using enzyme histochemistry, electron microscopy, and surface tensio
n measurement lung biopsies from 4 patients with pulmonary al- veolar proteinosis were studied using histochemical
me- thods, electron microscopy, and surface tension mea- surement
----------------------------------------------------------------------

Document 234
cortisone and atypical pulmonary /epithelial/ hyperplasia further studies including electron microscopy, tissue cu
lture, animal transplantation and long term observations
----------------------------------------------------------------------

Document 275
3075. vaccinia pneumonia in mice.  a light and electron microscopic and viral assay study swiss white mice between
2 and 4 days of age developed generalized vaccinia viral infection 2 to 7 days after intranasal inoculation
----------------------------------------------------------------------

Document 276
1161. electron microscopy of the bovine lungs  lattice and lamellar structures in the alveolar lumen in an electro
n microscopic study of samples from the lungs of 20 normal cattle, and from 4 with high mountain disease, lattice
and lamellar structures were obser- ved free in the alveolar lumens in 25% of the normal cattle and in 100% of tho
se with high mountain disease
```

Following this, the user is asked if they would like to give feedback



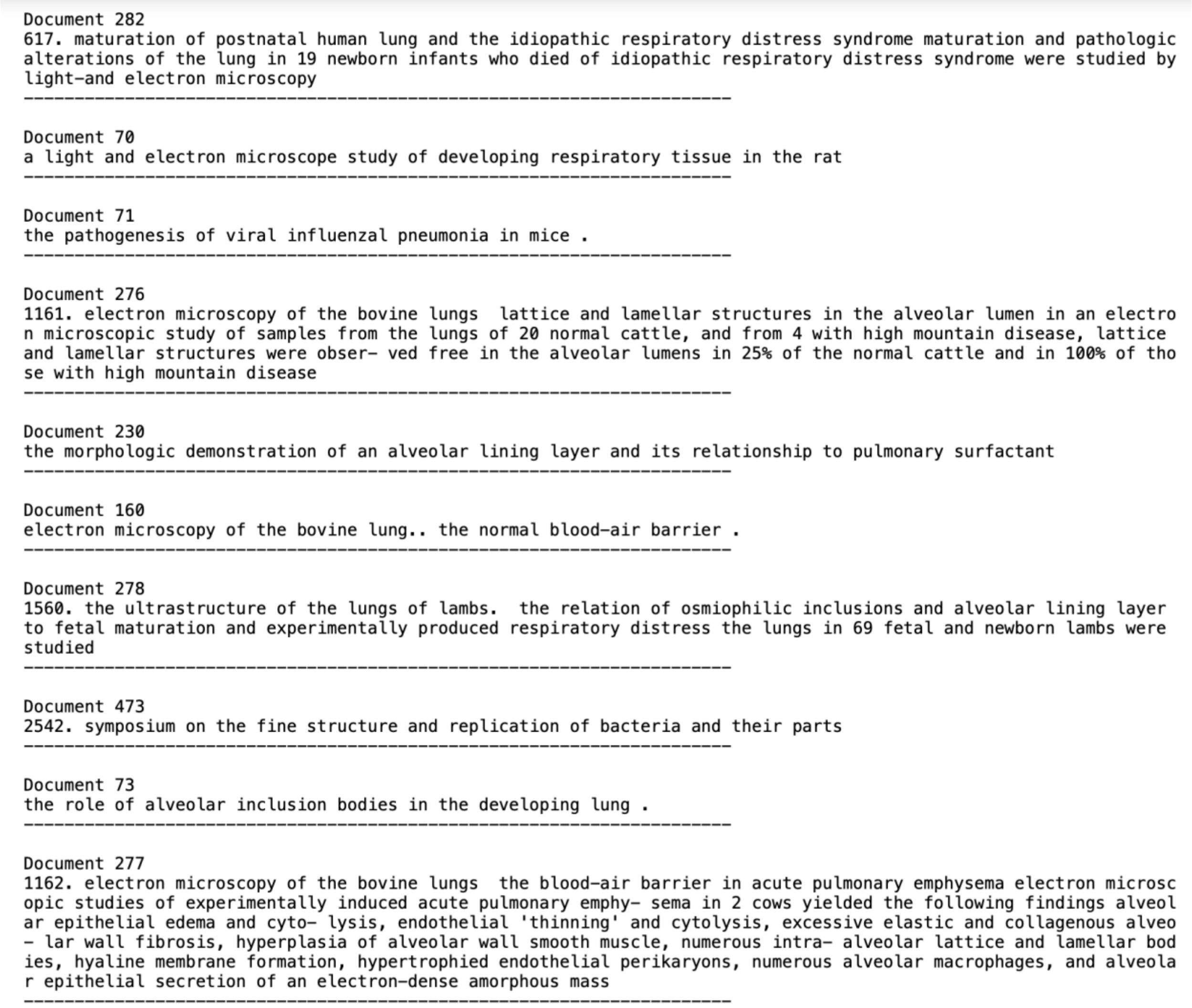
```
Would you like to provide feedback on the relevance of the documents?(Y/N)Y

Please type out the IDs of the relevant documents from above [docid1, docid2,..]:
 [70,71,230,160,282,234,276]
```

This is what our modified query term weights look like after Rocchio method: (part of the dictionary)



```
{'electron': 0.5952380952380953, 'microscopi': 0.6785714285714286, 'lung': 0.9166666666666666, 'bronchi': 0.0, 'co
rrel': 1.0714285714285714, 'plasma': 1.0714285714285714, 'free': 1.130952380952381, 'acid': 1.1428571428571428, 'd
etermin': 0.07142857142857142, 'line': 1.0476190476190474, 'appear': 0.13095238095238093, 'wherea': 0.071428571428
57142, 'chang': 0.0, 'phospholipid': 0.0, 'postnat': 0.07142857142857142, 'develop': 0.33333333333333337, 'follow'
: 0.07142857142857142, 'rat': 0.21428571428571427, 'remov': 0.0, 'day': 0.047619047619047616, 'observ': 0.25, 'fac
t': 0.0, '1': 0.14285714285714285, '5': 0.07142857142857142, '2': 0.0, 'period': 0.07142857142857142, '3': 0.14285
714285714285, 'low': 0.0, 'increas': 0.07142857142857142, 'birth': 0.07142857142857142, '4': 0.0, 'high': 0.059523
80952380952, 'throughout': 0.07142857142857142, '6': 0.14285714285714285, 'result': 0.0, 'us': 0.07142857142857142
, 'character': 0.0, 'three': 0.07142857142857142, 'stage': 0.42857142857142855, 'growth': 0.0, 'hyperplasia': 0.14
285714285714285, 'without': 0.07142857142857142, 'hypertrophi': 0.07142857142857142, 'cellular': 0.071428571428571
42, 'surfact': 0.047619047619047616, 'fluid': 0.0, 'section': 0.07142857142857142, 'intact': 0.07142857142857142,
'reveal': 0.05952380952380952, 'contain': 0.05952380952380952, 'materi': 0.0, 'surfac': 0.0357142857142857, 'activ
': 0.0, 'term': 0.07142857142857142, 'administr': 0.07142857142857142, '10': 0.0, 'per': 0.0, 'alter': 0.071428571
42857142, 'properti': 0.07142857142857142, 'lipid': 0.14285714285714285, '100': 0.07142857142857142, '40': 0.07142
857142857142, 'part': 0.07142857142857142, 'compon': 0.14285714285714285, 'report': 0.07142857142857142, 'research
': 0.07142857142857142, 'pathogenesi': 0.14285714285714285, 'hyalin': 0.14285714285714285, 'membran': 0.4285714285
7142855, 'diseas': 0.05952380952380952, 'blood': 0.0, 'total': 0.07142857142857142, 'tissu': 0.4761904761904763, '
one': 0.21428571428571427, 'similar': 0.14285714285714285, 'occur': 0.0, 'studi': 0.23809523809523808, 'infant': 0
```

Then the list of the top 10 retrieved documents for the Rocchio altered query is printed



```
Document 282
617. maturation of postnatal human lung and the idiopathic respiratory distress syndrome maturation and pathologic
alterations of the lung in 19 newborn infants who died of idiopathic respiratory distress syndrome were studied by
light-and electron microscopy
----------------------------------------------------------------------

Document 70
a light and electron microscope study of developing respiratory tissue in the rat
----------------------------------------------------------------------

Document 71
the pathogenesis of viral influenzal pneumonia in mice .
----------------------------------------------------------------------

Document 276
1161. electron microscopy of the bovine lungs  lattice and lamellar structures in the alveolar lumen in an electro
n microscopic study of samples from the lungs of 20 normal cattle, and from 4 with high mountain disease, lattice
and lamellar structures were obser- ved free in the alveolar lumens in 25% of the normal cattle and in 100% of tho
se with high mountain disease
----------------------------------------------------------------------

Document 230
the morphologic demonstration of an alveolar lining layer and its relationship to pulmonary surfactant
----------------------------------------------------------------------

Document 160
electron microscopy of the bovine lung.. the normal blood-air barrier .
----------------------------------------------------------------------

Document 278
1560. the ultrastructure of the lungs of lambs.  the relation of osmiophilic inclusions and alveolar lining layer
to fetal maturation and experimentally produced respiratory distress the lungs in 69 fetal and newborn lambs were
studied
----------------------------------------------------------------------

Document 473
2542. symposium on the fine structure and replication of bacteria and their parts
----------------------------------------------------------------------

Document 73
the role of alveolar inclusion bodies in the developing lung .
----------------------------------------------------------------------

Document 277
1162. electron microscopy of the bovine lungs  the blood-air barrier in acute pulmonary emphysema electron microsc
opic studies of experimentally induced acute pulmonary emphy- sema in 2 cows yielded the following findings alveol
ar epithelial edema and cyto- lysis, endothelial 'thinning' and cytolysis, excessive elastic and collagenous alveo
- lar wall fibrosis, hyperplasia of alveolar wall smooth muscle, numerous intra- alveolar lattice and lamellar bod
ies, hyaline membrane formation, hypertrophied endothelial perikaryons, numerous alveolar macrophages, and alveola
r epithelial secretion of an electron-dense amorphous mass
----------------------------------------------------------------------
```

Retrieved docs: Before Rochio 70,71,407,230,160,282,286,234,275,276

No of Relevant Docs: (7)
Retrieved Docs After Rochio : 282,70,71,276,230,160,278,473,73,277

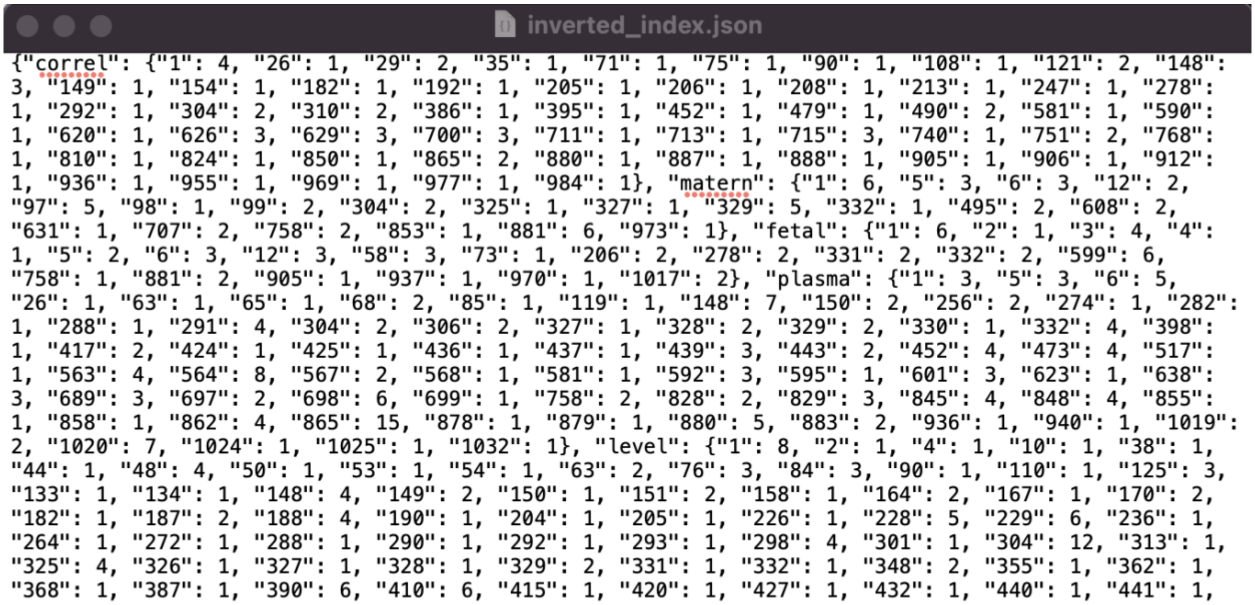
No of Relevant Docs: (8)

Relevant docs provided from the TestCollection (MED.REL)

[59,62,67,69,70,71,73,78,81,160,163,230,231,232,233,234,276,277,279,282,283,287]

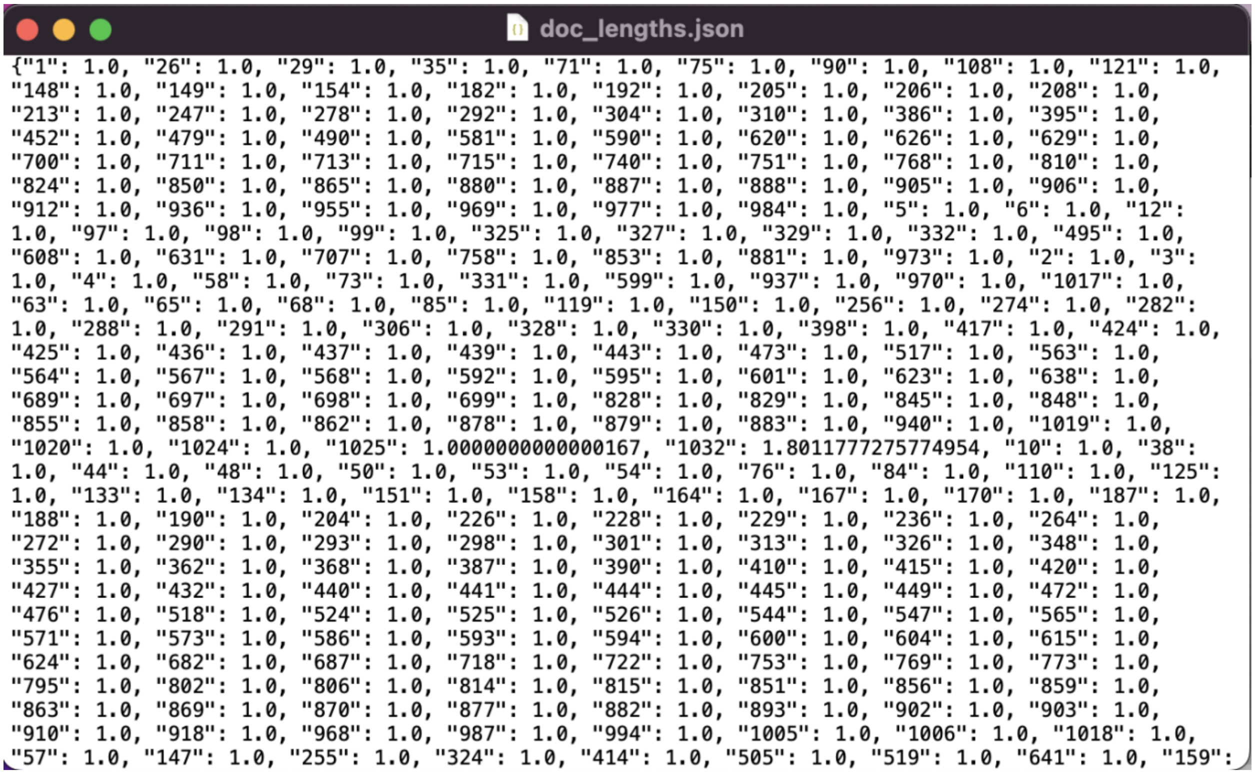
After Rochio, the Relevant retrieved docs are increased from 7 to 8.

INVERTED INDEX



```
{"correl": {"1": 4, "26": 1, "29": 2, "35": 1, "71": 1, "75": 1, "90": 1, "108": 1, "121": 2, "148":
3, "149": 1, "154": 1, "182": 1, "192": 1, "205": 1, "206": 1, "208": 1, "213": 1, "247": 1, "278":
1, "292": 1, "304": 2, "310": 2, "386": 1, "395": 1, "452": 1, "479": 1, "490": 2, "581": 1, "590":
1, "620": 1, "626": 3, "629": 3, "700": 3, "711": 1, "713": 1, "715": 3, "740": 1, "751": 2, "768":
1, "810": 1, "824": 1, "850": 1, "865": 2, "880": 1, "887": 1, "888": 1, "905": 1, "906": 1, "912":
1, "936": 1, "955": 1, "969": 1, "977": 1, "984": 1}, "matern": {"1": 6, "5": 3, "6": 3, "12": 2,
"97": 5, "98": 1, "99": 2, "304": 2, "325": 1, "327": 1, "329": 5, "332": 1, "495": 2, "608": 2,
"631": 1, "707": 2, "758": 2, "853": 1, "881": 6, "973": 1}, "fetal": {"1": 6, "2": 1, "3": 4, "4":
1, "5": 2, "6": 3, "12": 3, "58": 3, "73": 1, "206": 2, "278": 2, "331": 2, "332": 2, "599": 6,
"758": 1, "881": 2, "905": 1, "937": 1, "970": 1, "1017": 2}, "plasma": {"1": 3, "5": 3, "6": 5,
"26": 1, "63": 1, "65": 1, "68": 2, "85": 1, "119": 1, "148": 7, "150": 2, "256": 2, "274": 1, "282":
1, "288": 1, "291": 4, "304": 2, "306": 2, "327": 1, "328": 2, "329": 2, "330": 1, "332": 4, "398":
1, "417": 2, "424": 1, "425": 1, "436": 1, "437": 1, "439": 3, "443": 2, "452": 4, "473": 4, "517":
1, "563": 4, "564": 8, "567": 2, "568": 1, "581": 1, "592": 3, "595": 1, "601": 3, "623": 1, "638":
3, "689": 3, "697": 2, "698": 6, "699": 1, "758": 2, "828": 2, "829": 3, "845": 4, "848": 4, "855":
1, "858": 1, "862": 4, "865": 15, "878": 1, "879": 1, "880": 5, "883": 2, "936": 1, "940": 1, "1019":
2, "1020": 7, "1024": 1, "1025": 1, "1032": 1}, "level": {"1": 8, "2": 1, "4": 1, "10": 1, "38": 1,
"44": 1, "48": 4, "50": 1, "53": 1, "54": 1, "63": 2, "76": 3, "84": 3, "90": 1, "110": 1, "125": 3,
"133": 1, "134": 1, "148": 4, "149": 2, "150": 1, "151": 2, "158": 1, "164": 2, "167": 1, "170": 2,
"182": 1, "187": 2, "188": 4, "190": 1, "204": 1, "205": 1, "226": 1, "228": 5, "229": 6, "236": 1,
"264": 1, "272": 1, "288": 1, "290": 1, "292": 1, "293": 1, "298": 4, "301": 1, "304": 12, "313": 1,
"325": 4, "326": 1, "327": 1, "328": 1, "329": 2, "331": 1, "332": 1, "348": 2, "355": 1, "362": 1,
"368": 1, "387": 1, "390": 6, "410": 6, "415": 1, "420": 1, "427": 1, "432": 1, "440": 1, "441": 1,
```

DOCUMENT LENGTHS

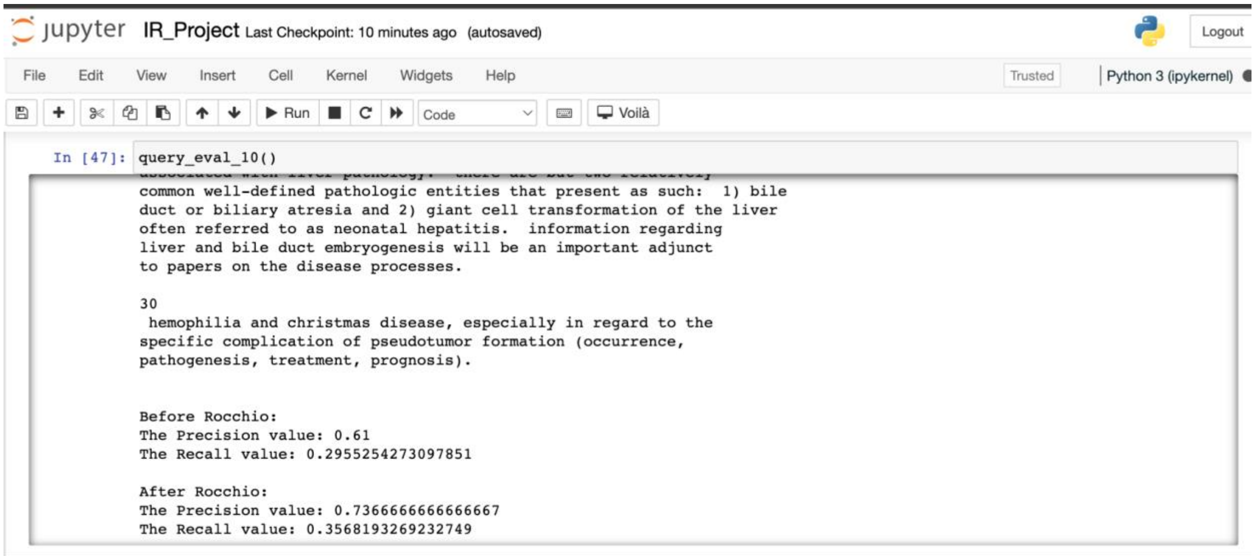


```
{"1": 1.0, "26": 1.0, "29": 1.0, "35": 1.0, "71": 1.0, "75": 1.0, "90": 1.0, "108": 1.0, "121": 1.0,
"148": 1.0, "149": 1.0, "154": 1.0, "182": 1.0, "192": 1.0, "205": 1.0, "206": 1.0, "208": 1.0,
"213": 1.0, "247": 1.0, "278": 1.0, "292": 1.0, "304": 1.0, "310": 1.0, "386": 1.0, "395": 1.0,
"452": 1.0, "479": 1.0, "490": 1.0, "581": 1.0, "590": 1.0, "620": 1.0, "626": 1.0, "629": 1.0,
"700": 1.0, "711": 1.0, "713": 1.0, "715": 1.0, "740": 1.0, "751": 1.0, "768": 1.0, "810": 1.0,
"824": 1.0, "850": 1.0, "865": 1.0, "880": 1.0, "887": 1.0, "888": 1.0, "905": 1.0, "906": 1.0,
"912": 1.0, "936": 1.0, "955": 1.0, "969": 1.0, "977": 1.0, "984": 1.0, "5": 1.0, "6": 1.0, "12":
1.0, "97": 1.0, "98": 1.0, "99": 1.0, "325": 1.0, "327": 1.0, "329": 1.0, "332": 1.0, "495": 1.0,
"608": 1.0, "631": 1.0, "707": 1.0, "758": 1.0, "853": 1.0, "881": 1.0, "973": 1.0, "2": 1.0, "3":
1.0, "4": 1.0, "58": 1.0, "73": 1.0, "331": 1.0, "599": 1.0, "937": 1.0, "970": 1.0, "1017": 1.0,
"63": 1.0, "65": 1.0, "68": 1.0, "85": 1.0, "119": 1.0, "150": 1.0, "256": 1.0, "274": 1.0, "282":
1.0, "288": 1.0, "291": 1.0, "306": 1.0, "328": 1.0, "330": 1.0, "398": 1.0, "417": 1.0, "424": 1.0,
"425": 1.0, "436": 1.0, "437": 1.0, "439": 1.0, "443": 1.0, "473": 1.0, "517": 1.0, "563": 1.0,
"564": 1.0, "567": 1.0, "568": 1.0, "592": 1.0, "595": 1.0, "601": 1.0, "623": 1.0, "638": 1.0,
"689": 1.0, "697": 1.0, "698": 1.0, "699": 1.0, "828": 1.0, "829": 1.0, "845": 1.0, "848": 1.0,
"855": 1.0, "858": 1.0, "862": 1.0, "878": 1.0, "879": 1.0, "883": 1.0, "940": 1.0, "1019": 1.0,
"1020": 1.0, "1024": 1.0, "1025": 1.0000000000000167, "1032": 1.8011777275774954, "10": 1.0, "38":
1.0, "44": 1.0, "48": 1.0, "50": 1.0, "53": 1.0, "54": 1.0, "76": 1.0, "84": 1.0, "110": 1.0, "125":
1.0, "133": 1.0, "134": 1.0, "151": 1.0, "158": 1.0, "164": 1.0, "167": 1.0, "170": 1.0, "187": 1.0,
"188": 1.0, "190": 1.0, "204": 1.0, "226": 1.0, "228": 1.0, "229": 1.0, "236": 1.0, "264": 1.0,
"272": 1.0, "290": 1.0, "293": 1.0, "298": 1.0, "301": 1.0, "313": 1.0, "326": 1.0, "348": 1.0,
"355": 1.0, "362": 1.0, "368": 1.0, "387": 1.0, "390": 1.0, "410": 1.0, "415": 1.0, "420": 1.0,
"427": 1.0, "432": 1.0, "440": 1.0, "441": 1.0, "444": 1.0, "445": 1.0, "449": 1.0, "472": 1.0,
"476": 1.0, "518": 1.0, "524": 1.0, "525": 1.0, "526": 1.0, "544": 1.0, "547": 1.0, "565": 1.0,
"571": 1.0, "573": 1.0, "586": 1.0, "593": 1.0, "594": 1.0, "600": 1.0, "604": 1.0, "615": 1.0,
"624": 1.0, "682": 1.0, "687": 1.0, "718": 1.0, "722": 1.0, "753": 1.0, "769": 1.0, "773": 1.0,
"795": 1.0, "802": 1.0, "806": 1.0, "814": 1.0, "815": 1.0, "851": 1.0, "856": 1.0, "859": 1.0,
"863": 1.0, "869": 1.0, "870": 1.0, "877": 1.0, "882": 1.0, "893": 1.0, "902": 1.0, "903": 1.0,
"910": 1.0, "918": 1.0, "968": 1.0, "987": 1.0, "994": 1.0, "1005": 1.0, "1006": 1.0, "1018": 1.0,
"57": 1.0, "147": 1.0, "255": 1.0, "324": 1.0, "414": 1.0, "505": 1.0, "519": 1.0, "641": 1.0, "159":
```

EVALUATION OF THE SYSTEM:

For evaluating our system, we used the MEDLINE which is a collection of articles from a medical journal to test the system (It can be found in the link https://ir.dcs.gla.ac.uk/resources/test_collections/medl/). We used the MED.ALL file which contains all the documents. This dataset has 1033 documents which goes through the crawler and undergoes preprocessing. Then the inverted index and document lengths are calculated for this set of documents and stored in the local directory as explained above. Then these two files are retrieved whenever we would like to run the system now.

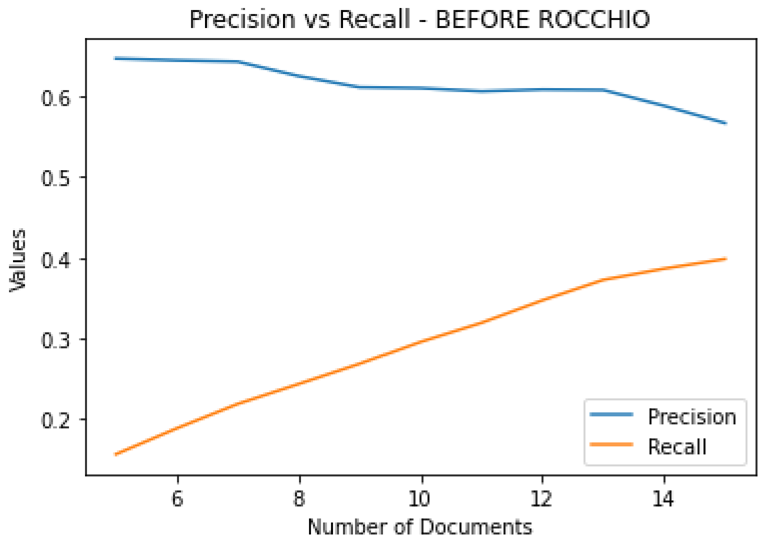
We run test queries which are also available as part of the dataset (https://ir.dcs.gla.ac.uk/resources/test_collections/medl/). This has 30 test queries and list of the relevant documents for each query. Using the function **query_eval_10()**, the system is tested on its precision and recall by running through all 30 queries and calculating the precision for each query and recall for each query using the function **precision_recall(retrievedDocs, relevantDocs).** The set of 10 retrieved document ids and the relevant document ids are passed to this function from query_eval_10 function. The query is then printed for our reference. The precision and recall for all the queries are calculated and the average precision and recall is returned. For this set of 30 queries, the average precision and recall was **0.61** and **0.295** which is pretty low. We also calculated in the same function with the same process as above, the precision and recall of these 30 queries after Rocchio relevance feedback was applied. Then the average precision and recall was calculated and the values are **0.737** and **0.357** respectively. The system definitely improved on doing relevance feedback.

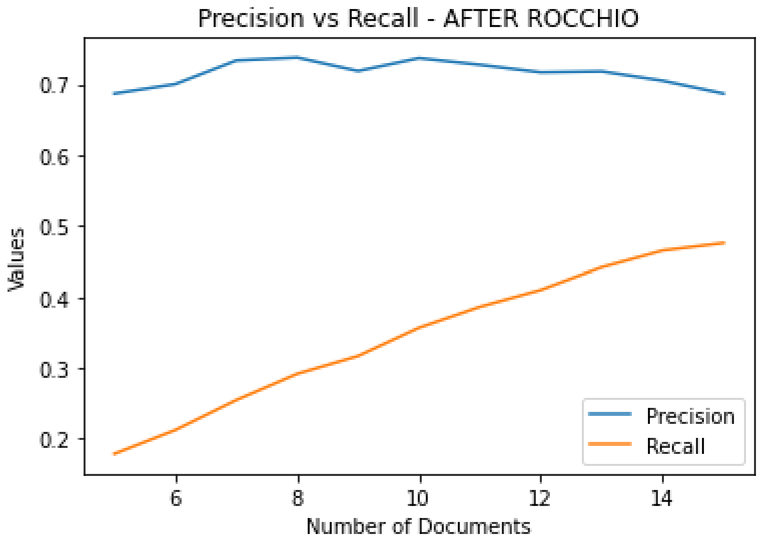


```
In [47]: query_eval_10()
```

```
common well-defined pathologic entities that present as such:  1) bile
duct or biliary atresia and 2) giant cell transformation of the liver
often referred to as neonatal hepatitis.  information regarding
liver and bile duct embryogenesis will be an important adjunct
to papers on the disease processes.

30
 hemophilia and christmas disease, especially in regard to the
specific complication of pseudotumor formation (occurrence,
pathogenesis, treatment, prognosis).


Before Rocchio:
The Precision value: 0.61
The Recall value: 0.2955254273097851

After Rocchio:
The Precision value: 0.7366666666666667
The Recall value: 0.3568193269232749
```

Apart from this, we also wanted to check what the ideal number of documents to retrieve i.e., the number of documents retrieved for which we get the highest possible precision and recall - the perfect tradeoff. So, using the function **precision_Recall_plot()**, we ran through the k values 5 to 15 where k is the number of retrieved documents. Then the average precision and recall for each k value for all queries are calculated and plotted to compare. We decided to stick with 10 as that retrieved the best results with respect to precision but still not too low on recall. There are two plots created here, one comparing precision and recall values before Rocchio and one comparing the values after Rocchio.


Precision vs Recall - BEFORE ROCCHIO
Values
0.2
0.3
0.4
0.5
0.6
6
8
10
12
14
Number of Documents
Precision
Recall


Precision vs Recall - AFTER ROCCHIO
Values
0.2
0.3
0.4
0.5
0.6
0.7
6
8
10
12
14
Number of Documents
Precision
Recall

REFERENCES:

1. Link for image
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.searchenginejournal.com%2Finformation-retrieval-seo%2F464164%2F&psig=AOvVaw3G8cRgbByrNfODkU84_EHR&ust=1678752790422000&source=images&cd=vfe&ved=0CBAQjhxqFwoTCMj61c_P1_0CFQAAAAAdAAAAABAE

2. Links useful when creating the system
https://medium.com/@janujaishree94/searchit-an-information-retrieval-system-33d2af956da4
Lecture 4 Notes - Implementation notes on Vector Space Retrieval
MEDLINE journals data - https://ir.dcs.gla.ac.uk/resources/test_collections/medl/
https://www.geeksforgeeks.org/python-stemming-words-with-nltk/
https://www.geeksforgeeks.org/python-remove-punctuation-from-string/
https://www.geeksforgeeks.org/reading-and-writing-json-to-a-file-in-python/
https://docs.python.org/3/library/tkinter.html

PROJECT VIDEO LINK: https://youtu.be/kzjI3PWg6D8