Data Collection:

I have choosen CISI test collection. Documents, queries, and query gudgments (qrels) were parsed from their respective files and converted into appropriate format to deal with them.

Preprocessing:

To prepare the text data for further worker , three functions were implemented

- 1. Function was implemented using the NLTK library to Steem The Text Firstly It Tokenize the passed text into parts then steem each part .

- 2.Function was implemented using regex module to ensure that there is no any sort of (Special Charachters , Tabs , Line Jumps , Extra White Spaces In the passed text

- 3.Function was implemented to traverse over the tokenized text and check if there any tokenized text existed in the stop words list defined from the NLTK if it exist it's removed immediately else pass .

```
def Stem_text(text):
    tokens = word_tokenize(text)
    stemmed_tokens = [stemmer.stem(word) for word in tokens]
    return ' '.join(stemmed_tokens)

def clean(text):
    text = re.sub(r"[\.\,\#_\|\:\?\?\/\=\@]", " ", text) # remove special characters
    text = re.sub(r"\t", ' ', text) # remove tabs
    text = re.sub(r"\n", ' ', text) # remove line jump
    text = re.sub(r"\s+", " ", text) # remove extra white space
    text = text.strip()
    return text

def remove_stopwords(text):
    tokens = word_tokenize(text)
    filtered_tokens = [word.lower() for word in tokens if word.lower() not in stop_words]
    return ' '.join(filtered_tokens)

def preprocess(sentence):
  sentence = clean(sentence)
  sentence = remove_stopwords(sentence)
  sentence = Stem_text(sentence)
  return sentence

[11] documents_df["Processed_text"]=documents_df['Text'].apply(preprocess).apply(clean)
     documents_df["docno"] = documents_df["ID"].astype(str)
     documents_df.drop(columns='ID',inplace=True)
```

Indexing:

This stage focused on creating an efficient index for document retrieval using the DFIndexer. Based on requirements i have developed Data Structure that map each term with it's postings and beside each doc id there is an assignment (number) this number refer to the frequency of the term inside this posting.

Built In Method To Map The Inverted Index .

Indexing

```
[13] indexer = pt.DFIndexer("./DatasetIndex", overwrite=True)
     index_ref = indexer.index(documents_df["Processed_text"], documents_df["docno"])
     index = pt.IndexFactory.of(index_ref.toString())
```

Built in Method

```
[14] for kv in index.getLexiconIterator(): print( kv.getKey()," -> " ,kv.getValue() )
     wolfe  -> term4557 Nt=1 TF=1 maxTF=1 @[0 92220 5]
     wolf  -> term4555 Nt=1 TF=1 maxTF=1 @[0 92223 3]
     wolter  -> term1004 Nt=1 TF=1 maxTF=1 @[0 92235 1]
     wole  -> term427 Nt=1 TF=1 maxTF=1 @[0 92237 1]
     women  -> term213 Nt=5 TF=8 maxTF=4 @[0 92238 3]
     wonder  -> term6813 Nt=3 TF=1 maxTF=1 @[0 92248 2]
     wont  -> term3921 Nt=1 TF=1 maxTF=1 @[0 92251 0]
     wood  -> term2379 Nt=9 TF=9 maxTF=3 @[0 92253 2]
     woodford  -> term6777 Nt=1 TF=1 maxTF=1 @[0 92268 6]
     woodward  -> term5174 Nt=1 TF=1 maxTF=1 @[0 92271 4]
     woolf  -> term6153 Nt=1 TF=1 maxTF=1 @[0 92274 0]
     woolgar  -> term6017 Nt=1 TF=1 maxTF=1 @[0 92276 4]
     wooster  -> term584 Nt=3 TF=1 maxTF=1 @[0 92279 2]
     wor  -> term5841 Nt=2 TF=2 maxTF=1 @[0 92280 2]
```

Adham From Scratch Method to map the terms

DS Map From Scratch

```
[15] TF = {}
     for row in documents_df[["docno", "Processed_text"]].values:
         for term in row[1].split():
             if not term in TF: TF[term] = dict()
             if not row[0] in TF[term]: TF[term][row[0]] = 0
             TF[term][row[0]] += 1

     for key in sorted(TF.keys()):
         print(f"{key} -> {TF[key]}")
```

```
     stone -> {'1042': 1, '1273': 1}
     stop -> {'204': 1}
     storag -> {'14': 1, '44': 1, '67': 2, '97': 1, '120': 1, '129': 2, '131': 1, '135': 1, '137': 1, '174': 2, '175': 1, '254': 2, '267': 1, '289': 1, '3
     store -> {'6': 1, '10': 1, '63': 1, '107': 1, '117': 1, '172': 1, '173': 1, '267': 1, '310': 1, '320': 2, '367': 1, '391': 1, '430': 1, '445': 2, '49
     storehous -> {'5': 1, '1373': 1}
     storer -> {'544': 1, '1330': 1}
     stori -> {'1': 1, '3': 1, '414': 1, '1051': 2}
     stow -> {'677': 1, '708': 1}
     straight -> {'750': 1, '765': 1, '860': 1}
     straightforward -> {'189': 1, '257': 1, '664': 1, '1387': 1}
     strain -> {'103': 1, '170': 1, '366': 1, '843': 1, '977': 1, '1203': 2, '1442': 1}
     strakhov -> {'1135': 1}
     strang -> {'180': 1, '401': 1, '1420': 1}
     strateg -> {'314': 1, '700': 1}
     strategi -> {'46': 1, '49': 1, '104': 1, '309': 1, '458': 1, '502': 1, '509': 5, '510': 1, '565': 2, '655': 1, '700': 1, '722': 1, '730': 2, '773': 1
     strati -> {'89': 5, '93': 1, '954': 1, '1337': 2, '1386': 1}
     stratifi -> {'1346': 1}
     strauss -> {'1318': 1}
```

Query processing:

- Received query from the CLI then processed it and called builtin method from pyterrier to map the passed query with relevant documents

```
get_query_ids(query)
```

	Needed_Document_IDS	Documents
0	15	Information Flow in Research and Development Laboratories Allen, Thomas J. Cohen, Stephen I. Technical communication patterns in two research and ...
1	16	Adapting the Library of Congress Classification System Mathis, R.E. This manual is designed to make it possible for any library to change efficie...
2	85	Information Science: Toward the Development of a True Scientific Discipline Yovits, M. C. It is pointed out that if information science is to be c...
3	95	Scientific Communication: Five Themes from Social Science Research Menzel, H. The recent upsurge of interest in the behavioral aspects of scienti...
4	96	Communication Patterns in Applied Technology Marquis, D.G. Allen, T.J. The preceding papers have addressed themselves to the study of information ...
5	97	Bibliographical Citation Characteristics of the Psychological Journal Network in 1950 and in 1960 Xhignesse, L.V. Osgood, C.E. The aforementioned ...
6	124	The User Interface in Interactive Systems Bennett, J.L. The digital computer has been used as a problem-solving tool for only two decades. During ...
7	132	Information Needs and Uses in Science and Technology Menzel, H. The way in which scientists and engineers make use of the information systems at t...
8	152	Organizational Aspects of Information Flow in Technology Allen, T.J. For about five years now a small group of us at MIT have been conducting a se...
9	166	The Future of Scientific and Technological Publications Smalen, A.A. I am sure you do not need reminding of the flood of scientific information t...

- Re ranked the same query using TFIDF

```
tfidf_retr = pt.BatchRetrieve(index , wmodel = "TF_IDF", controls = {"c": 0.15}, num_results = 100)
tfidf_results = tfidf_retr.search(query)

doc_descriptions = [documents_df.loc[int(doc_id) - 1, "Text"] for doc_id in tfidf_results['docno']]

result_df = pd.DataFrame({
    "Needed_Document_IDS": tfidf_results['docno'],
    "Score": tfidf_results['score'],
    "Documents": doc_descriptions})
result_df
```

	Needed_Document_IDS	Score	Documents
0	665	5.342459	Morphology of "Information Flow" Fairthorne, R.A. Such phrases as "information flow" may be purely metaphorical, or may refer to portage and sto...
1	1114	4.459758	Analysis of Information Flows in Shipbuilding and the Allied Fields Mitsevich, A.T. Parameters of information flows are identified as reflected in ...
2	1122	4.010387	Analysis of Some Regularities of the Flow of Engineering Information Mitsevich, A. T. Solov'ev, N. K. Some theoretical propositions are considered ...
3	786	3.944218	Documentation Notes Zipf, Library A recent article by Kozachkov and Khursin, entitled "The basic probability distribution in information flow syst...
4	1173	3.887322	A Probability Distribution in Information Flow Systems Kazachkov, I. S. Khursin, L. A. The paper treats the probability distribution in informatio...
5	96	3.626000	Communication Patterns in Applied Technology Marquis, D.G. Allen, T.J. The preceding papers have addressed themselves to the study of information ...
6	1125	2.968018	A Contribution to the Theory of the Systems of Information Flows Kozachkov, L. S. Certain structural properties of information distributions are e...

Query Expansion

Here we can find easily by focusing on the score of the screenshot and the above one the influence of the query expansion

```
[25] tfidf_results = tfidf_retr.search(expanded_q)
     results_documents = documents_df[documents_df["docno"].isin(tfidf_results["docno"])]
     doc_descriptions = [documents_df.loc[int(doc_id) - 1, "Text"] for doc_id in tfidf_results['docno']]

     result_df = pd.DataFrame({
         "Needed_Document_IDS": tfidf_results['docno'],
         "Score": tfidf_results['score'],
         "Documents": doc_descriptions})
     result_df
```

	Needed_Document_IDS	Score	Documents
0	1122	36.285919	Analysis of Some Regularities of the Flow of Engineering Information Mitsevich, A. T. Solov'ev, N. K. Some theoretical propositions are considered...
1	1340	10.147123	Social Theory and Social Structure Merton, R.K. Of the four chapters, added to this edition, two come from published symposia, one of which is out ...
2	1114	9.731981	Analysis of Information Flows in Shipbuilding and the Allied Fields Mitsevich, A.T. Parameters of information flows are identified as reflected in...
3	735	9.300457	Journals Most Cited by Chemists and Chemical Engineers Barrett, R.L. Barrett, M.A. The purpose of this paper is to present up-to-date material to ...
4	788	9.241198	Documentation Notes Zipf, Library A recent article by Kozachkov and Khursin, entitled 'The basic probability distribution in information flow syst...
...	...	...	...
95	1200	3.330770	On the Statistics of Individual Variations of Productivity in Research Laboratories Shockley, W. In the following pages a co-winner of the 1956 No...

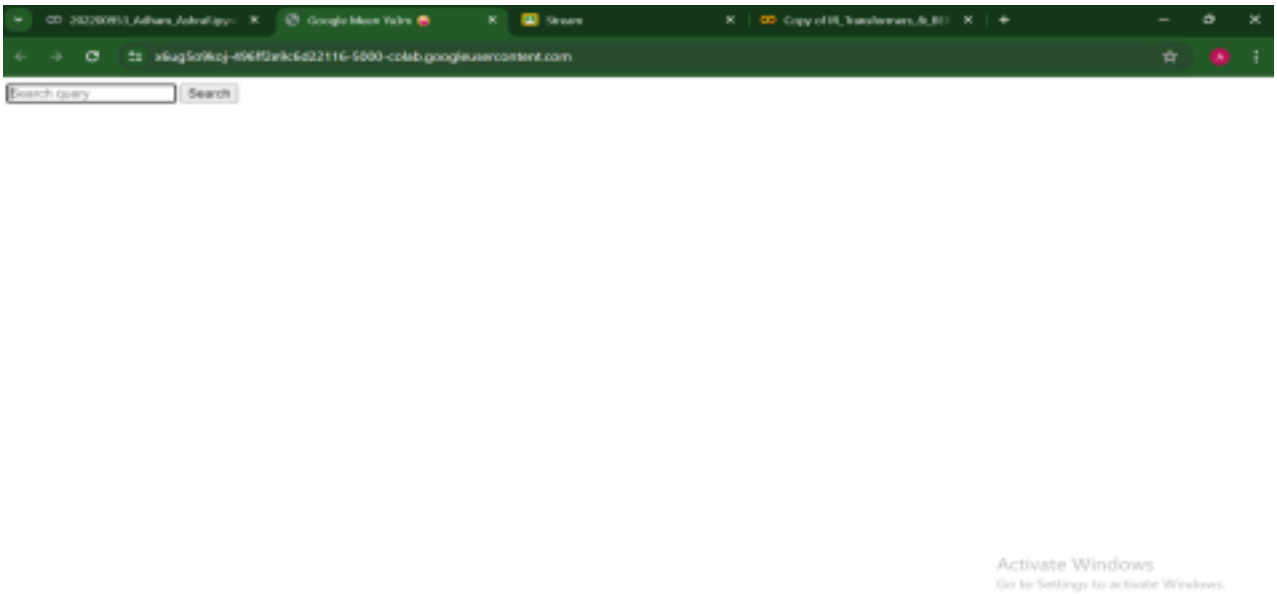
Then Reranked the docs using Elmo Embeedings

	score	docs
0	0.581055	Analysis of Some Regularities of the Flow of Engineering Information Mitsevich, A. T. Solov'ev, N. K. Some theoretical propositions are considered...
1	0.332099	Documentation Notes Zipf, Library A recent article by Kozachkov and Khursin, entitled 'The basic probability distribution in information flow syst...
2	0.242589	Analysis of Information Flows in Shipbuilding and the Allied Fields Mitsevich, A.T. Parameters of information flows are identified as reflected in...
3	0.176654	A Probability Distribution in Information Flow Systems Kazachkov, L. S. Khursin, L. A. The paper treats the probability distribution in informatio...
4	0.041194	Morphology of "Information Flow" Fairthorne, R.A. Such phrases as "information flow" may be purely metaphorical, or may refer to porterage and sto...

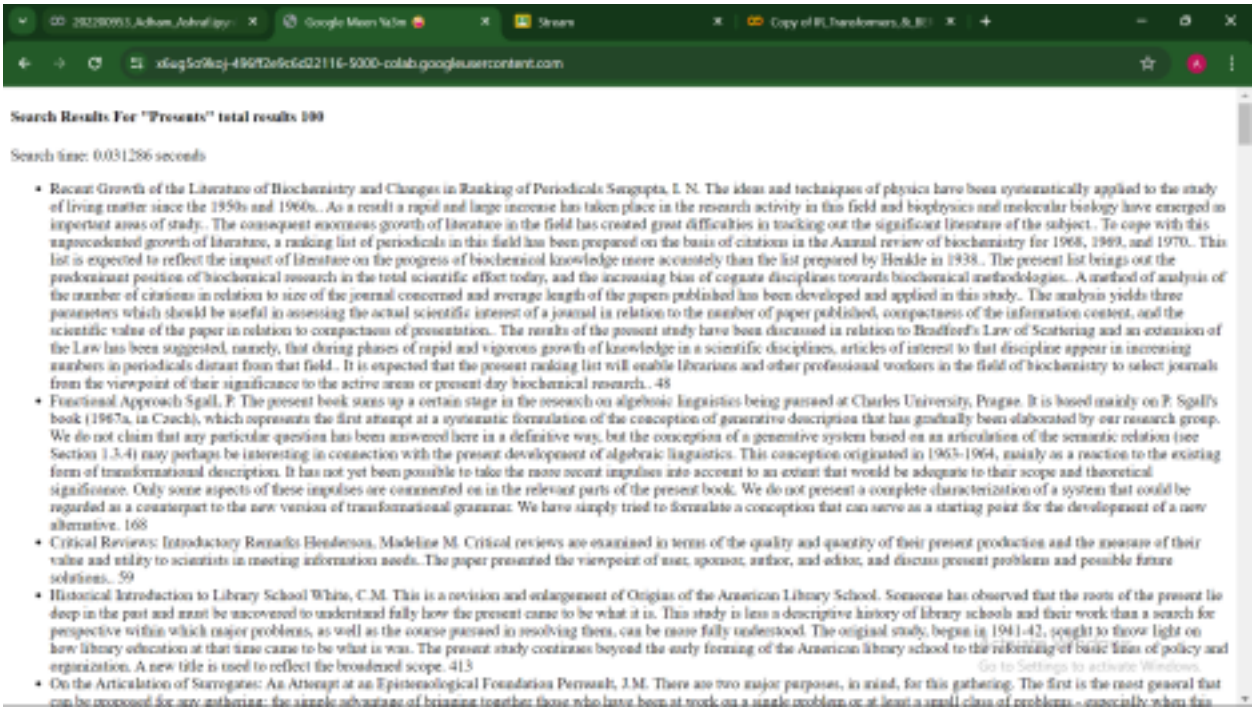
User Interface

Using Flask I have decided to built very basic web application to show off my work :

In the Search Bar I have Typed “ Presents “



As Shown In the screenshot {Total Num of results , Speed of the search engine }

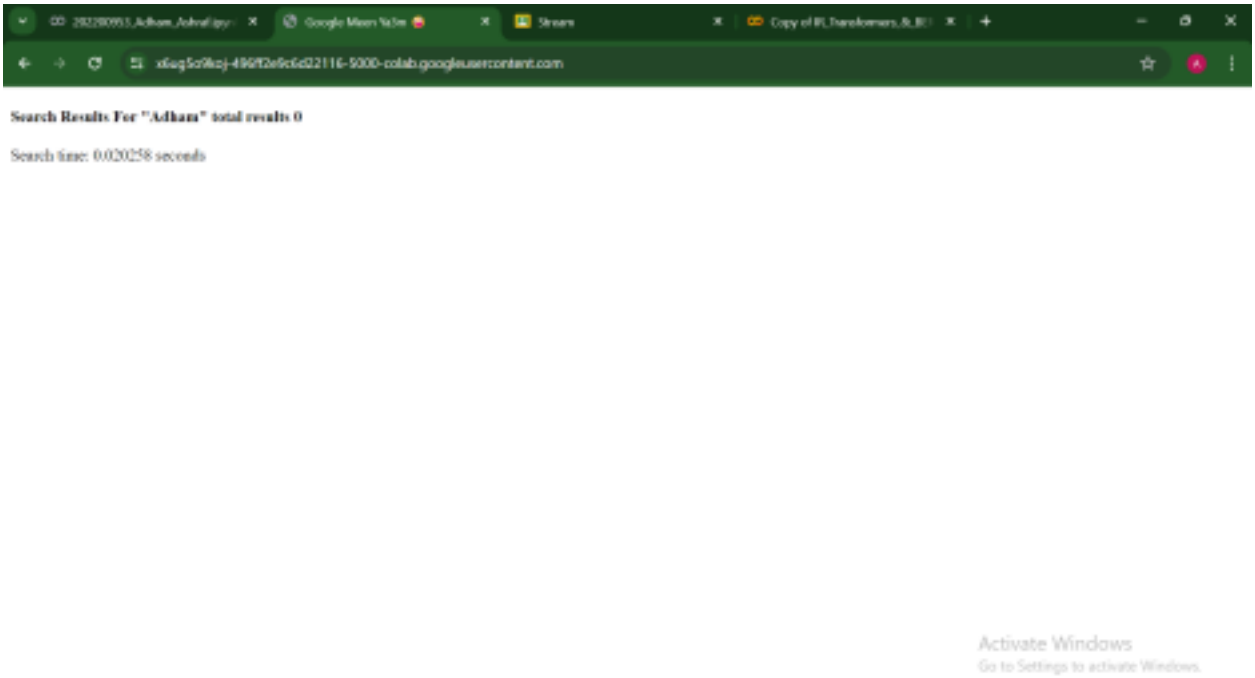


Search Results For "Presents" total results 100

Search time: 0.031286 seconds

- Recent Growth of the Literature of Biochemistry and Changes in Ranking of Periodicals Sengupta, I. N. The ideas and techniques of physics have been systematically applied to the study of living matter since the 1950s and 1960s.. As a result a rapid and large increase has taken place in the research activity in this field and biophysics and molecular biology have emerged as important areas of study.. The consequent enormous growth of literature in the field has created great difficulties in tracking out the significant literature of the subject.. To cope with this unprecedented growth of literature, a ranking list of periodicals in this field has been prepared on the basis of citations in the Annual review of biochemistry for 1968, 1969, and 1970.. This list is expected to reflect the impact of literature on the progress of biochemical knowledge more accurately than the list prepared by Henkle in 1938.. The present list brings out the predominant position of biochemical research in the total scientific effort today, and the increasing bias of cognate disciplines towards biochemical methodologies.. A method of analysis of the number of citations in relation to size of the journal concerned and average length of the papers published has been developed and applied in this study.. The analysis yields three parameters which should be useful in assessing the actual scientific interest of a journal in relation to the number of paper published, compactness of the information content, and the scientific value of the paper in relation to compactness of presentation.. The results of the present study have been discussed in relation to Bradford's Law of Scattering and an extension of the Law has been suggested, namely, that during phases of rapid and vigorous growth of knowledge in a scientific disciplines, articles of interest to that discipline appear in increasing numbers in periodicals distant from that field.. It is expected that the present ranking list will enable librarians and other professional workers in the field of biochemistry to select journals from the viewpoint of their significance to the active areas or present day biochemical research.. 48
- Functional Approach Sgall, P. The present book sums up a certain stage in the research on algebraic linguistics being pursued at Charles University, Prague. It is based mainly on P. Sgall's book (1967a, in Czech), which represents the first attempt at a systematic formulation of the conception of generative description that has gradually been elaborated by our research group. We do not claim that any particular question has been answered here in a definitive way, but the conception of a generative system based on an articulation of the semantic relation (see Section 1.3.4) may perhaps be interesting in connection with the present development of algebraic linguistics. This conception originated in 1963-1964, mainly as a reaction to the existing form of transformational description. It has not yet been possible to take the more recent impulses into account to an extent that would be adequate to their scope and theoretical significance. Only some aspects of these impulses are commented on in the relevant parts of the present book. We do not present a complete characterization of a system that could be regarded as a counterpart to the new version of transformational grammar. We have simply tried to formulate a conception that can serve as a starting point for the development of a new alternative. 168
- Critical Reviews: Introductory Remarks Henderson, Madeline M. Critical reviews are examined in terms of the quality and quantity of their present production and the measure of their value and utility to scientists in meeting information needs. The paper presented the viewpoint of user, sponsor, author, and editor, and discuss present problems and possible future solutions.. 59
- Historical Introduction to Library School White, C.M. This is a revision and enlargement of Origins of the American Library School. Someone has observed that the roots of the present lie deep in the past and must be uncovered to understand fully how the present came to be what it is. This study is less a descriptive history of library schools and their work than a search for perspective within which major problems, as well as the course pursued in resolving them, can be more fully understood. The original study, begun in 1941-42, sought to throw light on how library education at that time came to be what it was. The present study continues beyond the early forming of the American library school to the reforming of basic lines of policy and organization. A new title is used to reflect the broadened scope. 413
- On the Articulation of Surrogates: An Attempt at an Epistemological Foundation Perreault, J.M. There are two major purposes, in mind, for this gathering. The first is the most general that can be proposed for any gathering: the simple advantage of bringing together those who have been at work on a single problem or at least a small class of problems - especially when this

I have tried to search for query isn't found in the corpus here is the result



Here is Another Result For The Search Engine

Search Results For "Cost Accounting" total results 100

Search time: 0.014243 seconds

- Managerial Cost Accounting for a Technical Information Center Helmkamp, John G. The purpose of this paper is to describe a research project conducted at a technical center to test the hypothesis that: A theoretically-sound managerial cost-accounting system can be designed to meet the specific characteristics of a technical information center by revising and innovating systems utilized by other enterprises.. A computerized cost system was developed and operated for a three-month period to test this hypothesis.. The results of the study indicate that effective managerial cost accounting is possible for a technical information center. Relevant cost information was generated periodically to measure the operating performance of the center's production process.. A summary of the data that were reported regularly to management is presented in this paper.. 24
- Cost Accounting and Analysis for University Libraries Leimkohler, Ferdinand F. Cooper, Michael D. The approach to library planning studies in this paper is the use of accounting models to measure library costs and implement program budgets.. A cost-flow model for a university library is developed and tested with historical data from the General Library at the University of California, Berkeley.. Various comparisons of an exploratory nature are made of the unit costs and total costs for different parts of the Berkeley system.. 5
- Library Cost Analysis: A Recipe Kountz, J. Unforgivably, time has passed since the days when the library's patron was the local monarch and cost was no deterrent. Time's passage has replaced the monarch with taxpayers or stockholders, and, concurrently, sensitivity to cost has attained stellar importance. The causes for being unaware of costs may stem from a variety of reasons, but they cannot, in all fairness to the profession, belie an inability to perform the simple arithmetic of cost accounting. What is suspected is a lack of the few simple ground rules and the logical operations that bind them together, in short - a recipe for cost accounting and analysis. In the following is outlined one such set of ground rules and their related procedural requirements, which have evolved and been applied with success over the past few years. It is stressed that since this set represents the findings of one library, it may not fully satisfy the specific requirements of your own shop. Therefore, feel free to adapt the ground rules to your immediate requirements. With regard to discipline, it is pretty much summed up in the six steps and five resource requirements which follow. In addition to identifying steps, requirements, and the mysterious ways of cost analysis, these ingredients are blended together in a manner which will be meaningful for your internal operations and may be significant for your library's future. 90
- Cost Comparison of Manual and On-Line Computerized Literature Searching Elman, S.A. Cost and searching time comparisons are made between manual and on-line literature searches. The formula Cmtal = (T X Cmm) + P is presented which captures all on-line cost factors. A minimum cost of $1.00 per minute of on-line searching is derived. Average searching time for manual searching is 22 hours at a total cost of $250; for on-line it is 45 minutes at total cost of $47.00. It is pointed out that most reported low-on line search costs fail to account for all cost factors. Figures are those prevailing at the time of writing. 124
- Library Participation in a Biomedical Communication and Information Network Bridegam, Willis E., Jr. Meyerhoff, Erich The experience of two libraries participating in the SUNY Biomedical Communication Network is described.. The history of the Network if briefly given together with its original aims and their current status.. Use of the terminals and formulation of queries are explained.. Figures are given for total costs, number of searches performed, and cost per search.. There is a account of the internal structure of the administration of the Network.. 65
- Cost Accounting for the Library Bratcher, C. Gosford, G. Rinford, E. Increasingly, librarians have felt the need for more accurate cost data. The prime reason for this need has been in the development and presentation of the budget which is the instrument used to determine and obtain the funds for the library's forthcoming fiscal period. Since libraries do not charge for the service they render their users, they must derive the funds necessary for their operations and growth from supporting bodies such as federal, state, or local governments, private institutions, and industrial firms. 27
- The Shared Cataloging System of the Ohio College Library Center Kilgour, Frederick G. Long, Philip L. Lundgraf, Alan L. Wydoff, John A. Development and implementation of an off-line catalog card production system and on-line shared cataloging system are described.. In off-line production, average cost per card for 529,893 catalog cards in finished form and

Evaluation

Here We can find the evaluation for the search engine

```
[34] evaluation = pt.Evaluate(tfidf_retr.transform(topics), qrels, metrics=["map", "recall", "P"], perquery=True)
     evaluation_df = pd.DataFrame(evaluation).T
     evaluation_df
```

	map	P@5	P@10	P@15	P@20	P@50	P@100	P@200	P@500	P@1000	R@5	R@10	R@15	R@20	R@50	R@100	R@200	R@500	R@1000
1	0.373060	0.6	0.5	0.4	0.55	0.5	0.34	0.170	0.068	0.034	0.005217	0.100696	0.130435	0.23913	0.326087	0.73913	0.73913	0.73913	0.73913

Screenshots Purpose For Illustrating My steps and Showing My results

You can find the same results clearly in the [Note Book]