

**Department of Applied Computer Science**

**COMP 479 (6791) Project 2**

**Project Report**

**Name :Anish Chavan**

**Student id :6421180**

***bytes\_from\_file():***

***Parser(chunk):***

***tokenize():***

***ADDToDictionary:***

**class MAIN**

**class Tokenize**

***main ():***

***IDFokapibm25(self,k):***

1. **System Architecture/Data Structure :**

**Classes**

***MAIN class*:**

***main ():***

Main class consists of *main ()* fucntion which is responsible for reading all the .smg files from the directory .

It uses global list like Querytoken=list(),querysplit=list() which were used to process the query.

main() also implements **Rank Retrieval Inverted index based on Term frequencyonly.**

main retrieves the inverted index from different method in the class **tokenizer.**

main it implements rank retrival on term frequency as follows:

newdict is the dictionary formed by the common token between the global

invertedindex {term,{docid,termfrequency}} and token in the Querytoken

newdict is also a dictionary of format newdict{term,{docid,termfrequency}}

for token in Querytoken:

newdict[token]=self.invertedindex[token]

for doc2 in self.newdict:

print doc2

xtemp=self.newdict.get(doc2)

for docid in xtemp:

if docid in compare.keys():

score=compare.get(docid)

score=score+xtemp.get(docid)

compare[docid]=score

main ()prints the top **5 documents which are ranked by the term frequency only**

***IDFokapibm25(self,k):***

MAIN class also consists *of IDFokapibm25(self,k)* which is the function which calculates the idf for each query in the term and uses okapibm25 to rank all the documents.

IDF for a query q is stored in idf dictionary **idf=dict()**

num=(N-nq.get(t)+0.5)

denom=nq.get(t)+0.5

x=math.log(num/denom)

idf[t]=x

It conveys that idf[term]=(x)

**x=calculated idf value**

This function also calculates the rank for each document based on okapibm25 format and is responsible for creating a dictionary storing the document and its score.

for ta in self.newdict:

temp=self.newdict.get(ta)

for docid in temp:

tf=temp.get(docid)

numerator=tf\*(k+1)denominator=tf+k\*(1-b+(b\*(len(self.termlist[docid])/avgdl)))

qscore=idf[ta]\*(numerator/denominator)

if docid not in score.keys():

score[docid]=qscore

else:

tempscore=score.get(docid)

score[docid]=tempscore+qscore

**Class Tokenize:**

Class tokenize is responsible for following functions

*bytes\_from\_file():*

it reads all th files ending with .smg

*Parser(chunk):*

It parses the document and only consides the documents between <BODY> </BODY>

*tokenize():*

This fucntion removes all the space ,\n\t, and punctuation

*ADDToDictionary:*

This function is responsible for adding the new term to the dictionary

**Data Structures:**

dict(),list()

invertedindex=dict{key{key,dict}}

key =terms, innerdict{docid,termfrequency}

**2.Observation and Analysis of k in okapibm25:**

* **Influence of k**

Influence of different values of k on okapi bm25 principle provided the following observation

tf=temp.get(docid)

numerator=tf\*(**k**+1)denominator=tf+**k**\*(1-b+(b\*(len(self.termlist[docid])/avgdl)))

qscore=idf[ta]\*(numerator/denominator)

**query="Drug and bankruptcies"**

**Table1:***The following table gives the highest scored document/s or top scoreddocuement for different values of k .*

|  |  |  |  |
| --- | --- | --- | --- |
| *K* | *Okapiscore{D,Q}* | *Document length(words)* | *Documentid* |
| 1 | 14.99904830344621 | 1597 | 3735 |
| 1.2 | 16.021180256045838 | 1597 | 3735 |
| 2 | 20.52607839506184 | 455 | 14830 |
| 100 | 86.3805799125519 | 455 | 14830 |
| 1000 | 94.74718795401732 | 455 | 14830 |

***Analysis of above observation*:-**

* For lower values of **k** like between 1 and 1.2 the score of top document changes at fast rate

but as **k increases and becomes more than 1.2 the score rate increases but at very slow rate example:** k=100 score =86.3805799125519,k=1000=94.74718795401732 not much difference as k=1.2 and k=1 which is 15 and 17 respectively.

* The top 5 highly ranked documents revealed that the k affects the length of the documents
* *More the value of k, the documents with the less word length rank more. This is because of denominator of okapi formula which has document length as parameter. Thus as k increases the document length which had very less impact on score becomes more critical and thus becomes a factor due to which the documents with less document length has less denominator value this higher .Thus the ranking score then inclines more toward short length documents.*
* *Between k=1 and k=1.2 the the length of document is ineffective.*

***Conclusion:***

Between k=1 and k=1.2 since the length of document dose not affect ranking score much ,its more dominated by the idf value of the token which ranks the document with rare terms more .but as k increases and becomes hundred the length also starts to affect the score .***Thus as k increases inverse proportionality/dependency of the score on length of document also increases.Thus it is better to have k betweeen 1 and 1.2.***

**3.Diffferences betweeen Two ranking techniques:**

* Term frequency ranking:

It a ranking based on only term frequency. The document with highest term frequency is ranked more

* OkapiBM25:Given a query Q, containing keywords q_1, ..., q_n, the BM25 score of a document D is:  \text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},

\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},Reference:-( <http://en.wikipedia.org/wiki/Okapi_BM25>)

Both the techniques give very different results in the ranking of the document .

*note-Stopwords not removed to show how okapibm25 is better than only term frequency.*

Following are the observations according to three test queries .

.Top highly ranked documents **in term frequency rank retrieval**

1. "Democrats welfare and healthcare reform policies"

doc id=8807

Length of document =5226

term count

welfare 0

healthcare 0

and 50

reform  0

policies 0

1. “Drug company bankruptcies"

term count

drug 17

company 8

bankruptcies 0

1. “Dow jones great depression"

&lt;Jones International Ltd> saidit realigned several subsidiaries to accommodate corporate

strategy.Jones said ultimate control of the subsidiaries will remain unchanged under the realignment.

Jones said it will exchange approximately 97 pct of its holdings of the common stock of Jones Intercable Inc &lt;JOIN> for

Class A common stock of Jones Spacelink Ltd &lt;SPLK>. The company said the move will enable Spacelink to elect 75 pct of

Intercable's directors.In addition, Jones said 60 pct of the common stock of its

subsidiary, &lt;The Jones Group Ltd>, will be exchanged for Class A common stock of Jones Spacelink Ltd. Following the exchange, Jones International's ownership of Spacelink Class A voting stock will increase to approximately

89 pct. Jones International said it will retain approximately two pct of the common stock and approximately four pct of the Class A common stock of Jones Intercable as well as the remaining 40 pct of the Jones Group stock.Reuter

&#3;

**.Top highly ranked document Okapi Bm25 rank retrieval k=1.2**

1. "Democrats welfare and healthcare reform policies"

Healthvest said it acquired theEastwood Hospital medical complex in Memphis from Healthcare

International Inc for 50 mln dlrs cash.Healthcare International will continue to operate the

hospital under a lease agreement, the company said.

Reuter

&#3;

1. “Drug company bankruptcies"

doc id 3735 Length 1579

term count

bankruptcies 6

drug 0

company 1

1. “Dow jones great depression"

The New York Stock Exchange said it willhold a press conference at 16:20 EDT.The exchange released no further details.The Dow Jones industrial average fell over 500 points in

trading on Monday, exceeding the Oct 28, 1929 decline thatheralded the Great Depression.

Reuter

&#3;

**Analysis from above observation from test queries:**

* The main difference regardless of all the above queries is term frequency rank retrival

system does not considered the length of document. It can be observed that OkapiBM25 ranks the shorter document above the larger ones thus the document retrieved by okapi are short.

* Query "Dow jones great depression" shows that OkapiBM25 retrieves the document containing all the terms and ranks that document at the top. whereas the term frequency based system ranks the document which has no "Dow "term in it but on the bases of most number of jones.
* Query “Drug company bankruptcies" indicates that OkapiBM25 retrieves documents **having rare term**

**more number of times .**Where in normal term frequency rank retrieval system has no rare term

bankruptcies, but more of term drug and company which never occupies the relevance to the query

* Query "Democrats welfare and healthcare reform policies" reveals the biggest disadvantage in the

term frequency as it ranks the document having highest number of term and in it which makes the retrieved document irrelevant.

* Idf plays an important role as it is responisble for ranking the document with more rare terms than

document having higher number of frequent terms.

**Conclusion**:

Using the term frequency rank retrieval system the retrieved document may not contain the relevant content .It ignores the importance of the information provided by the rare terms .OkapiBm25 ranks the document having the rare terms more than regular terms thus using the information provided by the rare terms and also considering the length of the document .