Implementation of a Vector Space Model for Ranked Document Retrieval

Aryan Gupta	2018A7PS0017P
Aarnav Dhanuka	2017B5A70945P
Yash Agrawal	2018A7PS0251P
Ahmad Faraz	2017B4A70558P
Sujeet Srivastava	2017A4PS0503P

April 30, 2021

Submitted in partial fulfillment to the course
Information Retrieval (CS F469)

Submitted to

Vinti Agarwal

Birla Institute of Technology and Science, Pilani, Rajasthan
India

Table of Contents

Code Description

Here we are showing a very brief overview of the implementation. The detailed one is shown in the Readme file.
In the first part of our task we created an inverted index to retrieve the documents based on the free-text queries.We had created three dictionaries in python for this purpose:

1)Inverted_index - In this dictionary keys are the tokens of the vocabulary and it maps the key with a posting list.
This posting list is the list of tuples where each tuple contains DocID and the frequency of the key in that document.
2)Documents - In this dictionary keys are the docID and its value is the processed content of the document.
3)doc_names - This is the dictionary which maps docID to doc Title.

Using all these data structures we are able to calculate lnc values for the documents and ltc values for the queries. Then we can calculate the cosine scores and based on these scores we are retrieving the top k document which matches best with query.

The second part of the task was implemented by incorporating the function title_weighting for implementing the title weighting part and passing biword=True to the tokenizer for the biword implementation of the inverted index.Further details can be seen later in the report.

Evaluation of Part-1

Presented here are 10 sample queries and the results:

Query	Top 10 Documents	Score	Relevance
1 Supreme court (considered relevant if any information / cases related to supreme court is present)	Council of State	0.00394198495939 79525	Yes
	Court order	0.00150793903121 8836	No
	Campaign for Homosexual Law Reform	0.00117564641879 08436	Yes
	Ishim River	0.00106436260497 79272	No
	Keepie uppie	0.00090846321856 13118	No
	Bückeburg	0.00066753487819 801	No
	List of heads of state of Mexico	0.00058415070994 7764	No
	Holly Valance	0.00049455923858 77931	Yes
	Asylum and Immigration Tribunal	0.00048498396689 92405	No
	Bnetd	0.00048420531283 10443	No

Precision : 3/10

Query	Top 10 Documents	Score	Relevance
2 Lakes and rivers (Document considered relevant if	Diogenianus	0.00138151921741 97466	Yes
	Zwijndrecht,	0.00060133775062	No

there are information regarding both lakes and rivers in conjunction)	Netherlands	69293	
	Avon River (Western Australia)	0.00050472405615 85368	Yes
	Ishim River	0.00047504974138 033143	Yes
	Bluebell wood	0.00045610682960 155	No
	Cloppenburg	0.00042037390347 427454	No
	Leine Score	0.00038648665942 900284	No
	Avon River (Nova Scotia)	0.00032262502232 14655	Yes
	Woluwe-Saint-Pie rre	0.00029270327724 236097	No
	Mytholmroyd	0.00029058377237 03097	No

Precision : 4/10

Query	Top 10 Documents	Score	Relevance
3 Forest and wood (Document considered relevant if there are information regarding both forest and wood in conjunction) Example of a case where no relevant document was found	bluebell wood	0.0011363562341 51188	No
	Teutoburg Forest	0.0011266522169 014082	No
	Blithfield Hall	0.0010794152696 054168	No
	Woluwe-Saint-Pi erre	0.0009110514452 177639	No
	François Tourte	0.0008233912508 099171	No
	Tong (ward)	0.0008038469055 834198	No
			No

	Birdwatchers' Field Club of Bangalore	0.0007828156317273199	
	Hildesheim (district)	0.0006991457356507244	No
	Homograph	0.0006740851046945977	No
	Churches of Peace	0.000616933377520642	No

Precision : 0/10

Query	Top 10 Documents	Score	Relevance
4 state of Mexico (relevant if related to Mexico)	List of heads of state of Mexico	0.0026023100713273213	Yes
	Council of State	0.0024014301097456826	No
	Saint-Savin, Vienne	0.0016077526658316065	No
	John Skeaping	0.0011186090418731517	Yes
	Vechta (district)	0.001057997660716343	No
	Oldenburg (district)	0.0009305896633176421	No
	Ambassadors of the United States	0.0008854013582421764	No
	Charytín Goyco	0.0007925863269184329	Yes
	Our Fair City	0.0007846672804547261	No
	County of Bentheim (district)	0.0007787169581223695	No

Precision : 3/10

Query	Top 10 Documents	Score	Relevance
5 United states (in case the document size is too small irrelevant results can come up by achieving a better score, example: Doc id 152675)	List of festivals in the United Kingdom	0.0473092159689304	No
	Artists of stamps of the United States	0.009810909028492736	Yes
	Ambassadors of the United States	0.008815075116739108	Yes
	Democratic Alliance (Sweden)	0.006398518595467262	Yes
	Fowler's solution	0.004716599736622909 5	No
	McKinley, Kittson County, Minnesota	0.004505803487843235	Yes
	Arthur Middleton	0.004244702962577078	Yes
	Unity, Kennebec County, Maine	0.003768563515338904 3	Yes
	Blowing a raspberry	0.003762432445768284	No
	Watertown (town), Wisconsin	0.003739953634664883 4	Yes

Precision : 7/10

Query	Top 10 Documents	Score	Relevance
6 Democratic country (relevant if related to democratic country)	Malta (disambiguation)	0.007697287853011857	No
	Democratic Alliance (Sweden)	0.003056384465141616	Yes
	Council of State	0.00197180750097	No

		8659	
	Municipalities of Liechtenstein	0.00131598244570 74846	No
	List of mobile network operators	0.00112013399814 93392	No
	Alexis Herman	0.00098675320331 25048	Yes
	Reagan Democrat	0.00087095399169 27214	Yes
	Abraham Beame	0.00086627336123 44742	No
	Austrian People's Party	0.00070803785956 95309	Yes
	Footprints (album)	0.00059539092131 1283	No

Precision : 4/10

Query	Top 10 Documents	Score	Relevance
7 ethnic group (Document relevant if it has information about ethnic group) Example of a case where all top 5 high scoring documents were relevant.	Rơ Măm people	0.0106482623742 58689	Yes
	List of ethnic groups in Laos	0.0056219390808 52252	Yes
	Si La people	0.0036882752191 059387	Yes
	List of ethnic groups in Vietnam	0.0034481007795 17296	Yes
	O Du people	0.0030198915886 529543	Yes
	Castor Cracking Group	0.0016539925203 09048	No
	Dedekind group	0.0014389801164 269122	No
	Conjugate	0.0012774006525	No

	closure	97143	
	TWAIN	0.0009427114041959566	No
	Birdwatchers' Field Club of Bangalore	0.0009208863990338285	No

Precision : 5/10

Query	Top 10 Documents	Score	Relevance
8 people’s party	Nordic Reich Party	0.003970902061754117	No
	Rơ Măm people	0.003943848218111424	No
	Jette	0.002375412956120923	No
	Conjugate closure	0.002105173026552776	No
	Si La people	0.001693176644673862	No
	Leif Zeilon	0.001668572752741403 2	Yes
	TWAIN	0.001504209325588493	No
	Austrian People's Party	0.000993067774006597	Yes
	Phoenix Object Basic	0.000978477936633268	No
	Elegy	0.000960052482886505 4	No

Precision : 2/10

Query	Top 10 Documents	Score	Relevance
9 Feminism and women (document is relevant if it has information about feminism and women)	Si La people	0.0004478508617647479	No
	Jennie Kidd Trout	0.00032832438828233235	Yes
	Severino Antinori	0.00025272437400189745	Yes
	Unitard	0.0002458270200887283	No
	Jackie Frazier-Lyde	0.00023863993213654532	Yes
	Alexis Herman	0.0001330989093947552	Yes
	Spandex	0.00012566953793086933	No
	Skin-tight garment	0.00010977627044188616	No
	Naginata	9.237964220922855e-05	No
	Mackintosh	9.061744445168843e-05	No

Precision : 4/10

Query	Top 10 Documents	Score	Relevance
10 belgian municipalities (document is relevant if it has information about municipalities in belgium) An example of a case where 9 out of 10 retrieved documents were relevant.	Sint-Agatha-Berchem	0.00326975947730982	Yes
	Ganshoren	0.0026533998990163343	Yes
	Municipalities of Liechtenstein	0.0018474727884760665	No
	Beernem	0.0018036111630527726	Yes
	Evere	0.0016484176202947412	Yes

	Namur (province)	0.00161188680821 9215	Yes
	Koekelberg	0.00153995460398 79647	Yes
	Saint-Josse-ten-Noode	0.00118286376523 64917	Yes
	Municipalities of Belgium	0.00084060861844 79574	Yes
	Anderlecht	0.00083101294225 54693	Yes

Precision : 9/10

Part-2 Discussion

As seen above, the queries often do not represent the information needed, or the document corpus might not have any file relevant for retrieval.We have tried to incorporate improvements to solve some of the syntactic issues. A complete positional index has been avoided for the very same reason.

For a small IR retrieval system, we assumed that size of index is of utmost importance, Hence, we made both of the improvements such that the size of the posting list does not increase a lot , while our first improvement(biword indexing) does increase the size a little, the second(title weighting) doesn't have any effect on the size of the posting list whatsoever

Improvement 1**: Using Bigram Indexing for Multiword Queries**

1. What is the issue with the IR system built in part 1?

   Some words in the query should be together with the other consecutive words in query for the result to be relevant. As there is a possibility that independently those words mean different and together they mean different. For example "stanford university" together give different retrieved documents and independently they give different.

2. What improvement are you proposing?

   The Improvement is using the Biword indexing, where the key in this case is a bigram rather than a single term. This biword indexing is to be done at both index creation and on query for score calculation

3. How will the proposed improvement address that issue?

   As now the documents that are retrieved will be having both words together that are together present in the query. Thus these documents are more relevant comparing to what we are getting in part-1

4. A corner case (if any) where this improvement might not work or can have an adverse effect.

   This index will not work for single term queries Moreover the presence of more matching bigrams does not mean that the query will be better answered.

5. Demonstrate the actual impact of the improvement. Give three queries, where the improvement yields better results compared to the part 1 implementation.

Query 1 : White house

Biword

```
Rank :-  1 Doc ID -> 153854 Name:->  Unitard Score:->
     0.0016777307479069277
Rank :-  2 Doc ID -> 152799 Name:->  Alexis Herman Score:->
     0.0013958817050953166
Rank :-  3 Doc ID -> 153567 Name:->  Bishopthorpe Palace Score:->
     0.001378672478282959
Rank :-  4 Doc ID -> 153690 Name:->  Alton, Staffordshire Score:->
     0.0013711213082372338
Rank :-  5 Doc ID -> 152404 Name:->  Stanbury Score:->
     0.0010574573422637466
```

single word

```
Rank :-  1 Doc ID -> 153854 Name:->  Unitard Score:->
     0.0031582941804878723
Rank :-  2 Doc ID -> 152799 Name:->  Alexis Herman Score:->
     0.0027599481549437023
Rank :-  3 Doc ID -> 153690 Name:->  Alton, Staffordshire Score:->
     0.0026921775531224307
Rank :-  4 Doc ID -> 153567 Name:->  Bishopthorpe Palace Score:->
     0.002649187369982596
Rank :-  5 Doc ID -> 152404 Name:->  Stanbury Score:->
     0.002051143063841654
```

Query 2 : New york

Biword

```
Rank :-  1 Doc ID -> 153572 Name:->  Askham Bog Score:->
     0.007062212701951896
Rank :-  2 Doc ID -> 153567 Name:->  Bishopthorpe Palace
     Score:-> 0.0037978187292217326
Rank :-  3 Doc ID -> 153591 Name:->  Heslington Score:->
     0.002601141141365586
Rank :-  4 Doc ID -> 153670 Name:->  Nether Poppleton Tithebarn
     Score:-> 0.0025626184077854395
Rank :-  5 Doc ID -> 153658 Name:->  Wellman Braud Score:->
     0.0021198917945149344
```

single word

```
Rank :-  1 Doc ID -> 153572 Name:->  Askham Bog Score:->
     0.012959256274580975
Rank :-  2 Doc ID -> 153567 Name:->  Bishopthorpe Palace Score:->
     0.0072976965663868265
Rank :-  3 Doc ID -> 153591 Name:->  Heslington Score:->
     0.005195658827891314
Rank :-  4 Doc ID -> 153670 Name:->  Nether Poppleton Tithebarn
     Score:-> 0.005101588295761995
Rank :-  5 Doc ID -> 153289 Name:->  Abraham Beame Score:->
     0.004164926593448229
```

Query 3 : Mobile network

Biword

```
Rank :-  1 Doc ID -> 151800 Name:->  List of mobile network
     operators Score:-> 0.004308501544066709
Rank :-  2 Doc ID -> 152799 Name:->  Alexis Herman Score:->
     0.0004675695624007011
Rank :-  3 Doc ID -> 152106 Name:->  Inter-process
     communication Score:-> 0.00035872966630966495
Rank :-  4 Doc ID -> 153685 Name:->  Mow Cop Castle Score:->
     0.0002437176355157571
Rank :-  5 Doc ID -> 151967 Name:->  Windowing system Score:->
     0.00022873217869091644
```

single word

```
Rank :-  1 Doc ID -> 151800 Name:->  List of mobile network
     operators Score:-> 0.008474436333171492
Rank :-  2 Doc ID -> 152799 Name:->  Alexis Herman Score:->
     0.0009244821723396186
Rank :-  3 Doc ID -> 152106 Name:->  Inter-process communication
     Score:-> 0.0006693633645137671
Rank :-  4 Doc ID -> 151799 Name:->  T-Mobile Score:->
     0.0004758643365701154
Rank :-  5 Doc ID -> 153685 Name:->  Mow Cop Castle Score:->
     0.0004745822952750776
```

Improvement 2**: Adding weights to Documents with title words that match query words**

1. What is the issue with the IR system built in part 1?
   The document whose title is matching with the query should be ranked very high which is not case in part-1

2. What improvement are you proposing?
   For every word in the query that comes in the title of the retrieved documents, the score is increased by constant scale.

3. How will the proposed improvement address that issue?
   By providing an additional weight to documents that also contain query terms in the title, those documents will have higher scores and will rank higher in the rank list.

4. A corner case (if any) where this improvement might not work or can have an adverse effect.
   In some cases, word matching to the title is bad for the user ,as the user may want to query documents that are not so relevant by title with the same query words.
   It also may fail when the query/title has too many stop words as other documents with the same stop words may get unfairly weighted.

5. Demonstrate the actual impact of the improvement. Give three queries, where the improvement yields better results compared to the part 1 implementation.

Query 1 : **Supreme court**

With weight

```
Rank :-  1 Doc ID -> 153237 Name:->  Court order Score:->
     0.0031050911407953493
Rank :-  2 Doc ID -> 152752 Name:->  Council of State Score:->
     0.0011205743996111246
Rank :-  3 Doc ID -> 152973 Name:->  Ishim River Score:->
     0.0006962669499548685
Rank :-  4 Doc ID -> 152528 Name:->  Campaign for Homosexual
Law
     Reform Score:-> 0.0006317039399088495
Rank :-  5 Doc ID -> 152981 Name:->  Bückeburg Score:->
     0.0005593813528483971
```

Without weight

```
Rank :-  1 Doc ID -> 152752 Name:->  Council of State Score:->

     0.0039419849593979525
Rank :-  2 Doc ID -> 153237 Name:->  Court order Score:->
     0.001507939031218836
Rank :-  3 Doc ID -> 152528 Name:->  Campaign for Homosexual Law
     Reform Score:-> 0.0011756464187908436
Rank :-  4 Doc ID -> 152973 Name:->  Ishim River Score:->
     0.0010643626049779272
Rank :-  5 Doc ID -> 152646 Name:->  Keepie uppie Score:->
     0.0009084632185613118
```

Query 2 : State council

With weight

```
Rank :-  1 Doc ID -> 152752 Name:->  Council of State Score:->
     0.01769317182520361
Rank :-  2 Doc ID -> 152578 Name:->  List of heads of state of
     Mexico Score:-> 0.0018696984450467663
Rank :-  3 Doc ID -> 152011 Name:->  Gamesley Score:->
     0.0014738106438973437
Rank :-  4 Doc ID -> 152057 Name:->  Rodmell Score:->
```

```
      0.0014514504943829267
Rank :-  5 Doc ID -> 152934 Name:->  Benasque Score:->
      0.001436187483012367
```

Without weight

```
Rank :-  1 Doc ID -> 152752 Name:->  Council of State Score:->
      0.014503155578638287
Rank :-  2 Doc ID -> 152283 Name:->  Saint-Savin, Vienne Score:->
      0.004098881311755566
Rank :-  3 Doc ID -> 152579 Name:->  List of Prime Ministers of
      Italy Score:-> 0.0032719201218514
Rank :-  4 Doc ID -> 152491 Name:->  Vechta (district) Score:->
      0.0026317423380187795
Rank :-  5 Doc ID -> 152033 Name:->  Oldenburg (district) Score:->
      0.0022374580865278898
```

Query 3 : Political party

With weight

```
Rank :-  1 Doc ID -> 151952 Name:->  Nordic Reich Party
Score:->
      0.007719059074448545
Rank :-  2 Doc ID -> 152417 Name:->  Austrian People's Party
      Score:-> 0.0021650140378175016
Rank :-  3 Doc ID -> 153436 Name:->  Mario Party 3 Score:->
      0.0018369169075823108
Rank :-  4 Doc ID -> 153448 Name:->  Mario Party 2 Score:->
      0.0015670768589059127
Rank :-  5 Doc ID -> 152707 Name:->  Jan Narveson Score:->
      0.0015523799712873797
```

Without weight

```
Rank :-  1 Doc ID -> 151952 Name:->  Nordic Reich Party Score:->
      0.007834026129585517
Rank :-  2 Doc ID -> 151801 Name:->  Our Fair City Score:->
      0.0029796258691742166
Rank :-  3 Doc ID -> 151951 Name:->  Leif Zeilon Score:->
      0.0024652754533645714
Rank :-  4 Doc ID -> 152707 Name:->  Jan Narveson Score:->
      0.0024515019934627427
```

```
Rank :-  5 Doc ID -> 152579 Name:->  List of Prime Ministers of
     Italy Score:-> 0.002306757066901961
```

Conclusions

The ranked retrieval system developed can give us a list of documents based on score computations from the lnc.ltc scoring system. In this IR System the bag of words model is assumed giving no preference to position of words in the query. The bigram index, implemented as improvement, helped in maintaining the positional information to a small extent as compared to the unigram index in the model. Once the index has been created on running the code, it is used to query for relevant documents. The query results showed us the Top k documents along with their scores. As mentioned in the problem statement the stemming and lemmatization operations weren't performed.

In an attempt to improve the model the bigram indexing was added. This can help in giving more relevant outputs except for cases where a single word has been queried. Another improvement that was worked on, was to add weights for the document title. This resulted in increasing the score of documents by a constant in cases where the document title had the query term. The query in which weights were given to the title gave better results. for example: The query results of Political Party were more relevant when the weights were used.

Further possible improvement can be brought to the IR model by adding a Spelling correction feature. It can help the model to be robust against incorrect spelling input by the user. Index compression could have led to saving space but since the data set size wasn't that large it wasn't implemented.

Overall the IR system implemented resulted in good results for the query terms searched. The improvements implemented were found to be working as expected. The relevance of the output depended heavily on the type of query and if the relevant search terms were added to the query or not.

References :-

1) Nltk library :- https://www.nltk.org/
   For tokenization and making bi-word indexing

2) Pickle library :- https://www.tutorialspoint.com/python-pickling
   For storing and writing in files

3) Lnc.ltc :-
   https://nlp.stanford.edu/IR-book/html/htmledition/document-and-query-weighting-schemes-1.html