CA4009: Search Technologies
Laboratory Session 4 & 5
30th November 2021

Joanna Talvo - 18342523
Chloe Ward - 18302716

4 Experimenting with Query Expansion in IR

4.1 Initial Investigations

TOPIC TITLE: Poliomyelitis and Post-Polio (302)

Below we searched the title “Poliomyelitis and Post-Polio”. We used the BM25 ranking functions with **k=1.2** and **b=0.75**. We got the below results for the top 5 documents in the TREC Collection alongside with the top 5 terms that appear in each document.

	Documents	**Term Freq**
1	FBIS4-67701	case:11, 7.778768 polio:11, 3451.9934 year:11, 1.9854704 diseas:10, 70.675095 vaccin:9, 562.4654
2	LA043090-0036	polio:8, 3451.9934 case:6, 7.778768 year:5, 1.9854704 annual:3, 10.915225 caus:3, 19.695518
3	LA013089-0022	dai:3, 4.1854615 birthdai:2, 177.65053 dime:2, 1123.734 fdr:2, 11237.341home:2, 4.517582

4	FBIS4-30637	health:13, 15.862416 minist:7, 6.0379205 problem:7, 6.402888 region:5, 8.286994 bangkok:4, 163.71823
5	FBIS3-60403	china:9, 20.070492 polio:9, 3451.9934 year:7, 1.9854704 erad:5, 316.07123 million:5, 9.764555

We then analysed all the terms that appeared in the top 5 documents, we added these terms to a **words.txt** file to get the frequency of each word that appears across all documents to get the word that appears most often. It is a simple code that just manipulates the strings to get rid of ":","." etc to allow us to count the frequencies of the terms in the words.txt file.

The short code we created to do this is as follows:

```
import sys
import collections

count = dict()

f = open("words.txt", "r")
words = f.read()


li = [word for word in words if any(not char.isdigit() for char in word)]

list1 = []
for n in li:
    if n != ":":
        list1.append(n)

list2 = []
for b in list1:
    if n != ".":
        list2.append(n)

remove = ([s.replace(' ', '') for s in list1])
final = "".join(remove)
step2 = final.strip().replace(" ", "").split(".")
occurrences = collections.Counter(step2)
```

This gave the results in the terminal as follows:

```
root@LAPTOP-JGAUQ0TS:/mnt/c/Users/chloe/case4/ca4009/lab4-5# python3 script.py
Counter({'case,': 4, 'polio,': 4, 'diseas,': 4, 'children,': 4, 'erad,': 4, 'poliomyel,': 4, 'time,': 4, 'report,': 4, 'year,': 3, 'v
accin,': 3, 'declin,': 3, 'english,': 3, 'health,': 3, 'minist,': 3, 'nation,': 3, 'program,': 3, 'set,': 3, 'world,': 3, 'ag,': 3, '
countri,': 3, 'make,': 3, 'public,': 3, 'text,': 3, 'effort,': 3, 'offici,': 3, 'problem,': 3, 'govern,': 2, 'programm,': 2, 'area,':
 2, 'launch,': 2, 'state,': 2, 'childhood,': 2, 'dr,': 2, 'regist,': 2, 'success,': 2, 'control,': 2, 'death,': 2, 'due,': 2, 'effect
,': 2, 'elimin,': 2, 'end,': 2, 'goal,': 2, 'jpr,': 2, 'level,': 2, 'mr,': 2, 'administ,': 2, 'april,': 2, 'birthdai,': 2, 'bodi,': 2
, 'compar,': 2, 'crippl,': 2, 'foundat,': 2, 'gave,': 2, 'higher,': 2, 'includ,': 2, 'local,': 2, 'made,': 2, 'manag,': 2, 'march,':
2, 'million,': 2, 'oper,': 2, 'past,': 2, 'present,': 2, 'result,': 2, 'sentenc,': 2, 'south,': 2, 'special,': 2, 'ten,': 2, 'tep,':
2, 'women,': 2, 'column,': 2, 'metro,': 2, 'campaign,': 2, 'desk,': 2, 'edit,': 2, 'found,': 2, 'home,': 2, 'intern,': 2, 'introduc,'
: 2, 'live,': 2, 'major,': 2, 'mondai,': 2, 'page,': 2, 'part,': 2, 'region,': 2, 'thousand,': 2, 'unit,': 2, 'word,': 2, 'spring,':
2, 'earli,': 2, 'futur,': 2, 'hundr,': 2, 'januari,': 2, 'post,': 2, 'achiev,': 2, 'fight,': 2, 'plan,': 2, 'yesterdai,': 2, 'cent,':
 1, 'gujarat,': 1, 'mass,': 1, 'declar,': 1, 'district,': 1, 'grade,': 1, 'india,': 1, 'massiv,': 1, 'occurr,': 1, 'organis,': 1, 'ru
ral,': 1, 'singl,': 1, 'urban,': 1, 'anganwadi,': 1, 'bombai,': 1, 'centr,': 1, 'centuri,': 1, 'confid,': 1, 'deem,': 1, 'deform,': 1
, 'februari,': 1, 'fell,': 1, 'gohil,': 1, 'measl,': 1, 'media,': 1, 'medic,': 1, 'month,': 1, 'nathani,': 1, 'neonat,': 1, 'orthopae
d,': 1, 'paediatrician,': 1, 'percent,': 1, 'perform,': 1, 'period,': 1, 'physic,': 1, 'reach,': 1, 'reduct,': 1, 'scourg,': 1, 'surg
eon,': 1, 'survei,': 1, 'tetanu,': 1, 'till,': 1, 'trend,': 1, 'add,': 1, 'ago,': 1, 'ahmedabad,': 1, 'ashram,': 1, 'asia,': 1, 'asse
rt,': 1, 'associ,': 1, 'attain,': 1, 'aver,': 1, 'avert,': 1, 'back,': 1, 'begin,': 1, 'bhagat,': 1, 'bharat,': 1, 'camp,': 1, 'care,
': 1, 'citi,': 1, 'club,': 1, 'come,': 1, 'conduct,': 1, 'confin,': 1, 'conform,': 1, 'consult,': 1, 'contact,': 1, 'correct,': 1, 'c
ough,': 1, 'credit,': 1, 'cso,': 1, 'date,': 1, 'debil,': 1, 'deep,': 1, 'dent,': 1, 'diarroho,': 1, 'die,': 1, 'diptheria,': 1, 'doc
tor,': 1, 'dose,': 1, 'dread,': 1, 'drive,': 1, 'east,': 1, 'educ,': 1, 'enlist,': 1, 'epidemiolog,': 1, 'examin,': 1, 'expect,': 1,
'extens,': 1, 'fact,': 1, 'fashion,': 1, 'fast,': 1, 'feb,': 1, 'free,': 1, 'inaccess,': 1, 'instanc,': 1, 'interior,': 1, 'june,': 1
, 'kumar,': 1, 'lame,': 1, 'monsoon,': 1, 'mop,': 1, 'novemb,': 1, 'panchayat,': 1, 'physiotherapist,': 1, 'pox,': 1, 'practic,': 1,
'predecessor,': 1, 'print,': 1, 'privat,': 1, 'profess,': 1, 'put,': 1, 'record,': 1, 'regular,': 1, 'rise,': 1, 'rotari,': 1, 'scale
,': 1, 'school,': 1, 'shaktisinh,': 1, 'show,': 1, 'signific,': 1, 'small,': 1, 'specialist,': 1, 'specif,': 1, 'stage,': 1, 'stand,'
: 1, 'statist,': 1, 'suffer,': 1, 'surgic,': 1, 'system,': 1, 'threshold,': 1, 'total,': 1, 'treat,': 1, 'truste,': 1, 'tv,': 1, 'uni
```

This allowed us to pick the top words that appeared in all the documents by selecting the top 5. We then also wrote another short script to allow us to get the RW and OW values for each term that we chose.

```python
import math

# top terms from all 5 documents
terms = ["case", "children", "polio", "disease", "year"]

# number of documents in the collection
N = 500000

# number of known documents top 5 documents from the query
R = 5


# number of documents term t occurs in
n = [67897, 23678, 153, 7473, 266010 ]


r = [4, 4, 5, 4, 3]


for i in range(len(terms)):
    top = (r[i] + 0.5) * (N - n[i] - R + r[i] + 0.5)
    down = (n[i] - r[i] + 0.5) * (R - r[i] + 0.5)
    rw = math.log(top/down)
    ow = r[i] * rw

    print(str(terms[i]) + ", " + "rw= " + str(rw) + "  ow= " + str(ow))
```

The table below shows the values of RW and OW for the most significant terms that are present in the top 7 documents we retrieved from searching the topic title “**Poliomyelitis and Post-Polio**” as a query.

Term i	**OW(i)**
case	11.797339265271736
children	16.401227262196425

polio	52.59684322718993
disease	21.149329969756266
year	0.6246591047974912

Next, we added some terms from the above table to our original query which is "**Poliomyelitis and Post-Polio**" to see the changes in the ranked retrieval list. Below are the three cases we have. We cater cases for different numbers of expansion terms.

1. case
2. children, polio
3. disease, polio, case

* Expanded Query = Original Query + 1/2/3

Original top 5 queries	Expanded Query (1)	Expanded Query (2)	Expanded Query (3)
FBIS4-67701	**FBIS4-67701**	**FBIS4-67701**	**FBIS4-67701**
LA043090-0036	**LA043090-0036**	**LA043090-0036**	**LA043090-0036**
LA013089-0022	**FBIS4-30637**	**FBIS3-60403**	FR940620-2-00118
FBIS4-30637	**LA013089-0022**	FBIS3-60404	FR940620-2-00117
FBIS3-60403	**FBIS3-60403**	FR940126-2-00106	LA072890-0066

The original top 5 documents list we retrieved are clearly identical to the top 5 documents with the expanded query 1. The only difference we noticed is the ranking of the two documents. **LA013089-0022** decreased in ranking but **FBIS4-30637** increased in ranking in the expanded query 1. And the rest of the documents have a similar rank. The reason we think that the document **FBIS4-30637** increased in ranking in expanded query 1 is because it is more relevant to the expanded query 1.

However, with the expanded query 2 and 3, new documents were retrieved. Comparing the original documents retrieved to the expanded query 2, we noticed that rank 1 and 2 documents are similar documents. **FBIS3-60403** increased in ranking in the expanded query 2. Two new documents **FBIS3-60404** and **FR940126-2-00106** were not in the top 5 from the original query. Comparing the original retrieved documents to the expanded query 3, top 2 ranks are similar documents. Three new documents were retrieved from the expanded query 3.

4.2 Investigations using qrel data

Precision = Total number of documents retrieved that are relevant / Total number of documents that are retrieved

Topic 302: Poliomyelitis and Post-Polio

Original Unexpanded Query = Poliomyelitis and Post-Polio
Expanded Query = Poliomyelitis and Post-Polio disease polio case

Original Unexpanded Query (1 = relevant, 0 = not relevant)

Ranked results		**Relevance**
Rank	**Retrieved Documents**	**qrel**
1	FBIS4-67701	1
2	LA043090-0036	1
3	LA013089-0022	0
4	FBIS4-30637	1
5	FBIS3-60403	1

Document **LA013089-0022** is found to be not relevant in the qrel file. Precision Value is ⅘ or 0.8

Expanded Query (1 = relevant, 0 = not relevant)

Ranked results		**Relevance**
Rank	**Retrieved Documents**	**qrel**
1	FBIS4-67701	1
2	LA043090-0036	1
3	FR940620-2-00118	0
4	FR940620-2-00117	1
5	LA072890-0066	1

Document **FR940620-2-00118** is found to be not relevant in the qrel file. Precision Value is ⅘ or 0.8

For Topic 302, the precision values for the original and expanded query turned out to be similar (0.8). This precision value is good as it means that every result we retrieved by conducting the search was relevant and it is near the perfect precision score which is 1.0. In terms of ranking, the top 2 for each type of queries are similar (**FBIS4-67701** & **LA043090-0036**). Three new documents were retrieved from using the expanded query. We think that the new retrieved documents were more related to the expanded query. Since we added terms like disease, polio, and case, these new documents retrieved are more likely to contain more of these new terms.

Topic 310: Radio Waves and Brain Cancer

Original Unexpanded Query = Radio Waves and Brain Cancer
Expanded Query = Radio Waves and Brain Cancer brain human effect

Original Unexpanded Query (1 = relevant, 0 = not relevant)

Ranked results		**Relevance**
Rank	**Retrieved Documents**	**qrel**
1	FT931-11958	1
2	LA100889-0041	0
3	LA053090-0120	0
4	FT922-289	0
5	FT922-15435	1

Documents **FT931-11958** & **FT922-15435** are the only documents retrieved that are relevant. Precision value is ⅖ or 0.4

Expanded Query (1 = relevant, 0 = not relevant)

Ranked results		**Relevance**
Rank	**Retrieved Documents**	**qrel**
1	FT931-11958	1
2	LA053090-0120	0

3	LA100889-0041	0
4	FT922-15435	1
5	FT922-289	0

Documents **FT931-11958** & **FT922-15435** are the only documents retrieved that are relevant. Precision value is ⅖ or 0.4

For topic 310, the precision values for the original and expanded query turned out to be similar as well (0.4). This precision value is not as bad but it is below 0.5 which is the balanced range. The retrieved documents are all similar for both however there is a difference in ranking. Only **FT931-11958** are similar in rank (rank 1) and the rest are ranked differently. This is due to the addition of the terms brain, human, and effect. The added terms could have more term frequency / more relevant in the documents that increased in ranking in the expanded query rank list. Unlike topic 302, there are no new documents retrieved in the top 5 for the expanded query for this topic.

Topic 309: Rap and Crime

Original Unexpanded Query = Rap and Crime
Expanded Query = Rap and Crime desk edit page

Original Unexpanded Query (1 = relevant, 0 = not relevant)

Ranked results		**Relevance**
Rank	**Retrieved Documents**	**qrel**
1	LA040289-0172	0
2	LA090190-0055	0
3	LA083189-0049	0

4	LA123089-0064	0
5	LA082490-0024	0

All documents are found to be not relevant in the qrel file. Precision Value is 0.

Expanded Query (1 = relevant, 0 = not relevant)

Ranked results		**Relevance**
Rank	**Retrieved Documents**	**qrel**
1	LA040289-0172	0
2	LA083189-0049	0
3	LA093089-0007	0
4	LA052790-0015	0
5	LA082490-0024	0

All documents are found to be not relevant in the qrel file. Precision Value is 0.

For topic 309, the precision values for both are 0. This means that every result retrieved by the search we conducted was not relevant. In terms of rankings, the top 1 for each type of queries are similar (LA040289-0172) and the rest are ranked differently. There are also two new documents retrieved in the expanded query (**LA093089-0007** & **LA052790-0015**). This could be the same case as topic 302 wherein adding the terms desk, edit, and page were more related / have a higher term frequency in these two new retrieved documents.

4.3 Extended Investigation: Examining “true” relevance feedback

Topic 302: Poliomyelitis and Post-Polio

Original Unexpanded Query = Poliomyelitis and Post-Polio
Expanded Query = Poliomyelitis and Post-Polio disease polio case

Table 1

Original Unexpanded Query (1 = relevant, 0 = not relevant)

Ranked results		Relevance	
Rank	**Retrieved Documents**	**PSEUDO**	**TRUE**
1	FBIS4-67701	1	1
2	LA043090-0036	1	1
3	LA013089-0022	1	0
4	FBIS4-30637	1	1
5	FBIS3-60403	1	1

Table 2

Expanded Query (1 = relevant, 0 = not relevant)

Ranked results	Relevance

Rank	Retrieved Documents	PSEUDO	TRUE
1	FBIS4-67701	1	1
2	LA043090-0036	0	1
3	FR940620-2-00118	1	0
4	FR940620-2-00117	1	1
5	LA072890-0066	1	1

To get the pseudo relevance values for both types of queries, we read each document retrieved using the original query and the expanded query. From there, we based if the documents are relevant or not. This is truly based on our assumption that these documents are relevant to the query we did.

The results obtained in table 2 were not similar to table 1. We obtained 3 new documents in table 2 (**FR940620-2-00118**, **FR940620-2-00117** & **LA072890-0066**) when we searched using the expanded query (original query + “disease”, “polio”, “case”).

Though we think that the true relevance feedback is more accurate than the pseudo relevance feedback because true relevance feedback is based on the algorithm whereas pseudo feedback is based more on the human perspective while reading the documents and each person who reads it will come to a different conclusion. The algorithm for the true relevance feedback will always be the same as it is following a set of rules.

ow(i) values for original unexpanded query = Poliomyelitis and Post-Polio

Term i	OW(i)
case	11.797339265271736
children	16.401227262196425

polio	52.59684322718993
disease	21.149329969756266
year	0.6246591047974912

ow(i) values for expanded query = Poliomyelitis and Post-Polio disease polio case

Term i	**OW(i)**
case	21.24317421885095
year	11.34829401890836
time	7.243612945724478
disease	32.93375696823448
health	25.19144729364048

These are the ow(i) values that we got for an original query and an expanded query and we have different terms for each type of query. If we compare the values of the terms that appear on the same table (case & disease), the ow(i) values for the expanded query seems to be higher than the one with the original query.