Your Name: LIU Xi

Your Andrew ID: xiliu1

Homework 1

1 Collaboration and Originality

Your report must include answers to the following questions:

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? (It is not necessary to describe discussions with the instructor or TAs).
   No.
   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.
   N/A
2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?
   No.
   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.
   N/A
3. Are you the author of every line of source code submitted for this assignment (Yes or No)?
   Yes.
   If you answered No:
   a. identify the software that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

4. Are you the author of every word of your report (Yes or No)?
   Yes.
   If you answered No:
   a. identify the text that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

2 Structured query set

2.1 Summary of query structuring strategies

Briefly describe your strategies for creating structured queries. These should be general strategies, i.e., not specific to any particular query.

Strategies:

For two terms A, B:

1) If they are correlated, use #AND(A, B).
2) If they are not correlated or represent for two distinct meanings, use #OR(A, B)
3) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).
4) If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.
5) If the term is too short or ambiguous, search it in the fields of keyword, url, and title and merge the results with #OR().
6) If the term is potential keywords, search in the field of title and keyword too.

For more than two terms, apply the same rules above too.

2.2 Structured queries

List your structured queries. For each query, provide a brief (1-2 sentences) description that:

1. identifies which strategy (from Question 2.1) was used for that query,
2. any important deviations from your default strategies, and
3. your intent, i.e., why you thought that particular structure was a good choice.

10:#AND(internet #SYN(cheap inexpensive affordable))

Strategy:

1) If they are correlated, use #AND(A, B).
2) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).

Important deviations:

Added “inexpensive” and “affordable” as the synonyms for “cheap”, to enlarge the searching scope

Intent:

The query intends to search for cheap internet options, so the two terms (or synonyms for “cheap”) should appear in the same document. #AND(cheap internet) could satisfy this need. “cheap” could be expressed by other words, so simply search #AND(cheap internet) will miss them. #SYN(cheap inexpensive affordable) expands the searching scope to solve the problem.

12:#OR(djs.url djs.keywords djs.inlink djs)

Strategy:

1) If the term is too short or ambiguous, search it in the fields of keyword, url, and title and merge the results with #OR().

Important deviations:
Search in field title, inlink, and keywords because djs is too ambiguous and infrequent so searching in all field could retrieve more documents.
Intent:
The single term "djs" is too ambiguous and infrequent to search, so searching in all fields then merging the results by #OR could increase the number of relevant document retrieved, in order to increase the precision.

26:#AND(#SYN(lower low slow slower) #NEAR/4(heart rate))
Strategy:

1) If they are correlated, use #AND(A, B).
2) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).
3) If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.

Important deviations:
Added low slow slower to #SYN
Set n = 4 for #NEAR/n
Intent:
This query intends to find something could lower heart rate. So lower (or its synonyms) and "heart rate" should appear in the same document. #AND() could satisfy this need.
#SYN(lower low slow slower) expands the searching scope to solve the problem. n = 4 achieves best performance in the experiment.

29:#AND(#OR(ps2 ps #NEAR/4(play station) playstation) games)
Strategy:

1) If they are correlated, use #AND(A, B).
2) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).
3) If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.

Important deviations:
Added "playstation" and "play stations"
Set n = 4 for #NEAR/4(play station)
Intent:
This query intends to find playstation 2 games. These terms or its synonyms should appear in the same document. #AND() could satisfy this need.
#OR(ps2 ps #NEAR/4(play station) playstation) expands the searching scope to solve the problem.
#NEAR(play station) guarantees they are searched as phrases and n = 4 achieves best performance in the experiment.

33:#OR(#NEAR/4(elliptical trainer) trainer.title elliptical.title)
Strategy:

1) If the term is too short or ambiguous, search it in the fields of keyword, url, and title and merge the results with #OR().
2) If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.

Important deviations:
Intent:
This query intends to find elliptical trainers. Because the query is short, I extended the query with #OR, and also retrieves the documents with title containing the terms because they should be relevant too.

52:#OR(avp.url avp.inlink, avp.keywords avp)
Strategy:
Important deviations:
Intent:
This query intends to find “avp”. Because the query is short, ambiguous, and infrequent, I merged the searching result in all fields with #OR

71:#AND(#SYN(living life working) in india)
Strategy:

1) f they are correlated, use #AND(A, B).
2) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).

Important deviations:
Add life to #SYN

Intent:
This query intends to find information about living in india. So living (or its synonyms) and india should appear in the same document. #AND() could satisfy this need.
#SYN(life living) expands the searching scope to solve the problem.

102:#NEAR/9(fickle creek farm)
Strategy:
If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.

Important deviations:
Set n = 9
Intent:
“fickle creek farm” is a name of a farm so using #NEAR to eliminate documents about other creeks or farms to improve the precision.

149:#AND(uplift at #OR(#NEAR/9(yellowstone national park) yellowstone yellowstone.title yellowstone.keywords))

Strategy:

1) If they are correlated, use #AND(A, B).
2) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).
3) If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.

Important deviations:

Set n = 9

Intent:

This query intends to find information about "uplift in Yellowstone national park". So they should appear in the same document. #AND() could satisfy this need.

#SYN() expands the searching scope to solve the problem.

#NEAR() guarantees "Yellowstone national park" is searched as a phrase and n = 9 achieves best performance in the experiment.

190:#AND(#NEAR/8(brooks brothers) #SYN(coupon clearance sale bargain discount))

Strategy:

1) If they are correlated, use #AND(A, B).
2) If they represent for similar meaning, use #SYN(A, B). Or expand the term with its synonyms with #SYN(A1, A2) or #OR(A, B).
3) If term A and term B belongs to a phrase, use #NEAR/n to constrain searching scope. As for the value of n, I adjusted n to 2-3 times of the phrase length.

Important deviations:

Added "coupon sale bargain discount" to SYN

Set n = 8

Intent:

This query intends to find information about brooks brothers clearance. So they should appear in the same document. #AND() could satisfy this need.

#SYN() expands the searching scope to solve the problem.

#NEAR() guarantees "brooks brothers" is searched as a phrase and n = 9 achieves best performance in the experiment.

3 Experimental results

Present the complete set of experimental results. Include the precision and running time results described above. Present these in a tabular form (see below) so that it is easy to compare the results for each algorithm.

3.1 Unranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.0100	0.0400	0.0300
P@20	0.0050	0.0200	0.0450
P@30	0.0033	0.0433	0.0500
MAP	0.0010	0.0142	0.0076
Running Time	00:18	00:03	00:04

3.2 Ranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.1500	0.2500	0.1800
P@20	0.1800	0.2600	0.2000
P@30	0.1667	0.2767	0.1900
MAP	0.0566	0.0385	0.0904
Running Time	00:18	00:03	00:04

4 Analysis of results

Discuss your observations about the differences between the three different approaches to forming queries, and the two different approaches to retrieving documents in terms of their retrieval performance and running time.

	BOW #OR	BOW #AND	Structured
Retrieval Performance	Retrieved more documents than other two approaches, with lowest precision, and smallest MAP, that is the retrieval covers a larger but is not accurate enough.	Retrieved fewest documents than other two approaches, with higher precision and highest MAP, that is the retrieval covers a smaller scope so it is accurate but too strict.	Retrieved fewer documents than #OR, more than #AND, with highest precision but lower MAP, that is the retrieval covers a reasonable scope and high performance for the top documents.

Running Time	Took longest running time because more documents are retrieved.	Took shortest running time because fewest documents are retrieved.	Took longest running time because more documents are retrieved.

	UnrankedBoolean	**RankedBoolean**
Retrieval Performance	Got very low precision and MAP because the documents are regarded equally relevant so less relevant documents may appear at the top.	Got higher precision and MAP for any operators because the relevance are computed by term frequency so relevant documents are retrieved and listed at the top.
Running Time	The running time is very close to each other because the most time were spent on file IO (retrieving file name, indices, write outputs) instead of sorting.	

Discuss the effectiveness, strengths, and weaknesses of the query operators and fields, and your success and failure at using them in queries. Did they satisfy your expectations given in Section 3?

	#AND	**#OR**	**#SYN**	**#NEAR/n**	**Fields**
Effectiveness	Set up more strict criteria for document selection (select by intersections) so the higher precision could be achieved with more relevant documents at the top.	Set up more tolerant criteria for document selection (select by union) so more candidate documents could be retrieved. The relevance could be improved by selecting candidates from a larger candidate pool and searching field.	Merges the result for terms of similar meanings, the searching scope could be extended. With more documents retrieved, more relevant document could be found and listed.	Set up the strictest rule for document selection, that terms should all appear in the documents within a certain distance. It significantly narrows the searching scope and improves precision.	Improves retrieval performance by limit the searching in a certain field, so the retrieval could be more precise if we expect the term to appear in these fields not the others.
Strengths	Effectively narrow the searching scope and improve precision. Faster in running because less documents are retrieved.	Enlarge the searching scope so that some relevant documents missing one or two terms could be selected.	Enlarge the searching scope so that some relevant documents represented by synonyms could be selected.	Effectively narrow the searching scope and improve precision. Faster in running because less documents are retrieved.	Better precision could be achieved if we expect the term to appear in certain field. Also documents should be more relevant if the term appear in title or keywords.
Weaknesses	Many relevant	Too many	The relevance	Relevant	The searching

	documents are ignored if #AND were abused. Fewer documents are retrieved.	irrelevant documents are retrieved if #OR were abused. The precision may be lower because documents containing only one term match could be selected.	depends on the selection of synonyms. #SYN could harm the precision if wrong synonyms were included.	documents are ignored if they do not contain one or two terms. What is more, even if they contain all the terms, they could still be ignored if n is set to be too small.	is limited to a certain field so some relevant documents are missed.
Success	For a phrase or highly correlated terms, for example, “cheap Internet”, #AND could significantly improve the retrieval performance because only the documents contains all terms could be selected.	For ambiguous or short queries, using #OR to merge the result for different fields could retrieve more relevant documents, for example #OR(djs.url djs.keywords djs.inlink djs). Also for less correlated terms, more documents are retrieved with higher term frequency and better relevance are achieved.	For short queries, using #SYN could extend the query to get a larger searching scope so that more relevant documents could be retrieved. For example, #AND(#NEAR/8(brooks brothers) #SYN(coupon clearance sale bargain discount)) is better than #AND(brooks brothers clearance) because “brooks brothers sale” could be what the user is looking for too.	For special phrases, using #NEAR/n could eliminate the documents where the terms are separated far away. For example, using #NEAR/9(yellowstone national park) retrieves the document contains the phrases only, which brings higher precision.	If the term appears in the title or keywords, the document is very likely to be relevant. Also for short abbreviations, it may not appear in body or stands for different meaning as we expected. By using field search, the retrieval performance is improved.
Failure	Because the same meaning could be represented by several word, simply including one term in #AND makes the scope to small to retrieve enough relevant	For highly correlated terms or phrases, simply using #OR retrieves documents contains only one or two terms, which are irrelevant.	For some professional phrases, using #SYN may harm the relevance. For example, using #AND(heart #SYN(rate beat))” is worse than simply	For phrases like “Yellowstone national park”, sometimes people only represent them with shorter forms. Simply using #NEAR will miss these document, so	The default setting is to search in body field, which may miss the documents contains the term in its title or meta.

	documents		using #AND(heart rate) because “heart rate” should be joint together in medicial field.	did a small n. In these cases, the precision is harmed.	
Satisfy expectation	Yes. Using #AND guarantees some terms appears in the retrieved documents so the precision could be improved.	Yes. Using #OR merges searching results in different fields so more relevant documents are retrieved.	Yes. Using #AND(#NEAR/8(brooks brothers) #SYN(coupon clearance sale bargain discount)) retrieved more relevant documents because the terms in #SYN stand for very similar meanings.	Yes. Using NEAR/4(brooks brothers) retrieves documents contains “brook brothers” which is a brand, and eliminate documents talking about brooks and brothers. The precision is improved.	Yes. For abbreviations like “djs”, using field search could find the highly relevant documents with “djs” in the title or meta. Together with #OR we could expand the searching scope from body to all the fields so that more relevant documents could be retrieved.

Feel free to include other comments about what you observed

In this part of the report, do not just summarize the results from the previous section. We can see your results. You are expected to write your interpretation of the results based on what you learned in the lectures and readings. This is your chance to show what you learned from this homework assignment - take this section very seriously.