Homework 1

1 Introduction

1.1 Collaboration and Originality

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

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

   NO

3. Did you examine anyone else's software for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   NO

4. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   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.

   YES

5. Are you the author of every word of your report (Yes or No)?

   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.

   YES

2 Structured queries

2.1 Query structuring strategies

1. When adjective describes a noun or there is a phrase (with order), use NEAR around them.
2. When different terms describe a same concept, use SYN to put them together.
3. If all terms need to show up in the document to make a match, use AND on those terms.
4. If any of the terms showing up in the document could make a match, use OR on those terms. For example, use OR on all field so that as long as the term appears in the document, it will be matched.
5. Use #AND(#OR() #OR #OR) and #OR(#AND() #AND() #AND()) structure to express different way for the same meaning.
6. Delete unnecessary terms.

2.2 Queries

711: #AND(#SYN(station #NEAR/5(Train station)) #NEAR/5(security measures) measures)

RULE: 1, 2, 3

730: #AND(#OR(Gastric Gastric.title Gastric.keywords) #OR(bypass bypass.title bypass.keywords) #OR(complications complications.title complications.keywords))

RULE: 3,4,5

733: #AND(#OR(Airline Airline.title Airline.keywords Airline.url) #OR(overbooking overbooking.title overbooking.keywords overbooking.url))

RULE: 3,4,5

751: #OR(#NEAR/5(Scrabble Players) #NEAR/5(Scrabble.title Players.title) #NEAR/5(Scrabble.keywords Players.keywords) #NEAR/5(Scrabble.url Players.url))

RULE: 1, 4

758: #OR (#NEAR/5(#SYN(Embryonic stem) cells) #NEAR/5(#SYN(Embryonic.keywords stem.keywords) cells.keywords) #NEAR/5(#SYN(Embryonic.title stem.title) cells.title) #NEAR/5(#SYN(Embryonic.url stem.url) cells.url))

RULE: 1,2,4

764: #AND(Increase #NEAR/5(mass transit))

RULE: 1, 3, 6

802: #AND(#OR(Volcano Volcano.title Volcano.keywords) #OR(eruptions eruptions.title eruptions.keywords) #NEAR/5(global temperature))

RULE: 1, 3, 4

809: #AND(#OR(wetlands wetlands.title wetlands.keywords) #OR(wastewater wastewater.title wastewater.keywords) #OR(treatment treatment.title treatment.keywords))

RULE: 3,4,5

811: #NEAR/5(handwriting recognition)

RULE: 1

826: #AND(#OR(Florida Florida.title Florida.keywords Florida.url) #OR(Seminole Seminole.title Seminole.keywords Seminole.url) #OR(Indians Indians.title Indians.keywords Indians.url))

RULE:3,4,5

3 Experiment: Unranked Boolean

	BOW #OR (Exp-3a)	BOW #AND (Exp-3b)	Structured (Exp-3c)
P@10	0.0000	0.1100	0.1800
P@20	0.0000	0.1750	0.2400
P@30	0.0033	0.2000	0.2067
MAP	0.0000	0.0428	0.0618
Running Time	00:17	00:02	00:02

4 Experiment: Ranked Boolean

	BOW #OR (Exp-4a)	BOW #AND (Exp-4b)	Structured (Exp-4c)
P@10	0.1100	0.4300	0.6400
P@20	0.1200	0.4400	0.5000
P@30	0.1133	0.4000	0.4333
MAP	0.0090	0.1060	0.1298
Running Time	00:17	00:02	00:02

5 Analysis of results: Ranking algorithms

Running time:

Ranked Boolean Algorithm takes a little bit longer because it needs to get term frequency as score.

Accuracy

Overall, the Ranked Boolean algorithm performs better than the un-ranked Boolean algorithm.

In the OR experiment, un-ranked Boolean works very bad because there are many documents that could match one of the terms in the OR arguments. However, since there is no score provided, the algorithm could not distinguish which document matches better than others. Meanwhile, the rank Boolean works better.

In the AND experiment, the term-frequency score also helps to distinguish those matched documents from each other. Ranked Boolean works better.

In the Structured experiment, Ranked Boolean still works better. And we can see there is a huge improvement on P@10. CF score really help pick out those most suitable candidates from all matched documents. CF did reflect the importance of one term in the document.

6 Analysis of results: Query operators and fields

NEAR: It is very useful to group terms in a phrase by NEAR together. This will largely help the algorithm to match a more accurate document. After I group #NEAR/5(security measures) from #AND(security measures), the MAP for Rank Boolean improves 0.04 (From 0.1250 to 0.1298).

SYN: It is not so easy to use SYN. If we apply them on very different terms, it will harm the discrimination ability of the model, since some term could be missed when the document matches. In my experiments, I try apply SYN on Florida and Seminole, and it worsen the performance. But applying SYN on Embryonic and stem helps with the performance, because Embryonic cells and stem cells could be very similar in documents.

OR will match any term in its arguments. This very helpful when we want to match thing that are very similar but when we are not sure in which form or where it will appear. I used OR on a same term in different field in my experiment. By this way, I am able to match one term wherever it appears, whether it is title or body.

AND needs to match all terms in its arguments. Because of this property, I remove some unnecessary term in the argument to avoid missing some documents that should be matched. For example, I remove “use” in “Increase mass transit” query. Because the same meaning could be expressed in other ways in which “use” does not show. Also, I will use OR on each component within the AND, so that it can be more flexible to match one concept with different terms.

Keywords and title is very useful in some news type document. For example, adding keywords and title for matching “Volcano eruptions” improves the performance.

url field could be very useful when an institute name appears. For example, gov, edu, Wikipedia etc.