Information Retrieval: Assignment 2

Adrian Langseth

September 2019

Task 1: Relevance Feedback

Explain the difference between automatic local analysis and automatic global analysis

Local analysis bases itself on the expansion of the query by creating clusters. These are created from the thesaurus built from the documents which are returned by the initial query. After this expansion is done, the query is rerun with the query expansion.
Global analysis builds its thesaurus from looking at all available documents. This thesaurus is then used in the same manner as in local analysis.

What is the purpose of relevance feedback?

The purpose of relevance feedback is to improve the search and find documents similar to the identified relevant documents, by relying on the indication of relevance by the user which queried.

Explain the terms Query Expansion and Term Re-weighting.

Query expansion is to expand the query to improve the search, by adding similar and related words so as to find related documents which do not necessarily explicitly include the query terms. Term Re-weighting is when we adjust the weights of the query based on our feedback of relevance to create a more refined query with weights more suited to the query.

What separates the two?

Query expansion concerns itself with the expansion of a query to create better searches in terms of relevance of results. Term reweighting does not expand query but rather tunes the query to more suitably fit the relevance.

Task 2: Language Model

Explain the language model, what are the weaknesses and strengths of this model?

The language model creates a model for each document and assesses the likelihood of the query being generated from the model.

The weaknesses of this model is that it does not take into consideration the distance between the terms when there is a co-occurance.

The strengths of the model is that it can take idf into consideration, they are conceptually simple, and bases itself on the distribution of the term.

Given the following documents and queries, build the language model according to the document collection.

d1 = failure is the opportunity to begin again more intelligently.
d2 = intelligence is the ability to adapt to change.
d3 = lack of will power leads to more failure than lack of intelligence or ability.

q1 = failure
q2 = intelligence opportunity
q3 = intelligence failure

m_{d1} :

failure	0.111
is	0.111
the	0.111
opportunity	0.111
to	0.111
begin	0.111
again	0.111
more	0.111
intelligently	0.111

m_{d2} :

Intelligence	0.125
is	0.125
the	0.125
ability	0.125
to	0.25
adapt	0.125
change	0.125

m_{d3} :

lack	0.143
of	0.143
will	0.071
power	0.071
leads	0.071
to	0.071
more	0.071
failure	0.071
than	0.071
intelligence	0.071
or	0.071
ability	0.071

m_C :

failure	0.065
is	0.065
the	0.065
opportunity	0.032
to	0.129
begin	0.032
again	0.032
more	0.065
intelligently	0.032
intelligence	0.065
ability	0.065
adapt	0.032
change	0.032
lack	0.065
of	0.065
will	0.032
power	0.032
leads	0.032
more	0.032
than	0.032
or	0.032

Total no. of terms: 32

q_1:

doc	score	rank
d_1	0.087	1
d_2	0.032	3
d_3	0.068	2

q_2:

doc	score	rank
d_1	$2.224 * 10^{-3}$	1
d_2	$1.465 * 10^{-3}$	2
d_3	$1.046 * 10^{-3}$	3

q_3:

doc	score	rank
d_1	$2.713 * 10^{-3}$	3
d_2	$2.930 * 10^{-3}$	2
d_3	$4.484 * 10^{-3}$	1

Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask.

Smoothing saves our ranking from ties stemming from the absence of the word in our document. If we look at q_2 we see that since neither of our documents have contain both terms, we have will have a three-way tie if we only work with serial probability. However, with smoothing we take care of this problem as we now include the constant of the presence of the word within the collection. This means we now will only have non-zero probabilities per term. This further allows us to rank upon the probability of the presence of some of the terms in the query and its relation to the total presence within the collection. By using smoothing we can rank our query results despite non of them containing all terms within the query.

Task 3: Evaluation of IR Systems

Explain the terms Precision and Recall, including their formulas. Describe how differently these metrics can evaluate the retrieval quality of an IR system.

Precision is the ability to which only relevant documents are returned as results to the query. The formula is given by:

$$Precision = \frac{relevant\ \&\ retrieved}{retrieved}$$

and shows that recall is the portion of relevant documents which are retrieved.

Recall is the ability to which relevant documents are returned as results to the query. The formula is given by:

$$Recall = \frac{relevant\ \&\ retrieved}{relevant}$$

and shows that precision is given by the proportion of retrieved documents are relevant.

To describe how differently these metrics can evaluate the retrieval quality of an IR system, i will illustrate with an example. our example query is ”ventricular tachycardia”. This is quite a niche subject, so the amount of relevant documents might be low. If we have an IR system which returns all documents, it will score high on recall as it returns 100 of our relevant documents. However, its precision will be awful. These metrics are quite disagreeing on our IR systems quality, because they measure vastly different things.

Given the following set of relevant documents, and the set of retrieved documents, provide a table with the calculated precision and recall at each level.

Retrieved = {91, 21, 45, 56, 82, 221, 72, 215}
Relevant = {82, 21, 45, 271, 72, 300, 94, 56, 88, 150}
# Relevant = 10

Rank	Doc	Relevant	Recall	Precision
1	91		0	0
2	21	REL	$\frac{1}{10}$	$\frac{1}{2}$
3	45	REL	$\frac{2}{10}$	$\frac{2}{3}$
4	56	REL	$\frac{3}{10}$	$\frac{3}{4}$
5	82	REL	$\frac{4}{10}$	$\frac{4}{5}$
6	221			
7	72	REL	$\frac{5}{10}$	$\frac{5}{7}$
8	215			

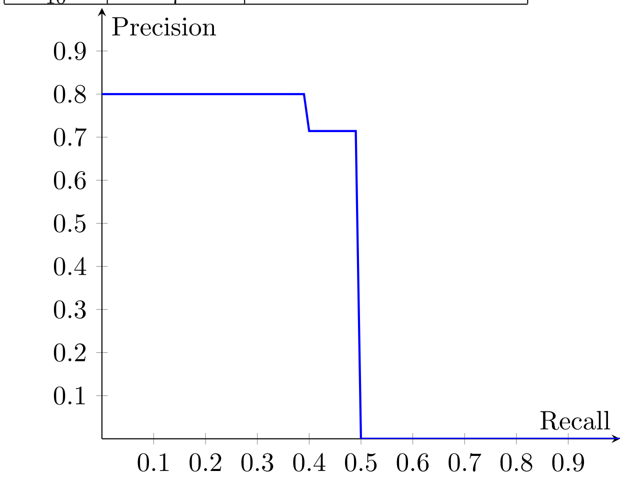
Task 4 - Interpolated Precision

What is interpolated precision?

Interpolate precision at a recall level is the maximum precision obtained for the topic at any actual recall level greater than or equal than itself. This means that if we see that a recall level larger that our current level has a higher precision, we interpolate our precision from higher-level recall.

Given the example in Task 3.2, find the interpolated precision and make a graph.

Recall	Precision	interpolated precision
$\frac{1}{10}$	$\frac{1}{2}$	0.8
$\frac{2}{10}$	$\frac{2}{3}$	0.8
$\frac{3}{10}$	$\frac{3}{4}$	0.8
$\frac{4}{10}$	$\frac{4}{5}$	0.8
$\frac{5}{10}$	$\frac{5}{7}$	0.714


Precision
Recall