CISC689/489-010 Information Retrieval Homework 2

Answer the following questions in your own words. You may discuss them with other students, but you must turn in your own work.

1. (30 points) This problem is about evaluating systems and determining whether a difference is statistically significant. Suppose you have two retrieval engines: one uses a standard “bag-of-words” model, and the other attempts to use additional semantic information about relationships between terms. You want to know whether there is a significant difference between the systems. You pick 10 queries randomly from a query log and run them on both systems. You calculate average precision for each query for each system:

   AP of system 1 = (0.1495, 0.3758, 0.8957, 0.2167, 0.1175, 0.3141, 0.2050, 0.2568, 0.2070, 0.8329)
   AP of system 2 = (0.1199, 0.3806, 0.8903, 0.1904, 0.1287, 0.3527, 0.1447, 0.2070, 0.1978, 0.7818)

   The second system appears to be worse than the first, but it is possible that is due to random chance—perhaps the queries we selected do not capture the ability of the second system.

   (a) Perform a paired t-test to determine whether the difference between these systems is statistically significant. You may use a spreadsheet or statistical software (recommended). Report the mean, variance, t-statistic, and p-value of the test. Be sure to show your work. Is the difference significant?

   (b) Now you sample 10 additional queries and run them on both systems. The 10 new average precisions are exactly the same as the first 10, so now you have 20 average precisions:

   AP of system 1 = (0.1495, 0.3758, 0.8957, 0.2167, 0.1175, 0.3141, 0.2050, 0.2568, 0.2070, 0.8329, 0.1495, 0.3758, 0.8957, 0.2167, 0.1175, 0.3141, 0.2050, 0.2568, 0.2070, 0.8329)
   AP of system 2 = (0.1199, 0.3806, 0.8903, 0.1904, 0.1287, 0.3527, 0.1447, 0.2070, 0.1978, 0.7818, 0.1199, 0.3806, 0.8903, 0.1904, 0.1287, 0.3527, 0.1447, 0.2070, 0.1978, 0.7818)

   Is the difference between systems significant by a paired t-test? Again, report the mean, variance, t-statistic, and p-value.

   (c) Which system do you prefer?

2. (40 points) Calculate $P(Q|D)$ in the inference networks below. Use $P(\text{president}|D) = 0.02$, $P(\text{abraham}|D) = 0.04$, and $P(\text{lincoln}|D) = 0.03$, and the belief combination operators presented in lecture 14.
Now pick a Wikipedia page and draw part of an inference network representation. The inference network should include representation nodes for the important markup in a Wikipedia page as detailed on http://en.wikipedia.org/wiki/Wikipedia:MARKUP such as titles, links, headings, bold text, italic text, etc. Include a few terms under each node. For example, an inference network for “Abraham Lincoln” that models only titles and links and uses some simple 2-word phrases might look like this:

For full credit pick your own page and model more than just three types of markup.

3. (30 points) In homework 1 we had a query “oil producing nations”. Suppose we have run that query and produced a ranked list of documents. Now we are going to apply relevance feedback using two of those documents.

The inverted lists for the original query terms are:

\[
I_k \rightarrow (df_k, ctf_k, (doc_i, tf_{ik}), ...)
\]

- oil \(\rightarrow (5, 18, (1, 4), (4, 3), (6, 1), (7, 2), (8, 8))\)
- producing \(\rightarrow (4, 20, (1, 6), (2, 2), (5, 4), (8, 8))\)
- nations \(\rightarrow (3, 11, (1, 1), (3, 2), (8, 8))\)

The two documents used for feedback are \(d_8\) and \(d_2\), both relevant. They contain the following terms:

- sahara \(\rightarrow (1, 3, (8, 3))\)
- gas \(\rightarrow (3, 6, (2, 2), (5, 2), (8, 2))\)
- iraq \(\rightarrow (4, 17, (6, 5), (8, 2), (9, 5), (10, 5))\)
- tradition \(\rightarrow (8, 25, (1, 3), (2, 2), (3, 1), (4, 5), (7, 2), (8, 4), (9, 6), (10, 2))\)
- nigeria \(\rightarrow (4, 12, (1, 3), (2, 3), (3, 3), (8, 3))\)
- east \(\rightarrow (3, 10, (4, 2), (8, 5), (9, 3))\)
- 2015 \(\rightarrow (4, 4, (2, 1), (5, 1), (7, 1), (9, 1))\)

Use the Rocchio formula to produce an expanded query. (Note that there are no nonrelevant documents.) Report the query and the weights for each of the 10 terms in it. (You may assume that the document term weight \(d_{ik} = \frac{tf_{ik}}{len_i} \log \frac{N+1}{df_k} \)). Also, \(len_8 = 423\) and \(len_2 = 627\), and \(N = 10\).