November 14, 2013

CSCI 2830
Solutions to Application 2, part 1: Building a Search Engine

The Assignment

1. Read and understand all of the above text.
2. Write a search engine script in Matlab.

   Here is a sample script that satisfies the specifications:

```
function indlist = search(a, q, tol)

% This script returns a list of the column indices of
% the columns of the term-by-document matrix a that
% make an angle with q having cosine greater than tol.

[m,n] = size(a);

% normalize q, leave ; off to repeat query

q = q/norm(q)

% normalize a

for j = 1:n
  na(:,j) = a(:,j)/norm(a(:,j));
end

% get cosines

cosvec = na'*q;

% make list of indices

indlist = [];
for j = 1:n
  if (cosvec(j) >= tol) indlist = [indlist j]; end
end

% print indlist

indlist
```

3. Check that you can reproduce all of the results in the above text using your search engine script. The queries you must try are

$$\begin{aligned} q^{(1)} &= (\,1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0\,)^T \\ q^{(2)} &= (\,1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\,)^T \end{aligned}$$

For q^1, the results should be $(\,0.8165, \quad 0, \quad 0, \quad 0.5774, \quad 0\,)$.

For q^2, the results should be $(\,0.5774, \quad 0, \quad 0, \quad 0.4082, \quad 0\,)$.

4. Experiment with the following two queries:

   (a) What is the difference between pastry and bread?

   (b) Show me all recipes for pastry

   Comment on the correctness of the vector space model for these queries. What is the best choice of cosine cutoff for these and the above queries?

   The two new query vectors are

   $q^3 = (\,0, \quad 0, \quad 1, \quad 0, \quad 1, \quad 0\,)$

   $q^4 = (\,0, \quad 0, \quad 1, \quad 0, \quad 0, \quad 0\,)$

   The two sets of cosines are

   For q^3, $(\,0.4082, \quad 0.7071, \quad 0, \quad 0.5774, \quad 0.5000\,)$.

   For q^4, $(\,0.4082, \quad 0.7071, \quad 0.7071, \quad 0.5774, \quad 1.0000\,)$.

   For query 3, all titles about bread and/or pastry rate nonzero cosines. Since D1, D2, and D5 cover only bread OR pastry, they all seem equally reasonable. For that reason, the test with query 3 suggests a cutoff of 0.4. This choice is consistent with the results for queries 1 and 2 where all correct documents have cosines greater than 0.4.

   For query 4, all titles produce nonzero cosines. D1 is not about pastry at all which suggests that a cutoff of 0.5 might be a good choice. Such a cutoff allows D2, D4, and D5 which are all about pastries (although D2 may not actually include recipes.) It, however, also allows D3: Numerical Recipes which is a plainly silly result.

   My choice for cutoff would be 0.4 because it gets all the good docs at the expense of a couple of wrong ones. You can make a different choice as long as your argument for it makes sense!

5. Turn in a listing of your script and the results of all of your tests including your comments from 4. (20 points).

   All of this stuff must be included in your submission for full credit.