CIS 833 – Information Retrieval and Text Mining **Name:_____**

Homework Assignment 3 [20 points] – due September 30th

Exercise 1 (Evaluation) [10 points] Suppose a retrieval system ranks a set of 50 documents and the 6 known relevant documents appear at the following ranks:

1, 2, 5, 7, 10, 22

First plot an exact recall/precision curve and then overlay it with a graph where the precision values are interpolated to the standard 11 points. Then, calculate the following evaluation measures for that ranked list or indicate that there is not sufficient information to calculate a particular measure:

Precision at rank 10
Precision when recall is 50%
Uninterpolated average precision
11-point interpolated average precision
Precision when recall is 25%
Uninterpolated average F1

Solution:

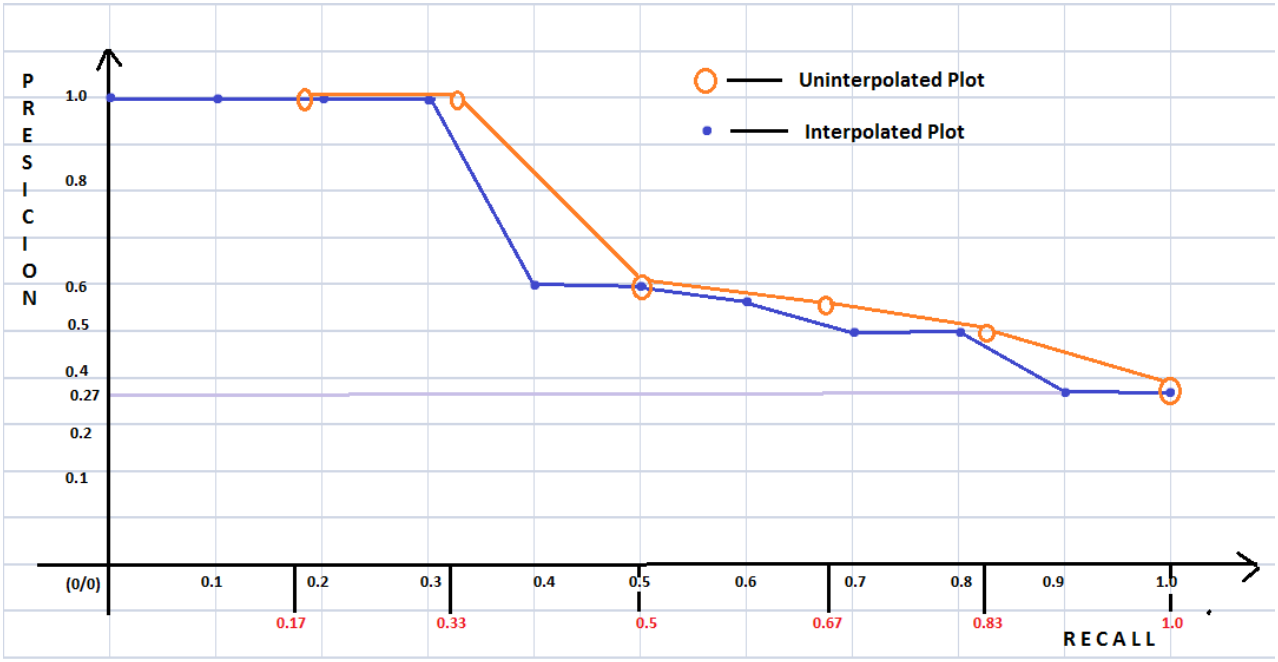
Exact precision/recall values at the given ranks 1,2,5,7,10,22:

Rank 1:	P = 1/1 = 1	R= 1/6 = 0.17
Rank 2:	P= 2/2 = 1	R= 2/6 = 0.33
Rank 5:	P= 3/5 = 0.60	R = 3/6 = 0.50
Rank 7:	P= 4/7 = 0.57	R = 4/6 = 0.67
Rank 10:	P = 5/10 = 0.5	R = 5/6 = 0.83
Rank 22:	P = 6/22 = 0.27	R= 6/6 = 1

Interpolated precision values at the 11 standard recall points:

R = 0.0	P= 1
R = 0.1	P = 1
R = 0.2	P = 1
R = 0.3	P = 1
R = 0.4	P = 0.6
R = 0.5	P = 0.6
R = 0.6	P = 0.57
R = 0.7	P = 0.5
R = 0.8	P = 0.5
R = 0.9	P = 0.27
R = 1	P = 0.27

Exact and interpolated precision/recall graphs:



Precision at rank 10 is 5/10 = 0.5
Precision at recall 50% is 3/5 = 0.6
Uninterpolated average precision = 0.656
11-point interpolated average precision = 0.66
Precision when recall is 25% = 1 (from the interpolated curve)
Uninterpolated average F1 = 0.4998

Exercise 2 (Latent Semantic Indexing) [10 points]

Consider the following document collection, where the index words are underlined.

1. Integer, any number that is a natural number (the counting numbers 1, 2, 3, 4, · · ·), a negative of a natural number (−1, −2, −3, −4, · · ·), or zero. A large proportion of mathematics has been devoted to integers because of their immediate application to real situations.

2. Any integer greater than 1 that is divisible only by itself and 1 is called a prime number (see Number Theory). Every integer has a unique set of prime factors, that is, a list of prime numbers that when multiplied together produce the integer concerned. For example, the prime factors of 42 are 2, 3 and 7.

3. All mesons must have spins equal to integers (0, 1, 2, and so on). Particles with spins equal to integers are called bosons. Bosons differ from particles with noninteger spins, called fermions, in that bosons do not obey a rule of physics called the Pauli exclusion principle.

The corresponding term-document matrix X is:

TERM	d_1	d_2	d_3
integer	2	3	2
natural number	2	0	0
mathematics	1	0	0
prime number	0	2	0
prime factor	0	2	0
Number Theory	0	1	0
meson	0	0	1
Boson	0	0	3
fermion	0	0	1
particle	0	0	2
physics	0	0	1
spin	0	0	2
Pauli exclusion	0	0	1
principle			

The singular value decomposition of the document-term matrix X is show below:

```
X = [T0,S0,D0]=svd(X)

T0 =

    0.6824    0.4111   -0.1760
    0.1073    0.1551   -0.7419
    0.0536    0.0776   -0.3710
    0.1909    0.3736    0.3461
    0.1909    0.3736    0.3461
    0.0954    0.1868    0.1730
    0.1444   -0.1523    0.0234
    0.4332   -0.4568    0.0702
    0.1444   -0.1523    0.0234
    0.2888   -0.3045    0.0468
    0.1444   -0.1523    0.0234
    0.2888   -0.3045    0.0468
    0.1444   -0.1523    0.0234
```

S_0 =

5.5182	0	0
0	3.9498	0
0	0	2.4390

D_0 =

0.2959	0.3063	-0.9048
0.5266	0.7379	0.4221
0.7969	-0.6014	0.0570

1. Construct a rank 2 approximation for matrix X. Show the reduced matrices T, S, D and calculate the approximation X^.

T =

0.6824	0.4111
0.1073	0.1551
0.0536	0.0776
0.1909	0.3736
0.1909	0.3736
0.0954	0.1868
0.1444	-0.1523
0.4332	-0.4568
0.1444	-0.1523
0.2888	-0.3045
0.1444	-0.1523
0.2888	-0.3045
0.1444	-0.1523

S =

5.5182	0
0	3.9498

D =

0.2959	0.3063
0.5266	0.7379
0.7969	-0.6014

D' =

0.2959	0.5266	0.7969
0.3063	0.7379	-0.6014

```
X^=T*S*D' =

    1.6116    3.1811    2.0243
    0.3628    0.7638    0.1034
    0.1814    0.3819    0.0514
    0.7637    1.6436   -0.0480
    0.7637    1.6436   -0.0480
    0.3818    0.8217   -0.0242
    0.0515   -0.0243    0.9968
    0.1547   -0.0725    2.9901
    0.0515   -0.0243    0.9968
    0.1032   -0.0483    1.9933
    0.0515   -0.0243    0.9968
    0.1032   -0.0483    1.9933
    0.0515   -0.0243    0.9968
```

2. Consider the query “"integer prime number." Show the graphical representation of the documents, terms and query in the 2-dimensional reduced vector space.

   The 2-dimensional coordinates for T are obtained from TS.
   The 2-dimensional coordinates for S are obtained from DS.
   The 2-dimensional vector corresponding to the query is

   $d_q = x'_q T S^{-1}$
   where x_q = [1 0 0 1 0 0 0 0 0 0 0 0 0]

   We have

```
TS =

    3.7656    1.6238 t1
    0.5921    0.6126 t2
    0.2958    0.3065 t3
    1.0534    1.4756 t4
    1.0534    1.4756 t5
    0.5264    0.7378 t6
    0.7968   -0.6016 t7
    2.3905   -1.8043 t8
    0.7968   -0.6016 t9
    1.5937   -1.2027 t10
    0.7968   -0.6016 t11
    1.5937   -1.2027 t12
    0.7968   -0.6016 t13
```

```
DS =

    1.6328    1.2098 d1
    2.9059    2.9146 d2
    4.3975   -2.3754 d3


dq =

    0.1583    0.1987 q
```

3. Rank documents in decreasing order of the similarity with the query.

cosine(q,d1) = 0.9662
cosine(q,d2) = 0.9938
cosine(q,d3) = 0.1764

Ranking: d2, d1, d3