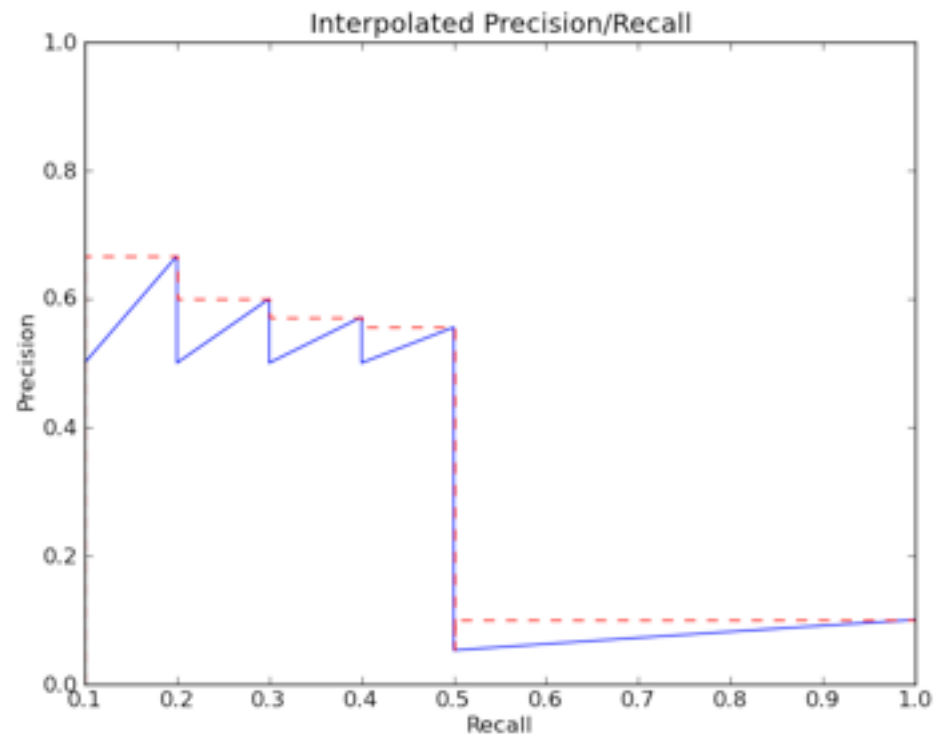


## Information Retrieval, Assignment-2

*Shreesh Ayachit [sa736]*

1.A)



11 point interpolated average precision : 0.44487

## Information Retrieval, Assignment-2

Shreesh Ayachit [sa736]

1.B)

a) Singular values:

*26.885, 23.858, 21.677, 3.876, 2.601, 0.805, 0.664, 0.497*

Since the last 4 values are comparatively less, I have retained the first 3 values, i.e  $k = 3$

b) Comparing  $(X_k - X)$ , I could make the following recommendations:

*User 5, Song 10 - Score 2.31654664*

*User 3, Song 8 - Score 2.23019539*

*User 10, Song 8 - Score 1.98958487*

c) Below are vectors for new users:

*Recommendation for n1: **Song 1, Song 8, , Song 10***

*[[ 2.475 -0.248 -0.168 -0.098 -0.163 2.475 -0.168 1.186 -0.141 1.009]]*

*Recommendation for n2: **Song 9, Song 10, Song4***

*[[ -0.22 2.435 -0.019 0.073 2.426 -0.22 -0.019 -0.125 1.221 1.08 ]]*

*Recommendation for n3: **Song 5, Song 2, Song 1, Song 6***

*[[ 0.434 1.146 -0.001 0.037 1.16 0.434 -0.001 0.213 0.572 0.754]]*

*Recommendation for new user: **Song 3, Song 5, Song 7***

*[[ 1.027 1.095 1.127 0.599 1.127 1.027 1.127 0.877 0.581 1.002]]*

## Information Retrieval, Assignment-2

Shreesh Ayachit [sa736]

1 .C)

For the LSI implementation I have used python to code and Scipy library for matrix and SVD calculations.

### Term Matrix and SVD

I created the Term-Document matrix [TM] for the corpus of 40 files[same as previous assignment] and calculated the tfidf weights and store it as a 2 dimensional Scipy array. I now prompt the user for the value of K and then do a SVD of TM to get the singular values. I reduce the concept space by retaining the k greatest singular values to get the reduced  $\Sigma$  matrix.

$$\text{SVD of TermMatrix } C \quad C = U \Sigma V^T$$

I then get the reduced concept space of TM by the above expression and store it as CM[Concept matrix] .

### Query vector and cosine similarity:

I tokenize the user input into a set of lower case, non-stop words delimited by space. Any query containing only stop words/ words not present in the corpus will not return any results.

The query token set is converted to a vector representation  $\vec{q}$  and cosine similarity is used to calculate the latent semantic relation to each of the document vectors  $\vec{d}$  in concept space. using the below expression (referred from Prof. Ginsparg's [note on LSI](#))

$$\text{Similarity}(\vec{q}, \vec{d}_{(j)}) = \cos(\theta) = \frac{\vec{q} \cdot \vec{d}_{(j)}}{|\vec{q}| |\vec{d}_{(j)}|} = \frac{\vec{q} \cdot C \cdot \vec{e}_{(j)}}{|\vec{q}| |C \vec{e}_{(j)}|} .$$

I then list the 4 top results based on greatest values for the similarity vector for each document

## Information Retrieval, Assignment-2

Shreesh Ayachit [sa736]

### Test Runs

**Note:** When I retrieve a document with the query term not present, I list the text snippet corresponding to the first word in the term list which matches

*I chose  $K=4$  arbitrarily.*

#### 1. Invalid words :

Words which don't appear in the corpus or which are stop words; are removed from the query before processing, hence these are gracefully handled.

Enter Query:and werwer  
No results Found

#### 2. Multiple valid words

Query vector is constructed with all the terms and similarity is calculated normally

Enter Query:planet earth  
Results:

Document: 1  
Score: 0.0693688190482  
Text: limited performance test and testing of the science instruments Following these tests and a spacecraft launch simulation the satellite will be fueled with its attitude control propellant NPP will be I

#### 3. Multiple words with one invalid word:

I ignore the invalid word and retrieve result for only valid words.

Enter Query:life and earth  
Results:

Document: 34  
Score: 0.058656591367  
Text: magnetic fields with atomic magnetometers Sensitive magnetometers could be used to measure electrical activity in the human brain and heart explains Professor Eugene Polzik

### Varying K Value

K value controls the number of dimensions we consider for retrieval, and the criteria we set for "noise". Reducing the dimension uncovers hidden or "latent" relations between documents by mapping related concepts.

For example when I searched for house with  $K = 4$ , I did not get document 8 as the top result which contains the word "house". This is because "house" was incorrectly mapped into reduced concept space. But with  $K=8$  and thereby with increased number of concepts, house as a concept was properly represented and now I got more relevant results.

## Information Retrieval, Assignment-2

*Shreesh Ayachit [sa736]*

### HOWTO RUN THE PROGRAM

★ Please Install Scipy package from <http://www.scipy.org/Download> for the appropriate OS. *(I am not aware of the OS environment it will run on, otherwise I would have included easy\_install scripts for easier installation)*

★ If the current directory does not have corpus files in folder “./test” it will retrieve the files from the course link

★ I am reading the stop list from the file “stopList.txt”, please ensure its present in the current directory.

>python LSI.py