

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

Exercise 1 (Evaluation) 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

Exercise 2 (Latent Semantic Indexing)

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:

$$X = [T_0, S_0, D_0] = \text{svd}(X)$$

$$T_0 =$$

| | | |
|--------|---------|---------|
| 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 \bar{X} .
2. Consider the query "integer prime number." Show the graphical representation of the documents, terms and query in the 2-dimensional reduced vector space.
3. Rank documents in decreasing order of the similarity with the query.