

Test 1

60 minutes. One page of handwritten notes is allowed.

1. [7 points] In many cases the relevance of a document can be estimated from two factors: its similarity to the query and the PageRank of the document itself. The relative importance of these two factors can, however, vary. For each of the following scenarios, indicate whether the contribution of PageRank should be high, medium, or low.
 - a. Ms. E searches on “Wells Fargo” expecting the top result to be the homepage of her bank.
 - b. Dr. W searches on “sprained knee” wanting to spend a few minutes learning about the common outcomes of this injury.
 - c. Mr. V searches on “UTEP Technical Report 2006 Medina Wavesurfer”, knowing that such a document exists.
 - d. Mr. H searches on “document camera”, wanting to learn about models, reputations, features, prices, and maybe find a store.
 - e. Mr. L searches on “Rebelde news” hoping to find anything new about the band.
 - f. Ms. B searches on “amusing” hoping to find a fun essay or jokelist to read.
 - g. Mr. G searches on “search engine test questions” hoping to find questions to use for studying.

2. [4] Pick two of the above scenarios where these factors may be insufficient, and explain what other factors should be considered.

6. [4] Various search engines allow you to can specify Boolean queries with AND and OR. For such queries is it still useful to consider PageRank (or another metric based on link analysis)? Why or why not?

7. [14] Define or explain:

a. precision

b. recall

c. metasearch

d. deep web

e. navigation

f. rank merging

g. hypertext

8. [10] Give pseudocode for using the vector space model to find the 10 documents most relevant to a query. Hint: you'll need at least one set of nested foreach loops.

9. [6] Your pseudocode undoubtedly referred to some data structure relating terms and documents. Considering this as an abstract data type, specify the operations it must support.

10. [8] Suppose UTEP Athletics reaches an arrangement with Search Engine X to advertise promotional tickets to Miners games. The agreement is to place the advertisement on all search results pages served to computers with an IP address identifying them as being from the El Paso area or served to users who are from El Paso with greater than .5 probability.

Suppose now that someone uses X from an unknown IP address. Company X knows from his search history that he has searched on “Don Haskins Center” within the past month, and it also knows that 20% of the people in El Paso have searched on “Don Haskins Center” within the past month, and that only one in 44,000 people in the world as a whole has done so.

Should X serve him the ad? Give quantitative justification, being explicit about any assumptions or estimates you use.