CS W186 Introduction to Database Systems
Fall 2019 Josh Hug

DIS 11

1 Text Search

(a) State if the following layers belong to Search Engine Architecture, DBMS Architecture, or both.

i Index Both

ii SQL DBMS

iii Buffer Both

iv Ranking Algorithm Search Engine

v Relational Operators DBMS

(b) What is an inverted file? Why are inverted files useful?

Alternative 3 B+ Tree with term as keys and the list doc_id where the term appears as the value. It is easy to find all of the documents related to a term.

(c) What is the query plan for boolean search (how do we scan and how do we join)?

Index scans followed by parallel merge joins

(d) Suppose we have 20 documents, and there are 300 distinct terms in all the documents. We want to find the document that is the most similar to the document "Database is fun. It is my favorite subject." Note, the multiple choice questions in this part could have more than one answer.

(i) If we use the vector space model, what would be the dimension of the vectors?

300

(ii) How could the dimension change if we duplicate an existing document?

(A) Increase

(B) Stay the same

(C) Decrease

Still 300 distinct terms

(iii) How could the dimension change if we add more terms to the longest document?

(A) Increase
(B) Stay the same
(C) Decrease

Might have more distinct terms

(iv) How could the dimension change if we remove a document?

(A) Increase
(B) Stay the same
(C) Decrease

Might have fewer distinct terms

(v) How could the dimension change if we add a term that is not in any of the documents?

(A) Increase
(B) Stay the same
(C) Decrease

More distinct terms

(e) Given 100 documents, consider a term t1 that appears 2 times in document d1. What is the DocTermRank of t1 in d1 if (assume base 10 logarithm):

(i) T1 appears in only d1? 4
(ii) T1 appears in 10 documents? 2
(iii) T1 appears in all documents? 0
(iv) What property of TF-IDF does this pattern show? Favors unusual words

(f) Given 100 documents, consider a term t2 that appears in 10 documents. What is the DocTermRank of t2 in d2 if (assume base 10 logarithm):

(i) T2 appears 2 times in d2? 2
(ii) T2 appears 4 times in d2? 4
(iii) T2 appears 6 times in d2? 6
(iv) What property of TF-IDF does this pattern show? Favors repeated words