technische universität
dortmund

Architecture & Implementation of DBMS
Summer semester 2019
Prof. Dr. Jens Teubner · DBIS Group

Exercise 12

Released: July 1, 2019 · Discussion: July 8, 2019

1 Search

1. Why are relational databases not suited for searching in unstructured text?
2. What is the goal of ranking? How can rankings be calculated?
3. What are Recall and Precision? How do they relate to each other?

2 tf/idf Ranking

tf/idf ranking is one way of ranking documents for a search query. In this assignment we will deepen our understanding of it.
Given the following four documents:

The European Union is a politico-economic union of 28 member states that are located primarily in Europe. It covers an area of $4,324,782km^2$, with an estimated population of over 508 million.	A union is a special class type that can hold only one of its non-static data members at a time.	Founded in 1815, the Cambridge Union is the oldest continuously running debating society in Europe and the world, and the largest and most famous society at the University of Cambridge.	The ethnic groups in Europe are the focus of European ethnology, the field of anthropology related to the various ethnic groups that reside in the nations of Europe.
doc_1	doc_2	doc_3	doc_4

We would like to search for the term *'european union'* in the four documents. For that we assume that search terms are normalized on word stems. Numbers and units are not counted as words. No stop words are eliminated by this ranking.

1. Calculate the *term frequency (tf)* for the search term *'european union'* and the four documents.

2. Calculate the *inverse document frequency (idf)* for the search term and the four documents.

3. Determine a ranking using the **vector space model** and the values from the previous assignments.

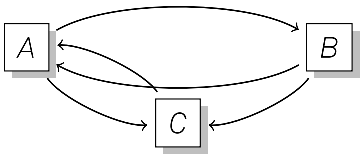
4. Sketch how an inverted file including the four documents would look like. It is sufficient to sketch the construction scheme and a small snipet of the file.

3 Page Rank

PageRank is a ranking method, popularized by Google, taking the reputation of a webpage into account by looking at its incoming and outgoing links.

1. Briefly describe the idea of PageRank.

2. Given the following graph showing the connection of three web pages.



3. Calculate the rank of each webpage for the first three iterations of the PageRank algorithm. Assume that λ equals 0.15.