Intelligent Data Management - Exercise 1

Christoph Prinz

October 19, 2017

Assignment 1

Exercise a

Term i appears in n_i of the N documents

Appears in 40 documents: $IDF_i = log_2(\frac{N}{n_i}) = log_2(\frac{10000000}{40}) = 18$

Appears in 10000 documents: $IDF_i = log_2(\frac{N}{n_i}) = log_2(\frac{10000000}{10000}) = 10$

Exercise b

Given the occurrence of a term i in document j is f_{ji} and $max_k f_{kj}$ the maximum number of occurrences of any term in this document, the term frequency TF is defined as:

$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$

Word w appears in 320 documents. In document d the maximum occurence of any word is 15.

a) w appears once:

$TF_{wd} = \frac{1}{15}$

$IDF_w = *log_2 \frac{10000000}{320} = 15$

$TF.IDF = \frac{1}{15} * 15 = 1$

b) w appears five times:

$TF_{wd} = \frac{5}{15} = \frac{1}{3}$

$IDF_w = *log_2 \frac{10000000}{320} = 15$

$TF.IDF = \frac{1}{3} * 15 = 3$

Exercise c