Solutions to Exercises

Chapter 1 – Information Retrieval Models

Djoerd Hiemstra

1.1(c) The Venn diagrams of Figure 1.2 show exactly 8 disjoint subsets of documents, including the area around the diagram. Whatever the final result of a Boolean query, each subset is either selected or not, so in total $2^8 = 256$ subsets can be defined.

1.2(b) Vector spaces are metric spaces, i.e., a set of objects equipped with a distance, where the distance from $\vec{q}$ to $\vec{d}$ is the same as from $\vec{d}$ to $\vec{q}$, so if we change $\vec{q}$ for $\vec{d}$ and $\vec{d}$ for $\vec{q}$, the distance should remain the same. In practice however, many practical term weighting algorithms do not use the same weights for queries and documents. In such a case, the similarity might not be equal to 0.08. One might argue that such a model is *not* a vector space model, though.

1.3(c) If we add a single document, the document frequencies of terms that occur in the added document will increase by 1. Furthermore, the number of documents N changes, which affects the *idf*s of all other terms.

1.4(b) Assuming that the *idf* is calculated using the number of documents N as shown in Equation (1.5), all weights need to change, as is the case in Exercise 3.

1.5(a) For each term, the probabilistic model considers either presence of the term in the document, or absence of the term in the document. So, per term, there are two cases, hence $2^3 = 8$ different scores in the case of three query terms.

1.6(c) As said above, the model only considers presence or absence, but in the case of presence it does not consider the number of occurrences of the term in the document. If the term is present in both D and E, then they will be assigned the exact same score. If the system needs to provide a total ranking, then the implementation has to determine which document is ranked first.

1.7(b) Without smoothing, the probability is a simple fraction of the number of occurrences, divided by the total number of terms in the document. Of course, language models do not need an additional term weighting algorithm.

1.8(c) If $\lambda = 0$, the model does not use smoothing. Without smoothing, terms that do not occur in the document are assigned zero probability. The score of a document is determined by multiplying the probabilities of the single terms. If one of them is zero, the final score of the document is zero.

Information Retrieval: Searching in the 21st Century Edited by A. Göker & J. Davies
 ISBN: 978-0-470-02762-2