Task Sheet 5
Vector Space Model

1. Which problem occurs when directly using the *term frequency* tf_{dk} for similarity calculation in the vector space model? How can the problem be reduced?
2. What is the meaning of the term *discriminatory power*? How is it considered in the *tf·idf formula*?
3. What is *Relevance Feedback* and how is it performed in the vector space model?
4. You are given a query vector q, the documents $D_1 - D_6$ (see Figure 2) and the mapping to relevant (F^+) and irrelevant (F^-) documents by the user (see Tables 1 and 2). Determine the modified query vector by sketching relevance feedback graphically, using the Ide(dec hi) technique! Also provide the formula for the technique and draw the resulting modified query vector q_{new}!

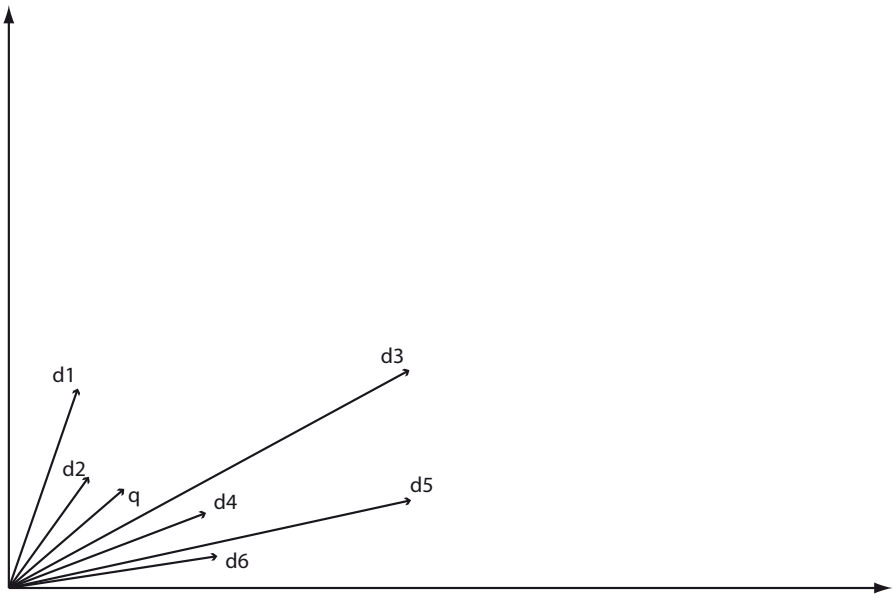

Figure 1: Query and document vectors

5. What is *Pseudo-Relevance Feedback*?
6. You are given an IR system that works internally with inverted lists and uses the vector space model for the search. Furthermore the following query vector and the necessary inverted lists are given.

Rank	Document
1.	D_1
2.	D_3
3.	D_5

Table 1: F^+

Rank	Document
1.	D_2
2.	D_4
3.	D_6

Table 2: F^-

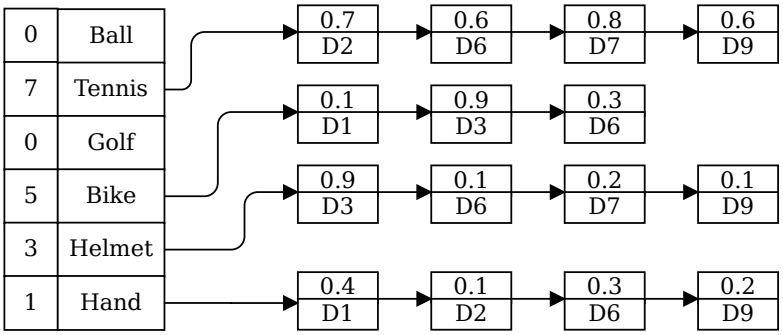

Figure 2: Query vector and inverted lists

Using the base algorithm of Buckley & Lewit presented in the lecture, determine the three most relevant documents for the query vector! Provide the content of the auxiliary structures after each step!