9 Information Retrieval (SHT)

The documents in the figure below are to be clustered according to their similarity in standard frequency-based vector space. The proximity metric to be used is the Manhattan distance $m(\vec{d_k}, \vec{d_j}) = \sum_i |d_{k,i} - d_{j,i}|$, where $\vec{d_k}$ and $\vec{d_j}$ represent the vectors assigned to documents k and j, and $d_{k,i}$ gives the frequency of term i in document k.

Doc 1: whale, sea, sea, whale, boat, boat, boat, boat, boat
Doc 2: whales, sea, sea, water
Doc 3: whale, water, water, whale, whale
Doc 4: whales, whales, whales

(*a*) Construct the term–document matrix under the assumption that the terms are not stemmed. [3 marks]

(*b*) Construct the corresponding document–document matrix. [3 marks]

(*c*) On the basis of the document–document matrix, perform complete-link clustering, showing the output as well as intermediate results. [6 marks]

(*d*) Starting from the situation in part (*c*), you now want to create a clustering which is guaranteed to be different from the one in (*c*). You are allowed to manipulate one of the following factors:

- the term weighting
- the proximity metric
- whether stemming is applied
- adding new terms to documents
- the similarity function (single-link instead of complete-link)

Which of the factors do you choose, and why? Demonstrate the changes affected. [8 marks]