8 Information Retrieval (RC)

Consider the following documents:

doc_1	phone ring person happy person
doc_2	dog pet happy run jump
doc_3	cat purr pet person happy
doc_4	life smile run happy
doc_5	life laugh walk run run

(*a*) (*i*) Construct the inverted index required for ranked retrieval for these five documents. Assume that no stemming or stop-word removal is required. [3 marks]

(*ii*) What is the complexity of processing a two-term conjunctive query using standard postings lists? Briefly describe one technique that can improve this efficiency. [2 marks]

(*iii*) Relating to the sample documents above, outline how the processing of the following Boolean query can be optimised:

happy AND run AND pet [2 marks]

(*iv*) What is the query-likelihood method in the language modelling approach to information retrieval? How does this differ conceptually from the measure of similarity used in the vector space model? [3 marks]

(*b*) (*i*) Smoothing is crucial in the language modelling approach to information retrieval. Why is smoothing important and how is it typically achieved? [2 marks]

(*ii*) Given the query {*happy person smile*}, show how a unigram language modelling approach would rank the documents outlined above. Choose a suitable form of smoothing and include all your workings. State any other assumptions made. [6 marks]

(*iii*) How might you relax the *term-independence* assumption in the unigram language model and how might it affect subsequent retrieval? [2 marks]