COMPUTER SCIENCE TRIPOS Part II – 2015 – Paper 8

7 Information Retrieval (SHT)

(*a*) Consider a standard bag-of-words model for the document retrieval problem.

(*i*) Give an expression for the tf-idf weighting scheme which assigns a weight to each term in a *document*. Motivate each part of your expression, using the notion of Zipf's law when appropriate. [3 marks]

(*ii*) How might you modify your tf-idf scheme for each term in a *query*? Why might you use different schemes for documents and queries? [3 marks]

(*b*) Edit distance can be used for spelling correction in search queries.

(*i*) Define edit distance. [1 mark]

(*ii*) As an example of how to calculate edit distance efficiently, show how dynamic programming can be used to calculate the edit distance between *able* and *belt*. [5 marks]

(*c*) The PageRank algorithm uses a model of a "random surfer" to calculate the importance or validity of a page. Describe how the random surfer can be modelled as an ergodic Markov chain, and how this leads to the PageRank values being calculated as a principal left eigenvector of the transition probability matrix. (You are not required to give a formal definition of an ergodic Markov chain; an informal description will suffice.) [8 marks]