8 Information Retrieval (RC)

(*a*) (*i*) Given the query **"indiana jones film"** and the following term-frequencies for the two documents doc_1 and doc_2:

	indiana	**jones**	**archaeologist**	**grail**	**film**	**crusade**
doc_1	5	4	3	3	0	5
doc_2	2	2	0	2	1	3

calculate the unsmoothed query-likelihoods for both documents. [2 marks]

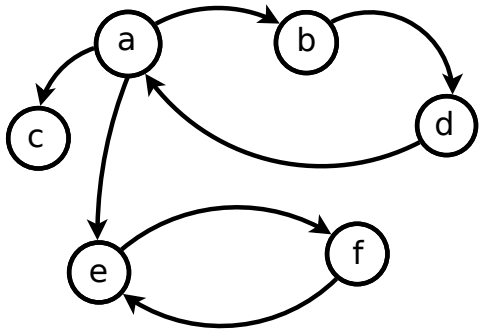
(*ii*) Describe two ways in which smoothing affects the retrieval of these documents. [2 marks]

(*iii*) Is smoothing more important for long or short queries? Justify your answer. [2 marks]

(*b*) (*i*) PageRank calculates a measure of *importance* for webpages. Give one high-level interpretation of this measure. [3 marks]

(*ii*) PageRank can be modelled as a Markov chain. What practical considerations must be addressed to ensure that the Markov chain has a stationary distribution? [3 marks]

(*iii*) Give the Markov transition matrix for the following graph assuming a teleportation probability of $\alpha = 0.5$. Discuss the suitability of this level (i.e. $\alpha = 0.5$) of teleportation for this graph. [5 marks]



(*iv*) Given the transition matrix from part (*b*)(*iii*), describe in detail how you would calculate the PageRank of each page. [3 marks]