CT102 Information Systems
Assignment 1

Date: Monday 15th November 2021
Due: Monday 29th November 2021
Total Marks: 15
Instructions:
Please submit via Blackboard in a SINGLE file in PDF or WORD format ONLY. Use MS Lens or a similar app to scan any hand-written pages; please include plagiarism declaration.

1. *(modified version of question from Summer 2021)*

In the context of web search engines, and given the following paragraph (taken from Wikipedia) with 52 words (see file on Blackboard for text):

> *The ethics of Artificial Intelligence is the branch of the ethics of technology specific to artificially intelligent systems. It is sometimes divided into a concern with the moral behaviour of humans as they design, make, use and treat, artificially intelligent systems, and a concern with the behaviour of machines, in machine ethics.*

(i) Show the resulting paragraph when the standard pre-processing steps are carried out, using a stemmer rather than a lemmatizer. Clearly state the steps taken and state what stop word list is used and what stemmer is used.

(ii) Using the pre-processed paragraph from part (i) show how the tf-idf (term frequency - inverse document frequency) weighting scheme is used to calculate a weight for the term 'ethic' given that there are 220 documents in the document collection and that the term occurs in 50 of them. You may assume that the word 'ethics' is stemmed to the term 'ethic'.

(6 marks)

2. *(modified version of question from Summer 2018)*

Given the following two vectors representing some of the text content of two web pages:

< 0.30, 0.25, 0.1, 0.02, 0.00, 0.11 >
< 0.35, 0.00, 0.3, 0.11, 0.02, 0.20 >

Calculate the similarity between the two given vectors using the cosine similarity (Euclidean dot product) to an accuracy of *at least* three decimal places (i.e. 3 digits after the decimal point). Clearly show your workings - you may take a picture of hand notes and include in your document **if legible**.

(3 marks)

3. *(modified version of question from 2021)*

Given the following six web pages (A, B, C, D, E and F), and the hyperlinks between them:

(A, B), (A, C),
(B, A), (B, D), (B, F),
(C, A), (C, B), (C, E),
(D, E), (D, F),
(E, A), (E, B), (E, C), (E, F),
(F, B), (F, D).

(i) Write the *PageRank* formula which can be used to calculate the *PageRank* scores for each web page, including all necessary information (This can be C code but ensure that the formula used for each web page is clear).

(ii) Calculate the page rank score of each of the web pages, showing the final scores of each page. Also include in your answer the number of iterations taken to find the scores.

(6 marks)

4.

******** Please include the following plagiarism declaration form in your solution: ********

Plagiarism Declaration:
"I am aware of what plagiarism is and include this here to confirm that this work is my own"

Please note that any suspected cases of plagiarism will not receive a mark until assurances can be given in person as to the origins of the solution.