CT102 Information Systems
Assignment 1

Date: Monday 16th October 2023
Due: Tuesday 31st October 2023
Total Marks: 15 (worth 5.5% of final mark)
Instructions:
Please submit via Canvas 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 Exam 2023)*

In the context of web search engines, and given the following paragraph (taken from a siliconrepublic.com article on 23rd January 2023) with 26 words:

> *ChatGPT is a powerful language model developed by OpenAI that has the ability to generate human-like text, making it capable of engaging in natural language conversations.*

(i) Show the resulting paragraph after the pre-processing techniques of case folding, punctuation and stop word removal, and stemming have been applied. Clearly explain each technique in your own words. 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 'language', given that there are 500 documents in the document collection and the term 'language' occurs in 50 of them.

(6 marks)

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

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

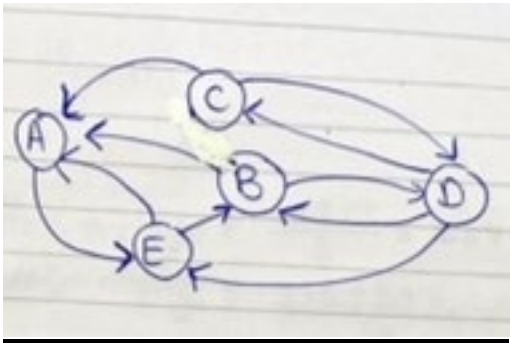
< 0.30, 0.25, 0.1, 0.02, 0.00, 0.11 >
< 0.35, 0.00, 0.3, 0.11, 0.02, 0.20 >

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 Summer Exam 2022)*

Given the following network between five web pages (A, B, C, D and E)



(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: *****

Declaration:

"I am aware of what plagiarism is and include this to confirm that this work is my own and, further, I confirm that this work was not, wholly or in part, produced by generative AI tools"

Please note that any suspected cases of plagiarism, or, of the use of generative AI tools will **not** receive a mark until assurances can be given in person as to the origins of the solution. Submissions will not be corrected if this declaration is absent.