1. (modified version of question from Summer Exam 2023) (6 marks)

> *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)

chatgpt power languag model develop openai abil gener human text make capabl engag natur languag convers

- Case folding is a process that converts each character to the same case, either lower or upper.
- Punctuation and stop word removal is a process that removes all punctuation from the text as well as stop words that are contained in a list used by the algorithm.
- stemming is a process that removes all prefixes, suffixes, and converts a word to the most basic form.

ii)

Frequency: 2

Number of terms in doc: 16

$$tf = \frac{Frequncy}{Number\ of\ terms\ in\ doc} = \frac{2}{16} = 0.125$$

N = 500

Df = 50

$$idf = \log_{10}(1 + \frac{N}{df_t}) = \log_{10}\left(1 + \frac{500}{50}\right) = 1.0414$$

$$tf * idf = 0.125 * 1.0414 = 0.13$$

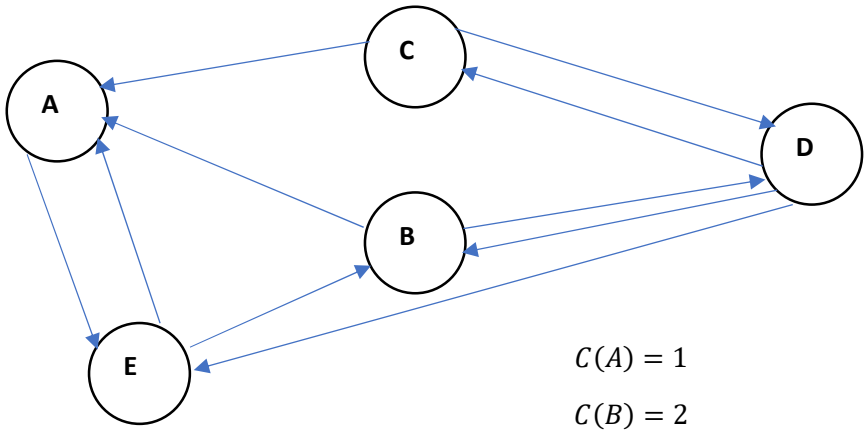
2. (modified version of question from Summer Exam 2018) (3 marks)

d=< 0.30, 0.25, 0.1, 0.02, 0.00, 0.11 >

q=< 0.35, 0.00, 0.3, 0.11, 0.02, 0.20 >

$$sim(d,q) = \frac{0{,}30 * 0.35 + 0.25 * 0.00 + 0.1 * 0.3 + 0.02 * 0.11 + 0.00 * 0.02 + 0.11 * 0.20}{\sqrt{0.30^2 + 0.25^2 + 0.1^2 + 0.02^2 + 0.00^2 + 0.11^2} + \sqrt{0.35^2 + 0.00^2 + 0.3^2 + 0.11^2 + 0.02^2 + 0.20^2}}$$
$$= \frac{0.1592}{0.4183 * 0.5148} \approx 0.739293$$

3.(modified version of question from Summer Exam 2022) (6 marks)



$$C(A) = 1$$
$$C(B) = 2$$
$$C(C) = 2$$
$$C(D) = 3$$
$$C(E) = 2$$

$$PR(A) = \frac{0.15}{5} + 0.85 * \left(\frac{PR(B)}{C(B)} + \frac{PR(C)}{C(C)} + \frac{PR(E)}{C(E)}\right)$$

$$PR(B) = \frac{0.15}{5} + 0.85 * \left(\frac{PR(D)}{C(D)} + \frac{PR(E)}{C(E)}\right)$$

$$PR(C) = \frac{0.15}{5} + 0.85 * \left(\frac{PR(D)}{C(D)}\right)$$

$$PR(D) = \frac{0.15}{5} + 0.85 * \left(\frac{PR(B)}{C(B)} + \frac{PR(C)}{C(C)}\right)$$

$$PR(E) = \frac{0.15}{5} + 0.85 * \left(\frac{PR(A)}{C(A)} + \frac{PR(D)}{C(D)}\right)$$

Calculated using Microsoft Excel (excel file will be attached

After 11 iterations page rank looks like this:

$$PR(A) = 0.276$$

$$PR(B) = 0.202$$

$$PR(C) = 0.071$$

$$PR(D) = \ 0.146$$

$$PR(D) = \ 0.305$$

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”