Question 1

Java Implementation:

A focused search engine that will look for documents inside the domain is implemented in *Crawl.java*

Question 2

Document/w	w_1	w_2	w_3	w_4	w_5
D_1	1	2	0	1	1
D_2	3	2	0	0	1
D_3	1	1	4	0	1
q	1	1	1	1	0

Compute the cosine similarity

$$tf - idf(w_1, D_1) = 1 \times \frac{3}{3} = 1$$
$$tf - idf(w_2, D_1) = 2 \times \frac{3}{3} = 2$$
$$tf - idf(w_3, D_1) = 0 \times \frac{3}{1} = 0$$
$$tf - idf(w_4, D_1) = 1 \times \frac{3}{1} = 3$$
$$tf - idf(w_5, D_1) = 1 \times \frac{3}{3} = 1$$

$\therefore\ V_{D1} = < 1, 2, 0, 3, 1 >$

$$tf - idf(w_1, D_2) = 3 \times \frac{3}{3} = 3$$
$$tf - idf(w_2, D_2) = 2 \times \frac{3}{3} = 2$$
$$tf - idf(w_3, D_2) = 0 \times \frac{3}{1} = 0$$
$$tf - idf(w_4, D_2) = 0 \times \frac{3}{1} = 0$$
$$tf - idf(w_5, D_2) = 1 \times \frac{3}{3} = 1$$

$\therefore\ V_{D2} = < 3, 2, 0, 0, 1 >$

$$tf - idf(w_1,)D_3 = 1 \times \frac{3}{3} = 1$$
$$tf - idf(w_2, D_3) = 1 \times \frac{3}{3} = 1$$
$$tf - idf(w_3, D_3) = 4 \times \frac{3}{1} = 12$$
$$tf - idf(w_4, D_3) = 0 \times \frac{3}{1} = 0$$
$$tf - idf(w_5, D_3) = 1 \times \frac{3}{3} = 1$$

$$\therefore\ V_{D3} = < 1, 1, 12, 0, 1 >$$

$$tf - idf(w_1, q) = 1 \times \frac{3}{3} = 1$$
$$tf - idf(w_2, q) = 1 \times \frac{3}{3} = 1$$
$$tf - idf(w_3, q) = 1 \times \frac{3}{1} = 3$$
$$tf - idf(w_4, q) = 1 \times \frac{3}{1} = 3$$
$$tf - idf(w_5, q) = 0 \times \frac{3}{3} = 0$$

$$\therefore\ V_q = < 1, 1, 3, 3, 0 >$$

$$\cos(\theta_{q,D_1}) = \frac{1+2+0+9+0}{\sqrt{1+4+0+9+1} \times \sqrt{1+1+9+9+0)}} = \frac{12}{\sqrt{15}\sqrt{20}} \approx 0.693$$

$$\cos(\theta_{q,D_2}) = \frac{3+2+0+0+0}{\sqrt{9+4+0+0+1} \times \sqrt{1+1+9+9+0)}} = \frac{5}{\sqrt{14}\sqrt{20}} \approx 0.299$$

$$\cos(\theta_{q,D_3}) = \frac{1+1+36+0+0}{\sqrt{1+1+144+0+1} \times \sqrt{1+1+9+9+0)}} = \frac{38}{\sqrt{147}\sqrt{20}} \approx 0.701$$

The decreasing order of cosine similarity with respect to the query: D3, D1, D2

1. D_3 with cosine similarity ≈ 0.701
2. D_1 with cosine similarity ≈ 0.693
3. D_2 with cosine similarity ≈ 0.299

Question 3

$r(i) = \sum_{j \in B(i)} \frac{r(j)}{N(j)}$

$r(1) = \frac{1}{2} r(3)$

$r(2) = \frac{1}{2} r(1)$

$r(3) = \frac{1}{2} r(5)$

$r(4) = r(2) + r(6)$

$r(5) = \frac{1}{2} r(1) + \frac{1}{2} r(3) + \frac{1}{2} r(4)$

$r(6) = \frac{1}{2} r(4) + \frac{1}{2} r(5)$

$$\mathbf{A} = \begin{array}{cccccc|c} 1 & 2 & 3 & 4 & 5 & 6 & \\ \hline & \frac{1}{2} & & & \frac{1}{2} & & 1 \\ & & & 1 & & & 2 \\ \frac{1}{2} & & & & \frac{1}{2} & & 3 \\ & & & & \frac{1}{2} & \frac{1}{2} & 4 \\ & & \frac{1}{2} & & & \frac{1}{2} & 5 \\ & & & 1 & & & 6 \end{array}$$

$\mathbf{A} =$ Adjacency matrix of the Web Graph $(r = A^T r)$
r is the eigenvector of A^T with eigenvalue 1
The Web graph is strongly connected. Under this assumption r exists and it is unique, if r is normalized, we have

$$\sum_{i=1}^{n} r(i) = 1$$

$\boldsymbol{r(1) = \frac{1}{2} r(3)}$

$r(2) = \frac{1}{2} r(1) = \frac{1}{4} r(3) \Rightarrow \boldsymbol{r(2) = \frac{1}{4} r(3)}$

$r(3) = \frac{1}{2} r(5) \Rightarrow \boldsymbol{r(5) = 2r(3)}$

$r(5) = \frac{1}{2} r(1) + \frac{1}{2} r(3) + \frac{1}{2} r(4) \Rightarrow 2r(3) = \frac{1}{4} r(3) + \frac{1}{2} r(3) + \frac{1}{2} r(4) \Rightarrow \boldsymbol{r(4) = \frac{5}{2} r(3)}$

$r(6) = \frac{1}{2} r(4) + \frac{1}{2} r(5) \Rightarrow r(6) = \frac{1}{2} . \frac{5}{2} r(3) + \frac{2}{2} r(3) \quad \therefore \boldsymbol{r(6) = \frac{9}{4} r(3)}$

$\because r(1) + r(2) + r(3) + r(4) + r(5) + r(6) = 1$

$\therefore \frac{1}{2} r(3) + \frac{1}{4} r(3) + r(3) + \frac{5}{2} r(3) + 2r(3) + \frac{9}{4} r(3) = 1 \Rightarrow r(3) = \frac{4}{34} \Rightarrow \boldsymbol{r(3) = \frac{2}{17}}$

$r(1) = \frac{1}{2} r(3) = \frac{1}{17} \qquad r(2) = \frac{1}{4} r(3) = \frac{1}{34} \qquad r(4) = \frac{5}{2} r(3) = \frac{5}{17}$

$r(5) = 2r(3) = \frac{4}{17} \qquad r(6) = \frac{9}{4} r(3) = \frac{9}{34}$

$\boldsymbol{Therefore,\ r(1) = \frac{1}{17}\ ,\ r(2) = \frac{1}{34}\ ,r(3) = \frac{2}{17}\ ,r(4) = \frac{5}{17}\ ,r(5) = \frac{4}{17}\ ,r(6) = \frac{9}{34}}$

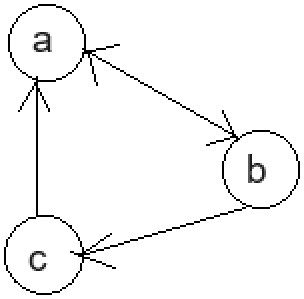
Question 4

a.

(i) Yes. The page rank always exists because it uniformly distributes the rank among all pages. All pages have the same rank $\frac{1}{n}$

(ii) This method does not differentiate between high-quality and low-quality pages. It assigns equal ranks to all pages regardless of their quality or the links they receive.

Example:



In this scenario, all pages would have the same rank ($\frac{1}{3}$) regardless of the links they receive.

This method fails to differentiate between high-quality and low-quality pages

b.

(i) Yes. The page rank always exists. In a strongly connected graph, there's always at least one incoming node, ensuring the existence of a rank for each page according to this method (r(i) = minimum{r(j) | j ∈ B(i)})

(ii) This method doesn't reliably depict page quality. It fails to ensure that a low-quality page has few references while a high-quality page receives many references. We can illustrate it using below two examples:

Examples:

(a) A page with fewer incoming links might obtain a higher rank if those few incoming links originate from highly ranked pages. This, however, might not truly reflect the quality of the page itself.

Or

(b) A page with numerous incoming links could potentially receive a lower rank if it happens to have only one very low-ranking incoming page among several high-ranking ones. In this scenario, this method selects the lowest-ranked link, which might not accurately represent the quality of the page itself.

c.

(i) Yes, the page rank always exists. By formulating the equation where a page's rank is the product of the ranks of pages linking to it, we can solve the system of equations using linear algebra methods to compute the rank for each node.

(ii) However, this method does not effectively achieve the goal of assigning high ranks to high-quality pages and low ranks to low-quality pages. The reliance on the product of incoming ranks leads to a reduction in a page's rank as it receives more incoming links. Since the sum of all node ranks equals 1, each node's rank must be less than 1. When multiplying numbers less than 1, the product becomes smaller. Therefore, a node with numerous incoming links will experience a decrease in its rank value. This reduction does not necessarily reflect the inherent quality of the page itself, undermining the accuracy of this ranking method.

Example:

Page A has three higher rank ($\frac{1}{3}$) incoming links, page B has one very lower rank ($\frac{1}{21}$) Incoming link. We get:

$$\mathrm{r(A)} = \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{27}$$

$$\mathrm{r(B)} = \frac{1}{21}$$

$$\mathrm{r(A)} < \mathrm{r(B)}$$

Therefore. In this case, the higher quality page gets the lower rank. While the lower quality page gets the higher rank.

(d)

(i) Yes. The page rank always exists. In a strongly connected graph, there's always at least one incoming node, ensuring the existence of a rank for each page according to this method ($\mathrm{r(i)} = \mathrm{maximum}\{\mathrm{r(j)} \mid \mathrm{j} \in B(\mathrm{i})\} + \mathrm{minimum}\ \{\mathrm{r(j)} \mid \mathrm{j} \in B(\mathrm{i})\}$))

(ii) This method doesn't reliably depict page quality. It fails to ensure that a low-quality page has few references while a high-quality page receives many references. We can illustrate it using below two examples:

Examples:

(a)

Page A receive 10 incoming links with rank: $\{\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{12}\}$

$r(A) = \frac{1}{2} + \frac{1}{12} = \frac{7}{12}$

Page B receive 2 incoming links with rank: $\{\frac{1}{2},\frac{1}{12}\}$

$r(B) = \frac{1}{2} + \frac{1}{12} = \frac{7}{12}$

r(A) = r(B)

Therefore, It fails to ensure that a low-quality page has few references while a high-quality page receives many references.

(b)

Page A receive 10 incoming links with higher rank: $\{\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{12}\}$

$r(A) = \frac{1}{2} + \frac{1}{12} = \frac{7}{12}$

Page B receive 10 incoming links with lower rank: $\{\frac{1}{2},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12}\}$

$r(B) = \frac{1}{2} + \frac{1}{12} = \frac{7}{12}$

r(A) = r(B)

Therefore, it fails to assign high ranks to high-quality pages and low ranks to low-quality pages.