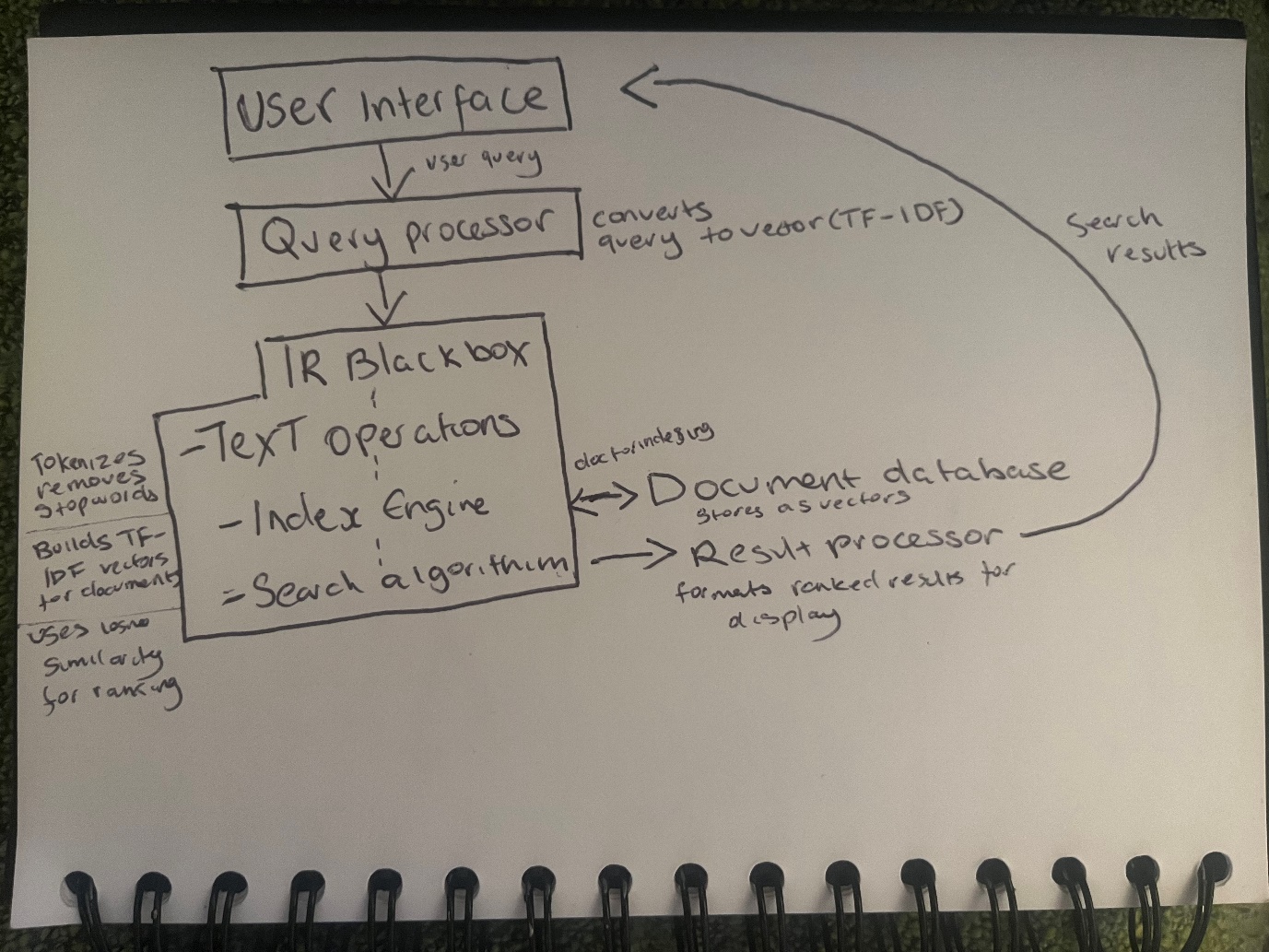
Question 1:

a and aii)



b)

D2 is the only relevant document because Document D1 doesn’t have "owl"; so fails the "(semantic AND owl)" condition. Document D3 has "owl" but doesn’t have "semantic" and "rdf". Document D2 has both "semantic" and "owl", meeting the "(semantic AND owl)" condition and has "rdf", meeting the "(rdf AND NOT(owl))" condition if "owl" were not present.

1c)

Query = [‘semantic’, ‘owl’] or [1,1]

D1 = [1, 0]

D2 = [1,1]

D3 = [0,2]

IDF (‘semantic’, ‘owl’)=log10​(3/2)=0.176

D1: [0.5+0.5×1/1 × 0.176, 0.5+0.5×0/0 × 0.176] = [0.176, 0]

D2: [0.5+0.5×1/1 × 0.176, 0.5+0.5 × 1/2 × 0.176] = [0.176, 0.088]

D3: [0.5+0.5×0/0 × 0.176, 0.5+0.5×2/2 × 0.176] = [0, 0.176]

D1: 1×0.176+0×0/√(12+02)(0.1762+02)=0.176

D2: 1×0.176+1×0.088/√(12+12)(0.1762+0.08822)​≈0.178

D3: 0×0+2×0.176/√ (02+22)(02+0.1762)=0.176

Ranking order: D2 , D1 and D3

Question 2:

a)

The Zero-Probability Problem occurs when a term from the query does not exist in the document. This results in a probability calculation of zero for the document. It causes the document to be considered irrelevant regardless of its content. To fix this a smoothing technique is used. Jelinek-Mercer is an example of a smoothing technique, it adds an alpha value that controls the balance between the document model and the collection model. The probability of seeing the query term in the document model is assessed if appears in the language collection, it adjusts the probability by considering the prevalence of the term across all documents, reducing the impact of the zero probabilities in individual documents.

bi)

p(‘Yoda’ | ‘Jedi’) = 0.26

p(‘is’ | ‘Yoda’) = 0.58

p(‘Jedi’ | ‘is’) = 0.52

p(‘master’ | Jedi’) = 0.12

0.26×0.58×0.52×0.12=0.00087808

bii)

p(‘Jedi’) = 0.092, p(‘Yoda’) = 0.055, p(‘is’) = 0.032, p(‘master’) = 0.019

p(‘Yoda’ | ‘Jedi’) = 0.6×0.26+0.4×0.055 = 0.1826

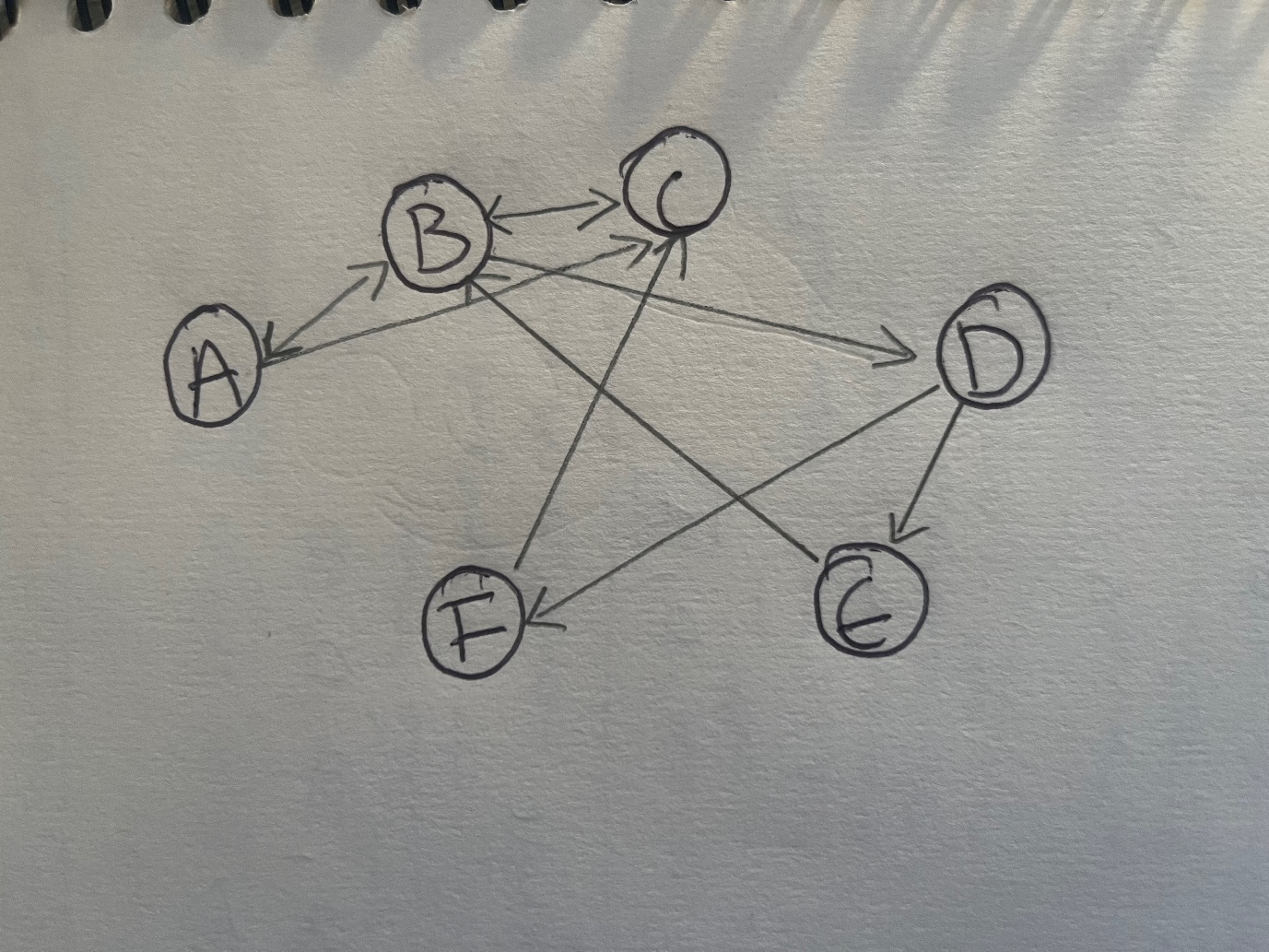
p(‘is’ | ‘Yoda’) = 0.6×0.58+0.4×0.032 = 0.3668

p(‘Jedi’ | ‘is’) = 0.6×0.52+0.4×0.092 = 0.3588

p(‘master’ | Jedi’) = 0.6×0.12+0.4×0.019 = 0.0874

0.1826×0.3668×0.3588×0.0874=0.002005

ci)



cii)

N(A) = 2 (links to B and C)

N(B) = 3 (links to A, C, and D)

N(C) = 1 (links to B)

N(D) = 2 (links to E and F)

N(E) = 1 (links to B)

N(F) = 1 (links to C)

Pagerank functions with the correct N(·) values:

PR(A) = (1 - d) + d (PR(B)/3 + PR(C)/1)

PR(B) = (1 - d) + d (PR(A)/2 + PR(D)/2)

PR(C) = (1 - d) + d (PR(B)/3)

PR(D) = (1 - d) + d (PR(E)/1)

PR(E) = (1 - d) + d (PR(B)/3 + PR(D)/2)

PR(F) = (1 - d) + d (PR(C)/1)

ciii)A screenshot of a computer

Description automatically generated

Question 3:

a)

230985072 – student number

𝑖=2,𝑗=3,𝑘=0,𝑙=9,𝑚=8,𝑛=5,𝑝=0,𝑞=7,𝑠=2,𝑡=3i=2,j=3,k=0,l=9,m=8,n=5,p=0,q=7,s=2,t=3

µi(Dj) = µ2(D3) = 0

µk(Dl) = µ0(D9) = 1

µm(Dn) = µ8(D5) = 1

µp(Dq) = µ0(D7) = 1

µs(Dt) = µ2(D3) = 0

3bi)

A close up of a sign

Description automatically generated

Bii)

Case (ws = 1, wo = 1, wr = 0):

Similarity = √ (1\*(1-1)2 + 1\*(1-0)2 + 0\*(1-0)2) = 1

Case (ws = 1, wo = 0, wr = 1):

Similarity = √ (1\*(1-1) 2 + 0\*(1-0) 2 + 1\*(1-0) 2) = 1

Case (ws = 1, wo = 0, wr = 0):

Similarity = √ (1\*(1-1) 2 + 0\*(1-0) 2 + 0\*(1-0) 2) = 0

Question 4:

ai)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | Doc | Relevant/Non-relevant | Precision | Recall |
| 1 | d5 | Non-relevant | 0/1 | 0/4 |
| 2 | d4 | Relevant | 1/2 | 1/4 |
| 3 | d1 | Relevant | 2/3 | 1/2 |
| 4 | d7 | Non-relevant | 1/2 | 1/2 |
| 5 | d20 | Non-relevant | 2/5 | 1/2 |
| 6 | d3 | Relevant | 1/2 | 3/4 |
| 7 | d8 | Non-relevant | 3/7 | 3/4 |
| 8 | d9 | Non-relevant | 3/8 | 3/4 |
| 9 | d10 | Relevant | 4/9 | 4/4 |
| 10 | d11 | Non-relevant | 4/10 | 4/4 |

aii)

Recall 0.0 - Precision = 0.67  
Recall 0.25 - Precision = 0.67  
Recall 0.5 - Precision = 0.67  
Recall 0.75 - Precision = 0.5  
Recall 1.0 - Precision = 0.44

A screen shot of a graph

Description automatically generated

bi)

The R-Precision is 0.5, where R is the number of relevant documents for the query. R = 4 (d1, d3, d4, d10). At the 4th position in result list, two relevant documents d1, d4 are found out of top four documents, giving an R-precision of 2/4 = 0.5.

bii)

rank 2 (d4): Precision = 0.5  
rank 3 (d1): Precision = 0.67  
rank 6 (d3): Precision = 0.5  
rank 9 (d10): Precision = 0.44

(0.5 + 0.67 + 0.5 + 0.44) / 4 = 0.4375

Mean Average Precision = 0.4375

biii)

E = 1 - (1 / ((b / P) + ((1 - b) / R)))

(P = 0.4)

recall (R = 1.0) at rank 10

 b = 0.3

E = 1 - (1 / ((0.3 / 0.4) + (0.7 / 1))) = 0.4878

E measure = 0.4878

c)

A close up of a paper

Description automatically generated

α = 1.0, β = 0.5, γ = 0.5

Positive Document Vector (d\_pos): [3, 1, 0, 0, 5, 0, 2, 6, 3, 0]

Negative Document Vector (d\_neg):[0, 2, 4, 3, 0, 0, 4, 6, 0, 3]

New Query Vector (q\_new): [2, 0, 2, 1, 1, 2, 3, 4, 5, 1]

q\_0 =[0.5, 0.5, 4, 2.5, -1.5, 2, 4, 4, 3.5, 2.5]