**ΕΡΓΑΣΙΑ 4**

**ΣΤΟ ΜΑΘΗΜΑ: ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ**

**1.**

Έχουμε:

Q1 = 50

Q2 = 20

Q3 = 200

S1 50 έγγραφα

S2 200 έγγραφα

Βρέθηκαν:



Για το Q1:

**S1**

R1 = 30 / (30+20) = 0,6

P1 = 30 / (30+20) = 0,6

**S2**

R1 = 50 / (50+0) = 1

P1 = 50 / (50+150) = 0,25

Για το Q2:

**S1**

R2 = 3 / (3+17) = 0,15

P2 = 3 / (3+47) = 0,06

**S2**

R2 = 20 / (20+0) = 1

P2 = 20 / (20+180) = 0,1

Για το Q3:

**S1**

R3 = 50 / (50+150) = 0,25

P3 = 50 / (50+0) = 1

**S2**

R3 = 150 / (150+50) = 0,75

P3 = 150 / (150+50) = 0,75

**A)**

Τα F με macro-averaging είναι:

Το R θα γίνει:

**S1**

R = (0,6 + 0,15 + 0,25) / 3 = 1 / 3 = **0,333**

**S2**

R = (1 + 1 + 0,75) / 3 = **0,916**

Το P θα γίνει:

**S1**

P = ( 0,6 + 0,06 + 1) / 3 = **0,553**

**S2**

P = (0,25 + 0,1 + 0,75) / 3 = **0,366**

Άρα τα Fείναι:

**S1**

**F = 3 / (0,333 + 0,553) = 3,38**

**S2**

**F = 3 / (0,916 + 0,366) = 2,34**

**S1,F>S2, F άρα το σύστημα S1 είναι προτιμότερο.**

**B)**

Κάνουμε την ίδια δουλειά απλά με micro-averaging

Το P θα γίνει:

**S1**

P= (30 + 3 + 50) / (30 + 3 + 50 + 20 + 17 + 150) = **0,307**

**S2**

P = (50 + 20 + 150) / (50 + 20 + 150 + 0 + 0 +50) = **0,814**

Το R θα γίνει:

**S1**

R = (30 + 3 + 50) / (30 + 3 +50 + 20 + 47 + 0) = **0,553**

**S2**

R = (50 + 20 + 150) / (50 + 20 + 150 + 150 + 180 + 50) = **0,366**

Άρα τα F είναι:

**S1  
F = 3 / (0,307 + 0,553) = 3,488**

**S2**

**F = 3 / (0,814 + 0,366) = 2,542**

**S1,F>S2,F άρα το σύστημα S1 είναι προτιμότερο.**

**2.**

**Α)**

**To Precision@5 είναι:**

**S1**

q1,P=3/5

q2,P=2/5

άραaverage = (3/5 + 2/5) / 2 = **1/2**

**S2**

q1,P=3/5

q2,P=3/5

άραaverage = (3/5 + 3/5) / 2 = **3/5**

**S2>S1 Άρα το σύστημα S2 είναι προτιμότερο.**

**Η ακρίβεια-R είναι:**

**S1**

q1,R= 6/10

q2,R=4/10

άραaverage = (6/10 + 4/10) / 2 = **1/2**

**S2**

q1,R=6/10

q2,R= 4/10

άραaverage = (6/10 + 4/10) / 2 = **1/2**

**S2=S1 Άρα τα συστήματα είναι προτιμότερο.**

**Β)**

Έχουμε:

**S1**

|  |  |
| --- | --- |
| Q1 | |
| Precision | Recall |
| 0,5 | 0,16 |
| 0,66 | 0,33 |
| 0,75 | 0,5 |
| 0,5 | 0,66 |
| 0,55 | 0,83 |
| 0,6 | 1 |

|  |  |
| --- | --- |
| Q2 | |
| Precision | Recall |
| 1 | 0,25 |
| 0,66 | 0,5 |
| 0,42 | 0,75 |
| 0,5 | 1 |

**S2**

|  |  |
| --- | --- |
| Q1 | |
| Precision | Recall |
| 1 | 0,16 |
| 1 | 0,33 |
| 0,6 | 0,5 |
| 0,66 | 0,66 |
| 0,71 | 0,83 |
| 0,6 | 1 |

|  |  |
| --- | --- |
| Q2 | |
| Precision | Recall |
| 1 | 0,25 |
| 0,5 | 0,5 |
| 0,6 | 0,75 |
| 0,44 | 1 |

Σύμφωνα με τα παραπάνω:

**S1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Q1 | Q2 | M |
| 0 | 1 | 1 | 1 |
| 0,1 | 0,75 | 1 | 0,875 |
| 0,2 | 0,75 | 1 | 0,875 |
| 0,3 | 0,75 | 0,66 | 0,705 |
| 0,4 | 0,75 | 0,66 | 0,705 |
| 0,5 | 0,75 | 0,66 | 0,705 |
| 0,6 | 0,6 | 0,5 | 0,55 |
| 0,7 | 0,6 | 0,5 | 0,55 |
| 0,8 | 0,6 | 0,5 | 0,55 |
| 0,9 | 0,6 | 0,5 | 0,55 |
| 1 | 0,6 | 0,5 | 0,55 |

**S2**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Q1 | Q2 | M |
| 0 | 1 | 1 | 1 |
| 0,1 | 1 | 1 | 1 |
| 0,2 | 1 | 1 | 1 |
| 0,3 | 1 | 0,6 | 0,8 |
| 0,4 | 0,71 | 0,6 | 0,655 |
| 0,5 | 0,71 | 0,6 | 0,655 |
| 0,6 | 0,71 | 0,6 | 0,655 |
| 0,7 | 0,71 | 0,6 | 0,655 |
| 0,8 | 0,71 | 0,44 | 0,575 |
| 0,9 | 0,6 | 0,44 | 0,52 |
| 1 | 0,6 | 0,44 | 0,52 |

**Άρα βλέπουμε πως το σύστημα S2 είναι προτιμότερο.**

**Γ)**

**S1**

Rank1 = (1/2 + 2/3 + 3/4 + 4/8 + 5/9 + 6/10) / 6 = 0,578

Rank2 = (1 + 2/3 + 3/7 + 4/8) / 4 = 0,648

**MAP = (0,714 + 0,648) / 2 = 0,613**

**S2**

Rank1 = (1 + 1 + 3/5 + 4/6 + 5/7 +6/10) / 6 = 0,763

Rank2 = (1 + 1/2 + 3/5 + 4/9) / 4 = 0,636

**MAP = (0,763 + 0,636) / 2 = 0,699**

**Άρα το σύστημα S2 είναι καλύτερο.**

**3.**

**A)**

**S1**

**MAP =** (1/2 + 2/3 + 3/4 + 4/6 + 5/8 + 6/9) / 6 **= 0,645**

**S2**

**MAP = (**1 + 1 + 3/5 + 4/6 + 5/7 + 6/10) / 6 **= 0,763**

**S1,Map>S1, Map Άρα το σύστημα S2 είναι προτιμότερο.**

**B)**

**S1**

**DCG =** 0 + 3/log2 + 1/log3 + 2/log4 + 0 + 3/log6 + 0 + 1/log8 + 2/log9 + 0 = **6,758**

**S2**

**DCG =** 2 + 1/log2 + 0 + 0 + 3/log5 + 3/log6 + 1/log7 + 0 + 0 + 2/log10 = **6,043**

**S1,DCG>S2,DCG Άρα το σύστημα S1 είναι προτιμότερο.**

**Γ)**

Βρίσκουμε τα βέλτιστα DCGγια να βρούμε τα NDCG.



Οι relείναι ίδιοι άρα θα χρειαστεί να βρούμε το DCGμόνο μια φορά.

DCG = 3 + 3/log2 + 2/log3 + 2/log4 + 1/log5 + 1/log6 + 0 + 0 + 0 + 0 = **9,08**

**Άρα**

**S1**

**NDGC = 6,758 / 9,08 = 0,744**

**S2**

**NDGC = 6,043 / 9,08 = 0,665**

**ΕίναιS1,NDGC>S2,NDGC άρα το σύστημα S1 είναι προτιμότερο.**

**4.**

**Α)**

Μετά από 3 επαναλήψεις φτάσαμε εδώ:

|  |  |  |
| --- | --- | --- |
| Hubs | Pages | Authorities |
| 0,61 | 1 | 0,62 |
| 1 | 2 | 1 |
| 1 | 3 | 0,62 |
| 0,07 | 4 | 1 |
| 0,61 | 5 | 0,05 |

**Τα καλύτερα Hubsείναι οι σελίδες 2 και 3 (score 1)**

**Τα καλύτερα Authoritiesείναι οι σελίδες 2 και 4(score 1).**

**B)**

Είναι:

W1 = W3 / 2

W2 = W2 / 2 + W1

W3 = W2 / 2

W4 = W3 / 2 + W5

W5 = W4

Και ξεκινάμε με:

W1 = W2 = W3 = W4 = W5 = 1 / 5

Μετά από 3 επαναλήψεις φτάσαμε:

W1 = 0,075

W2 = 0,175

W3 = 0,125

W4 = 0,375

W5 = 0,25

**Άρα ο σημαντικότερος κόμβος είναι ο 4.**

**Γ)**

Ξεκινώντας πάλι με δεδομέναW1 = W2 = W3 = W4 = W5 = 1 / 5, β = 0,8

Και (1-β) \* 1/n= 0,2 \* 1/5 = 1/25

Έχουμε:

**Επανάληψη 1**

W1 = 0,8\* ((1/5)/2) + 1/25 = 0,12

W2 = 0,8 \* ((1/5)/2 + 1/5) + 1/25 = 0,28

W3 = 0,8 \* ((1/5)/2) + 1/25 = 0,12

W4 = 0,8 \* ((1/5)/2 + 1/5) + 1/25 = 0,28

W5 = 0,8 \* (1/5) + 1/25 = 0,2

**Επανάληψη 2**

W1 = 0,8\* ((3/25)/2) + 1/25 = 4/5 \* 1,5/25 + 1/25 = 1,2/25 + 1/25 = **2,2/25**

W2 = 0,8\* ((7/25)/2 + 3/25) + 1/25 = 4/5 \* 6,5/25 + 1/25 = **6,2/25**

W3 = 0,8\* ((7/25)/2) + 1/25 = 4/5 \* 3,5/25 + 1/25 = **3,8/25**

W4 = 0,8\* ((3/25)/2 + 5/25) + 1/25 = 4/5 \* 6,5/25 + 1/25 = **6,2/25**

W5 = 0,8\* (7/25) + 1/25 =**6,6/25**

**Επανάληψη 3**

W1 = 0,8\* ((3,8/25)/2) + 1/25 = 4/5 \* 1,9/25 + 1/25 = **0,1**

W2 = 0,8 \* ((6,2/25)/2 + 2,2/25) + 1/25 = 4/5 \* 5,3/25 + 1/25 = **0,21**

W3 = 0,8\* ((6,2/25)/2) + 1/25 = 4/5 \* 3,1/25 + 1/25 = **0,14**

W4 = 0,8\* ((3,8/25)/2 + 6,6/25) + 1/25 = 4/5 \* 8,5/25 + 1/25 = **0,31**

W5 = 0,8\* (6,2/25) + 1/25 = **0,23**

**Άρα ο σημαντικότερος κόμβος είναι ο 4.**

**Δ)**

ΣτονHitsοι σημαντικότεροι κόμβοι είναι οι 2 και 4,αυτό σημαίνει ότι στους κόμβους 2 και 4 δείχνουν πολλές σελίδες. Από την άλλη με τον PageRank και Pagerankwithjumps οι κόμβοι 4 και 5έχουν το υψηλότερο rank, δηλαδή οι πιθανότητες να είσαι στις σελίδες 4 και 5 είναι οι μεγαλύτερες.

Εάν συνεχίζαμε τα randomwalks:

Με τον Hitsθα μεγάλωνε περισσότερο ο κόμβος 2.

Και με τους PageRankο κόμβος 4.

**Ε)**

Θα χρησιμοποιούσα τονPageRank με Jumbsγιατί υπάρχει ένα SpiderTrapστους κόμβους 4 και 5. Άρα θα επιλυθεί αυτό το πρόβλημα. Ο PageRank, αν δεν υπάρχουν αδιέξοδα και SpiderTraps, είναι πιο γρήγορος, αφού δεν θα πηγαίνουμε σε τυχαίους. Με τον Hitsδεν έχουμε τα προβλήματα των αδιεξόδων και SpiderTraps.

**ΑΠΟ ΤΟΝ ΦΟΙΤΗΤΗ: Τσιαούση Σταύρο (dai17173)**