**Information retrieval HW 2.**

**Group: 404 Brain Not Found.**

**Submmiters:**

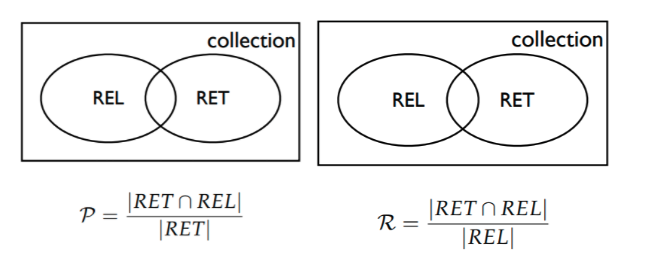
Tomer Lazarovitch 316377100  
Itamar Kraus 318304763  
Guy Pariente 206132250  
Chay Fadida 318302783

שאלה 1:

1.



פתרון:



חישוב Precision:

מספר המסמכים שהוחזרו ע״י המערכת והם רלוונטיים.

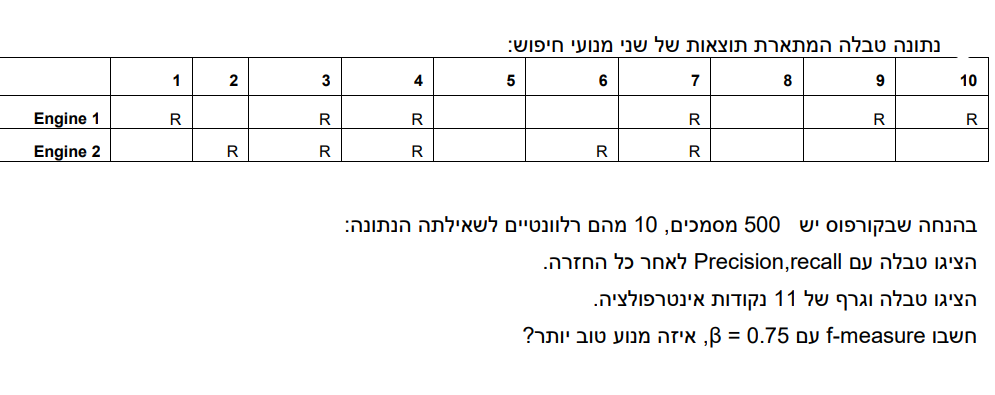
סה״כ המסמכים שהוחזרו על ידי המערכת

סה״כ המסמכים הרלוונטיים.

מספר המסמכים שהוחזרו ע״י המערכת והם רלוונטיים.

חישוב Recall:

2.



פתרון:

נציג טבלה עם Precision & Recall לאחר כל החזרה:

מנוע 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recall | Precision | סה״כ רלוונטיים | סה״כ הוחזרו | מסמך |
|  |  | 1 | 1 | 1 |
|  |  | 1 | 2 | 2 |
|  |  | 2 | 3 | 3 |
|  |  | 3 | 4 | 4 |
|  |  | 3 | 5 | 5 |
|  |  | 3 | 6 | 6 |
|  |  | 4 | 7 | 7 |
|  |  | 4 | 8 | 8 |
|  |  | 5 | 9 | 9 |
|  |  | 6 | 10 | 10 |

מנוע 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recall | Precision | סה״כ רלוונטיים | סה״כ הוחזרו | מסמך |
|  |  | 0 | 1 | 1 |
|  |  | 1 | 2 | 2 |
|  |  | 2 | 3 | 3 |
|  |  | 3 | 4 | 4 |
|  |  | 3 | 5 | 5 |
|  |  | 4 | 6 | 6 |
|  |  | 5 | 7 | 7 |
|  |  | 5 | 8 | 8 |
|  |  | 5 | 9 | 9 |
|  |  | 5 | 10 | 10 |

כעת נציג טבלה וגרף של 11 נקודות אינטרפולציה:

ניצור נקודות אינטרפולציה ע״י בחירת 11 נקודות Recall באופן קבוע: [0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0]

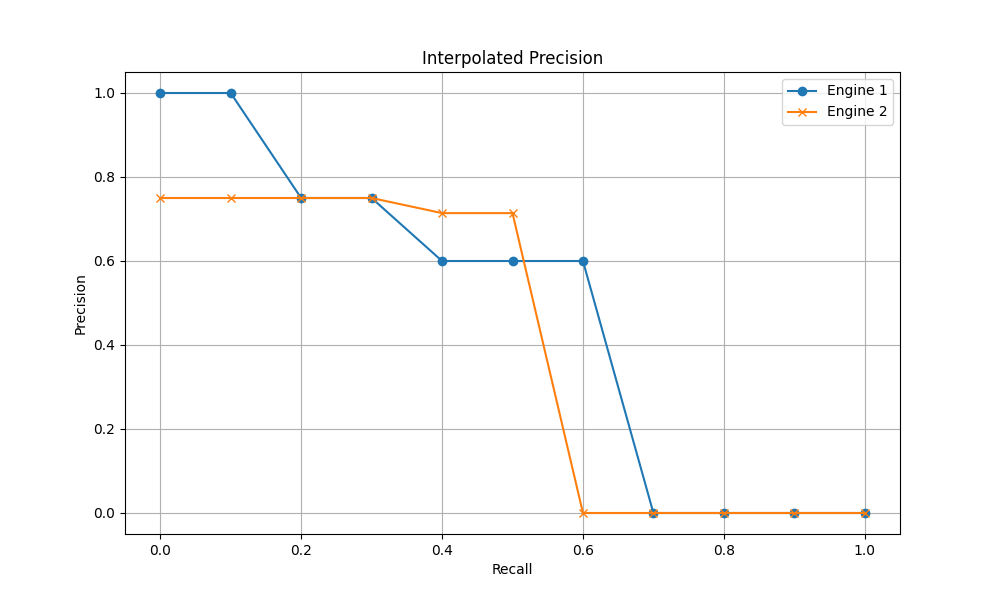
באינטרפולציה לוקחים את ערכי ה Precision הגבוהים ביותר עבור כל נקודת Recall קבועה.

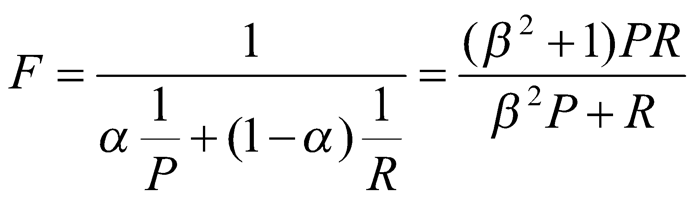
מנוע 1:

|  |  |
| --- | --- |
| Precision | Recall |
| 1.0 | 0.0 |
| 1.0 | 0.1 |
| 0.75 | 0.2 |
| 0.75 | 0.3 |
| 60. | 0.4 |
| 0.6 | 0.5 |
| 0.6 | 0.6 |
| 0 | 0.7 |
| 0 | 0.8 |
| 0 | 0.9 |
| 0 | 1.0 |

מנוע 2:

|  |  |
| --- | --- |
| Precision | Recall |
| 0.75 | 0.0 |
| 0.75 | 0.1 |
| 0.75 | 0.2 |
| 0.75 | 0.3 |
| 0.714 | 0.4 |
| 0.714 | 0.5 |
| 0 | 0.6 |
| 0 | 0.7 |
| 0 | 0.8 |
| 0 | 0.9 |
| 0 | 1.0 |



 נחשב את ה f-measure עם ערך β = 0.75:

על מנת לקבוע איזה מנוע טוב יותר נחשב את הממוצע של ערך ה f-measure בכל אחד המנועים ונבחר במנוע עם הממוצע הגבוה יותר:

עבור מנוע 1 :

Avg precision:

Avg recall:

f-measure:

עבור מנוע 2 :

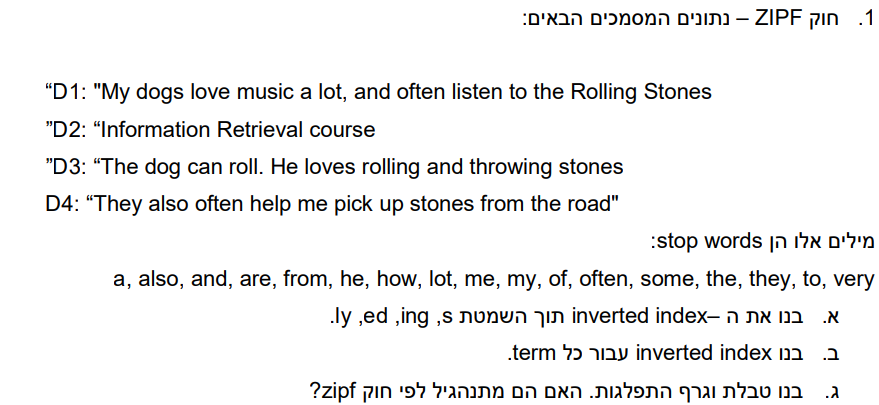
Avg precision:

Avg recall:

f-measure:

**מאחר וה- f-measure של מנוע 1 גבוה יותר נוכל לקבוע גם כי הוא המנוע הטוב יותר.**

שאלה 2:



פתרון:

1. נבנה את ה inverted index תוך השמטת ly,ed,ing,s:

|  |  |
| --- | --- |
| Frequency | Term |
| D1,D3 | Dog |
| D1,D3 | Love |
| D1 | Music |
| D1 | Listen |
| D1,D3 | Roll |
| D1,D3,D4 | Stone |
| D2 | Information |
| D2 | Retrieval |
| D2 | Course |
| D3 | Can |
| D3 | Throw |
| D4 | Help |
| D4 | Pick |
| D4 | Up |
| D4 | Road |

1. נבנה inverted index עבור כל term:

|  |  |
| --- | --- |
| Frequency | Term |
| D1 | Dogs |
| D1 | Love |
| D1 | Music |
| D1 | Listen |
| D1,D3 | Rolling |
| D1,D3,D4 | Stones |
| D2 | Information |
| D2 | Retrieval |
| D2 | Course |
| D3 | Dog |
| D3 | Can |
| D3 | Roll |
| D3 | Loves |
| D3 | Throwing |
| D4 | Help |
| D4 | Pick |
| D4 | up |
| D4 | Road |

1. טבלת התפלגות מתאימה:

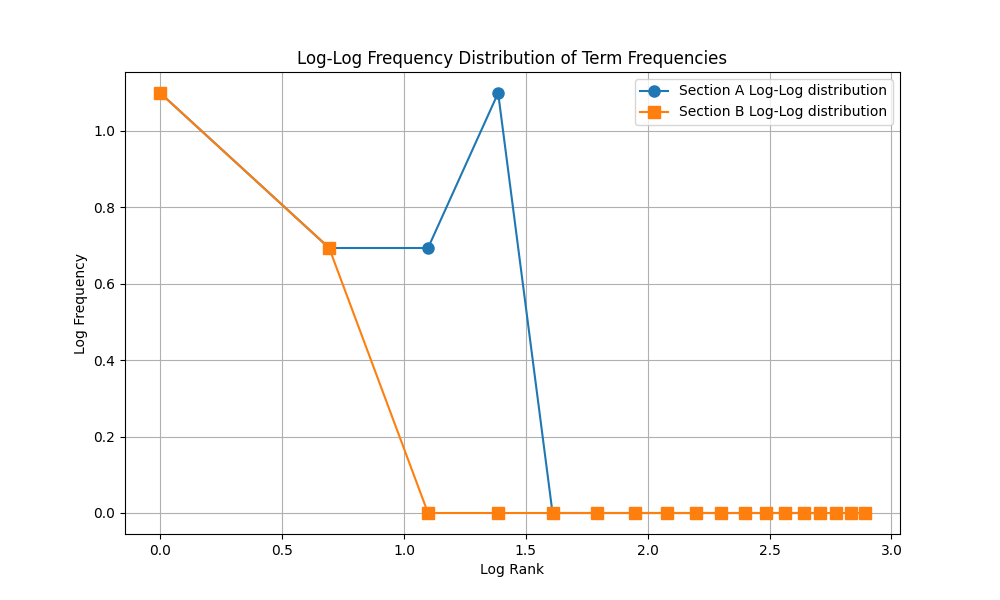
עבור סעיף א׳ (נבנה את ה inverted index תוך השמטת ly,ed,ing,s):

|  |  |
| --- | --- |
| Frequency | Term |
| 3 | Stone |
| 3 | Roll |
| 2 | Dog |
| 2 | Love |
| 1 | Music |
| 1 | Listen |
| 1 | Information |
| 1 | Retrieval |
| 1 | Course |
| 1 | Can |
| 1 | Throw |
| 1 | Help |
| 1 | Pick |
| 1 | Up |
| 1 | Road |

עבור סעיף ב:

|  |  |
| --- | --- |
| Frequency | Term |
| 3 | Stones |
| 2 | Rolling |
| 1 | Help |
| 1 | Pick |
| 1 | up |
| 1 | Road |
| 1 | Dog |
| 1 | Can |
| 1 | Roll |
| 1 | Loves |
| 1 | Throwing |
| 1 | Information |
| 1 | Retrieval |
| 1 | Course |
| 1 | Dogs |
| 1 | Love |
| 1 | Music |
| 1 | Listen |

גרף התפלגות מתאים:



גרף התפלגות התדירות הנתונים עבור שני הסעיפים מציג מאפיינים התואמים חלקית עם חוק ZIPF. מערך הנתונים של סעיף ב מראה ירידה טיפוסית יותר של ZIPF, ואילו מערך הנתונים של סעיף א סוטה בשיא, מה שמצביע על כמה סטיות מהתנהגות ZIPF מושלמת. סטיות אלו נובעות ככל הנראה מהגודל המוגבל או המאפיינים הספציפיים של מערך הנתונים. לכן, בעוד שהנתונים תואמים חלקית את חוק ZIPF, הם אינם תואמים אותו באופן מושלם.

שאלה 3: קדם פרויקט –בניית זחלן



פתרון:

1. [https://www.premierleague.com](https://www.premierleague.com/stats/top/players/appearances?se=-1)

התגיות הרלוונטיות:

Word: emirates, Class ID: ['match-fixture\_\_stadium-name'], Page: fixtures

Word: goodison, Class ID: ['match-fixture\_\_stadium-name'], Page: fixtures

Word: anfield, Class ID: ['match-fixture\_\_stadium-name'], Page: fixtures

Word: emirates, Class ID: ['top-stats\_\_row-team-name'], Page: stats

Word: emirates, Class ID: ['top-stats\_\_hero-ground'], Page: stats

Word: anfield, Class ID: ['top-stats\_\_row-team-name'], Page: stats

Word: anfield, Class ID: ['top-stats\_\_row-team-name'], Page: stats

Word: goodison, Class ID: ['top-stats\_\_row-team-name'], Page: stats

Word: goodison, Class ID: ['top-stats\_\_row-team-name'], Page: stats

Word: anfield, Class ID: ['js-lazy-load', 'u-observed', 'lazy-image-wrapper'], Page: video

Word: anfield, Class ID: ['media-thumbnail\_\_title'], Page: video

Word: emirates, Class ID: ['club-filter-promo\_\_content-title'], Page: tickets

Word: goodison, Class ID: ['club-filter-promo\_\_content-title'], Page: tickets

Word: anfield, Class ID: ['club-filter-promo\_\_content-title'], Page: tickets

Word: emirates, Class ID: ['club-card\_\_stadium'], Page: clubs

Word: emirates, Class ID: ['team-index\_\_stadium-name', 'u-show-mob'], Page: clubs

Word: emirates, Class ID: ['team-index\_\_stadium-name'], Page: clubs

Word: goodison, Class ID: ['club-card\_\_stadium'], Page: clubs

Word: goodison, Class ID: ['team-index\_\_stadium-name', 'u-show-mob'], Page: clubs

Word: goodison, Class ID: ['team-index\_\_stadium-name'], Page: clubs

Word: anfield, Class ID: ['club-card\_\_stadium'], Page: clubs

Word: anfield, Class ID: ['team-index\_\_stadium-name', 'u-show-mob'], Page: clubs

Word: anfield, Class ID: ['team-index\_\_stadium-name'], Page: clubs

שאילתות:

כל העמודים באתר המכילים את אחד משמות האצטדיונים הבאים:

* Emirates
* Anfield
* Goodison
* Stamford

ב.

BASE\_URL = "https://www.premierleague.com/"

PAGES = {

1: "fixtures",

2 :"results",

3 :"tables",

4 :"transfers",

5 :"stats",

6 :"news",

7 :"video",

8 :"tickets",

9 :"clubs",

10 :"players",

11 :"awards?at=1&aw=-1&se=578",

12 :"about",

13 :"about/what-we-do",

14 :"about/governance",

15 :"about/statement-of-principles",

16 :"about/inclusion",

17 :"about/publications",

18 :"about/partners",

19 :"footballandcommunity/wider-football/overview",

20 :"wider-football/stadium-fund",

21 :"wider-football/womens-and-girls-football",

22 :"wider-football/academies-and-player-welfare",

23 :"history/dashboard",

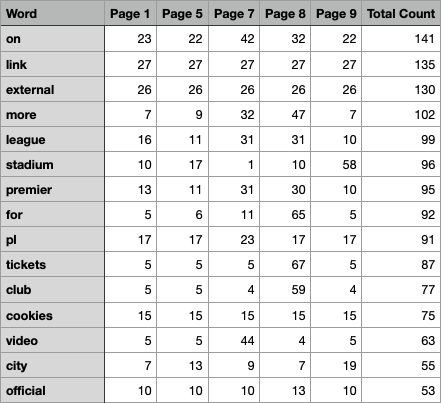
24 :"history/season-reviews",

25 :"history/kits?se=719&cl=-1",

26 :"history/origins"

{

1. טבלת inverted index ל- 15 המילים הנפוצות ביותר שהוחזרו:



2.

3.

התוצאות שחזרו מהשאילתה:

נקבע כי page 1 – ‘fixtures’ , page 7 – ‘video’ , page 8 – ‘tickets’ , page 9 – ‘clubs’   
הינם דפים רלוונטיים עבור השאילתה (לצורך חישוב precision & recall)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Recall | Precision | Page 9 | Page 8 | Page 7 | Page 5 | Page 1 | User |
| 1 | 4/5 | R | R | R |  | R | a |
| 3/4 | 3/5 | R | R |  |  | R | b |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W-d9 | W-d8 | W-d7 | W-d5 | W- d1 | IDF | D/dfi | Dfi | Page 9 | Page 8 | Page 7 | Page 5 | Page 1 | Word |
| 3\*0.511 = 1.533 | 1\*0.511=0.511 | 1\*0.511=0.511 | 2\*0.511 = 1.022 |  |  |  | 8 | 3 | 1 | 1 | 2 | 1 | Anfield |
| 3\*0.569 = 1.707 | 1\*0.569=0.569 | 0 | 2\*0.569 = 1.138 | 1\*0.569=0.569 |  |  | 7 | 3 | 1 | 0 | 2 | 1 | Goodison |
| 3\*0.636 = 1.908 | 0 | 0 | 2\*0.636 = 1.272 | 1\*0.636=0.636 |  |  | 6 | 3 | 0 | 0 | 2 | 1 | Stamford |
| 3\*0.569 = 1.707 | 1\*0.569=0.569 | 0 | 2\*0.569 = 1.138 | 1\*0.569=0.569 |  |  | 7 | 3 | 1 | 0 | 2 | 1 | Emirates |

