**תרגיל בית 2:**

**מגישים:**

ישראל שושן 318740834

עמית פרץ 209054303

דוד אסולין 205908569

יובל הילאי 318609237

קישור לגיט: [AmitPerets/Index\_Invincibles\_HW\_2 (github.com)](https://github.com/AmitPerets/Index_Invincibles_HW_2)

**שאלה 1:**



נשתמש בנוסחאות המתאימות לprecision ול-recall:

Precision מייצג את היחס בין המסמכים הרלוונטים שהוחזרו לבין כלל המסמכים שהוחזרו.

כלומר במקרה שלנו:

מסמכים רלוונטים שהוחזרו – 8

כלל המסמכים שהוחזרו – 8 (רלוונטים שהוחזרו) + 10 (לא רלוונטים שהוחזרו)

Precision = 8/18 = 0.444

Recall מייצג את היחס בין המסמכים הרלוונטים שהוחזרו לבין כלל המסמכים הרלוונטים באוסף.

כלומר במקרה שלנו:

מסמכים רלוונטים שהוחזרו – 8

כלל המסמכים הרלוונטים באוסף – 20

Recall = 8/20 = 0.4

**שאלה 2:**

A close up of a number

Description automatically generated

Engine 1:

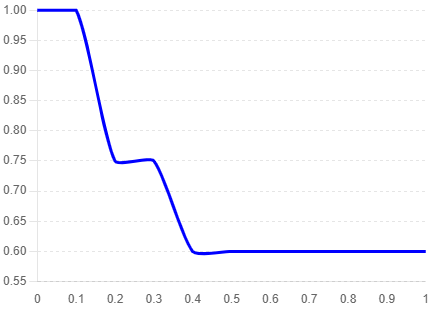
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | Relevant | Cumulative Relevant | Percision | Recall |
| 1 | R | 1 | 1/1 = 1.0 | 1/10 = 0.1 |
| 2 |  | 1 | 1/2 = 0.5 | 1/10 = 0.1 |
| 3 | R | 2 | 2/3 = 0.67 | 2/10 = 0.2 |
| 4 | R | 3 | 3/4 = 0.75 | 3/10 = 0.3 |
| 5 |  | 3 | 3/5 = 0.6 | 3/10 = 0.3 |
| 6 |  | 3 | 3/6 = 0.5 | 3/10 = 0.3 |
| 7 | R | 4 | 4/7 = 0.57 | 4/10 = 0.4 |
| 8 |  | 4 | 4/8 = 0.5 | 4/10 = 0.4 |
| 9 | R | 5 | 5/9 = 0.56 | 5/10 = 0.5 |
| 10 | R | 6 | 6/10 = 0.6 | 6/10 = 0.6 |

Engine 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | Relevant | Cumulative Relevant | Percision | Recall |
| 1 |  | 0 | 0/1 = 0.0 | 0/10 = 0.0 |
| 2 | R | 1 | 1/2 = 0.5 | 1/10 = 0.1 |
| 3 | R | 2 | 2/3 = 0.67 | 2/10 = 0.2 |
| 4 | R | 3 | 3/4 = 0.75 | 3/10 = 0.3 |
| 5 |  | 3 | 3/5 = 0.6 | 3/10 = 0.3 |
| 6 | R | 4 | 4/6 = 0.67 | 4/10 = 0.4 |
| 7 | R | 5 | 5/7 = 0.71 | 5/10 = 0.5 |
| 8 |  | 5 | 5/8 = 0.62 | 5/10 = 0.5 |
| 9 |  | 5 | 5/9 = 0.56 | 5/10 = 0.5 |
| 10 |  | 5 | 5/10 = 0.5 | 5/10 = 0.5 |

חישוב אינטרפולציה ל- 11 נקודות - Engine 1

|  |  |
| --- | --- |
| Recall | Percision |
| 0.0 | 1.0 |
| 0.1 | 1.0 |
| 0.2 | 0.75 |
| 0.3 | 0.75 |
| 0.4 | 0.6 |
| 0.5 | 0.6 |
| 0.6 | 0.6 |
| 0.7 | 0.6 |
| 0.8 | 0.6 |
| 0.9 | 0.6 |
| 1.0 | 0.6 |



חישוב אינטרפולציה ל- 11 נקודות - Engine 2

|  |  |
| --- | --- |
| 0.0 | 0.75 |
| 0.1 | 0.75 |
| 0.2 | 0.75 |
| 0.3 | 0.75 |
| 0.4 | 0.71 |
| 0.5 | 0.71 |
| 0.6 | 0.71 |
| 0.7 | 0.71 |
| 0.8 | 0.71 |
| 0.9 | 0.71 |
| 1.0 | 0.71 |

A green line on a black background

Description automatically generated

חישוב F-measure עם **β = 0.75 על פי הנוסחה:**

**Engine 1:**

Recall ממוצע: (0.1+0.1+0.2+0.3+0.3+0.3+0.4+0.4+0.5+0.6) / 10 = 0.32

Precision ממוצע: 0.625 (1.0+0.5+0.67+0.75+0.6+0.5+0.57+0.5+0.56+0.6) / 10 = ​

**F\_β** ≈

**Engine 2:**

Recall ממוצע: (0+0.1+0.2+0.3+0.3+0.4+0.5+0.5+0.5+0.5) / 10 = 0.33

Precision ממוצע: (0.5+0.5+0.67+0.75+0.6+0.67+0.71+0.62+0.56+0.5) / 10 = 0.558

**F\_β** ≈

כלומר Engine 1 הוא המנוע הטוב יותר על פי החישובים.

**שאלה:**

A screenshot of a text

Description automatically generated

1. נבצע נורמליזציה של הטקסט על ידי הסרת המילים המוגדרות כ-stop words והסרת הסיומות ly, ed, ing, s מהמילים.

הסרת stop words:

|  |  |
| --- | --- |
| מסמך | טקסט לאחר הסרת stop words |
| D1 | dog love music, listen Rolling Stones |
| D2 | Information Retrieval course |
| D3 | dog can roll. loves rolling throwing stones |
| D4 | help pick up stones road |

הסרת סיומות ly, ed, ing, s :

dog -> dog

love -> love

music -> music

listen -> listen

Rolling -> Roll

Stones -> Stone

Information -> Information

Retrieval -> Retrieval

course -> course

can -> can

roll -> roll

loves -> love

rolling -> roll

throwing -> throw

stones -> stone

help -> help

pick -> pick

up -> up

stones -> stone

road -> road

נבנה אינדקס הפוך שמכיל את כל המילים שנותרו לאחר הנורמליזציה והמסמכים שבהם הן מופיעות.

|  |  |
| --- | --- |
| **Term** | **Documents** |
| dog | D1, D3 |
| love | D1, D3 |
| music | D1 |
| listen | D1 |
| roll | D1, D3 |
| stone | D1, D3, D4 |
| information | D2 |
| retrieval | D2 |
| course | D2 |
| can | D3 |
| throw | D3 |
| help | D4 |
| pick | D4 |
| up | D4 |
| road | D4 |

1. נבנה Inverted Index עבור כל המונחים ללא הסרת סיומות וללא הסרת stop words

|  |  |
| --- | --- |
| **Term** | **Documents** |
| My | D1 |
| dog | D1, D3 |
| love | D1 |
| music | D1 |
| a | D1 |
| lot | D1 |
| and | D1, D3 |
| often | D1, D4 |
| listen | D1 |
| to | D1 |
| the | D1, D3, D4 |
| rolling | D1, D3 |
| stones | D1, D3, D4 |
| information | D2 |
| retrieval | D2 |
| course | D2 |
| can | D3 |
| roll | D3 |
| He | D3 |
| loves | D3 |
| throwing | D3 |
| They | D4 |
| also | D4 |
| help | D4 |
| me | D4 |
| pick | D4 |
| up | D4 |
| from | D4 |
| road | D4 |

טבלת התפלגויות:

|  |  |
| --- | --- |
| Term | Frequency |
| dog | 2 |
| love | 2 |
| music | 1 |
| listen | 1 |
| roll | 2 |
| stone | 3 |
| information | 1 |
| retrieval | 1 |
| course | 1 |
| can | 1 |
| throw | 1 |
| help | 1 |
| pick | 1 |
| up | 1 |
| road | 1 |

גרף:

A graph with numbers and lines

Description automatically generated

הגרף המוצג מראה את התפלגות התדירויות של המילים במסמכים לפי חוק Zipf. על פי הגרף, ניתן לראות שההתפלגות היא בערך לינארית בסקאלה לוגריתמית, מה שמעיד על כך שהתפלגות המילים אכן מתנהגת לפי חוק Zipf , כלומר, המילה הכי נפוצה מופיעה בערך פי שניים יותר מהשנייה הכי נפוצה, וכן הלאה.

**שאלה 3:**

A white background with black text

Description automatically generated

**חלק א: הגדרת הנתונים והשאילתות**

**האתר שלנו: https://www.ea.com/news**

**שאילתות/פילטרים**:

**שאילתה ראשונה**:

site:ea.com "news" OR "event" OR "update" OR "details"

חיפוש בעמודים המכילים חדשות, אירועים, עדכונים או פרטים כלליים על EA SPORTS

תגיות רלוונטיות:

"news"

"event"

"update"

"details"

**שאילתה שניה:**

site:ea.com "update" OR "game" OR "performance" OR "patch"

משמעות השאילתה: חיפוש בעמודים המכילים עדכוני משחק וביצועים.

תגיות רלוונטיות:

"update"

"game"

"performance"

"patch"

**שאילתה שלישית:**

site:ea.com "player" OR "statistics" OR "ranking" OR "top"

משמעות השאילתה: חיפוש בעמודים המכילים נתונים על שחקנים, סטטיסטיקות, דירוגים ושחקנים מובילים.

תגיות רלוונטיות:

"player"

"statistics"

"ranking"

"top"

**חלק ב:**

1.

A table with numbers and text

Description automatically generated

2. Term Frequencies:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **details** | **update** | **event** | **news** |  |
| 0 | 2 | 0 | 0 | **Doc1** |
| 0 | 0 | 0 | 3 | **Doc2** |
| 0 | 1 | 0 | 1 | **Doc3** |
| 0 | 0 | 0 | 1 | **Doc4** |
| 5 | 0 | 0 | 2 | **Doc5** |
| 0 | 0 | 2 | 0 | **Doc6** |
| 0 | 0 | 0 | 1 | **Doc7** |
| 0 | 0 | 0 | 1 | **Doc8** |
| 0 | 0 | 0 | 0 | **Doc9** |
| 0 | 0 | 0 | 0 | **Doc10** |
| 0 | 0 | 0 | 3 | **Doc11** |
| 1 | 0 | 0 | 0 | **Doc12** |
| 0 | 0 | 0 | 0 | **Doc13** |
| 6 | 0 | 2 | 0 | **Doc14** |
| 0 | 2 | 0 | 0 | **Doc15** |
| 0 | 0 | 0 | 3 | **Doc16** |
| 0 | 5 | 0 | 0 | **Doc17** |
| 3 | 0 | 0 | 0 | **Doc18** |
| 0 | 0 | 0 | 0 | **Doc19** |
| 0 | 0 | 0 | 0 | **Doc20** |

|  |  |
| --- | --- |
| **Number of terms** | **Docs** |
| 762 | 1 |
| 130 | 2 |
| 251 | 3 |
| 69 | 4 |
| 239 | 5 |
| 681 | 6 |
| 69 | 7 |
| 69 | 8 |
| 179 | 9 |
| 310 | 10 |
| 130 | 11 |
| 34 | 12 |
| 59 | 13 |
| 3284 | 14 |
| 125 | 15 |
| 130 | 16 |
| 3466 | 17 |
| 690 | 18 |
| 157 | 19 |
| 70 | 20 |

TF: החישוב מתבצע לפי Term Frequencies / Number of terms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **details** | **update** | **event** | **news** |  |
| 0 | 2/762 | 0 | 0 | **Doc1** |
| 0 | 0 | 0 | 3/130 | **Doc2** |
| 0 | 1/251 | 0 | 1/251 | **Doc3** |
| 0 | 0 | 0 | 1/69 | **Doc4** |
| 5/239 | 0 | 0 | 2/239 | **Doc5** |
| 0 | 0 | 2/681 | 0 | **Doc6** |
| 0 | 0 | 0 | 1/69 | **Doc7** |
| 0 | 0 | 0 | 1/69 | **Doc8** |
| 0 | 0 | 0 | 0 | **Doc9** |
| 0 | 0 | 0 | 0 | **Doc10** |
| 0 | 0 | 0 | 3/130 | **Doc11** |
| 1/34 | 0 | 0 | 0 | **Doc12** |
| 0 | 0 | 0 | 0 | **Doc13** |
| 6/3284 | 2/3284 | 2/3284 | 0 | **Doc14** |
| 0 | 0 | 0 | 0 | **Doc15** |
| 0 | 0 | 0 | 3/130 | **Doc16** |
| 3/3466 | 5/3466 | 0 | 0 | **Doc17** |
| 0 | 0 | 0 | 0 | **Doc18** |
| 0 | 0 | 0 | 0 | **Doc19** |
| 0 | 0 | 0 | 0 | **Doc20** |

IDF:

|  |  |
| --- | --- |
| 0.798508 | news |
| 1.897120 | event |
| 1.609438 | update |
| 1.386294 | details |

3.

הגדרנו ש2 משתמשים יחליטו עבור כל תוצאה מ-10 התוצאות הראשונות האם היא רלוונטית. עבור היוזר הראשון: user1\_relevant = [1, 1, 0, 1, 1, 0, 1, 0, 1, 1]

עבור היוזר השני:         user2\_relevant = [1, 0, 1, 1, 1, 0, 1, 0, 0, 1]

(1 רלוונטי, 0 לא רלוונטי)

קיבלנו את התוצאות הבאות:

User1:

User2: