אחזור מידע – תרגיל בית 2 – אביב התשפד

שם קבוצה: Data Diggers

מגישים:

יובל זוהר 318159373

אבישי הרשקוביץ 209460443

אביחי חדד 209286665

שאלה 1: מודלים לאחזור מידע

שאלה 1:



**תשובה שאלה 1:**

**A white background with black and white clouds

Description automatically generated**

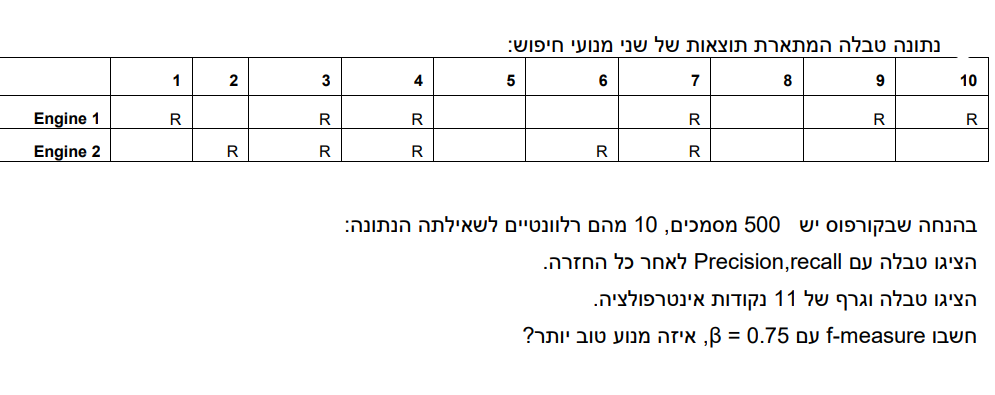
**Precision:**

Number of relevant Documents retrieved / Total number of Documents retrieved = 8/18 ≅ **0.444**

**Recall:**

Number of relevant Documents retrieved / Total number of relevant documents = 8/20 = **0.4**

שאלה 2:



**תשובה לשאלה 2:**

**Engine 1:**

**Relevant documents retrieved: 1, 3, 4, 7, 9, 10 (6 relevant documents)**

**Total retrieved documents: 10**

**Engine 2:**

**Relevant documents retrieved: 2, 3, 4, 6, 7 (5 relevant documents)**

**Total retrieved documents: 10**

**Precision:**

|  |  |  |
| --- | --- | --- |
| Engine 2 | Engine 1 | Rank |
| 0/1 = 0 | 1/1 = 1 | **1** |
| 1/2 = 0.5 | 1/2 = 0.5 | **2** |
| 2/3 = 0.666 | 2/3 = 0.666 | **3** |
| 3/4 = 0.75 | 3/4 = 0.75 | **4** |
| 3/5 = 0.6 | 3/5 = 0.6 | **5** |
| 4/6 = 0.66 | 3/6 = 0.5 | **6** |
| 5/7 = 0.71 | 4/7 = 0.57 | **7** |
| 5/8 = 0.625 | 4/8 = 0.5 | **8** |
| 5/9 = 0.55 | 5/9 = 0.55 | **9** |
| 5/10 = 0.5 | 6/10 = 0.6 | **10** |

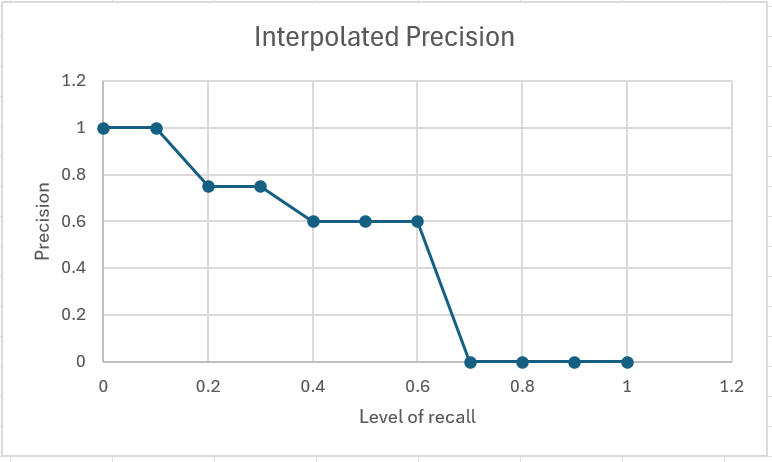
**Recall:**

|  |  |  |
| --- | --- | --- |
| Engine 2 | Engine 1 | Rank |
| 0/10 = 0 | 1/10 = 0.1 | **1** |
| 1/10 = 0.1 | 1/10 = 0.1 | **2** |
| 2/10 = 0.2 | 2/10 = 0.2 | **3** |
| 3/10 = 0.3 | 3/10 = 0.3 | **4** |
| 3/10 = 0.3 | 3/10 = 0.3 | **5** |
| 4/10 = 0.4 | 3/10 = 0.3 | **6** |
| 5/10 = 0.5 | 4/10 = 0.4 | **7** |
| 5/10 = 0.5 | 4/10 = 0.4 | **8** |
| 5/10 = 0.5 | 5/10 = 0.5 | **9** |
| 5/10 = 0.5 | 6/10 = 0.6 | **10** |

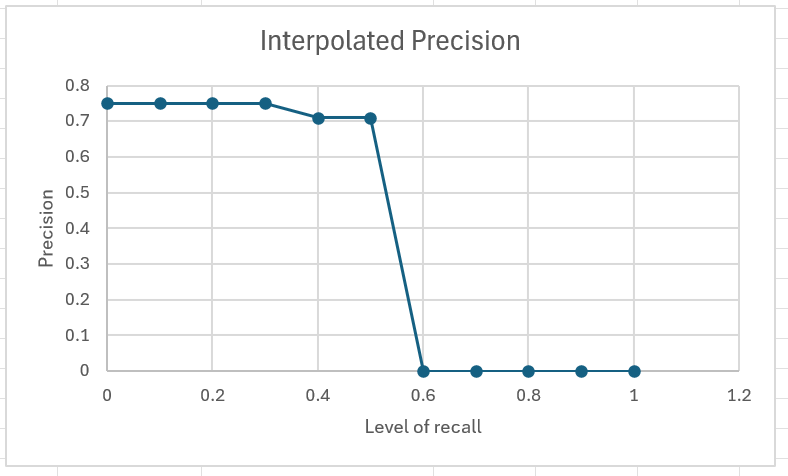
**Interpolated precision:**

**We take each level of recall between 0,0.1…..1 and find the maximum precision value that is related to the greater or equal recall level.**

**Engine 1 – Table and Graph:**

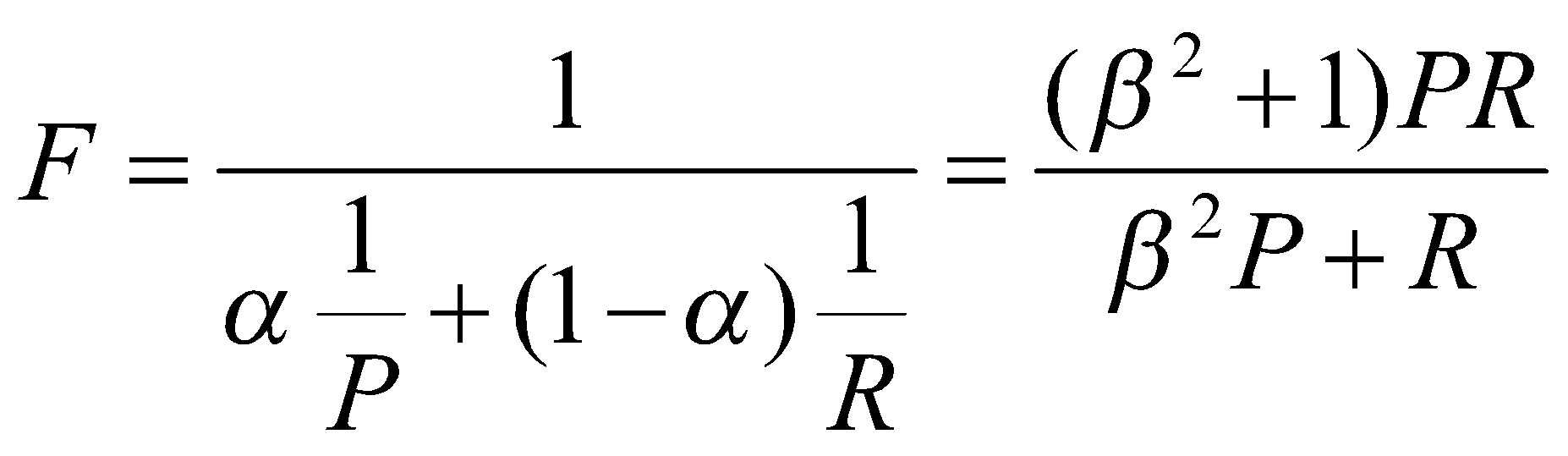
****

|  |  |
| --- | --- |
| Interpolated Precision | Level of recall |
| 1.0 | 0.0 |
| 1.0 | 0.1 |
| 0.75 | 0.2 |
| 0.75 | 0.3 |
| 0.6 | 0.4 |
| 0.6 | 0.5 |
| 0.6 | 0.6 |
| 0 | 0.7 |
| 0 | 0.8 |
| 0 | 0.9 |
| 0 | 1.0 |

**Engine 2 – Table and Graph:**

|  |  |
| --- | --- |
| Interpolated Precision | Level of recall |
| 0.75 | 0.0 |
| 0.75 | 0.1 |
| 0.75 | 0.2 |
| 0.75 | 0.3 |
| 0.71 | 0.4 |
| 0.71 | 0.5 |
| 0 | 0.6 |
| 0 | 0.7 |
| 0 | 0.8 |
| 0 | 0.9 |
| 0 | 1.0 |

**F-measure formula:**



**Engine 1 calculations:**

**Average Precision:**

**Average Recall:**

**F-measure:**

**Engine 2 calculations:**

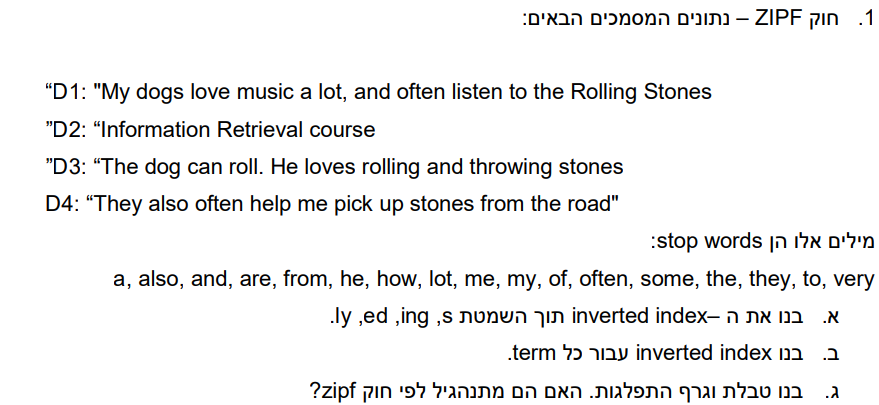
**Average Precision:**

**Average Recall:**

**F-measure:**

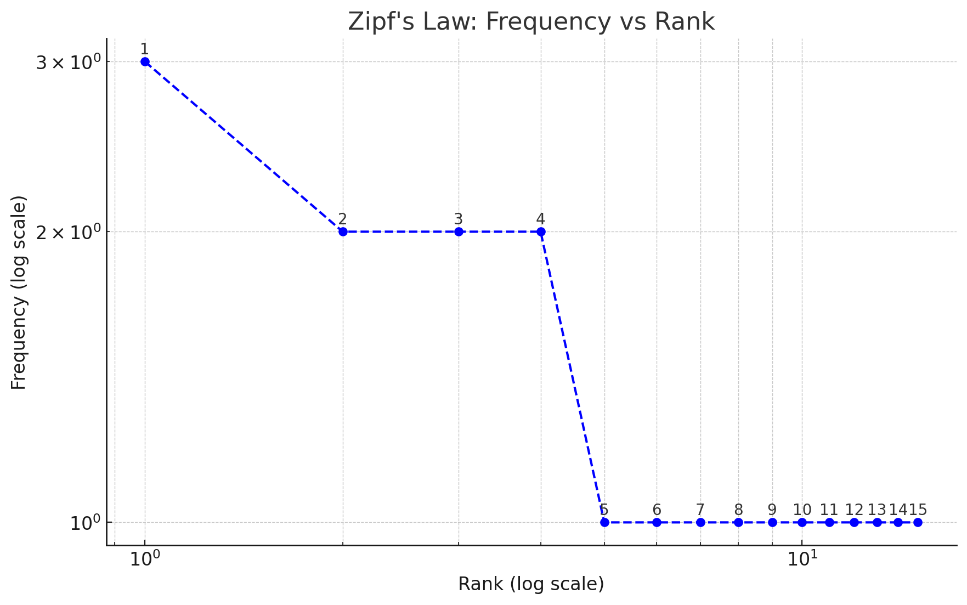
**According to the F-measure results, Engine 1 is slightly better.**

שאלה 3



|  |  |
| --- | --- |
| Term | Documents |
| dog | 1, 3 |
| love | 1, 3 |
| music | 1 |
| listen | 1 |
| roll | 1, 3 |
| stone | 1, 3, 4 |
| information | 2 |
| retrieval | 2 |
| course | 2 |
| can | 3 |
| throw | 3 |
| help | 4 |
| pick | 4 |
| up | 4 |
| road | 4 |

|  |  |
| --- | --- |
| Term | Documents |
| 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 |

לאחר מיון הערכים נקבל:

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

על מנת שהגרף יתנהג לפי חוק Zipfאנו צריכים לראות קו ישר על הגרף, כלומר הקשר בין הדירוג לתדירות המילים הוא קו ישר בגרף לוגריתמי.

אם יש סטיות משמעותיות מהקו הישר אז זה עלול להוביל שהנתונים לא תואמים לחוק Zipf.  
בשאלה זו ישנם סטיות, אך בהינתן עוד terms ייתכן כי הגרף אכן יתנהג לפי חוק Zipf.

שאלה 3: קדם פרויקט –בניית זחלן



הזחלן שאנחנו נבנה הוא לאתר audible, אתר מאוד פופולארי להשמעת ספרים, אנחנו נחקור עם הזחלן נתונים הלא מופיעים בצורה אינטואיטיבית באתר עצמו, אלא נחלץ את נתונים אלה ע"י הזחלן.

**פתרון חלק א':**

השאילתות בהן עסקנו (שאילתות אלו שונו ממטלה מס' 1 משום שכעת על מנת לענות על שלוש השאלות הבאות נדרשנו להביא שאילתות שמניבות תוכן יותר רלוונטי):

1. בהינתן חיפוש שם של ספר באתר, מה תיאור הספרים (Books description) מבין כלל הספרים שחוזרים בשאילתת החיפוש, כאשר ספרים אלו מכילים את כל השדות הקיימים (By, Narrated By, Series, Length, Release date, Language).

הסבר לשאילתא:

בurl הבא באתר:audible <https://www.audible.com/search>

ניתן לערוך חיפוש עפ"י שם של ספר. בעת חיפוש ספר, מגוון ספרים עם מילים דומות בשם הספר ומאפיינים דומים לספר שחיפשנו יחזרו לשאילתת החיפוש.

לחלק מהספרים, חסרים שדות בתיאור הספר (אחד או יותר מהשדות: By, Narrated By, Series, Length, Release date, Language).

שאילתא זו מחזירה את התיאור (Description) של כלל הספרים שחזרו בחיפוש שמכילים את כלל השדות.

1. בהינתן חיפוש שם של ספר באתר,

מהם ערכי השדות [Title, author, overall rating, review count, description]

של הספרים מבין כלל הספרים שחוזרים בשאילתת החיפוש, כאשר ספרים אלו עם דירוג (Rating) גדול או שווה ל4.5 .

הסבר לשאילתא:

בurl הבא באתר:audible <https://www.audible.com/search>

ניתן לערוך חיפוש עפ"י שם של ספר. בעת חיפוש ספר, מגוון ספרים עם מילים דומות בשם הספר ומאפיינים דומים לספר שחיפשנו יחזרו לשאילתת החיפוש.

כל ספר מכיל דירוג (Rating) בין 1-5 שדורג ע"י הקוראים.

שאילתא זו מחזירה את השדות שנרשמו לעיל של כלל הספרים שחזרו בחיפוש שדירוגם גדול או שווה ל4.5 .

1. בהינתן חיפוש שם של ספר באתר,

מהם ערכי השדות [Title, author, overall rating, review count, description]

של הספרים מבין כלל הספרים שחוזרים בשאילתת החיפוש, כאשר ספרים אלו נכתבו ע"י הסופר שכתב את כמות הספרים המרבית ביותר מבין הספרים שחזרו בשאילתת החיפוש .

הסבר לשאילתא:

בurl הבא באתר:audible <https://www.audible.com/search>

ניתן לערוך חיפוש עפ"י שם של ספר. בעת חיפוש ספר, מגוון ספרים עם מילים דומות בשם הספר ומאפיינים דומים לספר שחיפשנו יחזרו לשאילתת החיפוש.

לכל ספר ישנו שדה "שם הסופר" אשר מעיד על כותב הספר.

מבין כלל הספרים שחוזרים בשאילתת החיפוש, מבוצעת בדיקה מי הסופר שכתב את מירב הספרים מבין הספרים שחזרו לשאילתת החיפוש.

שאילתא זו מחזירה את השדות שנרשמו לעיל של כלל הספרים שחזרו בחיפוש אשר נכתבו ע"י הסופר שכתב את מירב הספרים כפי שתואר לעיל.

**פתרון חלק ב':**

נשתמש בBeautifulSoup בPython על מנת לממש את הזחלן.

ישנו EXCEL מצורף להגשה עם Sheets המכיל את התוצאות לשאילתא שביצענו.

את הקוד ביצענו בGoogle Colab: [קישור](https://colab.research.google.com/drive/1aDMceE2Epe3FOM00NNLpi6XUwp4gD1GM?usp=sharing)

\*\* tf-idf חושב עבור שאילתא 1.

\*\* חישוב הPrecision, Recall עבור 2 משתמשים נמצא בEXCEL וחושב עבור שאילתא מספר 1.

קישור לריפו: <https://github.com/yuvalzohar36/DataDiggersHw2>