אחזור מידע – תרגיל בית 2

**מגישים:**

**שרה סאלח 318107448**

**יובל סבאטו 318474269**

**שחף ישראל 209288414**

**שי אוסמו 318370285**

**שאלה 1:**

נתונים:

Retrieved documents (P): 15 (5 relevant + 10 non-relevant)

Relevant documents retrieved (A): 20

נחשב את הprecision ואת הrecall:

Precision = (relevant documents retrieved)/(retrieved documents) = (5/15) = 0.333 = 33.3%

Recall = (relevant documents retrieved)/(total relevant documents) = (5/25) = 0.2=20%

**שאלה 2:**

בהתאם לנתונים:

* **מנוע חיפוש 1**: התוצאות הרלוונטיות מופיעות במיקומים 1, 3, 4, 7, 9, 10 מתוך 10 התוצאות שהוצגו.
* **מנוע חיפוש 2**: התוצאות הרלוונטיות מופיעות במיקומים 2, 3, 4, 6, 7 מתוך 10 התוצאות שהוצגו.

בטבלה, כל שורה מייצגת "חזרה" אחת, כלומר, רשימת תוצאות מסוימת לכל מנוע חיפוש (10 תוצאות בכל פעם).

נציג את התוצאות של and percision recall עבור כל מנוע חיפוש:

Engine 1:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Relevant Retrieved | Retrieved documents | |  | | --- | | **Precision** |  |  | | --- | |  | | | **Recall** | | --- |  |  | | --- | |  | |
| 1 | **1** | **1** | **1/1 = 1** | **1/10 = 0.10** |
| 2 | **1** | **2** | **1/2 = 0.5** | **1/10 = 0.10** |
| 3 | **2** | **3** | **2/3 = 0.67** | **2/10 = 0.2** |
| 4 | **3** | **4** | **3/4 = 0.75** | **3/10 = 0.3** |
| 5 | **3** | **5** | **3/5 = 0.6** | **3/10 = 0.3** |
| 6 | **3** | **6** | **3/6 = 0.5** | **3/10 = 0.3** |
| 7 | **4** | **7** | **4/7 = 0.57** | **4/10 = 0.4** |
| 8 | **4** | **8** | **4/8 = 0.5** | **4/10 = 0.4** |
| 9 | **5** | **9** | **5/9 = 0.56** | **5/10 = 0.5** |
| 10 | **6** | **10** | **6/10 = 0.6** | **6/10 = 0.6** |

Engine 2:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Relevant Retrieved | Retrieved documents | |  | | --- | | **Precision** | | | **Recall** | | --- | |
| 1 | **0** | **1** | **0/1 = 0** | **0/10 = 0** |
| 2 | **1** | **2** | **1/2 = 0.5** | **1/10 = 0.1** |
| 3 | **2** | **3** | **2/3 = 0.67** | **2/10 = 0.2** |
| 4 | **3** | **4** | **3/4 = 0.75** | **3/10 = 0.3** |
| 5 | **3** | **5** | **3/5 = 0.6** | **3/10 = 0.3** |
| 6 | **4** | **6** | **4/6 = 0.67** | **4/10 = 0.4** |
| 7 | **5** | **7** | **5/7 = 0.71** | **5/10 = 0.5** |
| 8 | **5** | **8** | **5/8 = 0.63** | **5/10 = 0.5** |
| 9 | **5** | **9** | **5/9 = 0.56** | **5/10 = 0.5** |
| 10 | **5** | **10** | **5/10 = 0.5** | **5/10 = 0.5** |

על מנת לחשב את ה- Interpolated Precision:

נגדיר 11 רמות Recall קבועות: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.

לאחר מכן, נחשב את ערך ה- Interpolated Precision עבור כל אחת מרמות ה-Recall האלה.

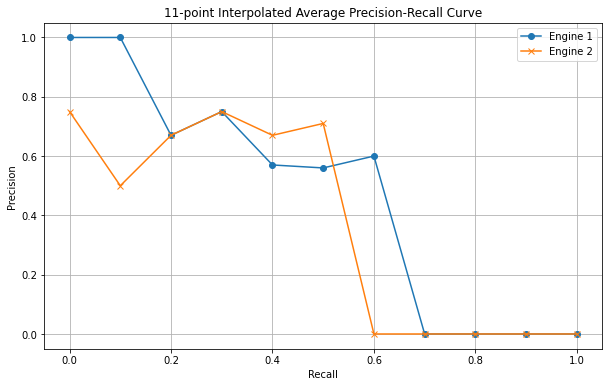
Engine1:

|  |  |  |  |
| --- | --- | --- | --- |
| Interpolated Percision | | **Recall** | | --- |  |  | | --- | |  | |
| 1 | **0** |
| 1 | **0.1** |
| 0.67 | **0.2** |
| 0.75 | **0.3** |
| 0.57 | **0.4** |
| 0.56 | **0.5** |
| 0.6 | **0.6** |
| 0 | **0.7** |
| 0 | **0.8** |
| 0 | **0.9** |
| 0 | **1.0** |

Engine2:

|  |  |  |  |
| --- | --- | --- | --- |
| Interpolated Percision | | **Recall** | | --- |  |  | | --- | |  | |
| 0.75 | **0** |
| 0.5 | **0.1** |
| 0.67 | **0.2** |
| 0.75 | **0.3** |
| 0.67 | **0.4** |
| 0.71 | **0.5** |
| 0 | **0.6** |
| 0 | **0.7** |
| 0 | **0.8** |
| 0 | **0.9** |
| 0 | **1.0** |

**הגרף שהתקבל:**



**חישוב הf-measure:**

**Engine1:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Relevant Retrieved | Retrieved documents | |  | | --- | | **Precision** |  |  | | --- | |  | | | **Recall** | | --- |  |  | | --- | |  | | f-measure |
| 1 | **1** | **1** | **1/1 = 1** | **1/10 = 0.10** | **0.2358** |
| 2 | **1** | **2** | **1/2 = 0.5** | **1/10 = 0.10** | **0.2049** |
| 3 | **2** | **3** | **2/3 = 0.67** | **2/10 = 0.2** | **0.3629** |
| 4 | **3** | **4** | **3/4 = 0.75** | **3/10 = 0.3** | **0.4870** |
| 5 | **3** | **5** | **3/5 = 0.6** | **3/10 = 0.3** | **0.4412** |
| 6 | **3** | **6** | **3/6 = 0.5** | **3/10 = 0.3** | **0.4032** |
| 7 | **4** | **7** | **4/7 = 0.57** | **4/10 = 0.4** | **0.4944** |
| 8 | **4** | **8** | **4/8 = 0.5** | **4/10 = 0.4** | **0.4587** |
| 9 | **5** | **9** | **5/9 = 0.56** | **5/10 = 0.5** | **0.5368** |
| 10 | **6** | **10** | **6/10 = 0.6** | **6/10 = 0.6** | **0.6** |

**Engine2:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Relevant Retrieved | Retrieved documents | |  | | --- | | **Precision** | | | **Recall** | | --- | | f-measure |
| 1 | **0** | **1** | **0/1 = 0** | **0/10 = 0** | **NaN** |
| 2 | **1** | **2** | **1/2 = 0.5** | **1/10 = 0.10** | **0.2049** |
| 3 | **2** | **3** | **2/3 = 0.67** | **2/10 = 0.2** | **0.3629** |
| 4 | **3** | **4** | **3/4 = 0.75** | **3/10 = 0.3** | **0.4870** |
| 5 | **3** | **5** | **3/5 = 0.6** | **3/10 = 0.3** | **0.4412** |
| 6 | **4** | **6** | **4/6 = 0.67** | **4/10 = 0.4** | **0.5390** |
| 7 | **5** | **7** | **5/7 = 0.71** | **5/10 = 0.5** | **0.6167** |
| 8 | **5** | **8** | **5/8 = 0.63** | **5/10 = 0.5** | **0.5761** |
| 9 | **5** | **9** | **5/9 = 0.56** | **5/10 = 0.5** | **0.5368** |
| 10 | **5** | **10** | **5/10 = 0.5** | **5/10 = 0.5** | **0.5** |

הממוצע של מדד ה- F-Measure עבור מנוע חיפוש 1 הוא 0.42249, בעוד שבמנוע חיפוש 2 הממוצע הוא 0.47384. נתונים אלה מראים בבירור שמנוע חיפוש 2 מספק ביצועים טובים יותר ממנוע חיפוש 1.

א. לשם בניית Inverted index נרצה קודם להסיר את הstop words מהמילים במסמכים הנתונים:

שלב 1: הסרת Stop Words

לאחר הסרת מילות הקישור והתוספות הלא נחוצות:

D1: "dogs love music listen Rolling Stones"

D2: "Information Retrieval course"

D3: "dog roll loves rolling throwing stones"

D4: "help pick up stones road"

שלב 2: הסרת סיומות (s, ing, ed, ly)

התוצאה לאחר הסרת הסיומות:

D1: "dog love music listen Roll Stone"

D2: "Information Retrieval course"

D3: "dog roll love roll throw stone"

D4: "help pick up stone road"

בשלב הבא, נבנה את ה-**Inverted Index** בצורה הבאה:

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]

ב. בגלל שפה הדרישה היא לבנות Inverted index לכל term נעשה את אותו התהליך רק בלי הסרת הסיומות s, ing, ed, ly:

dogs: [1]

love: [1]

music: [1]

listen: [1]

rolling: [1, 3]

stones: [1, 3, 4]

information: [2]

retrieval: [2]

course: [2]

dog: [3]

can: [3]

roll: [3]

loves: [3]

throwing: [3]

help: [4]

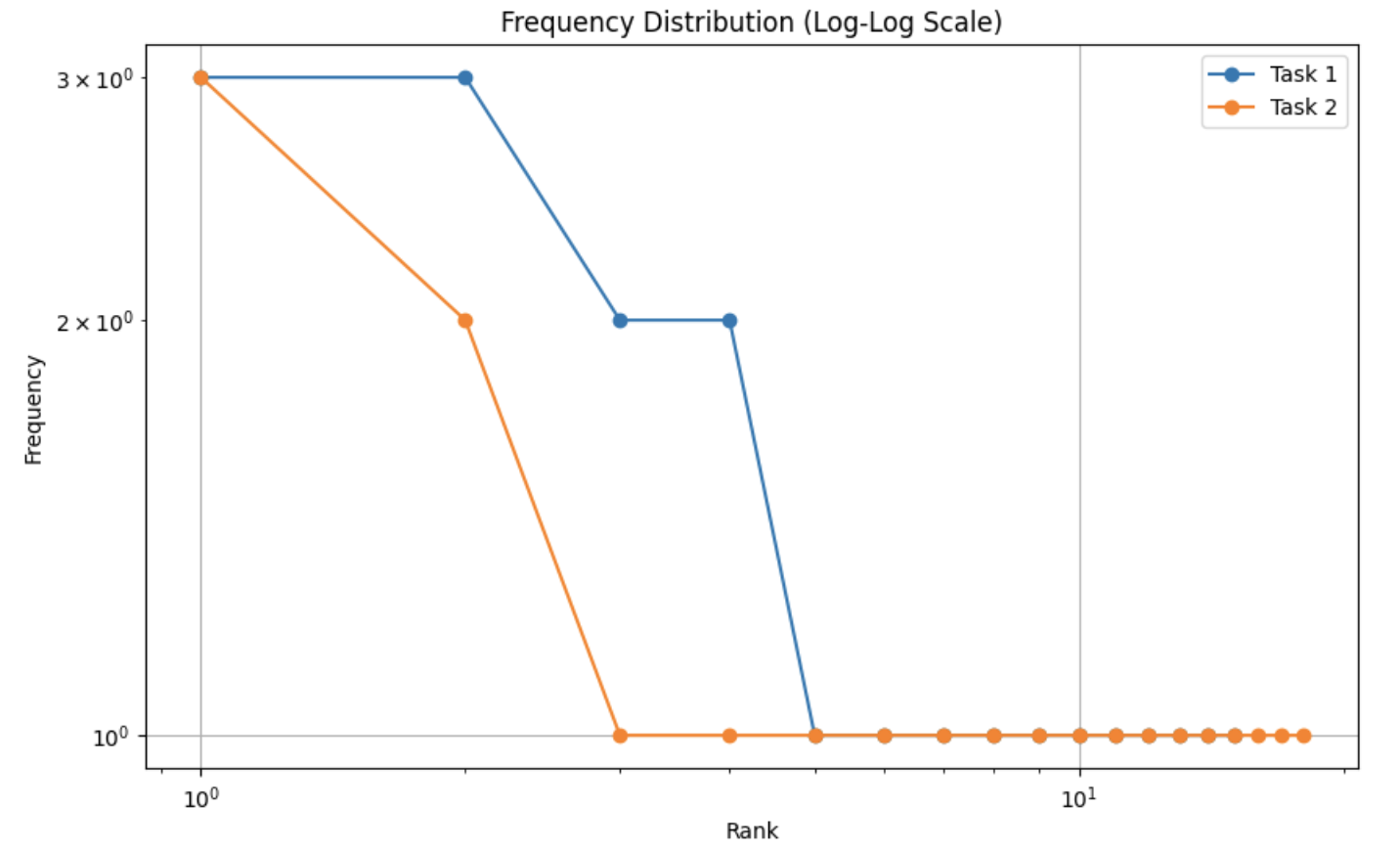
pick: [4]

up: [4]

road: [4]

ג.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | משימה 2  מילים | משימה 2  תדירות | משימה 1  מילים | | **משימה 1 תדירות** | | --- | |
| 1 | **stones** | **3** | **roll** | **3** |
| 2 | **rolling** | **2** | **stone** | **3** |
| 3 | **dogs** | **1** | **dog** | **2** |
| 4 | **love** | **1** | **love** | **2** |
| 5 | **music** | **1** | **music** | **1** |
| 6 | **listen** | **1** | **listen** | **1** |
| 7 | **information** | **1** | **information** | **1** |
| 8 | **retrieval** | **1** | **retrieval** | **1** |
| 9 | **course** | **1** | **course** | **1** |
| 10 | **dog** | **1** | **can** | **1** |
| 11 | **can** | **1** | **throw** |  |
| 12 | **roll** | **1** | **help** |  |
| 13 | **loves** | **1** | **pick** |  |
| 14 | **throwing** | **1** | **up** |  |
| 15 | **help** | **1** | **road** |  |
| 16 | **pick** | **1** |  |  |
| 17 | **up** | **1** |  |  |
| 18 | **road** | **1** |  |  |



תשובה:

לא, הם אינם עומדים בחוק zipf בגלל שהגרף לא מציג קו ישר על סקאלת לוג-לוג.

במקום זאת, ההתפלגות מציגה ירידות חדות ולא אחידות בתדירות המילים.

**שאלה 3:**

הקוד עם קובץ אקסל בתיקייה.

פתרנו כל סעיף בקוד, כאשר פתרון כל סעיף הוא לשונית באקסל.

קישור לגיט:

https://github.com/shahaf5641/IR---Project