אחזור מידע – תרגיל בית 2

עומר זומרשטיין 316439876

ניצן עזרא 208697334

ליאור בוזגלו 314774787

דניאל פריימוביץ 315880963

<https://github.com/NitzanEz/InformationRecovery>

שאלה 1:

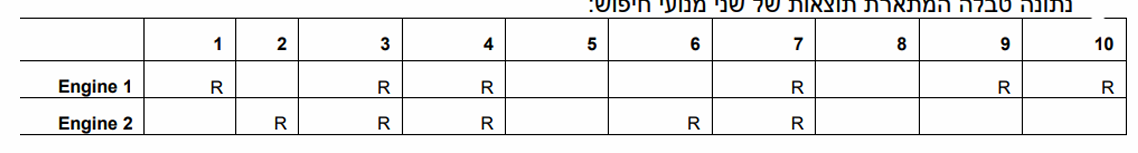
-**precision: Precision יחס המסמכים הרלוונטיים שהמערכת החזירה מתוך כלל המסמכים שהמערכת החזירה:**

**Precision = 0.444**

-**recall יחס המסמכים הרלוונטיים שהמערכת החזירה מתוך כלל המסמכים הרלוונטיים באוסף:**

**Recall = 0.4**

שאלה 2:



נחשב recall ע"פ הנוסחה: Recall = TP / (TP + FN) עבור כל החזרה K.

|  |  |
| --- | --- |
| Engine 1 | Rank |
| P = 1/1 = 1.0  R = 1/10 = 0.1 | **1** |
| P = 1/2 = 0.5  R = 1/10 = 0.1 | **2** |
| P = 2/3 = 0.666  R = 2/10 = 0.2 | **3** |
| P = 3/4 = 0.75  R = 3/10 = 0.3 | **4** |
| P = 3/5 = 0.6  R = 3/10 = 0.3 | **5** |
| P = 3/6 = 0.5  R = 3/10 = 0.3 | **6** |
| P = 4/7 = 0.57  R = 4/10 = 0.4 | **7** |
| P = 4/8 = 0.5  R = 4/10 = 0.4 | **8** |
| P = 5/9 = 0.55  R = 5/10 = 0.5 | **9** |
| P = 6/10 = 0.6  R = 6/10 = 0.6 | **10** |

|  |  |
| --- | --- |
| Engine 2 | Rank |
| P = 0/1 = 0.0  R = 0/10 = 0.0 | **1** |
| P = 1/2 = 0.5  R = 1/10 = 0.1 | **2** |
| P = 2/3 = 0.666  R = 2/10 = 0.2 | **3** |
| P = 3/4 = 0.75  R = 3/10 = 0.3 | **4** |
| P = 3/5 = 0.6  R = 3/10 = 0.3 | **5** |
| P = 4/6 = 0.66  R = 4/10 = 0.4 | **6** |
| P = 5/7 = 0.71  R = 5/10 = 0.5 | **7** |
| P = 5/8 = 0.625  R = 5/10 = 0.5 | **8** |
| P = 5/9 = 0.55  R = 5/10 = 0.5 | **9** |
| P = 5/10 = 0.5  R = 5/10 = 0.5 | **10** |

11 נקודות אינטרפולציה:  
עבור כל רמה של Recall (בין 0.0 ל 1.0), נמצא את הערך Precision הגבוה ביותר שמשויך לRecall שגדול/שווה לאותה רמה.

מנוע 1:

|  |  |
| --- | --- |
| Interpolated Precision | Recall Level |
| 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 | 0.7 |
| 0.0 | 0.8 |
| 0.0 | 0.9 |
| 0.0 | 1.0 |

הגרף של מנוע 1:

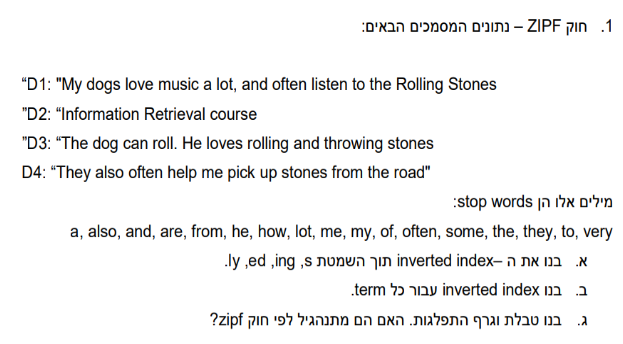
מנוע 2:

|  |  |
| --- | --- |
| Interpolated Precision | Recall Level |
| 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 | 0.6 |
| 0.0 | 0.7 |
| 0.0 | 0.8 |
| 0.0 | 0.9 |
| 0.0 | 1.0 |

הגרף של מנוע 2:

f-measure עם :

ניתן לראות כי ע"פ חישובי F-measure, מנוע 1 עדיף על פני מנוע 2.



פתרון:

א+ב.

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

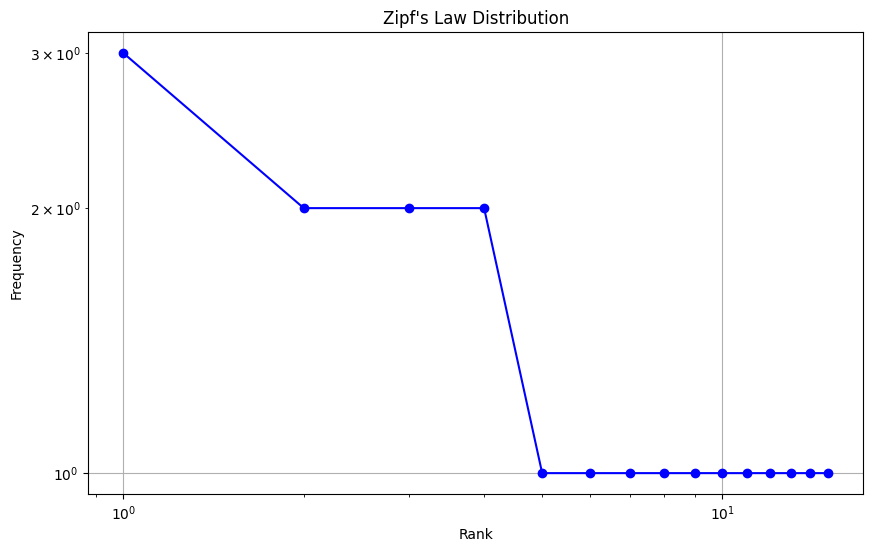
ג. בדיקת חוק Zipf:

טבלת תדירות בהתבסס על הinverted index:

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

לאחר מיון:

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



ניתן לראות כי הנקודות יוצרות קו ישר ומכך ניתן להסיק כי **הנתונים עוקבים אחרי חוק Zipf**.

שאלה 3 – קדם פרויקט בניית זחלן

**(התשובות מופיעות באקסל Q1- answers )**

אנו מתמקדים בבניית זחלן לאתר האינטרנטי **Reddit**, אתר פורומים מוביל עבור מגוון רחב של נושאים. כל אדם שנכנס יוצר לעצמו משתמש, יכול לבצע חיפושים ולהגיב על פוסטים, בהתבסס על מידע זה אנו ניצור שאילתות.

קבצי הקוד עם השאילתות, קבצי האקסל עם תוצאות השאילתות וקובץ המפרט את המענה על שאר השאלות נמצאות בתיקייה.

כעת נגיד במדויק את השאילתות אשר עליהן נעבוד ונחקור:

* איזה פוסטים על חתולים מצחיקים קיימים מהחודש האחרון תחת קטגוריה HOT. (בreddit אין אופציה לבצע את החיפוש הזה באופן ישיר)
* מה הם הפוסטים על חתולים מצחיקים שעוררו הכי הרבה מחלוקות היום. (בreddit אין אופציה לבצע את החיפוש הזה באופן ישיר)