

אחזור מידע- תרגיל בית 2

אחזור מידע תרגיל 2- Query Kings

קישור לגיטהאב: [**https://github.com/guybanbo/IR-Query-Kings.git**](https://github.com/guybanbo/IR-Query-Kings.git)

שאלה 1:

יש 5 מסמכים רלוונטיים ו10 מסמכים לא רלוונטיים שהוחזרו . יש סה"כ 20 מסמכים רלוונטיים.

אז TP=5

FP=10

FN=15

כי יש 15 רלוונטיים שלא הוחזרו

recall=TP/(TP+FN)=5/(5+15)=5/20=0.25

precision=TP/(TP+FP)=5/15=1/3

שאלה 2:

א.

מנוע 1:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| precision | 1/1 | 1/2 | 2/3 | 3/4 | 3/5 | 3/6 | 4/7 | 4/8 | 5/9 | 6/10 |
| recall | 1/10 | 1/10 | 2/10 | 3/10 | 3/10 | 3/10 | 4/10 | 4/10 | 5/10 | 6/10 |

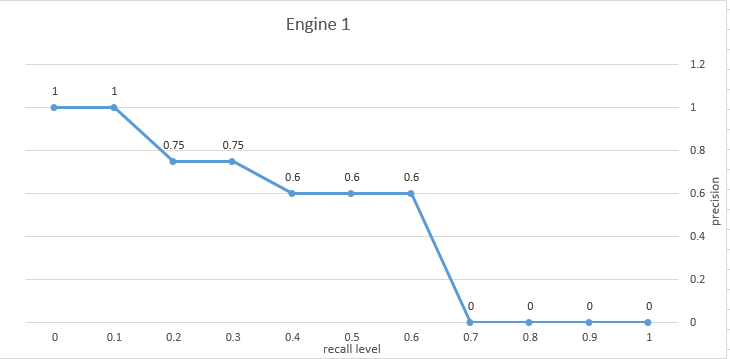
מנוע 2:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| precision | 0 | 1/2 | 2/3 | 3/4 | 3/5 | 4/6 | 5/7 | 5/8 | 5/9 | 5/10 |
| recall | 0 | 1/10 | 2/10 | 3/10 | 3/10 | 4/10 | 5/10 | 5/10 | 5/10 | 5/10 |

אינטרפולציה:

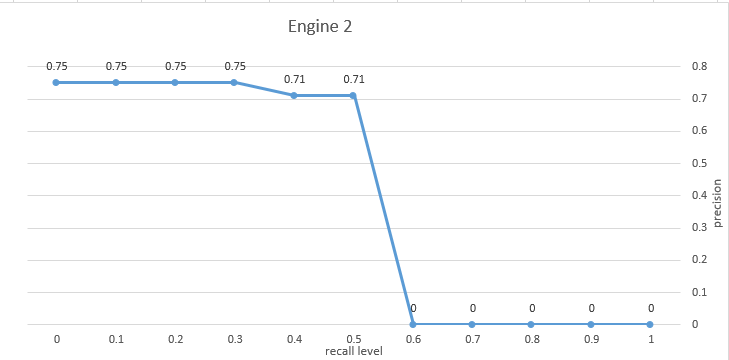
מנוע 1:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | 0 | recall level |
| 0 | 0 | 0 | 0 | 0.6 | 0.6 | 0.6 | 0.75 | 0.75 | 1 | 1 | interpolated precision |



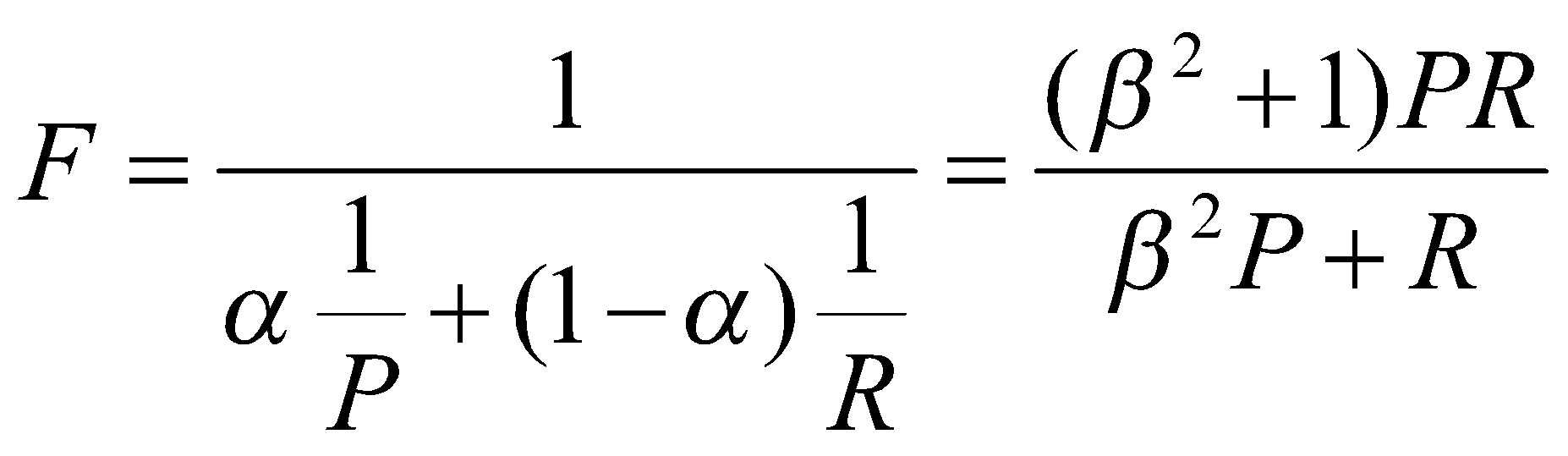
מנוע 2:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | 0 | recall level |
| 0 | 0 | 0 | 0 | 0 | 0.71 | 0.71 | 0.75 | 0.75 | 0.75 | 0.75 | interpolated precision |



F-measure:

מנוע 1:



p=r=6/10

B=0.75

F=((0.75^2+1)\**0.6\**0.6)/(0.6\*0.75^2+0.6)=0.56/0.94=0.6

מנוע 2:

p=r=5/10

B=0.75

F=((0.75^2+1)\**0.5\**0.5)/(0.5\*0.75^2+0.5)=0.39/0.78=0.5

מנוע 1 טוב יותר כי הf-measure שלו יותר גבוה.

**סעיף ב':**

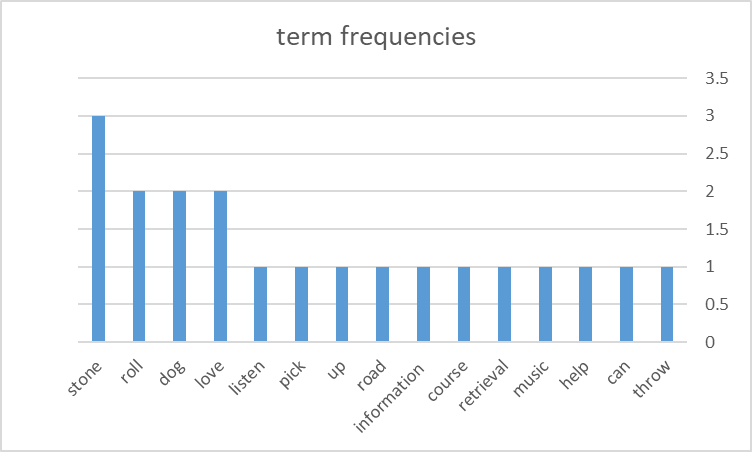
ב.א

האינדקס:

|  |  |
| --- | --- |
| 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 |
| 2 | can |
| 3 | throw |
| 4 | help |
| 4 | pick |
| 4 | up |
| 4 | road |

ג. בנו טבלת תדירות לכלל המילים. האם ההתפלגות מתנהגת לפי חוק zipf?

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



ניתן לראות שיש מילה אחת (stone) המופיעה בתדירות גבוהה יחסית ואפשר לראות שרוב המילים המופיעות בתדירות אחת, ויש מספר בינוני של מילים המופיע בתדירות בינונית. מתוך הנתונים ניתן להסיק שזה עומד בחוק zipf

**שאלה 3- זחלן**

בנינו זחלן בפייתון שסקר את האתר<http://www.kizur.co.il/home.php>, החל מדף הבית. במהלך הפעולה נאספו למעלה מ-600 קישורים, כולל דפי קטגוריות ודפי קיצורים שונים.

הגענו לעומק מסוים שאפשר לנו לקבל תשובות מספקות לשאילתות השונות. לאחר מכן, ביצענו סינון ממוקד לקישורים שזוהו כרלוונטיים לכל שאילתא, מתוכם חילצנו את הקיצורים וההגדרות השונות.

**הקודים של הזחלן והשאילתות מצורפים במחברת .Google colab**

**השאילתות שבחרנו:**

**שאילתה 1- מהו הממוצע של מספר המילים להגדרה של קיצור בקטגוריה של ארגונים ומוסדות?**

אופן הביצוע- סיננו את הקישורים שהזחלן החזיר ובחרנו קישורים של הדפים של הקטגוריה הנ"ל, אלו דפים שמתחילים ב: [**http://www.kizur.co.il/search\_group.php?group=10**](http://www.kizur.co.il/search_group.php?group=10)

מתוך הדפים חילצנו את ההגדרות והקיצורים שנמצאים בתגיות HTML רלוונטיות.

**שאילתה 2- מהם המילים הכי נפוצות בהגדרות של קיצורים עבור הקטגוריה כלכלה?**

אופן הביצוע- סיננו את הקישורים שהזחלן החזיר, ובחרנו קישורים של הדפים של הקטגוריה הנ"ל, אלו דפים שמתחילים ב:

<http://www.kizur.co.il/search_group.php?group=14>

מתוך הדפים חילצנו את ההגדרות והקיצורים שנמצאים בתגיות HTML רלוונטיות.

**שאילתה 3- מה מספר ההגדרות הממוצע לקיצור**

אופן הביצוע:

בחרנו מתוך הקישורים שהזחלן אסף דפים של קיצורים. הקישורים של דפים מסוג זה מתחילים ב<http://www.kizur.co.il/search_word.php?abbr>,

דפים אלה מכילים גם הגדרות שלא מסווגות לקטגוריות ובאמצעותם אפשר לקבל תוצאה יותר מדויקת עבור השאילתה.

1.חישבנו אינדקס עבור כל שאילתא, הדפים שנסרקו וההגדרות שחולצו מהם עבור כל שאילתא נמצאים בקבצי אקסל מצורפים Query 1, Query 2, Query 3.

בכל קובץ יש גיליון של inverted index.

עבור 2 השאילתות הראשונות מספר מזהה של מסמך זה המספר שלו בתוך הקטגוריה באתר, ועבור השאילתא השלישית המסמכים ממוספרים לפי סדר החזרה שלהם עבור השאילתא.

צירפנו גם קובץ אקסל pages עם כל הקישורים שהזחלן אסף.

2.בחרנו לחשב TF-IDF עבור שאילתא 1. התוצאות מצורפות בקובץ Query1.xlsx בגיליון TF-IDF

3. עבור שאילתה 1:  
10 התוצאות שהוחזרו רלוונטיות כי מכילות רשימות של קיצורים בקטגוריה הרלוונטית.  
נתנו לקרובי משפחה להסתכל על הדפים, הם החליטו על הרלוונטיות שלהם לפי הקיצורים שנאספו מהדפים.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **9** | **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** |  |
|  |  | R | R | R | R | R | R |  | R | 1 |
| R | R | R | R | R | R |  |  | R | R | 2 |

**עבור משתמש 1:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **9** | **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** |  |
| 7/10 | 7/9 | 7/8 | 6/7 | 5/6 | 4/5 | 3/4 | 2/3 | 1/2 | 1/1 | **Precision** |
| 7/10 | 7/10 | 7/10 | 6/10 | 5/10 | 4/10 | 3/10 | 2/10 | 1/10 | 1/10 | **Recall** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
| 8/10 | 7/9 | 6/8 | 5/7 | 4/6 | 3/5 | 2/4 | 2/3 | 1/2 | 1/1 | Precision |
| 8/10 | 7/10 | 6/10 | 5/10 | 4/10 | 3/10 | 2/10 | 2/10 | 1/10 | 1/10 | Recall |

**עבור משתמש 2:**