אחזור מידע – תרגיל בית 2 – Optimizers Outlaws

שאלה 1:

נתון בשאלה כי המערכת החזירה 8 מסמכים רלוונטיים וגם 10 מסמכים שלא רלוונטיים. אוסף כל המסמכים במאגר הוא 100 שמתוכם גם 20 מסמכים רלוונטיים בשבילנו.

המערכת החזירה ס"ה מסמכים. נחשב את ה-precision ו-recall.

​

נחשב לפי הנוסחאות ונקבל:

שאלה 2:

חלק א:

לפי הטבלה המתוארת בשאלה, כל אחד ממנועי החיפוש החזיר תוצאות שונות.

מנוע ראשון החזיר את המסמכים 1,3,4,7,9,10 כלומר 6 מסמכים רלוונטיים מתוך 10

מנוע שני החזיר את המסמכים 2,3,4,6,7 כלומר 5 מסמכים רלוונטיים מתוך 10.

נתחיל בלחשב את ה- Precision, Recall לכל שורה בטבלה (כלומר עבור כל מנוע חיפוש בנפרד).

:Precision, Recallחישוב

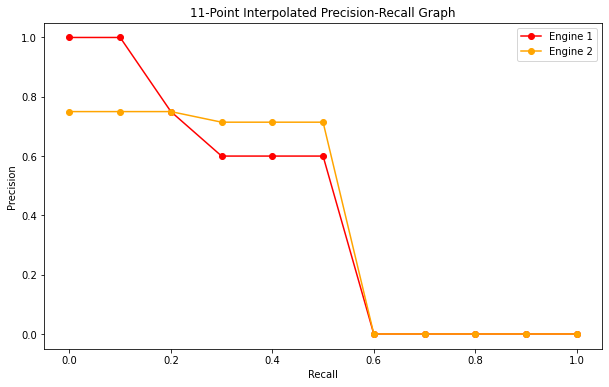
Engine 1:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Engine 2:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

נציג גרף לנקודות שחישבנו עבור שני המנועים:



נחשב כעת את ה-MEASURE -F:

Engine 1:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Engine 2:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

תשובה:

ניתן לראות כי מנוע 2 הוא יותר טוב ממנוע הראשון כי לפי החישובים של (MAP – Mean Average Precision) ו- F-Measure, קיבלנו תוצאות יותר טובות ממנוע 1 ומרות שהחישובים יותר מידי קרובים אך החישובים היו לטובת מנוע 2.

חלק ב:

חוק ZIPF:

The given documents are:

D1: "My dogs love music a lot, and oftren listen to the Rolling Stones"

D2: "Information Retrieval course"

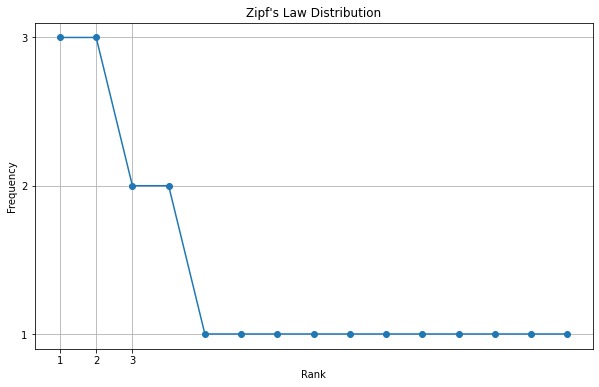
D3: "The dog can roll. He loves rolling and throwing stones"

D4: "They also often help me pick up stones from the road"

The terms in each document is:

The inverted index after removing (ly, ed, ing, s) and removing stopwords:

Term frequency:



חוק Zipf אינו מתאים למערך נתונים זה מכיוון שהתדרים מקובצים, ואין ירידה ברורה ותלולה בתדירות ככל שהדרגות עולות. כמו כן, מערך הנתונים קטן מכדי שהחוק יופיע בבירור מכיוון שכדי שחוק Zipf יעבוד טוב יותר, נצטרך קורפוס גדול יותר עם מגוון רחב יותר של תדירויות מונחים, כאשר המונחים השכיחים ביותר עולים בבירור על המונחים הפחות תכופים באופן יחסי. .

שאלה 3:

אנחנו רוצים להחזיר את כל המשחקים המתאימים לדרישות המשתמש לפי סוגים שונים של חיפוש לדוגמה:

מפרט מחשב מומלץ, דירוג המשחק, קטגוריית משחק וכו..

דוגמאות לשאילתות שרוצים לעסוק בהם הם:

1. משחקים מומלצים לכרטיס מסך RTX 4070 TI
2. משחקים לפי דירוג. כמו משחקים עם דירוג POSITVEומעלה או דירוג שלילי
3. משחקי FPS עם דירוג VERY POSITIVE

בינתיים אנחנו חושבים לעשות את הפילטרינג בטבלה כך שבכל שורה ישמר לנו כמפתח שם המשחק ואז לעבד את המסמך של המשחק ולשמור את כל הדאטא הרלוונטי כמו מפרט מומלץ/מפרט מינימלי, דירוג, קטגוריה וכו...

דוגמה לשאילתה מסוג 1:

כאן לדוגמאה, אנחנו מבצעים חיפוש לפי סוג מערכת הפעלה וכמות זיכרון פנויה וסוג כרטיס מסך , אנחנו מצפים מהזחלן שיחזיר לנו את כל המשחקים שמתאימים למפרט הנ"ל.

דוגמה לשאילתה מסוג 2:

בשאילתה הזו אנחנו מבצעים חיפוש לפי דירוג המשחק באתר שיחזיר את כל המשחקים עם רמת דירוג שבשאילתה .

דוגמה לשאילתה מסוג 3:

בשאילתה הזו אנחנו מבצעים חיפוש לפי קטגוריית משחק בליווי של רמת דירוג, כאן בעצם המטרה שלנו לעשות חיפוש כפול של סוג משחק רצוי בנוסף לדירוג גבוה של המשחק.

\*\*הערה:

* הקוד מבצע חישוב ל-TF-IDF לכל חיפוש רלוונטי שהוא מצא ומציג את הערך. ניתן לראות את הערכים בקובץ ה-Excel שהקוד מפיק אחרי כל הרצה.
* בקובץ ה-Excel שמור את כל התוצאות עבור החיפוש וגם התוצאות מוצגים ב-output של הקוד.

**Precision & Recall for the first 10 relevant results:**

הרצנו את השאילתה הבאה:

חישוב ה- Precision & Recall ניתן על שני סטודנטים אשר סיווגו את התוצאות הרלוונטיות של הזחלן בצורה הבאה:

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

חישוב Recall&Precision :

Student 1:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Student 2:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

קישור לתקיה של גיט: [GitHub Repository For HW2](https://github.com/oRABiiA/Data-Retrieval)