אחזור מידע – תרגיל בית 2 – חורף התשפ"ה

Leonid Vyushkov 327312526

Yotam Gilad 315819599

Tomer Ben Lolo 318318987

Yaacov Shitrit 304959422

שאלה 1:

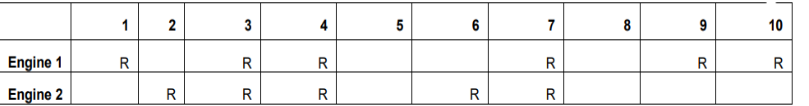
מערכת יש מחיפוש 5 מסמכים רלוונטיים ו-10 מסמכים לא רלוונטיים.

ידוע כי באוסף ישנם עוד מסמכים, מתוכם 20 מסמכים רלוונטיים.

מהו precision ומה recall?

שאלה 2:

א. נתונה טבלת תוצאות של שני מנועי חיפוש:



בהנחה שבקורפוס יש 500 מסמכים, 10 מהם רלוונטיים לשאילתה הנתונה.

הציגו טבלה עם Precision,recall לאחר כל חזרה.

הציגו טבלה וגרף של 11 נקודות אינטרפולציה.

חשבו f-measure עם β = 0.75, איזה מנוע טוב יותר?

טבלה מנוע 1:

|  |  |  |  |
| --- | --- | --- | --- |
| Retrieved | Relevant | Precision | Recall |
| 1 | 1 | 1.0 | 0.1 |
| 2 | 1 | 0.5 | 0.1 |
| 3 | 2 | 0.667 | 0.2 |
| 4 | 3 | 0.75 | 0.3 |
| 5 | 3 | 0.6 | 0.3 |
| 6 | 3 | 0.5 | 0.3 |
| 7 | 4 | 0.571 | 0.4 |
| 8 | 4 | 0.5 | 0.4 |
| 9 | 5 | 0.556 | 0.5 |
| 10 | 6 | 0.6 | 0.6 |

טבלה מנוע 2:

|  |  |  |  |
| --- | --- | --- | --- |
| Retrieved | Relevant | Precision | Recall |
| 1 | 0 | 0.0 | 0.0 |
| 2 | 1 | 0.5 | 0.1 |
| 3 | 2 | 0.667 | 0.2 |
| 4 | 3 | 0.75 | 0.3 |
| 5 | 3 | 0.6 | 0.3 |
| 6 | 4 | 0.667 | 0.4 |
| 7 | 5 | 0.714 | 0.5 |
| 8 | 5 | 0.625 | 0.5 |
| 9 | 5 | 0.556 | 0.5 |
| 10 | 5 | 0.5 | 0.5 |

*ניתן לראות כי המנוע חיפוש 1 יותר טוב*   *כי הf-measure שלו יותר גבוהה.*

*נציג כעת טבלה טבלה וגרף עבור 11 נקודות אינטרפולציה :*

*כדי לבנות את הטבלה נבחר נקודות קבועות של RECALL באופן קבוע [0.1,0.2,. . . ,1.0]*

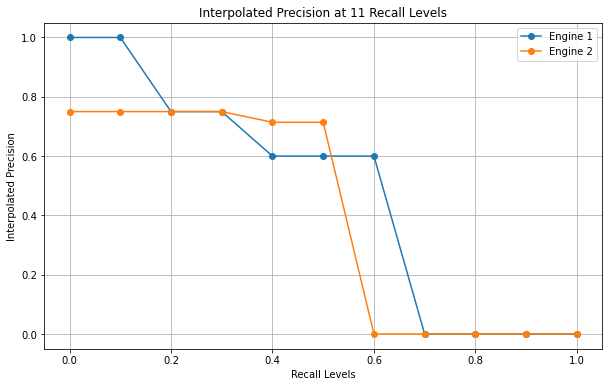
*עבור האינטרפולציה של נקודות הPrecision נבחר רת הגבוהה ביותר עבור כל Recall קבוע.*

*עבור מנועה 1:*

|  |  |
| --- | --- |
| Recall | Precision |
| 0 | 1 |
| 0.1 | 1 |
| 0.2 | 0.75 |
| 0.3 | 0.75 |
| 0.4 | 0.6 |
| 0.5 | 0.6 |
| 0.6 | 0.6 |
| 0.7 | 0 |
| 0.8 | 0 |
| 0.9 | 0 |
| 1 | 0 |

*עבור מנוע 2:*

|  |  |
| --- | --- |
| Recall | Precision |
| 0 | 0.75 |
| 0.1 | 0.75 |
| 0.2 | 0.75 |
| 0.3 | 0.75 |
| 0.4 | 0.714 |
| 0.5 | 0.714 |
| 0.6 | 0 |
| 0.7 | 0 |
| 0.8 | 0 |
| 0.9 | 0 |
| 1 | 0 |



ב. חוק ZIPF - נתונים המסמכים הבאים:

D1: "My dogs love music a lot, and often 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"

מילים אלו הן stop words:

a, also, and, are, from, he, how, lot, me, my, of, often, some, the, they, to, very

א. בנו את ה-inverted index תוך השמטת s, ing, ed, ly.

|  |  |
| --- | --- |
| Term | Document IDs |
| dog | D1, D3 |
| love | D1, D3 |
| music | D1 |
| listen | D1 |
| roll | D1, D3 |
| stone | D1, D3, D4 |
| information | D2 |
| retrieval | D2 |
| course | D2 |
| can | D3 |
| throw | D3 |
| help | D4 |
| pick | D4 |
| up | D4 |
| road | D4 |

שימו לב - בנו inverted index עבור כלל terms (החליטו מהם ה -terms וכיצד לשמור אותם)

|  |  |
| --- | --- |
| Term | Document IDs |
| Dogs | D1 |
| Love | D1 |
| Music | D1 |
| Listen | D1 |
| Rolling | D1,D3 |
| Stones | D1,D3,D4 |
| Information | D2 |
| Retrieval | D2 |
| Course | D2 |
| Dog | D3 |
| Can | D3 |
| Roll | D3 |
| Loves | D3 |
| Throwing | D3 |
| Help | D4 |
| Pick | D4 |
| up | D4 |
| Road | D4 |

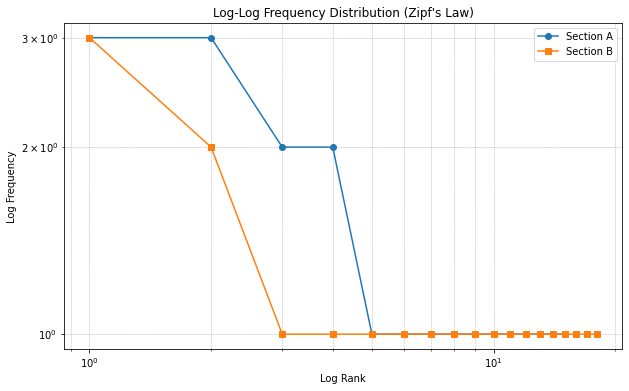
ג. בנו טבלת תדירות לכלל המילים. האם ההתפלגות מתנהגת לפי חוק zipf?

בלי הסיומות

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

**עם הסיומות:**

|  |  |
| --- | --- |
| Term | Frequency |
| Stones | 3 |
| Rolling | 2 |
| Help | 1 |
| Pick | 1 |
| up | 1 |
| Road | 1 |
| Dog | 1 |
| Can | 1 |
| Roll | 1 |
| Loves | 1 |
| Throwing | 1 |
| Information | 1 |
| Retrieval | 1 |
| Course | 1 |
| Dogs | 1 |
| Love | 1 |
| Music | 1 |
| Listen | 1 |

ניתן לראות מהגרף כי מגרף התדירות של שני הסעיפים הוא מתנהג באופן מעט דומה לzipf law. ניתן לראות כי לסעיף השני בו לקחנו את כל הterms ההתנהגות קרובה יותר לzipf law אך לעומת זאת מהסעיף בו מורידים סיומות יש יותר סטיות מהzipf law.

שאלה 3: קדם פרויקט –בניית זחלן

הריצו את החיפוש והחזירו את התוצאות בקובץ אקסל, עם כל השדות הרלוונטיים.

כעת, עליכם לבנות אינדקס לתוצאות שחזרו.

1. בנו inverted index ל-15 מילים הנפוצות ביותר שחזרו. במידה והמילים הנפוצות מופיעות ביותר מ-20 דפים, התייחסו רק ל- 20 הדפים הראשונים שחזרו.
2. בחרו את אחת השאילתות בה עשיתם שימוש, וחשבו tf-idf של המושגים בשאילתא, על פני כל הדפים שחזרו. פרטו את כל שלבי החישוב.
3. הציגו את 10 התוצאות הראשונות שחזרו עבור השאילתה ל-2 משתמשים שונים. בקשו מכל משתמש להחליט עבור כל תוצאה האם היא רלוונטית. חשבו precision ו-recall.

את כל התשובות עבור שאלה זו ניתן לראות במחברת colab:

<https://colab.research.google.com/drive/1DR543ZUsVgeIPpm6QYsnGnfT9BPG7lua?usp=sharing>

קישור עבור הריפו:

<https://github.com/MrLegiongame/Royal-Searchers>