****

מטלה 1-Query Kings

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

שאלה 1: מודלים לאחזור מידע

* (4a).a. Draw the term-document incidence matrix for this document collection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Doc1** | **Doc2** | **Doc3** | **Doc4** |
| **breakthrough** | 1 | 0 | 0 | 0 |
| **drugs** | 1 | 1 | 0 | 0 |
| **for** | 1 | 0 | 1 | 1 |
| **schizophrenia** | 1 | 1 | 1 | 1 |
| **new** | 0 | 1 | 1 | 1 |
| **approach** | 0 | 0 | 1 | 0 |
| **treatment** | 0 | 0 | 1 | 0 |
| **of** | 0 | 0 | 1 | 0 |
| **hopes** | 0 | 0 | 0 | 1 |
| **patients** | 0 | 0 | 0 | 1 |

* (4a).b Draw the inverted index representation for this collection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **breakthrough** | 1 |  | | |
| **drugs** | 1 | 2 |  | |
| **for** | 1 | 3 | 4 |  |
| **schizophrenia** | 1 | 2 | 3 | 4 |
| **new** | 2 | 3 | 4 |  |
| **approach** | 3 |  | | |
| **treatment** | 3 |  | | |
| **of** | 3 |  | | |
| **hopes** | 4 |  | | |
| **patients** | 4 |  | | |

(4b)For the document collection shown in Exercise 4a, what are the returned results for these queries?

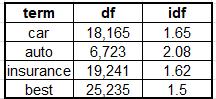
* “schizophrenia and drugs”: doc1 +doc2
* “for and not (drugs or approach)”: doc4

**4CA.**

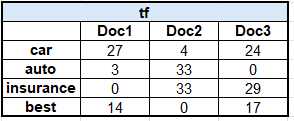
כן, אפשר לעבור על הרשימות של 2 המונחים ולדלג על המסמכים המשותפים ל2 הרשימות.

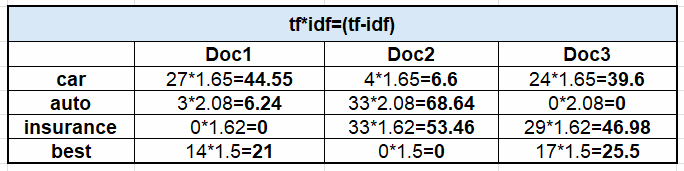
**4CB**. לא, אי אפשר מכיוון שכדי לקבל את החלק של not caesar בשאילתת OR נצטרך לעבור במקרה הגרוע על כל רשימות המסמכים ואז הסיבוכיות תהיה O(N) כאשר N זה מספר המסמכים הכולל.

שאלה 2:



Total documents: 806,791





**שאלה 3:**

כתובת האתר: [http://www.kizur.co.il/home.php\](http://www.kizur.co.il/home.php%5C)

העיסוק המרכזי של אתר קיצור הוא קיצורים בשפה העברית. המידע הזמין באתר כולל חלוקה של קיצורים לקטגוריות, קיצורים שונים וראשי תיבות בשפה העברית וההגדרות השונות שלהם. בנוסף, גם ראשי תיבות מלועזית נכללים באתר.

שאילתות מעניינות:

1. **מהו הממוצע של מספר המילים להגדרה של קיצור עבור כל קטגוריה?**

פרטי המידע הנדרשים:

● אורך הטקסט של כל הגדרה באתר (במילים).

● מספר ההגדרות של קיצורים הזמינים באתר.

● קטגוריות או נושאים של התקצירים.

**ב. מהם המילים הכי נפוצות בהגדרות של קיצורים עבור כל קטגוריה?**

**פרטי המידע הנדרשים:**

● רשימת כל הקיצורים באתר.

● ההגדרות המלאות של כל קיצור.

● קטגוריות או נושאים של התקצירים.

**ג. מה מספר ההגדרות הממוצע לקיצור?**

פרטי המידע הנדרשים:

1. רשימת כל הקיצורים באתר.

2. מספר ההגדרות לכל קיצור.

3. מספר הקיצורים הכולל באתר.