אחזור מידע – תרגיל בית 1

קבוצה – Optimizer Outlaws

שאלה 1:

1. המודל הבוליאני:

Question 4a:

The given documents in the exercise are:

Doc1: breakthrough drugs for schizophrenia.  
Doc 2: new schizophrenia drugs.  
Doc 3: new approach for treatment of schizophrenia.  
Doc 4: new hopes for schizophrenia patients.

First we identify all the terms in all the documents:

Terms breakthrough, drugs, for, schizophrenia, new, approach, treatment, of, hopes, patients.

The term-document matrix is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Doc 1** | **Doc 2** | **Doc 3** | **Doc 4** |
| **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 |

Now we implement the inverted index for the collection:

Question 4b:

1. For the given Query "schizophrenia and drugs", the returned results are:
2. For the given Query " for and not (drugs or approach)", the returned results are:

Question 4c:

1. Query: Brutus AND NOT Caesar

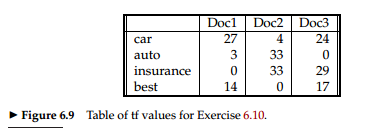
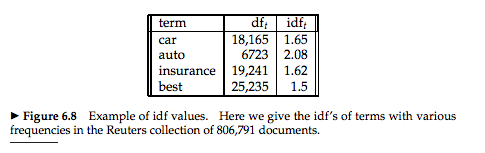
**Yes**, the intersection can be processed in time.

Explanation: We retrieve the postings lists for **Brutus** and **Caesar**. Then we can traverse both lists simultaneously and for every document d in Brutus, check if it is not in Caesar and include it in the result.   
In the worst case we traverse both lists entirely, so this operation still takes  
 time.

1. Query: Brutus OR NOT Caesar

**No,** the operation cannot be achieved in .

Explanation: We retrieve the postings lists for **Brutus** and **Caesar**. Then we compute the set of all documents in the collection (D). We calculate NOT Caesar as (all documents except those in P2). We take the union of P1 and .  
Calculating requires scanning through all documents, which depends on the total number of documents in the collection is (|D|), then the time complexity will be .

1. tf-idf:  
   נתון:

נחשב ערכי TF-IDF לכל TERM בכל Document:

חישובנו את הערכים של בעזרת הנוסחה:

|  |  |  |  |
| --- | --- | --- | --- |
| **Doc3** | **Doc2** | **Doc1** | **Term** |
|  |  |  | **car** |
|  |  |  | **auto** |
|  |  |  | **insurance** |
|  |  |  | **best** |

שאלה 3:

1.     מהו תחום העיסוק המרכזי של האתר? מהו המידע הזמין למשתמשי האתר? ענו בפסקה אחת. צרפו את הקישור לאתר.

**STEAM.com** <https://store.steampowered.com/>

אתר משחקי מחשב ותוכנות שונות, המידע הזמין הוא הכל על המשחק: פירוט, קטגוריה, דרישות…  
האתר מציע שלל משחקים שונים ומציג את המידע בצורה יפה אבל יש כמה דברים שחסרים במנוע החיפוש של האתר כמו: חיפוש לפי כרטיס מסך מסוים, חיפוש לפי מעבד וכו...

2.     רשמו שלוש שאילתות מעניינות שהייתם רוצים לקבל עליהן תשובה באתר, והאתר אינו **עונה עליהן** כעת. עבור כל שאילתא כזו, ציינו מהם פרטי המידע הנדרשים לצורך מענה על השאילתא.

שאילתה 1: משחקים עבור כרטיס מסך RTX-2080TI – חיפוש לפי חומרה

שאילתה 2: משחקי MMORPG (קטגוריה למשחקי מחשב)

שאילתה 3: משחקים שהדרוגים שלהם הוא "VERY POSITIVE"

שאלה 2:

קישור לתקיה של Github: <https://github.com/oRABiiA/Data-Retrieval-Course>