**תרגיל בית 1:**

**ישראל שושן 318740834**

**דוד אסולין 205908569**

**עמית פרץ 209054303**

**יובל הילאי 318609237**

[**GitHub**](https://github.com/AmitPerets/Index_Invincibles_HW_1)

**שאלה 1 – סעיף א :**

1. **4a. Term-Document Incidence Matrix:**

נתונים:

* **Doc 1:** breakthrough drugs for schizophrenia
* **Doc 2:** new schizophrenia drugs
* **Doc 3:** new approach for treatment of schizophrenia
* **Doc 4:** new hopes for schizophrenia patients

ראשית, נרשום את כל המונחים הייחודיים באוסף המסמכים:

* breakthrough
* drugs
* for
* schizophrenia
* new
* approach
* treatment
* hopes
* patients

נבנה מטריצת שכיחות מונח במסמך.

שורות מייצגות מונחים ועמודות מייצגות מסמכים. ערך 1 מציין את נוכחותו של המונח במסמך, ו-0 מציין את היעדרו.

A table with numbers and letters

Description automatically generated

**4b. Inverted Index Representation:**

נבנה את ייצוג האינדקס ההפוך, כל מונח ישויך לרשימה של מזהי מסמכים שבהם הוא מופיע.

* breakthrough: [1]
* drugs: [1, 2]
* for: [1, 4]
* schizophrenia: [1, 2, 3, 4]
* new: [2, 3, 4]
* approach: [3]
* treatment: [3]
* hopes: [4]
* patients: [4]

תוצאות שהוחזרו עבור שאילתות:

עבור אוסף המסמכים המוצג לעיל, נמצא את התוצאות שהוחזרו עבור השאילתות הנתונות:

א. שאילתה: schizophrenia and drugs

השאילתה schizophrenia and drugs דורשת מסמכים המכילים גם את המונחים " schizophrenia " וגם " drugs". מהאינדקס ההפוך, יש לנו:

* schizophrenia: [1, 2, 3, 4]
* drugs: [1, 2]

החיתוך של רשימות אלה נותן לנו:

[1, 2]

אז המסמכים שמוחזרים הם: Doc1 וDoc2

ב. שאילתה: בעד ולא (סמים או גישה)

השאילתה עבור ולא (סמים או גישה) דורשת מסמכים המכילים את המונח "עבור" אך אינם מכילים את המונחים "סמים" או "גישה". מהאינדקס ההפוך, יש לנו:

* for: [1, 4]
* drugs: [1, 2]
* approach: [3]

ראשית, אנו מוצאים את האיחוד של "סמים" ו"גישה":

* drugs or approach: [1, 2, 3]

לאחר מכן, אנו שוללים רשימה זו כדי למצוא מסמכים שאינם מכילים "סמים" או "גישה":

* not (drugs or approach): [4] (since only Doc 4 is not in [1, 2, 3])

לבסוף, אנו עושים את החיתוך בין התוצאה הזו עם רשימת המסמכים המכילה "עבור":

* for: [1, 4]
* not (drugs or approach): [4]

החיתוך נותן לנו

[4]

אז המסמך המוחזר הוא **Doc 4**

**שאלה 1 – סעיף ב :**

שלבי חישוב :

* 1. סה''כ Terms ב Doc1 : 27+3+14 = 44
  2. סה''כ Terms ב Doc2 : 4 + 33 + 33 = 70
  3. סה''כ Terms ב Doc3: 24+29+17 = 70

**TF :**

|  |  |  |  |
| --- | --- | --- | --- |
| Doc3 | Doc2 | Doc1 |  |
|  |  |  | car |
| 0 |  |  | auto |
|  |  | 0 | insurance |
|  |  |  | best |

TF-IDF:

|  |  |  |  |
| --- | --- | --- | --- |
| Doc3 | Doc2 | Doc1 |  |
|  |  |  | car |
| 0 |  |  | auto |
|  |  | 0 | insurance |
|  |  |  | best |

3.האתר שבחרנו הוא האתר הרשמי של EA Sports, הנמצא בקישור: [EA Sports](https://www.ea.com/sports).

האתר של EA Sports מתמקד במידע על משחקי הספורט השונים שהחברה מפתחת ומשווקת. המידע הזמין למשתמשי האתר כולל פרטים על משחקי הספורט הפופולריים כמו FIFA, Madden NFL, UFC, NFL, NBA וכו, האתר מציע חדשות ועדכונים על המשחקים, תאריכי השקה, תכונות חדשות, סרטוני טריילרים, מידע על גרסאות המשחקים השונות ואפשרויות לרכישתן. בנוסף, ישנם פורומים וקהילות של שחקנים, תמיכה טכנית, ומידע על טורנירים ואירועים שונים.

**שלוש שאילתות מעניינות והפרטים הנדרשים:**

**שאילתא 1:** מהו אחוז השחקנים בכל אזור גיאוגרפי שמשחקים במשחקי ספורט שונים של EA Sports?  
פרטי מידע נדרשים: נתוני משתמשים לפי אזור גיאוגרפי, נתוני משחקים נפוצים לפי אזור, אחוזי שימוש במשחקים שונים.

**שאילתא 2:** מהן השעות הפופולריות ביותר לשחק במשחקי הספורט השונים של EA במהלך השבוע?  
פרטי מידע נדרשים: זמני כניסה ופעילות של משתמשים, דפוסי פעילות יומיים ושבועיים, פילוח לפי סוג המשחק.

**שאילתא 3:** מהם המשחקים הפופולריים ביותר בקרב שחקנים חדשים המצטרפים ל EA Sports  
פרטי מידע נדרשים: נתוני שימוש במשחקי EA, פידבקים ודירוגים משחקנים חדשים וכמות מצטרפים חדשים ביחס לקהילה עצמה.