**Information retrieval HW 1.**

**Group: 404 Brain Not Found.**

**Submmiters:**

Tomer Lazarovitch 316377100  
Itamar Kraus 318304763  
Guy Pariente 206132250  
Chay Fadida 318302783

**Question 1.1**

4a)

a.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **D1** | **D2** | **D3** | **D4** |
| **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 |

b.

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)

1. Schizophrenia and drugs – [1,1,1,1] & [1,1,0,0] = [1,1,0,0] = Doc1 and Doc2
2. For and not (drugs and approach) – [1,0,1,1] & ! ([1,1,0,0] & [0,0,1,0]) =

[1,0,1,1] & !([0,0,0,0]) = [1,0,1,1] & [1,1,1,1] = [1,0,1,1] = Doc1 and Doc3 and Doc4.

**Question 1.2**

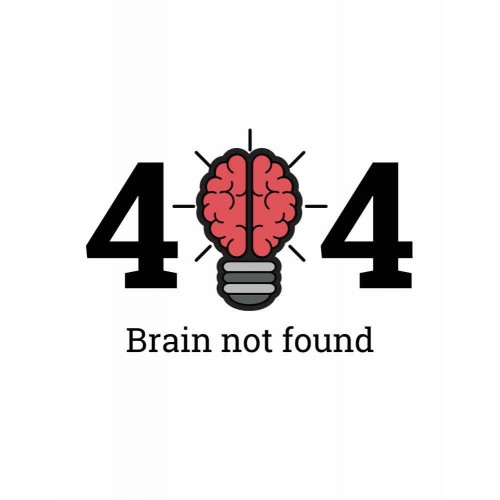
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Term | Doc1 | Doc2 | Doc3 | Df | Idf | Doc1 tf-idf | Doc2 tf-idf | Doc3 tf-idf |
| Car | 27 | 4 | 24 | 18165 | 1.65 | 27\*1.65 = 44.55 | 4 \* 1.65 = 6.6 | 24 \* 1.65 = 39.60 |
| Auto | 3 | 33 | 0 | 6723 | 2.08 | 3 \* 2.08 = 6.24 | 33 \* 2.08 = 68.64 | 0 \* 2.08 = 0 |
| Insurance | 0 | 33 | 29 | 19241 | 1.62 | 0 \* 1.62 = 0 | 33 \* 1.62 = 53.46 | 29 \* 1.62 = 46.98 |
| best | 14 | 0 | 17 | 25235 | 1.5 | 14 \* 1.5 = 21 | 0 \* 1.5 = 0 | 17 \* 1.5 = 25.5 |

**3**.

בחרנו את האתר של הליגה האנגלית בכדורגל ( הפרמייר ליג ): <https://www.premierleague.com/>

1. תחום העיסוק המרכזי של האתר הוא הליגה האנגלית בכדורגל, הוא מרכז מידע רלוונטי על הקבוצות השונות, מציג חדשות על הקבוצות השונות ואפילו מציג סטטיסטיקות על השחקנים והקבוצות המשתתפים בליגה.
2. 3 שאילתות מעניינות שאין עליהן תשובה באתר הליגה האנגלית שחשבנו עליהן:

* מיהו השחקן שהיה הכי הרבה מעורב בשערים שכבשה קבוצתו (גולים + אסיסטים)? המידע שנצטרך הוא הכובש והמוסרים המצטיינים של כל קבוצה.
* מיהי הקבוצה המסוקרת ביותר? משמע הקבוצה שיש עליה הכי הרבה כתבות, נצטרך את מאגר הכתבות הקיים ורשימת הקבוצות בליגה.
* מיהי הקבוצה שהוציאה הכי הרבה כסף בחלון ההעברות? לשם כך נצטרך את היסטוריית ההעברות שקרו בליגה.

****

קישור לתיקיית הגיט:

[**https://github.com/ChayFadida/scrapingPL**](https://github.com/ChayFadida/scrapingPL)