**פרויקט:**

**מגישים:**

ישראל שושן 318740834

עמית פרץ 209054303

דוד אסולין 205908569

יובל הילאי 318609237

1. **האתר שבחרנו – EA SPORTS.**

האתר EA SPORTS עוסק בעיקר בפיתוח והפצת משחקי וידאו בתחום הספורט, עם דגש על משחקי כדורגל.

באתר ניתן למצוא מידע רחב על המשחקים המובילים של החברה, כגון EA SPORTS FC™ 24, FIFA, Madden NFL, NHL, UFC, ו-Apex Legends.

המשתמשים יכולים לקבל עדכונים על גרסאות חדשות, תכונות ומצבי משחק שונים, חדשות על תחרויות, פרטים על שחקנים וקבוצות, הוראות רכישה, ומבצעים מיוחדים.

בנוסף, האתר מציע גישה לפורומים, עדכוני דוא"ל, ופרטי תמיכה טכנית.

קישור לאתר: [www.ea.com](http://www.ea.com)

1. **השאילתות:**

**שאילתה:** מידע על האירועים הקהילתיים המתוכננים על ידי EA SPORTS  
**מידע נדרש:** פרטי אירועים קהילתיים כגון תאריכים, מיקומים, פעילויות מתוכננות, והוראות השתתפות.

**שאילתה:** השפעת העדכונים השונים על ביצועי המשחק  
**מידע נדרש:** השוואה בין גרסאות המשחק השונות, כולל הבדלים בביצועים, תכונות חדשות, ותיקונים שבוצעו.

**שאילתה:** נתונים סטטיסטיים על השחקנים הטובים ביותר במשחקי EA SPORTS  
**מידע נדרש:** רשימות וסטטיסטיקות של השחקנים המובילים בכל משחק, כולל נתוני ביצועים ותחזיות לשיפור עתידי.

1. קוד הזחלן מצורף.
2. **טכנולוגיות בשימוש:**

ספריות Python:

* + requests: לשליחת בקשות HTTP כדי למשוך דפי אינטרנט.
  + BeautifulSoup: לניתוח HTML וחילוץ מידע.
  + pandas: לניהול נתונים ולייצוא תוצאות ל-Excel.
  + openpyxl: לעבודה עם קבצי Excel ולעיצוב תאים.
  + nltk: לעיבוד שפה טבעית כולל הסרת מילים לא חשובות.

ניהול נתונים:

* + Counter לספירת תדירות מילים.
  + defaultdict לבניית אינדקס הפוך.

חישובים מתמטיים:

* + math לחישוב ציוני IDF ו-TF – IDF.

טכנולוגיות אינטרנט:

* + urljoin ליצירת כתובת URL מלאה ממסלולים יחסיים.

1. **זמן ריצה של השאילתות:**

****

זמן הריצה תלוי בגורמים הבאים:

* גודל הדף: דפים ארוכים לוקחים יותר זמן לעיבוד.
* מספר הדפים: מספר הדפים המתקבלים משפיע על הזמן הכולל.
* תשובות מהשרת: זמן תגובה של השרת.
* קוד: יעילות הקוד ואופטימיזציה.

שיפורים:

* שימוש בטכניקות caching לשמור תוצאות של דפים שנבדקו.
* הגבלת מספר השאילתות: כדי למנוע עיכובים מהשרת.
* אופטימיזציה של קוד: שיפור אלגוריתמים לניקוי טקסט והפקת מילות מפתח.

1. **Inverted Index:**

A table of numbers and symbols

Description automatically generated

1. **חישוב TF-IDF:**

TF:

A table with numbers and letters

Description automatically generated

IDF:

A list of information on a white background

Description automatically generated

1. **Hubs ו- Authorities:**

לכל דף שהוחזר חושבו כל הלינקים שיש בדף הזה:  
A blue and white text

Description automatically generated

עם קביעת Nodes ו- Edges:



ועבור כל דף שהוחזר חושבו הHubs ו- Authorities:

A screenshot of a computer

Description automatically generated

אם יש ערכים גבוהים בשדה ה- hubs זה מצביע על כך שישנם דפים שמרכזים קישורים רבים, מה שהופך אותם ל- hubs חשובים.

אם יש ערכים גבוהים בשדה ה- authorities זה מצביע על כך שישנם דפים שמקבלים קישורים רבים, מה שהופך אותם ל- authorities חשובים.

1. **pageRank:**

נתנו לזחלן 10 מאמרים וקיבלנו את ה-pageRanks הבאים:

* https://www.ea.com/games - PageRank: 0.1763
* https://www.ea.com/news/2024-ea-sports-latest-tech-innovations - PageRank: 0.1601
* https://www.ea.com/sports - PageRank: 0.1450
* https://www.ea.com/ea-play - PageRank: 0.1450
* https://www.ea.com/playtesting - PageRank: 0.1130
* https://www.ea.com/careers - PageRank: 0.1130
* https://www.ea.com/news/hiring-our-heroes - PageRank: 0.0369
* https://www.ea.com/news/how-the-community-helps-shape-the-sims - PageRank: 0.0369
* https://www.ea.com/legal - PageRank: 0.0369
* https://www.ea.com/store - PageRank: 0.0369

קיבלנו את הגרף הבא:

A diagram of a network

Description automatically generated

1. **דירוג המשתמשים:**

|  |  |  |  |
| --- | --- | --- | --- |
| page | PageRank | User1 | User2 |
| 1 | 0.1763 | רלוונטי | רלוונטי |
| 2 | 0.1601 | רלוונטי | רלוונטי |
| 3 | 0.145 | רלוונטי | רלוונטי |
| 4 | 0.145 | רלוונטי | רלוונטי |
| 5 | 0.1130 | לא רלוונטי | רלוונטי |
| 6 | 0.1130 | רלוונטי | לא רלוונטי |
| 7 | 0.0369 | לא רלוונטי | לא רלוונטי |
| 8 | 0.0369 | לא רלוונטי | לא רלוונטי |
| 9 | 0.0369 | לא רלוונטי | לא רלוונטי |
| 10 | 0.0369 | לא רלוונטי | לא רלוונטי |

ניתן לראות כי ארבעת הדפים הראשונים רלוונטיים עבור שני המשתמשים, הדף החמישי והשישי חלוקים בעיניי שניהם וארבעת הדפים האחרונים הוגדרו בתור לא רלוונטיים.

**Precision:**

הדפים שסומנו כ"רלוונטים" גם על ידי User1 וגם על ידי User2 הם 1,2,3,4, כלומר .TP = 4

הדפים שסומנו כ"רלוונטים" על ידי User1 אבל כ"לא רלוונטים" על ידי User2 הוא 6, כלומר .FP = 1

Precision = TP/(TP+FP) = 4/(4+1) = 0.8

**Recall:**

הדפים שסומנו כ"לא רלוונטים" על ידי User1 אבל כ"רלוונטים" על ידי User2 הוא 5, כלומר .FN = 1

Recall = TP/(TP+FN) = 4/(4+1) = 0.8

**Git:** [AmitPerets/Index\_Invincibles\_final\_project (github.com)](https://github.com/AmitPerets/Index_Invincibles_final_project)