אחזור מידע – הגשת פרויקט – *Optimizer Outlaws*

**Steam Search Engine**

* **שאלה 1: תחום העיסוק של האתר  
  STEAM.com** [**https://store.steampowered.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…
* **שאלה 2: שאילתות מעניינות  
  השאילתות שבעיקר האתר שלנו מתמקד בהם הן:**
  1. ***Hardware Search***: המשחקים בפלטפורמה שמתעסקים בה יש להם דרישות חומרה שצריך להתקיים לפני קניה והתקנת התוכנה. רצינו לתת את האפשריות הזו למשתמש שיחפד את התוכנות לפי רכיבי חומרה (מעבד, כרטיס מסך וכו..).
  2. ***Category Search***: חיפוש לפי קטגוריה של משחקים. לדוגמה: משחקי מחשב מסוג MMORPG.
  3. ***Search via Review***: חיפוש לפי דירוג של משחקים. לדוגמה: "Very Positive"
* **שאלה 3: הזחלן  
  הזחלן שלנו נבנה במחברת של Google Colab**קישור לזחלן: [Steam Search Crawler](https://colab.research.google.com/drive/1vOlLUR0kOr68qZ23S9m_4tiW71CbKB9N#scrollTo=kDx-tk6j66uL)
* **שאלה 4: טכנולוגיות שונות בפרויקט  
  רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.**בפרויקט שלנו השתמשנו בשלוש טכנולוגיות מרכזיות אשר תרמו משמעותית לשיפור איכות ומהירות החיפוש:
  1. ***Multithreading Search***: טכנולוגיה זו מאפשרת ביצוע חיפושים מרובים במקביל באמצעות חלוקת המשימות בין מספר תהליכונים (*threads*). כך, מתקצר זמן העיבוד הכולל ומשתפרת היעילות של מנוע החיפוש, במיוחד עבור שאילתות מורכבות או מסדי נתונים גדולים.
  2. ***Phrase Handling Search***: יישמנו טכניקה שמאפשרת זיהוי והתמודדות עם ביטויים מורכבים במהלך החיפוש. הדבר מאפשר למנוע החיפוש להחזיר תוצאות מדויקות יותר כאשר המשתמשים מחפשים ביטויים מרובי מילים או מונחים מחוברים.
  3. ***Synonyms Recognition***: שילבנו מערכת לזיהוי מילים נרדפות (סינונימים) כדי להרחיב את טווח התוצאות הרלוונטיות. טכנולוגיה זו מסייעת למנוע החיפוש להבין כוונות משתמשים גם כאשר הם מבטאים את חיפושיהם בדרכים שונות או משתמשים במילים דומות עם משמעות דומה.
* **שאלה 5: זמני ריצה של השאילתות**

**כמה זמן רצו השאילתות שלכם? במה זה תלוי? האם לדעתכם ניתן לשפר זמן זה?**

ברוב המקרים, זמן הרצת השאילתות עמד על טווח של כמה מילישניות עד כמה מאות מילישניות, תלוי במורכבות השאילתה, גודל מאגר הנתונים ומספר התהליכונים (*threads*) שהוקצו.

**שיפור זמני החיפוש אפשרי** בכמה דרכים עיקריות:

1. **שימוש במטמון (*Caching*)**: שמירת תוצאות של שאילתות נפוצות.
2. **אופטימיזציה של אלגוריתם החיפוש**: שימוש במבני נתונים מתקדמים.
3. **ניהול *Multithreading* יעיל**: ניצול מיטבי של תהליכונים.

### שאלה 6: Inverted Index For The Most Frequent 15 Words In The Search

### שאילתת חיפוש: "Very Positive"

**ה-inverted index עבור 15 המילים הנפוצות ביותר:**

| **Word** | **Frequency** |
| --- | --- |
| review | 701 |
| steam | 608 |
| game | 453 |
| view | 253 |
| 1 | 236 |
| community | 223 |
| support | 201 |
| sport | 198 |
| add | 198 |
| new | 195 |
| strategy | 195 |
| language | 189 |

### שאלה 7: חישוב TF-IDF עבור השאילתה "Very Positive"

**בחרנו ב-20 הדפים הראשונות שבהם הערך של TF-IDF היה הגבוה ביותר משאר הדפים הרילוונטים שהזחלן החזיר ורק הראינו אותם בסדר ממוין:**

| **Score** | **Topic** |
| --- | --- |
| 0.006464585 | Black\_Myth\_Wukong\_Benchmark\_Tool |
| 0.005712751 | Farmer\_Against\_Potatoes\_Idle |
| 0.005359385 | Predecessor |
| 0.005211633 | KILL\_KNIGHT |
| 0.005075935 | Revolution\_Idle |
| 0.004858508 | The\_Matchless\_Kungfu |
| 0.004740368 | Shin\_chan\_Shiro\_and\_the\_Coal\_Town/?snr=1\_4\_661\_\_145 |
| 0.004740368 | Shin\_chan\_Shiro\_and\_the\_Coal\_Town |
| 0.004593759 | Frostpunk\_2 |
| 0.004556848 | Blood\_Strike |
| 0.00440871 | Ready\_or\_Not |
| 0.004198065 | Street\_Fighter\_6 |
| 0.004198065 | 1364780/?snr=1\_5\_9\_\_205 |
| 0.004045606 | Age\_of\_Empires\_II\_Definitive\_Edition |
| 0.003778513 | Star\_Wars\_Outlaws |
| 0.003778513 | 2842040/?snr=1\_5\_9\_\_205 |
| 0.003650278 | God\_of\_War\_Ragnark |
| 0.003188349 | Royal\_Quest\_Online |
| 0.002546463 | Age\_of\_Empires\_II\_Definitive\_Edition\_\_Lords\_of\_the\_West |
| 0.001664885 | Age\_of\_Empires\_II\_Definitive\_Edition\_\_The\_Mountain\_Royals/?snr=1\_5\_9\_\_405 |

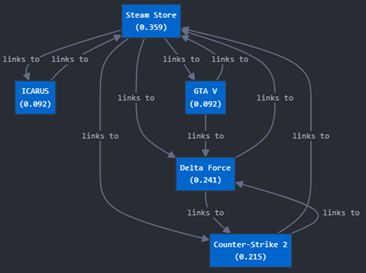
### שאלה 8: Hubs ו-Authorities

**האם בדפים שהוחזרו קיימים *hubs? Authorities*? נמקו.**

בדפים שלנו יש *Related Pages* ו-*Similar Pages*, ולכן ניתן להסיק שהם כוללים *Hubs* ו-*Authorities*:

* *Hubs*: הדפים כוללים קישורים ל-*Related Pages*, אשר מפנים למקורות נוספים ורלוונטיים לנושא. בכך הם מתפקדים כמרכזים שמחברים למידע נוסף.
* *Authorities*: הדפים כוללים גם *Similar Pages*, המשקפים תוכן סמכותי שאוסף הפניות מדפים אחרים ומעיד על מהימנותם בנושאים ספציפיים.

### שאלה 9: PageRank של הדפים



| **URL** | **PageRank** |
| --- | --- |
| https://store.steampowered.com/ | 0.359 |
| https://store.steampowered.com/app/2507950/Delta\_Force/ | 0.241 |
| https://store.steampowered.com/app/1149460/ICARUS/ | 0.092 |
| https://store.steampowered.com/app/730/CounterStrike\_2/ | 0.215 |
| https://store.steampowered.com/app/271590/Grand\_Theft\_Auto\_V/ | 0.092 |

**הערה: באישור מנועמי התאפשר לנו לעשות את חישוב ה-*pagerank* בשיטת *hard coded* כי היה לנו בעיה לחשב את ה-*pagerank* ב-*steam* כי הזחלן לא היה מוציא את הקישורים לדפים אחרים כמו הצורך.**

### שאלה Relevance Feedback:10

**Relevance Feedback ניתן על ידי שני סטודנטים אשר סיווגו את התוצאות הרלוונטיות של הזחלן בצורה הבאה:**

**עבור השאילתה: "Very Positive"**

כמות התוצאות שהזחלן נכנס להם: 132 דפים.  
הזחלן החזיר 97 דפים רלוונטיים מתוך 132 הדפים.  
מתוך המאגר, הערכנו שקיים 220 דפים רלוונטיים באמת עבור השאילתה "Very Positive" מתוך כל המאגר שקיים ב-Steam.

נחשב כעת את ה-Precision ו- Recall.

**חישוב Recall & Precision :**

**סטודנט 1:**

סטודנט ראשון סימן שיש 56 דפים רלוונטים מתוך ה-97 דפים שהזחלן החזיר.

##### 

##### 

**סטודנט 2:**

סטודנט ראשון סימן שיש 65 דפים רלוונטים מתוך ה-97 דפים שהזחלן החזיר.

##### 

##### 

Final Results:

| **Student** | **Precision in %** | **Recall in %** |
| --- | --- | --- |
| Student 1 | 42.42% | 25.45% |
| Student 2 | 49.24% | 29.55% |

התייחסות למשובים אחרי ההצגה:

| איזה שינוי הוצע? | האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו |
| --- | --- |
| הוצע לשנות את החישוב של precision & recall בשאלה 10, החישוב הקודם היה מחשב רק עבור 10 התוצאות הראשונות ומתעלם משאר התוצאות שהזחלן מחזיר. | שיפרנו את החישוב כך שיוצג את החישוב עבור כל התוצאות הרלוונטיים שחזרו ולא רק עבור ה-10 הראשונים. |
| "הציגו את התדירות של המילים ולא את הinverted index" | אנו יכולים לשנות את החלק בקוד שמציג את המילים הנפוצות ביותר ובמקום זאת להציג את ה-inverted index (שמראה באילו דפים מופיעה כל מילה). |
| "לא ראיתי משהו שמצריך שיפור, אולי כן אפשר לבנות שאילתות שהן מדוייקות יותר" | אנו יכולים לשלב היגיון חיפוש בוליאני מתקדם יותר כגון AND, OR, ולא כדי לצמצם את תוצאות החיפוש. |
| "להגדיל את הדפים recall" | ביצענו את השינוי בחישוב כך שיחשב מסך כל התוצאות שחזרו ולא רק מ10 הראשונים |
| "אולי כן יש דרך לבצע PageRank לא בשיטת Hard Coded." | במקום לעשות Hard Coded של PageRank, נוכל לאסוף אותו באופן דינמי תוך כדי סריקה ואינדקס של דפים. יש לבנות את מבנה הקישורים על סמך הדפים בפועל והקישורים היוצאים שלהם שנמצאו במהלך תהליך הסריקה. |
| "לשלב עיבוד שפה טבעית להבנת השאילתה בצורה טובה יותר" | אנו יכולים לפרק את השאילתה של המשתמש לאסימונים משמעותיים ולהחיל גזירה או למטיזציה כדי לצמצם מילים לצורות הבסיס שלהן, מה שעוזר בסטנדרטיזציה של שאילתות ושיפור דיוק ההתאמה. |
| "להציג את האתר קצת יותר צבעוני ונראה מושך את העין" | אנו יכולים לעבוד על העיצוב יותר ולעשות את האתר יותר מושך לעיין. לקשר את העיצוב יותר לעולם המשחקים. |
| "אולי לחפש דרך להתמודד עם התוצאות הדומות של שמות המשחקים" | אנו יכולים ליישם אלגוריתמים של התאמה מטושטשת שיכולים לחשב את הדמיון בין שתי מחרוזות על סמך כמה עריכות נדרשות כדי להפוך את המחרוזות לשוות. |

* מהו היו האתגרים בקורס / פרויקט כצוות?

בעבודה שלנו נתקלנו בכמה בעיות שהפך את העבודה להיות יותר מאתגרת:

1. להבטיח שכולם נמצאים באותו עמוד לגבי משימות, מועדים ואחריות.
2. ניפוי בעיות מורכבות בזחלן, כגון טיפול בהתאמות דפים שגויות.
3. ה-Pagerank לא עבד איתנו כצפוי, נאלצנו להשתמש במנגנון אחר.
4. עבור הסטודנטים שביקשנו מהם להעריך אם הדפים שחזרו היו רלוונטים או לא, היו להם פרשנויות שונות של המשמעות של "רלוונטי" עבור אותה שאילתה.
5. החלטה כיצד להשתמש במשוב משתמשים כדי לשפר את דיוק החיפוש.
6. מיזוג יישומים שונים ממספר חברי צוות ללא התנגשויות.

למרות האתגרים שעמדנו בפניהם, הצוות שלנו שיתף פעולה בהצלחה, פתר קונפליקטים ושיכלל את הגישה שלנו לשיפור הדיוק של הזחלן. באמצעות תקשורת יעילה, חלוקת משימות מאוזנת ושיפורים איטרטיביים המבוססים על משוב רלוונטי, השגנו את התוצאות הצפויות שלנו.

—————————————————————————————————————

**קישור לתקיה של גיט:** [**Github Repository**](https://github.com/oRABiiA/Data-Retrieval/tree/main)

**קישור לקוד של הזחלן:** [**Steam Crawler**](https://colab.research.google.com/drive/1vOlLUR0kOr68qZ23S9m_4tiW71CbKB9N#scrollTo=njQcEc902GUe)

**קישור לדף גיט של שאלה 11:** [**Github Pages**](https://orabiia.github.io/Data-Retrieval/)