אחזור מידע – פרויקט להגשה – אביב התשפ"ד

*מגישים:*

*אביחי חדד 209286665*

*אבישי הרשקוביץ 209460443*

*יובל זוהר 318159373*

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

האתר Audible מתמחה במכירת והשכרת ספרים מוקלטים (אודיובוקס) ופודקאסטים.

הוא מספק למשתמשים גישה למגוון רחב של תוכן מוקלט, הכולל ספרי קריאה,

תכנים חינוכיים, והפקות מקוריות.

קישור לאתר: <https://www.audible.com>

1. 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.

* What are the annual sales of books in Israel?
* Sales Records divided to regions
* What is the most popular book in Israel?

במטלה 2 שאילתות אלה שונו לשאילתות שרלוונטיות ומחזירות תוצאות מרובות על מנת לעבוד איתם (מופיעים בשאלה 5 )

1. בנו זחלן המחזיר את התוצאות לשאילתות שהגדרתם. מומלץ (אך לא חובה) לבנות את הזחלן בפייתון. הציגו את קוד הזחלן.

<https://github.com/yuvalzohar36/DataDiggersHw2/blob/main/HW2/WebScraperAudible.ipynb>

1. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.

במהלך בניית הזחלן השתמשנו בשני טכנולוגיות מעניינות – BeautifulSoup, Pandas.

**BeautifulSoup –**

BeautifulSoup היא ספריית Python המשמשת לניתוח מסמכי HTML ו-XML. ספרייה זו מאפשרת ביצוע של תהליך של web scraping בכך שהיא מאפשרת לחלץ ולתפעל בקלות נתונים מדפי אינטרנט (HTML). בזחלן שבנינו לשליפת נתונים מאתר Audible, הספרייה אפשרה לנו "לחפש" ספרים ספציפיים ולאסוף עליהם מידע כמו כותרות של הספרים, שמות הסופרים, דירוגים הספרים ותיאוריהם על ידי איתור רכיבי HTML ספציפיים והתוכן שלהם.

**Pandas –**

Pandas היא ספריית Python שמאפשרת מניפולציות שונות וניתוח נתונים בכמויות גדולות. ספרייה זו מספקת מבני נתונים כמו DataFrames המקלים על טיפול וניתוח מאגרי נתונים גדולים. בזחלן שאנו בנינו, Pandas משמשת להמרת רשימות של נתוני ספרים ל-DataFrames, מה שמאפשר לנו לפשט פעולות כמו סינון, מיון ואיסוף של הנתונים. בנוסף, ספרייה זו מאפשרת בייצוא נתונים לקבצי Excel בצורה קלה.

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



כאשר ביצענו מדידה של הזמנים, הזמן שלקח לכלל השאילתות לרוץ כולל יבוא של הנתונים הוא 7.74 שניות.

כעת נציג במה תלוי זמן הריצה וכיצד ניתן לשפר (אם ניתן) את זמן הריצה לכל גורם:

1. **הפעולות שמבצעים על הנתונים** – כמובן שזמן הריצה מושפע ישירות מהפעולות אותם מבצעים על הנתונים שאותם ייבאנו. במידה ועושים פעולות מסובכות בכמויות גדולות, כמובן שהדבר עלול להשפיע על זמני הריצה משום שחישובים אלו עלולים לקחת זמן.

במקרה שלנו ביצענו פעולות כמו סינון, מיון ואיסוף של הנתונים.

לפעולות אלו דאגנו להשתמש בספריית Pandas אשר ידוע שזו ספרייה יעילה מבחינת ביצועי זמן ריצה המבצעת את הפעולות הנדרשות ביעילות מרבית.

1. **ביצועי הרשת** – כמובן שזמן הריצה מושפע ישירות מהביצועים של הרשת בין המחשב לשרת שממנו אנו מביאים את הנתונים (מהירות הרשת וזמן התגובה של השרת). במקרים כאלו יותר קשה לשפר את הביצועים, אמנם ניתן לחסוך ככל הניתן בבקשות לשרת שממנו מביאים את הנתונים על מנת לחסוך את הזמן שכרוך בהבאתם.
2. **כמויות הנתונים –** עיבוד כמות גדולה של נתונים תגדיל את זמן הייבוא שלהם וזמן הניתוח שמבצעים עליהם. בשאילתות שביצענו, כמות הנתונים תלויה באופן ישיר משם הספר אותו אנו מחפשים. עבור ספרים מסוימים, כמות התוצאות שהוחזרו גדולה משמעותית מספרים אחרים ובכך יש השפעה ישירה על זמן ביצוע השאילתה.

יתר על כן, שימוש במבני נתונים יעילים (כמו DataFrames) ואלגוריתמים יעילים מקצרים את זמן ביצוע השאילתה.

1. **משאבי המערכת –** בדומה ל-"ביצועי הרשת", זמן הניתוח והעיבוד על הנתונים אותם ייבאנו תלוי באופן ישיר ממשאבי המערכת כגון: סוג המעבד, כמות הזיכרון וכו' .

אנו הרצנו את קוד הזחלן בGoogle Colab אשר משתמש בשרתים של גוגל על מנת להריץ את הקוד. כמובן שניתן להשתמש ב"מחשבים" חזקים יותר בעלי משאבים איכותיים יותר על מנת לשפר את זמן ביצוע השאילתה.

1. בנו inverted index ל- 15 המילים הנפוצות שחזרו (מספיק להתיחס ל -20 הדפים הראשונים שחזרו)

**Query1: the descriptions of fully valid books related to 'It Ends with Us'**

|  |  |  |
| --- | --- | --- |
| **inverted index** |  |  |
| **terms** | **docs** |  |
| world | [1, 6, 10, 14, 20] |  |
| one | [1, 4, 10, 13, 14] |  |
| years | [2, 4, 7, 17, 19] |  |
| us | [1, 6, 8, 15] |  |
| love | [1, 10, 15] |  |
| means | [1, 11, 12] |  |
| often | [1, 5, 8] |  |
| follow | [4, 6, 7] |  |
| knows | [4, 8, 14] |  |
| new | [5, 9, 20] |  |
| life | [7, 10, 17] |  |
| every | [8, 12, 13] |  |
| good | [8, 12, 13] |  |
| book | [8, 12, 20] |  |
| dont | [9, 13, 17] |  |
|  |  |  |

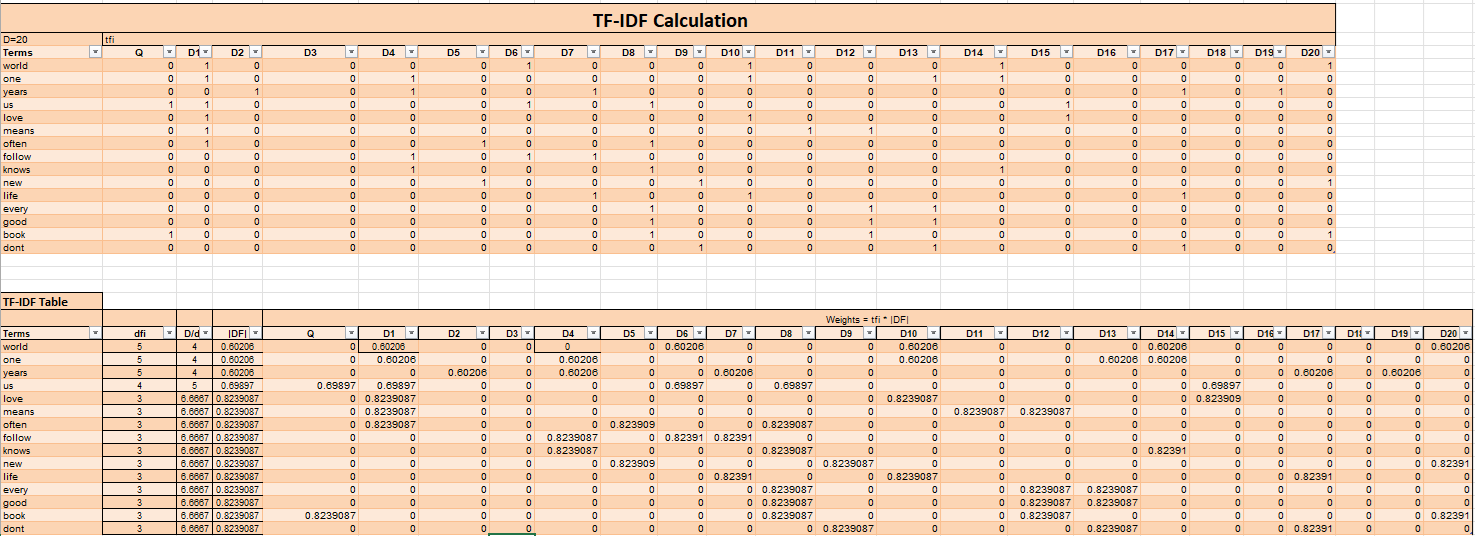
**Query2: the descriptions of highly rated books related to 'It Ends with Us'**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **inverted index** |  |  |
| **terms** | **docs** |  |
| new | [1, 4, 7] |  |
| dont | [1, 2] |  |
| matter | [2, 7] |  |
| goals | [2, 7] |  |
| ones | [2, 3] |  |
| york | [4, 7] |  |
| life | [5, 7] |  |
| know | [1] |  |
| relationship | [1] |  |
| women | [1] |  |
| experience | [1] |  |
| menstrual | [1] |  |
| abnormalities | [1] |  |
| pelvic | [1] |  |
| issues | [1] |  |
|  |  |  |

**Query3: the descriptions of books by the most frequently mentioned author related to 'It Ends with Us'**

|  |  |
| --- | --- |
|  |  |
| **inverted index** |  |
| **terms** | **docs** |
| lily | [2, 4, 7, 8] |
| ryle | [2, 4, 7, 8] |
| love | [1, 2, 5] |
| two | [1, 2, 5] |
| atlas | [2, 7, 8] |
| years | [2, 3, 6] |
| one | [2, 3, 5] |
| life | [2, 3, 4] |
| doesnt | [1, 6] |
| first | [1, 2] |
| time | [1, 2] |
| could | [1, 5] |
| long | [1, 4] |
| much | [2, 5] |
| hate | [2, 5] |
|  |  |
|  |  |

1. בחרו את אחת השאילתות, וחשבו tfIdf של המושגים בשאילתא.

**Query1: most common words in the descriptions of fully valid books related to 'It Ends with Us"**

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

בכלל התוצאות של השאילתות, אנו מבצעים בקשה מהאתר Audible , כאשר האתר כשלעצמו מהווה Hub לספרים משום שהוא מאגד ומקשר למספר רב של ספרים, כולל ספרים ממקורות שונים ומספק נתונים ומידע עליהם. כאשר אנו מקבלים מספר ספרים בתור תוצאה לשאילתה, כל ספר בפרט הוא Authority משום שהוא מהווה מקור חשוב ומוסמך בתחום שלו, עם דירוגים, ביקורות ומידע איכותי שמצביע על הרלוונטיות והסמכותיות שלו בתחום הנושא שבו עוסק הספר.

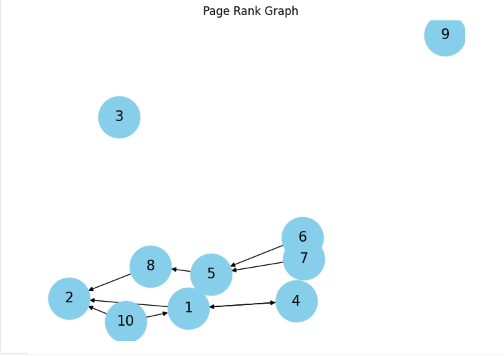
1. בחרו 10 דפים שהזחלן החזיר, ואשר יש ביניהם קישורים .חשבו pageRank לכל דף, מומלץ להעזר בסקריפט מוכן. הציגו את החישובים ואת הדירוג הסופי.

**10 ספרים אשר חזרו לנו לשאילתה כאשר חיפשנו את הספר "It ends with us" הם:**

|  |
| --- |
| **Book name** |
| A Stolen Life |
| Bad Mormon |
| Healing the Shame That Binds You |
| Unf--k Your Brain |
| It Happened One Summer |
| George Orwell’s 1984 |
| It Ends with Violence |
| A Court of Thorns and Roses |
| James Patterson, David Ellis |
| Think Twice |

כעת, עבור כל ספר נגיד כמה קישורים יוצאים ממנו וכמה ספרים (מאלו שצוינו למעלה) מצביעים עליו.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Book Number** | **Book name** | **Total Number of out-links** | **Number of points to me** | **Books number I link to** |
| 1 | A Stolen Life | 2 | 3 | [2, 4] |
| 2 | Bad Mormon | 0 | 3 | [] |
| 3 | Healing the Shame That Binds You | 0 | 0 | [] |
| 4 | Unf--k Your Brain | 1 | 1 | [1] |
| 5 | It Happened One Summer | 2 | 2 | [1, 8] |
| 6 | George Orwell’s 1984 | 2 | 1 | [7, 5] |
| 7 | It Ends with Violence | 2 | 1 | [6, 5] |
| 8 | A Court of Thorns and Roses | 1 | 1 | [2] |
| 9 | James Patterson, David Ellis | 0 | 0 | [] |
| 10 | Think Twice | 2 | 0 | [1, 2] |







1. הראו לשני משתמשים שונים את הדירוג מהסעיף הקודם, בקשו מהם לסמן relevance feedback .

חשבו precision ו - recall.האם ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות?

על מנת לענות על סעיף זה, בחרנו 2 משתמשים אשר נתנו להם 10 מסמכים והם התבקשו לבחור אילו מסמכים רלוונטיים עבורם ואילו לא רלוונטיים עבורם.

המסמכים אשר ניתנו להם הם:

|  |
| --- |
| A Stolen Life |
| Bad Mormon |
| Healing the Shame That Binds You |
| Unf--k Your Brain |
| It Happened One Summer |
| George Orwell’s 1984 |
| It Ends with Violence |
| A Court of Thorns and Roses |
| James Patterson, David Ellis |
| Think Twice |

המשתמש הראשון בחר בדפים הרלוונטיים : 1,2 – סה"כ 2 מסמכים רלוונטיים.

המשתמש השני בחר בדפים הרלוונטיים :1,4,7 – סה"כ 3 מסמכים רלוונטיים .

**Precision & Recall:**

על מנת לחשב את הPrecision אנו נצטרך לקחת את כמות הדפים הרלוונטיים (של כל משתמש) ולחלק בכמות הספרים שניתנו למשתמש (10).

על מנת לחשב את הRecall אנו נבדוק כמה ספרים בממוצע לערך קיימים עבור תוצאה של שאילת מסוימת באתר, כמות תוצאות השאילתות נעות בין 55 ל30 לכן נגדיר את כמות הספרים כ43 ספרים הקיימים במאגר הרלוונטיים לשאילתת החיפוש עבור הספר Its end with us.

|  |  |  |
| --- | --- | --- |
| **User 2** | **User 1** |  |
| 3/10 | 2/10 | **Precision** |
| 3/43 | 2/43 | **Recall** |

על מנת לשפר את התוצאות נבחן בנפרד Precision וRecall.

על מנת לשפר את הPrecision, כאשר מספר הספרים שנתן כל פעם למשתמש הוא 10, מספיק פשוט לחפש בשאילתא ספר "שיותר קרוב" לקטגוריה ולסוג הספרים שאותם המשתמש מעדיף ומעוניין למצוא באתר. כך ה"מונה" יגדל, והמכנה נשאר קבוע (10) ובכך הPrecision יגדל.

לעומת זאת, על מנת לשפר את הRecall, נרצה גם להגדיל את המונה .

כלומר במקרה כזה, נרצה לחפש ספר אשר יחזיר מספר תוצאות מקסימלי שחזרו שרלוונטיות למשתמש (ספר שכנראה מאותה הקטגוריה שהוא מחפש) מה שיגדיל את המונה .

1. הציגו את כל התוצרים של סעיפים א-ו בדף HTML יחיד. הדף יכול להיות דינמי או סטטי (כלומר מציג צילומי מסך של התשובות לשאלות). העלו את הדף לgit pages. צרפו קישור לדף שהעליתם.

אתר הפרויקט זמין בgit page בקישור :

<https://avishayhersh18.github.io/DataDiggerAudibleProject/>

1. בתאריך 12.8 תציגו את הדף מסעיף 11, הכולל את סעיפים 1-10. תקבלו משוב מחבריכם. ענו בטבלה:

לקחנו את כלל המשובים, וענינו על המשובים הרלוונטים המציעים שינוי / שיפור, המשובים החיוביים לא נלקחו ידי חשבון.

|  |  |
| --- | --- |
| **איזה שינוי הוצע?** | **האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו** |
| לא היה כל כך ברור מה הוחזר מהזחלן ומה היו השאילתות בסוף. | כלל הביקורות פרט לזו, ציינו שהשאילתות היו ברור, ייתכן כי היה ניתן להמחיש בצורה גרפית יותר מה המטרה של השאילתות |
| הזכירו את השאילתות רק פעם אחת, לא כלכך ברור מה התוצאה הסופית. | ייתכן כי הסטודנט פספס חלק זה בהצגה, כן הוצגו השאילתות ומדוע הם קיימות ובנוסף התוצרים של הזחלן |
| לא לקחת את כל המידע אלא רק דפים רלוונטים | המידע היחיד שנלקח זה הדפים עצמם, התיאור של הספרים המתאימים, ולא נלקח מידע לא רלוונטי מתוך audible |
| לצמצם את הscope של השאלות להחזיר פחות תוצאות | לנושא זה התייחסנו בסוף ההצגה, ע"מ לשפר את הprecision אכן ניתן לכתוב שאילתות ספציפיות יותר ע"מ להקטין את כמות הדפים שחוזרים |
| זמן ריצה | לא ניתן לשפר בצורה משמעותית, השתמשנו בטכנולוגיות וספריות מתקדמות מאוד בתעשייה (pandas, BSoup), בנוסף הגענו ל7.4 שניות עבור 3 שאליתות, שזה יחסית מהיר |
| אולי להשתמש באלגוריתם ריקו כדי לשפר את השאילתות | ניתן היה לשפר ולהשתמש באלגוריתם זה, ניתן להתחיל משאילתה מסוימת ולהריץ עליה את האלגוריתם ע"מ להגיע לשאילתה אופטילית יותר מבחינת התוצאות הרלוונטיות שיוחזרו בשאילתה |
| אפשר היה להעמיק יותר אולי בתחום של מכירות ספרים ואיזה תחומים או רעיונות הם בולטים היום בניהול מידע בתחום המכירות של ספרים דיגיטליים. | ניתן לשפר, רעיון זה הוא טוב, אך התבקשנו לעשות רק 3 שאילתות, ומצאנו כי השאילתות שניסחנו הם הרלוונטיות ביותר |
| אולי סדר רלוונטיות של לינקים עם מענה של פידבק רלוונס | התייחסו לfeedback relevance בסוף הצגתנו, לא ברור מה כוונתו של הסטודנט |
| הצגה מעניינת יותר להצגה של המשתמש | אין פירוט של הסטודנט מה השינוי שהוא מעוניין שנבצע בנראות למשתמש, אך סך הכל האתר עוצב ברמה גבוהה ע"י CSS. קשה להתייחס להערה כללית זו משום שאין אידנקציה ספציפית למה בדיוק אפשר לשפר בניראות. |
| שיפור אלגוריתמים של חיפוש: פיתוח אלגוריתמים חכמים יותר שמבינים את כוונת המשתמש ואת ההקשר של השאילתה, תוך התבססות על למידת מכונה ובינה מלאכותית. | ניתן לשפר, יובל התייחס לנושא זה בסוף ההצגה שניתן לשפר את הזחלן ע"י שימוש במנועי LLM |
| אפשר אולי להכניס לai שיוסיף עוד רקע ויתן את חוות דעתו על חלקים מסוימים באחזור ולהתייחס לתוצאה כדירוג שאפשר לשפר. | כפי שענינו בשיפור מעל - ניתן לשפר, התייחסנו לנושא |
| להגדיל טווחים וקטגוריות מסויימות | השינוי המוצע לא ברור כל כך, אבל כמות הדפים שהשתמשנו בהם היא בהתאם למה שהתבקשנו במטלות השונות, לא הוספנו דפים נוספים מעבר למה שנדרש |
| אולי לטעון יותר מידע, נשמע קצת קצר | 3 השאילתות שלנו מתייחסות לספרים מסויימים (אשר עונים לשאילתה) ולתיאור שלהם. את תיאור הספרים ניתחנו ולשם כך משכנו רק את המידע הנדרש ולא יותר על מנת לא לפגוע בביצועים ולא לעשות עבודה מיותרת. לפיכך, בהינתן שאלו השאילתות שלנו אנו לא רואים טעם למשוך עוד מידע |
| למצוא פתרון לעיבוד של נתונים גדולים יותר ושימוש במבני נתונים טובים יותר | כפי שהוצג במהלך ההצגה, השתמשנו בספריית PANDAS אשר ידועה ביעילותה ובנוחיות שלה בניתוח כמויות גדולות של נתונים. יתר על כך השתמשנו במבנה נתונים שהספרייה מספקת בשם DataFrame שהוא אכן מבנה נתונים יעיל. לפיכך אנו לא רוצים בהערה זו רלוונטית משום שכן התייחסנו לדבר והשתמשנו ככל הניראה מבנה נתונים המתאים ביותר לניתוח כמויות גדולות של מידע כמו שאנו קיבלנו. |
| לספרים ללא תוצאות, ניתן לבדוק את התגובות של המשתמשים אולי כדי לקבל עוד מידע | יש מקום לשיפור בנושא זה, ניתן "להבין" ע"י LLM האם זה תוצאה רלוונטית ע"י ניתוח ביקורות מפורטות המדברות על הנושא. אך ישנם ספרים גם ללא ביקורות ודבר זה לא יהיה אפשרי במקרה זה. |
| שאילתא 3 | לא ברור מה השיפור שמתבקש. |
| לנסות לבצע פילטור על מה שחוזר מהזחלן ולא להציג ספרים לא קשורים כמו שהוצג. | ב3 השאילתות שלנו התייחסנו למספר הספרים הראשונים שעונים לשאילתא ואותם החזרנו. כלומר אין צורך לבצע פילטור נוסף כי זה הספרים שעונים לשאילתא. לדוגמה לשאילתא שמחזירה את הספרים שעליהם מוצג כל המידע בלבד (כלומר לא חסר להם מידע) החזרנו רק את ספרים אלו (ה15 הראשונים). לפיכך לא רואים כי יש משהו לשפר בטענה זו |
| לא כלכך ברור מה בכלל היה התוצאה הסופית, אולי פספסתי. | התייחסנו לתוצאה הסופית שיצאה מתוך השאילתות בהצגה, בנוסף הוצג הTF-IDF של הנתונים והPAGE-RANK של אותם ספרים |
| אפשר להשתמש ב-LLM כמו שהציעו בסוף ההצגה | כפי שענינו בשיפורים השונים בנוגע לLLM/AI - ניתן באמת להוסיף יכולת זו |
| מה שנעמי ציינה לגבי השאילתה הלא נכונה | אכן מה שנעמי ציינה זה הערה רלוונטית ואנו נבחן ותנקן זו בהגשה הסופית |
| הייתי מגדיל קצת את התמונה באתר | ניתן לשפר, ניתן לעצב את האתר שיהיה יותר נוח לעין, ברור ומפורט. |
| ההסברים היו מאוד מפורטים וברורים, הייתי משנה את העיצוב של הדף (לא שחור על כיתוב לבן) | רצינו שיהיה שחור על גבי לבן על מנת לא ליצור עומס חזותי על המשתמש, אך כמובן שאפשר לשפר זאת. |
| דף ההצגה טיפה מבלבל. | לא מפורט מה בדיוק מבלבל, ממבט שערכנו באתר ומתגובות אחרות שהתקבלו לנו על האתר ניראה כי הוא אכן היה ברור. ייתכן והסטודנט לא שם לב לתוכן ולכלל ניראות האתר |

13. מהו היו האתגרים בקורס/ פרויקט כצוות?

* + - 1. **שימוש בספריית Pandas** - השתמשנו בספרייה זו על מנת לנתח את כלל הנתונים שהוחזרו לנו מהאתר Audioble. זו פעם ראשונה שאנו עובדים עם ספרייה וזו ובהתחלה השימוש היה מאתגר עד שהבנו כיצד משתמשים בספרייה זו.
      2. **שימוש בספרייה BeautifulSoup** - על מנת להביא את כלל הנתונים מהאתר Audioble ועל מנת לחלץ רכיבי HTML מדפי הHTML של הספרים השתמשנו בספרייה זו. בספרייה זו השתמשנו בפעם הראשונה דבר שלקח לנו זמן רב לפענח כיצד בדיוק להשתמש ואיך לחלץ כל רכיב HTML שאותו רצינו להביא.
      3. **WEB CRAWLER** - זו הפעם הראשונה שאנו בונים WEB CRAWLER, דבר מאתגר לבנייה בפעם הראשונה משום שיש צורך להבין כיצד בדיוק הוא עובד ואיך בדיוק מחלצים רכיבי HTML מהאתר.
      4. **GOOGLE COLAB** - פלטפורמה זו הכרנו עוד מלפני אך לא יצאנו לנו שימוש רב בה. דבר זה בהתחלה היווה אתגר משום שלא הצלחנו להסתדר לעבוד במקביל על הGOOGLE COLAB. בשלב מאוחר יותר הבנו כי לעבוד על אותם התאים בGOOGLE COLAB הוא לא דבר מומלץ ועדיף להפריד את העבודה ולעבוד במקביל על תאים שונים.
      5. **עבודה בצוות** - אנו נמצאים בסמסטר האחרון כאשר כל אחד מאיתנו כבר עובד בתעשייה ובנוסף לכך נמצא בלימודים במהלך הסמסטר. דבר זה בהתחלה היווה קושי למצוא זמנים פנויים אשר נוכל לשבת ולהשקיע במטלות. בהמשך הקורס למדנו כיצד להתנהל שלושתנו יחד וידענו למצוא את הזמנים המתאימים. יתר על כן, הבנו כי אפשר לפצל הרבה פעמים את העבודה ולאחר מכן לאחד הכל על מנת להגיע לתוצאות טובות יותר ובזמן טוב יותר.

יש להגיש קישור לריפו שלכם ב -GIT, הכולל את כל קבצי האתר, ותיקייה בשם project שבה קובץ וורד ובו מענה על השאלות.

ניתן לפנות אליי בכל שאלה.

בהצלחה!