אחזור מידע – פרויקט להגשה – אביב התשפ"ד

**Team:** Crawling in My Skin

**Members:**

Dmytri Kislitsyn 323349738

Liel Pasman 323068502

Maxim Lebedinsky 318171485

GitHub Link: <https://github.com/MaximL98/CrawlingInMyProtein.github.io>

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

[MyProtein](https://www.myprotein.com/) is a company website that sells sports nutrition and equipment, a variety of protein powders, workout supplements, athletic equipment and more. The users can use the website to get access to different types of supplements, their description, suggestion for usage and more relative data. Their catalogue also includes gym equipment, clothes, and foods. For each item they can see the price, user reviews about the products and suggested on relative products.

<https://www.myprotein.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.

**Queries:**

1. “sucralose”, we would like to find specific ingredients contained in a product description. Each item has a description, each food item has an ingredient list. From which we will pull the data.

2. “Gym equipment”, users need to search “Gym” and by looking through different related and unrelated items, to find what they are looking for. Each item has item details, for example gym equipment labelled as “Hard accessory”, with that we will know to access the right products.

3. “price: 40 - 70”, there is a specific price range on the website, which we would like to make more flexible. Give the user the option to set the range. Each item has a price, when the user specifies price or price range, sort all the items in that price range and display them to the user.

1. בנו זחלן המחזיר את התוצאות לשאילתות שהגדרתם. הציגו את קוד הזחלן.

**Google Colab links to all 3 crawler codes:**

Main crawler code:

https://colab.research.google.com/drive/1yqArFeSeDVwN0JxcNIhm6J95wcrL9RHj

Authority, hub and page rank calculations:

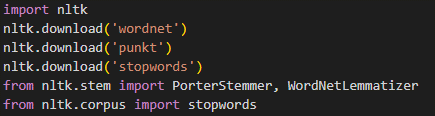
https://colab.research.google.com/drive/1954WQ2ew7y6GjxjcivEDj3\_MWctbaYxJ

Hard and soft accessory retrieval crawler:

https://colab.research.google.com/drive/1YXGn\_4YNiHQ5gXQKtLcbIrtVUHLMlCmK

1. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.

For the project, Python was selected as the primary programming language due to its extensive ecosystem of libraries, facilitating efficient development. To reduce computational constraints and enhance code organization and collaboration, Google Colaboratory was utilized as the development environment. Web scraping was performed using the BeautifulSoup4 library. Data extracted from the MyProtein website was stored in JSON format for subsequent integration into the project website. JSON's structure and readability make it suitable for data management and retrieval.

Also, we have added query pre-processing such as removing stop words, lemmatizing, tokenizing and removing punctuations. This was done using the library nltk.  


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

**Theme 1: Ingredient Retrieval**

• Product Link Extraction: 1.5269 seconds.  
• Ingredient Data Retrieval: 32.7 seconds (total for 30 products).

**Theme 2: Product Categorization**

• Product Link Extraction: 4.45 seconds  
• Product Detail Extraction: 77 seconds (total for approximately 150 products).

**Theme 3: Price Extraction**

• Price Data Extraction: 5.717 seconds.

**Can it be improved? What is it depending on?**

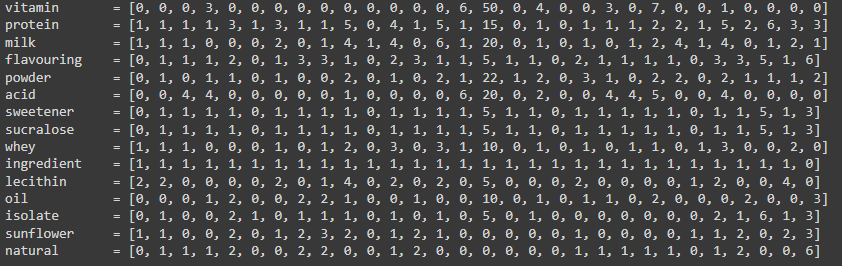
Execution time is primarily influenced by the quantity of products accessed during data retrieval. For instance, a topic with a smaller product count exhibits significantly faster processing times compared to one with a larger product count. A second critical factor influencing data extraction efficiency is the depth of the target data within the website's HTML structure. Data residing in prominent elements such as <h1> headings are typically readily accessible. Contrarily, data encapsulated within nested <div> elements pose a greater degree of complexity in the extraction process.

**Possible improvements**

Execution time can be reduced through the implementation of parallel processing techniques. This involves utilizing multiple threads concurrently on multiple processors or employing more computationally powerful hardware. Or identifying the optimal data extraction strategy to minimize the volume of processed elements.

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

<https://docs.google.com/spreadsheets/d/1qQnpFjCVNhV5qzZt2tZOkw3QCKtpw7MXxiG17YgeIao/edit?usp=sharing>.

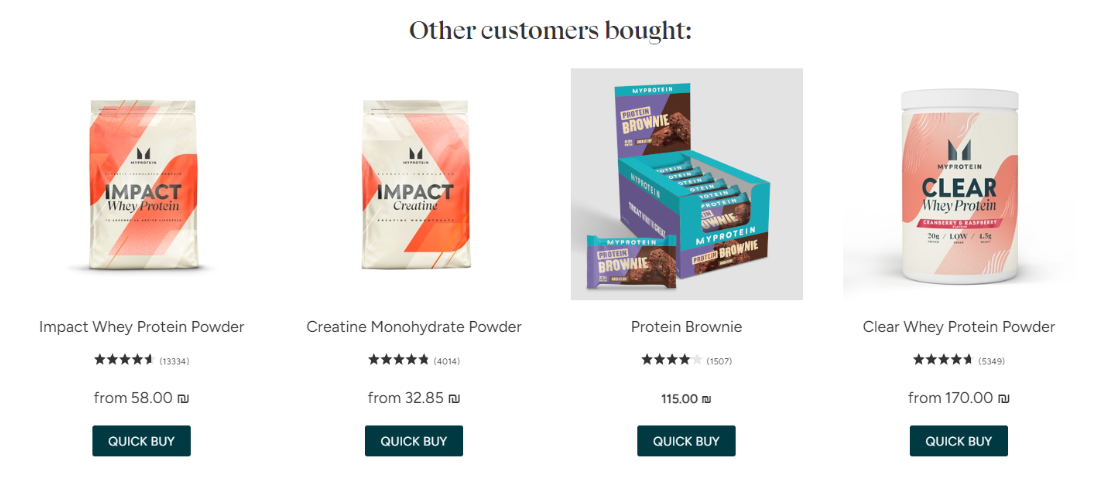


1. בחרו את אחת השאילתות, וחשבו tfIdf של המושגים בשאילתא.

The following text is the result of the TF\_IDF calculations:   
'vitamin': 1.66015273117091  
'protein': 0.1301845467101477  
'milk': 0.4879228193870323  
'acid': 1.1870448277518928  
'powder': 0.5336478845679888  
'flavouring': 0.21008467286988267  
'whey': 0.6244458112713444  
'sweetener': 0.199419071947844  
'sucralose': 0.199419071947844  
'ingredient': 0.0  
'lecithin': 0.9798866806379186  
'isolate': 0.6637654004650809  
'oil': 0.9675369173943055  
'sunflower': 0.7109509202358797  
'soy': 0.9450106896028247

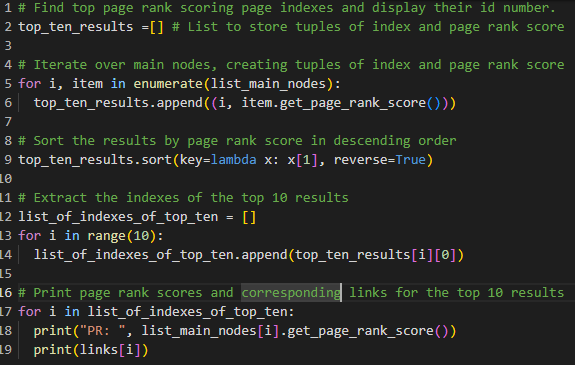
1. האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.

Each product page points to 4 other recommended pages. That is why, we can build a graph of hubs and authorities and calculate their scores. The scores are presented in the HTML page.



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

**The calculations were done by code and not by hand**:

Then, the 10 top pages, based on the highest page rank are returned.

**PR: 0.3072475127646222**

<https://www.myprotein.co.il/sports-nutrition/clear-whey-protein-powder/12081395.html>

**PR: 0.27951246918333494**

<https://www.myprotein.co.il/sports-nutrition/impact-whey-isolate-powder/10530911.html>

**PR: 0.1777605054993788**

<https://www.myprotein.co.il/sports-nutrition/impact-whey-protein-powder/10530943.html>

**PR: 0.03164354967948718**

<https://www.myprotein.co.il/sports-nutrition/vegan-protein-blend/11776868.html>

**PR: 0.021651089075854703**

<https://www.myprotein.co.il/sports-nutrition/soy-protein-isolate/10529701.html>

**PR: 0.019082790464743593**

<https://www.myprotein.co.il/sports-nutrition/clear-vegan-protein/12360400.html>

**PR: 0.011375**

<https://www.myprotein.co.il/sports-nutrition/impact-diet-whey/10530657.html>

**PR: 0.010560416666666668**

<https://www.myprotein.co.il/sports-nutrition/clear-vegan-protein-sample/12360405.html>

**PR: 0.006416666666666667**

<https://www.myprotein.co.il/sports-nutrition/impact-weight-gainer/10529988.html>

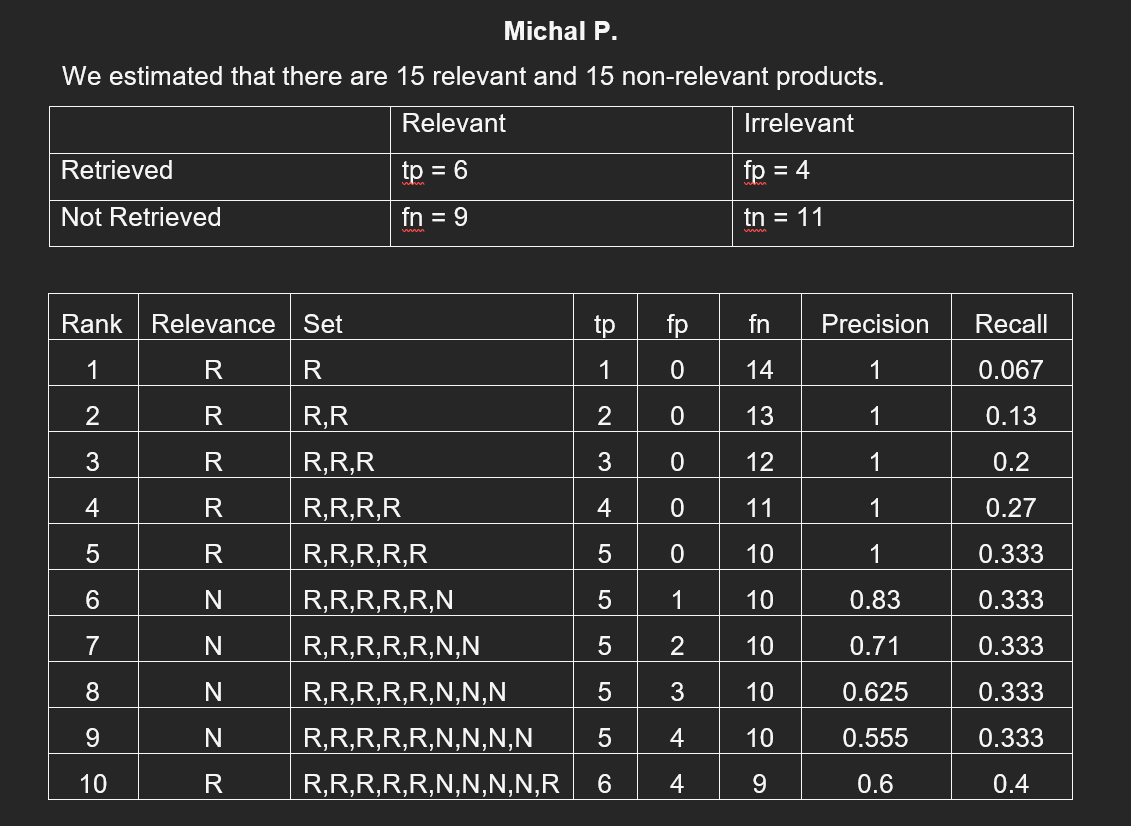
**PR: 0.006416666666666667**

<https://www.myprotein.co.il/sports-nutrition/soy-protein-isolate-sample/11332868.html>

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?

To improve query accuracy, we propose calculating tf-idf values for words in documents relevant to the user. These high-scoring terms can then be added to the original query.



A black and white chart with white text

Description automatically generated

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.

<https://maximl98.github.io/CrawlingInMyProtein.github.io/MP_Retrieval.html>

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

לאחר מעבר על קובץ המשוב, מחקנו את המשובים הלא רלוונטיים והריקים, וסיכמנו את השינויים הרלוונטיים שהוצעו.

|  |  |
| --- | --- |
| **איזה שינוי הוצע?** | **האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו** |
| שיפור זמני ריצת השאילתות. | יש מקום לשיפור, ניתן לממש תכנות מקבילי עבור ביצוע ייבוא המידע. |
| החלפת השימוש בRegex ושימוש בטכנולוגיות אחרות שBeautiful Soup מציע. | לדעתנו אין מקום לשינוי זה. השימוש בRegex בקוד שלנו היה הכי נכון עבורנו בשל פשטותו, והשימוש שלו בוצע בפונקצייה יעודית לפילטור השאילתות. |
| במקום לתת למשתמש לרשום שאילתה בפורמט String, לאפשר למשתמש לסמן את האופציות הרלוונטיות לו (Option box). | יש מקום לשיפור, נוכל להוסיף זאת בנוסף למימוש הקיים עכשיו על מנת להקל על נוחות השימוש עבור המשתמש. |
| שילוב TF-IDF. | בפרויקט הנוכחי, השימוש בTF-IDF מתבצע בחיפוש עבור שאילתת מוצרים לפי הרכיבים (השאילתה המרכזית שלנו). ההוספה שיפרה את סדר הלינקים שחזרו למשתמש (הרלוונטיים קודם) אך לא את הביצועים הכוללים של האלגוריתם. ניתן להוסיף את השילוב הזה עבור שאילתות אחרות שאותן נרצה לבצע בעתיד, ונעבוד על פי אותה השיטה של חיפוש לפי המרכיבים של מוצר. |
| "חיפוש של המילה ויטמין היא כיוון נכון אבל עדיף לסווג לויטמינים ספציפיים שמגיעים ממקורות שונים בטבע וממזון ומשמשים למטרות שונות ולפונקציונליות שונה של תהליכים בגוף. יותר הגיוני לחפש על פי ויטמינים ספציפיים" | אין מקום למימוש הצעה זאת כיוון שמטרת החיפוש שלנו היא לפי מרכיבי המוצר שאותם המשתמש מחפש. כמובן, אם ברכיבי המוצר מופיע ויטמין מסוג מסויים, המשתמש יוכל לחפש ולמצוא את כל המוצרים. |
| עקיפת קובץ Robots.txt שחוסם את החיפוש. | אין מקום למימוש ההצעה, עם זאת שרשום בקובץ שהחיפוש הוא אסור, לא נתקלנו בבעיה בביצוע Crawling באתר. אין בכוונותינו לחפש באתר בצורה חופשית. מטרת הפרויקט הזה הייתה לקחת דף עם הרבה מוצרים, ולהחזיר שאילתה מעניינת שהאתר לא מספק. |
| שיפור נראות אתר ההצגה:   * המראה עמוס * הצגת לינקים במקום מידע בפידבק * פרטים שונים מטאבים אחרים * שיפור UX | יש מקום לשיפור, עיצבנו את האתר בצורה מינימליסטית (טקסט לבן על שחור) ונקייה לטעמנו, על מנת להציג את החומר הנלמד והתוצאות בצורה ברורה. אך, אכן ניתן להפחית בפרטים (כגון בתמונות הקוד) ולבצע שינויים נוספים בעיצוב האתר על מנת להנגיש אותו בצורה ברורה יותר לקהל רחב יותר. |

1. מהו היו האתגרים בקורס / פרויקט כצוות?

**אתגר 1:** זמני עבודה משותפים על הפרויקט.

* היה לנו את הקושי להפגש יחד בזמנים מתואמים, לכן השתדלנו להפגש פעם בכמה זמן, לחלק את המטלות ולהתעדכן על הביצויים והקשיים שפגשנו בטלפון.

**אתגר 2:** חישובי Auth&Hub.

* לצורך החישובים הללו, נתקלנו בקושי בבניית הגרף המקשר בין מוצרי המזון באתר.
* נדרשנו ליצור גרף מקושר ולממש את אלגוריתם החישוב לפי הנלמד בהצראה ובתרגולים. בניית הגרף והחישובים ארכה כ-7 שעות.

**אתגר 3:** הבנת דרישות הקורס והתנעת הפרויקט.

* בגלל קובץ Robots.txt שחוסם את החיפוש באתר, נדרשנו להעלות רעיון יצירתי של חיפוש לפי שאילתה שהאתר לא נותן (לפי מרכיבים).
* למרות שהפרויקט היה מפוצל לתת תרגילי בית שהיו מפוצלים לתתי משימות, לא תמיד ידענו אם בין כל משימה יש קשר ועלינו להתייחס אליו במשימה הבאה, או שזוהי משימה בפני עצמה.
* יכלנו לממש את החיפוש במגוון דרכים שונות. לא ידענו מהי הדרך הנכונה והמהירה ביותר להחזרת הדפים\מידע לשאילתה.