

אחזור מידע-פרויקט:

צוות מגיש: Query Kings

קישור לגיטהאב: <https://github.com/guybanbo/IR-Query-Kings.git>

1.כתובת האתר שבחרנו: [http://www.kizur.co.il/home.php\](http://www.kizur.co.il/home.php%5C)

העיסוק המרכזי של אתר קיצור הוא קיצורים בשפה העברית.

המידע הזמין באתר כולל חלוקה של קיצורים לקטגוריות, קיצורים שונים וראשי תיבות בשפה העברית וההגדרות השונות שלהם. בנוסף, גם ראשי תיבות מלועזית נכללים באתר.

2. שאילתות מעניינות:

**א.** מהו הממוצע של מספר המילים להגדרה של קיצור בקטגוריה של ארגונים ומוסדות**?**

פרטי המידע הנדרשים:

1. אורך הטקסט של כל הגדרה בקטגוריה (במילים).

2. מספר ההגדרות של קיצורים הזמינים בקטגוריה.

3. רשימת הקיצורים בקטגוריה

**ב.** מהם המילים הכי נפוצות בהגדרות של קיצורים עבור הקטגוריה כלכלה**?**

1. רשימת כל הקיצורים בקטגוריה.

2. ההגדרות המלאות של כל קיצור בקטגוריה.

**ג. מה מספר ההגדרות הממוצע לקיצור?**

פרטי המידע הנדרשים:

1. רשימת כל הקיצורים באתר.

2. מספר ההגדרות לכל קיצור.

3. מספר הקיצורים הכולל באתר

3. קישור לקולאב עם כל קבצי הזחלן:

<https://colab.research.google.com/drive/1o4bGox8oPtPh2EE9vJ_1oElKjfbp9Vm_?usp=sharing>

4.את הזחלן והשאילתות מימשנו בפייתון והרצנו בGoogle Colab.

בפרויקט זה השתמשנו במספר טכנולוגיות מעניינות שתרמו לפיתוח ולמימוש הפתרון:

### **א) Firebase**

השתמשנו במסד נתונים **Firebase** כדי לשמור את הדפים שהזחלן אסף.

לצורך ביצוע השאילתות, שלפנו משם את הדפים הרלוונטיים לכל שאילתא.

### **ב)BeautifulSoup**

שימוש ב-BeautifulSoup אפשר לנו לחלץ את המידע שאנחנו רוצים מדפי HTML.

הזחלן השתמש בזה כדי לחלץ קישורים ואת הכותרות של דפים.

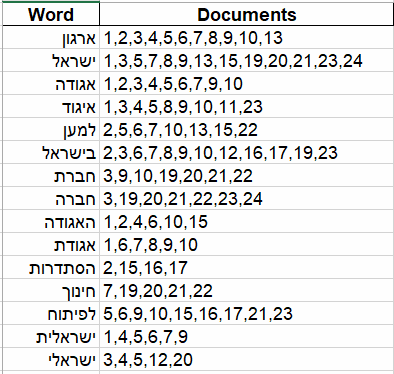
בכל שאילתא השתמשנו בספרייה זו כדי לחלץ את הקיצורים ואת המידע הרלוונטי לה.

ג) **openpyxl**- ספריית פייתון ליצירת קבצי אקסל. השתמשנו בספריה זו כדי לייצא לאקסל את הדפים שהזחלן אסף ואת כל המידע הרלוונטי והחישובים שנאספו ונעשו בשאילתות.

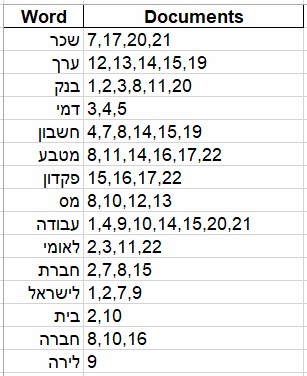
5.שאילתות 1-2 רצו כחצי דקה ושאילתא 3 רצה 12.5 דקות. הזמן תלוי במספר הדפים בכל שאילתא וגם בכמה זמן לוקח לכל בקשה לגשת לדף להחזיר תשובה. בנוסף הוספנו בקוד שלנו השהייה(sleep) של שנייה לפני שמנסים לגשת לדף ולחלץ ממנו את התוכן כדי להקטין את הסיכוי שהזחלן שלנו ייחסם וזה גם משפיע על זמן הריצה. ניתן לשפר את זמן הריצה ע"י הקטנת הזמן של הsleep או לא להשתמש בsleep בכלל.

6.חישבנו אינדקס עבור כל שאילתא:

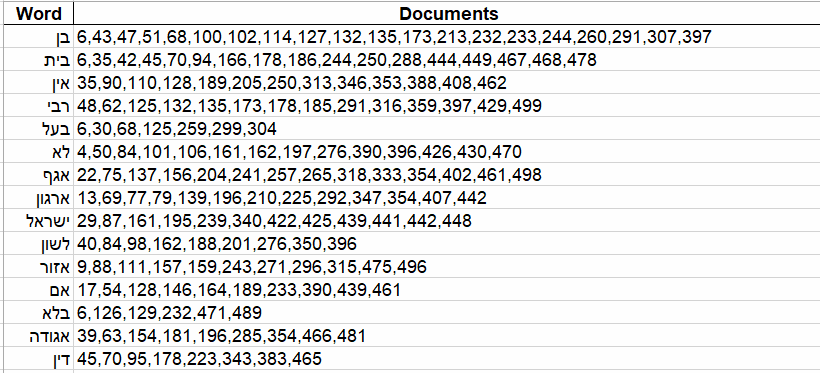
שאילתא 1:



שאילתא 2:



שאילתא 3:



7.בחרנו לחשב TF-IDF עבור שאילתא 1: באקסל בקובץ Quer1.xlsx

8.הזחלן החזיר לנו דפים שהם דפי קטגוריות מהם ניתן להגיע לקיצורים והחזיר לנו גם את הדפים של הקיצורים. אז בדפים שהוחזרו דפי הקטגוריות הם Hubs כי הם מפנים להרבה קישורים אחרים.

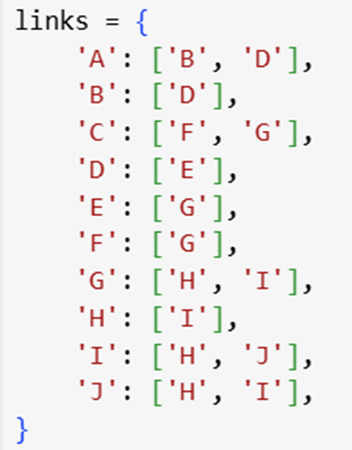
הדפים של הקיצורים הם authorities כי יותר נכנסים אליהם מדפים אחרים מאשר שהם מפנים לדפים אחרים בהשוואה לדפי קטגוריות- יכול להיות דרך דף של קיצור אחר(יש בתחתית הדפים האלה קישורים לערכים סמוכים) וגם דרך מספר דפי קטגוריות אם הקיצור שייך ליותר מקטגוריה אחת.

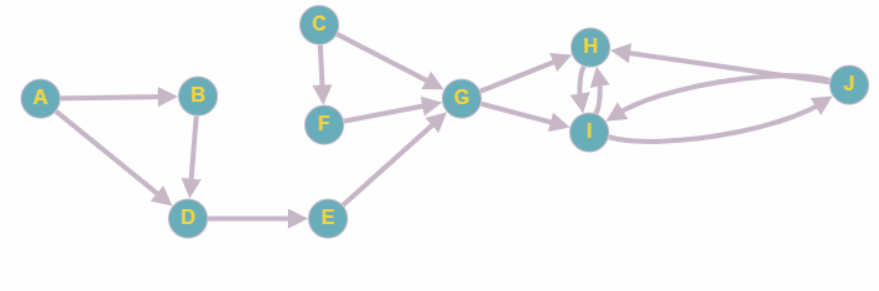
9.

בחרנו 10 דפים של קיצורים משאילתא 3:



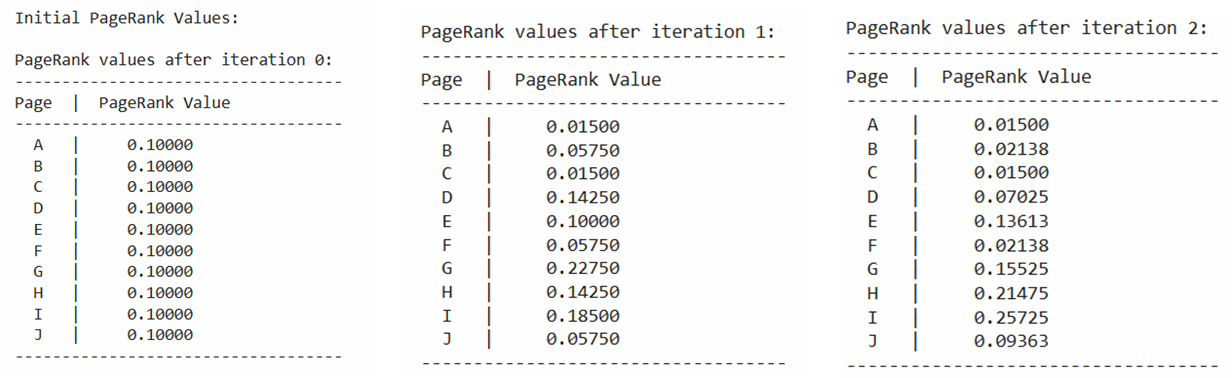
פה ניתן לראות את הקישורים בין הדפים:

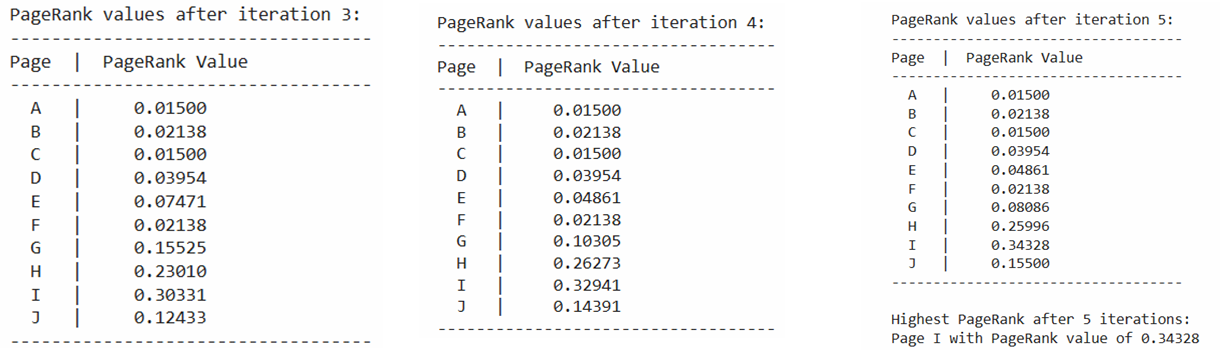




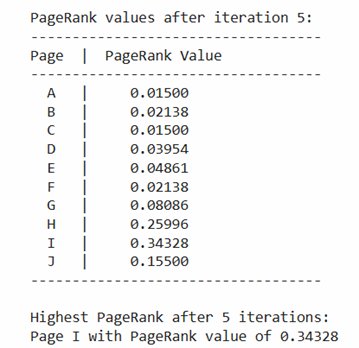
הרצנו סקריפט כדי לחשב את הpage rank עם Damping factor.

ההרצה:





התוצאה הסופית:



10. **משתמש 1:**

דפים רלוונטיים: דפים **H**, **I**, **J**.

דפים לא רלוונטיים: דפים **A**, **B**, **C**, **D**, **E**, **F**, **G**.

#### **משתמש 2:**

דפים רלוונטיים: דפים **I**, **H**, **G**, **J**.

דפים לא רלוונטיים: דפים **A**, **B**, **C**, **D**, **E**, **F**.

נבצע relevance feedback לפי המשוב של המשתמשים באופן הבא: נסתכל על איזה דפים המשתמשים הסכימו עליהם כרלוונטיים וניתן להם יותר משקל- בחישוב הממוצע של מספר הגדרות לקיצור,במקום לעשות ממוצע רגיל נעשה ממוצע משוקלל כך שלהגדרות בדפים שסומנו כרלוונטיים יהיה יותר משקל בממוצע(למשל, משקל של פי 2 מההגדרות בדפים שסומנו כלא רלוונטיים).

RECALL וPRECISION:

בשאילתא הוחזרו 500 דפים של קיצורים. כיוון שמשתמשים יכולים להוסיף הגדרות חדשות וקיצורים ולא ברור עד כמה בודקים את הקיצורים החדשים, נעריך כי רק כ70% מהדפים הם רלוונטיים(בויקיפדיה למשל יש אמינות מוערכת ב80%-<https://www.researchgate.net/publication/249362832_Comparison_of_Wikipedia_and_other_encyclopedias_for_accuracy_breadth_and_depth_in_historical_articles>).

אז יש כ350 דפים רלוונטיים.

משתמש 1:

precision=3/10=0.3

recall=3/350=0.0085

משתמש 2:

precision=4/10=0.4

recall=4/350=0.011

שיפור שאילתא- ננסה להבין מה מאפיין את הדפים שסומנו כרלוונטיים לעומת לא רלוונטיים- למשל ייתכן כי לקיצור פחות מוכר/לא בשימוש תהיה הגדרה אחת בלבד ואז הוא כנראה יהיה פחות רלוונטי לשאילתא כי הוא מוריד את הממוצע של מס' הגדרות לקיצור. אז כדי לייעל את השאילתה נשנה את השאילתה כך שתסנן קיצורים כאלה ולא תיקח אותם בחשבון בחישוב.

11. קישור לאתר: <https://guybanbo.github.io/IR-Query-Kings/SITE/>

12.

|  |  |
| --- | --- |
| **איזה שינוי הוצע?** | **האם לדעתכם יש מקום לשיפור זה? אם כן - הסבירו כיצד ניתן לשפר. אם לא, נמקו** |
| להגביר את מהירות ההרצה | כן, התייחסנו לזה בפרויקט. אפשר לשפר באמצעות הקטנת/הורדת השימוש בsleep ואולי גם באמצעות הקטנת מספר הדפים שמשתמשים בהם לשאילתא 3 שלוקחת יותר זמן |
| שימוש בcache | לא, זה פחות מתאים למבנה של המערכת, התוצאות כבר שמורות בקבצי אקסל ואין צורך בcache |
| הייתי בוחר שאילתות קצת יותר מורכבות שבאמת יכולות לתת ערך גדול יותר | כן, ניתן להפוך את שאילתות 1 ו2 לקצת יותר מורכבות כך שהן יענו על אותן שאלות ויעשו אותם חישובים אבל על כל הקטוגריות ולא רק עבור קטגוריה ספציפית. |
| אולי להוריד את המילים שדומות במשמעות | לא, כיוון שהמילים שדומות במשמעות הן חלק מהגדרות של ראשי תיבות ויש חשיבות לכל וריאציה/נטייה של המילה בחיפוש. |
| להחזיר פחות דפים וככה לקצר את זמן הריצה | כן, ניתן בשאילתא 3 למשל לסרוק דפים של הקטגוריות במקום דפים של הקיצורים עצמם וככה לאסוף אותה כמות הגדרות עם פחות דפים ובפחות זמן. |
| שימוש בחישוב ספציפי להערכת מבנה הקישורים על מנת לחשב את הAutorities hubs | כן, אפשר לעשות את החישוב הזה באמצעות קוד מוכן שמחלץ את הקשרים בין דפים ומחשב את הערכים לפי הגדרה. |
| לנסות לשפר את ה- stemming בעברית אני יודע שזה קשה אבל במידה ותצליחו יהיה לכם אתר מושלם | לא, אין מקום לשיפור הזה כיוון שכמו שכתבנו מקודם יש חשיבות לצורות השונות של אותה מילה ובכל מקרה stemming בעברית זה משהו מאוד מסובך. |
| אפשר להוסיף קישורים בhub להמחשה | כן, היינו מוסיפים דפי קטגוריות להמחשה וגם חישוב של הערכי hub שלהם כדי להראות שהם באמת hub |
| צורת חישוב האינדקס | לא, אין מקום לשיפור ופחות מתאים ואפשרי להוסיף stemming למשל מהסיבות שאמרנו מקודם, החישוב של האינדקס מספיק ומתאים לאתר שלנו. |
| להוריד SLEEP | כן, אפשר אולי להוריד לגמרי את השימוש בsleep כיוון שאין על אתר קיצור הרבה אבטחה ואז אולי אין סיכון שהזחלן ייחסם- אפשר לנסות להריץ בלי sleep ולראות מה קורה. |
| שיפור השאילתה מאוד מוזר | לא, המשוב הזה לא היה מספיק ברור ואין מה לעשות עם זה. |

13. היו לנו מספר אתגרים בפרויקט:

* בחירת שאילתות מעניינות לענות עליהן- כיוון שהאתר קיצור מוגבל יחסית מבחינת סוגי המידע שהוא מכיל ויש בו רק הגדרות של קיצורים וראשי תיבות ודפים של קטגוריות היה קצת מאתגר לחשוב על שאילתות מעניינות לענות עליהן.
* בניית זחלן עבור אתר בעברית- כיוון שאתר קיצור הוא אתר בעברית היה קצת מאתגר בהתחלה לבנות אותו כך שיקודד וישמור את התוכן בעברית בצורה תקינה. בנוסף יש גם מילים בעברית בפרמטרים של הקישורים עצמם והיה צריך גם למצוא דרך להתמודד עם זה כדי שהזחלן יוכל להמשיך לקישורים הבאים ולסרוק עוד דפים ולא להיתקע.
* תיאום מפגשים בין חברי הצוות- כיוון שלכל חבר צוות יש לו"ז קצת אחר וחלק גם עובדים היה קצת מאתגר לקבוע מפגשים שיתאימו לכולם כדי לעבוד על הפרויקט אך הצלחנו כן לקבוע ימים קבועים להיפגש בהם בשעות הערב בזום.