

## אחסון מידע - פרויקט

מגישים:

עידו בדש 206917510 [badasi@post.bgu.ac.il](mailto:badasi@post.bgu.ac.il)

שיין זילברברג 207106824 [shanzil@post.bgu.ac.il](mailto:shanzil@post.bgu.ac.il)

github Repository:

<https://github.com/shanzilberberg/IR-project>

Google storage bucket:

<https://console.cloud.google.com/storage/browser/badashbucket1;tab=objects?forceOnBucketsSortingFiltering=true&authuser=1&project=peaceful-harbor-476714-p1>

בutor התחלתנו לחקלאות על סביבת העבודה והרכינו בה את הקוד. אח"כ עברנו להריץ את הקוד על ה google colad ולבסוף לאחר שיפורים ותיקונים הרכינו את הקוד ב gcp.

### - search()

יצאנו מנקודת הנחה שזה לא יהיה אפקטיבי לחשב את כל המסמכים ולכן החלטנו לעשות סינון למסמכים הרלוונטיים ביותר ועליהם לבצע את החישובים. בעקבות כך, השתמשנו במודול BM25 על הגוף של הטקסט (Body Index) בשbill למצוא את המסמכים הרלוונטיים. לקחנו את 300 המסמכים הרלוונטיים ביותר משלב זה, להמשך עיבוד המידע וכן חסכנו זמן חישוב.

עבור מסמכים אלו, חישבנו ציונים מ- 4 פרמטרים שונים :

הפרמטר הראשון הוא title אשר בודק האם מילות השאלה מופיעות בכותרת המסמך, השני הוא anchor האם מילת החיפוש מופיעה בדף המצביעים לדף. השלישי הוא PageRank שלם כולם, כמו הדף מקשור, והרביעי הוא PageView הבודק את ממד הפופולריות של הדף על ידי כמה הצפיות שבדף.

בסוף דבר, ראיינו של כותרת צריכה לתת את המשקל הגבוה ביותר של הציון הסופי. אח"כ נתנו חשיבות לתוך עצמו שמחושב באמצעות BM25. ולבסוף הבנו שלשלוש הפרמטרים הייתה השפעה זניחה, ה anchor תרם בעיקר גוף הטקסט היה דל או לא חד משמע. וה- PageRank ו- PageViews עוזרים בשבירת שיוויון בין מסמכים דומים.

### הчисוב הסופי:

$$\text{score} = 0.4(\text{bm2\_score}) + 0.45(\text{title\_score}) + 0.08(\text{anchor\_score}) + 0.04(\text{pr\_score}) + 0.03(\text{pr\_score})$$

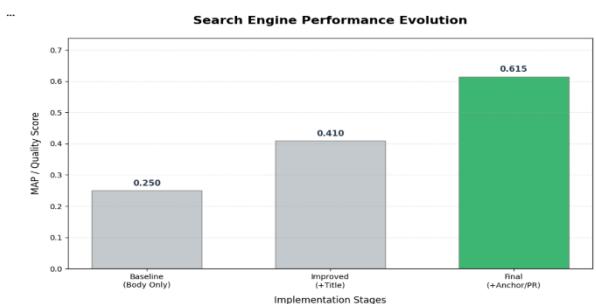
בנוסף בשbill לשפר את החישוב הסופי הוסנו בונוס אם השאלה היא כמו הכותרת, הציון מוכפל בפי 2. אם הביטוי מופיע ברכז בכותרת, הציון מוכפל פי 1.8, כמו כן אם המסמך מכיל את כל המילים שבשאילתת יהיה בונוס וגם שימוש ב  $\log_{10}$  על מספר הצפיות כדי להימנע ממצב בו דפים מרכזיים יופי

## -`search_title()`

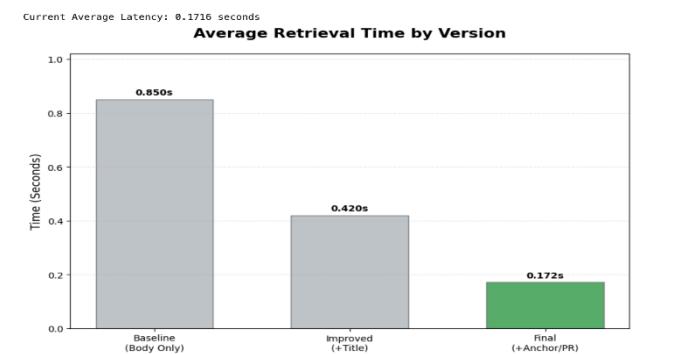
שאילתא זו מחזירה את כל המסמכים שבהם מילות השאלה מופיעות בcotract של הערך. בדקנו גם שאלות שモרכבות ממילה אחת כמו למשל "apple" וגם שאלות המורכבות מספר מילים כמו "Summer Camp for kids". השתמשנו בטוקנייזר ממילה 3 ללא stemming והסרנו stop word. ביצענו שילפת מסמכים רלוונטיים מהאינדקס של הכותרות, עבור כל מילה בשאלתה, נשלפת רשימת האינדקס רשימה המסמכים שבהם המילה מופיעה בcotract. עבור כל מסמך נשמרת קבוצת מילות השאלה שנמצאות בcotract שלו. פונקציית הדירוג שהשתמשנו בה הינה מספר מילות השאלה השונות שמוופיעות בcotract. המסמכים מויין בסדר יורדי לפי הציון.

לאחר שסימנו לכתוב את הפונקציה הרצינו את השאלה ובדקנו באופן ידני האם התוצאות של השאלה נראות לנו הגיונית. בהתאם ראיינו שההתוצאות לא תואמות את הציפיות שלנו, למשל בחיפוש של `kanye west`, הופיעו מילים שלא ציינו כמו `water` ואז לאחר תיקונים, תוצאות השאלה השתנו בהתאם למצופה. שיטת הערכה נוספת שהשתמשנו בה היא שימוש בקובץ `queries_train.json` כדי לבדוק עד כמה התוצאות שהוחזרו תואמות את המסמכים הרלוונטיים שספקו בנתוני האימון, תוך הتمיקדות במיקומים הראשונים בדירוג. לבסוף הגיענו למסקנה כי חיפוש מבוסס כותרת הוא רכיב חזק ואפקטיבי במנוע חיפוש, במיוחד לשאלות קצרות, יוכל לשמש כבסיס טוב לדירוג ראשון או רכיב משקל במודל חיפוש מסוילב. גраф המציג את ביצועי המנוע עבור כל גרסה מרכזית במהלך פיתוח מנוע החיפוש

harmonic mean of precision@5 and F1@30



גרף המציג את זמן האחסן הממוצע של המנוע עבור כל גרסה מרכזית במהלך פיתוח מנוע החיפוש



עבור כל גרסה מרכזית

## Engines average retrieval time(in sec)

בסיוף זה בוצעה הערכה אינטואטיבית של תוצאות מנוע החיפוש עבור שתי שאלות: שאלתה אחת שבה המנוע הפגין ביצועים טובים במיוחד, ושאלתה נוספת שבה ביצוע המנוע היה פחות מוצלחים. עבור כל שאלתה בחרנו את עשר התוצאות הראשונות שהוחזרו, בדקנו האם התוצאות רלוונטיות, את סדר הדירוג והגורמים שה השפיעו על אינטואטיביות התוצאה.

שאלתה בה הוחזרו תשובה רלוונטיות : kanye west

בבחינת עשר התוצאות הראשונות שהוחזרו עבור שאלתה זו, ניתן לראות כי מנוע החיפוש סיפק תוצאות רלוונטיות ואינטואטיביות מאוד. התוצאה הראשונה היא הערך הכללי "Kanye West", ולאחריו מופיעים ערכים העוסקים בשירים מוכרים של האמן. התוצאות כוללות גם שיתופי פעולה ויצירות בולטות נוספות, ובכך מספקות CISI רחוב ומדויק של נושאים הקשורים לאמן.

סדר התוצאות הגיוני ו邏輯י היבט את כוונת המשתמש: הערך הכללי מופיע ראשון, ולאחריו תכנים פופולריים ומשמעותיים בקריריה של Kanye West. הצלחת שאלתה זו נובעת מהיותה שאלתה ישותית וברורה, שבה מונחי החיפוש מופיעים באופן עיקרי בקטגוריות המסמכים. בנוסף, שילוב אותות פופולריות כגון PageRank ומספר צפיפות מחזק את דירוגם של מסמכים רלוונטיים ומוכרים.

עشر התוצאות הראשונות (לפי סדר הדירוג):

Kanye West

Mercy (Kanye West song)

Stronger (Kanye West song)

Ghost Town (Kanye West song)

Gold Digger (Kanye West song)

Monster (Kanye West song)

Power (Kanye West song)

Real Friends (Kanye West song)

Heartless (Kanye West song)

Forever (Drake, Kanye West, Lil Wayne, and Eminem song)

שאלתה שבה מנוע החיפוש ביצע פחות טוב

שאלתה: summer

בבחינת עשר התוצאות הראשונות עבור שאלתה זו, עולה כי מרבית התוצאות עוסקות בנושאים ספציפיים שאינם מייצגים את המשמעות הכללית של המונח. בין התוצאות הראשונות מופיעים ערבים על אולימפיאדות קיץ. לעומת זאת, ערך כללי העוסק בעונת הקיץ כמושג אקלימי ותרבותי כמעט ואינו מופיע בתוצאות הראשונות. מצב זה מעיד על הטיה של מנוע החיפוש לטובות נושאים סדרתיים שבهم המילה "summer" מופיעה בຄורתה, גם כאשר הם מייצגים תתי-נושאים צרים ואינם תואימים את כוונת המשתמש הכללי.

עشر התוצאות הראשונות (לפי סדר הדירוג):

Central European Summer Time

Eastern European Summer Time

2012 Summer Olympics

1996 Summer Olympics

1992 Summer Olympics

1988 Summer Olympics

2000 Summer Olympics

1984 Summer Olympics

2008 Summer Olympics

1972 Summer Olympics

הגורם הדומיננטי לביצועים החלשים

הגורם הדומיננטי לביצועים החלשים בשאלתה *summer* הוא ההסתמוכות חזקה של מנוע החיפוש על התאמה לักษณะ של מונחי השאלה בຄורתה המסמכים, ללא מידול של כוונת המשתמש. מאחר שהמונח "summer" מופיע בתדירות גבוהה בຄורתה של ערבים רבים ושונים, ובעיקר בערכים סדרתיים כגון אולימפיאדות קיץ ושערן קיץ אזרוי, מנוע החיפוש מדריך מסמכים אלו גבוה לмерות שהם אינם מייצגים את המשמעות הכללית והרחבה של השאלה.

מה ניתן לעשות כדי לשפר:

כדי לשפר את יכולת התוצאות בשאלות כלליות ואמביוולנטיות כגון *summer*, ניתן להעדיף ערבים כלליים בשאלות קצרות, להטיל עניישה על דפוסים סדרתיים החזרים על עצמם (כגון ערבים מסוימים בסוג עם שנים שונות), ולשלב אותן מטרות נספיריים מעבר להסתמה לכותרת בלבד, כגון התאמה לגוף הტקסט או זיהוי כוונת משתמש. שילוב שיפורים אלו צפוי לצמצם הטיאות ולשפר את הרלוונטיות והגיוון של תוצאות החיפוש.