

אחזור מידע - פרויקט

מגשים:

עידו בדש 206917510 badasi@post.bgu.ac.il

שני זילברברג 207106824 shanizil@post.bgu.ac.il

github Repository:

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

Google storage bucket:

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

בתור התחלה עבדנו לוקאלית על סביבת העבודה והרצנו בה את הקוד. אח"כ עברנו להריץ את הקוד על ה google colad ולבסוף לאחר שיפורים ותיקונים הרצנו 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 במודל 25BM על הגוף של הטקסט (Body Index) בשביל למצוא את המסמכים הרלוונטיים. לקחנו את 300 המסמכים הרלוונטיים ביותר משלב זה, להמשך עיבוד המידע וכך חסכנו זמן חישוב. עבור מסמכים אלו, חישבנו ציונים מ - 4 פרמטרים שונים :

הפרמטר הראשון הוא title אשר בודק האם מילות השאילתא מופיעות בכותרת המסמך, השני הוא Anchor האם מילת החיפוש מופיעה בדפים המצביעים לדף. השלישי הוא PangeRank כלומר, כמה הדף מקושר, והרביעי הוא PageView הבודק את מדד הפופולאריות של הדף על ידי כמות הצפיות שבדף.

בסופו של דבר, ראינו שלכותרת צריך לתת את המשקל הגבוה ביותר של הציון הסופי. אח"כ נתנו חשיבות לתוכן עצמו שמחושב באמצעות ה 25BM. ולבסוף הבנו שלשאר הפרמטרים הייתה השפעה זניחה, ה Anchor תרם בעיקר כאשר גוף הטקסט היה דל או לא חד משמעי. וה- PageRank ו- PageViews עוזרים בשבירת שיוויון בין מסמכים דומים.

החישוב הסופי:

$$score = 0.4(bm2_score) + 0.45(title_score) + 0.08(anchor_score) + 0.04(pr_score) + 0.03(pr_score)$$

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

search_title()-

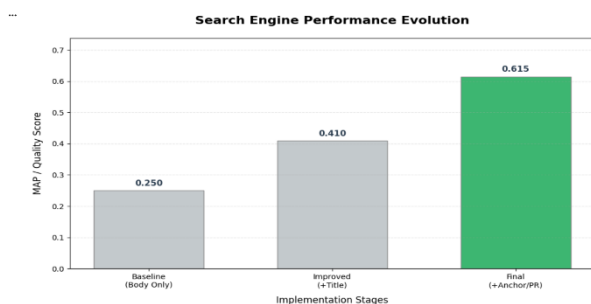
שאלתא זו מחזירה את כל המסמכים שבהם מילות השאלתה מופיעות בכותרת של הערך. בדקנו גם שאלות שמורכבות ממילה אחת כמו למשל "Apple" וגם שאלות המורכבות ממספר מילים כמו "Summer Camp for kids". השתמשנו בטוקניזר ממטלה 3 ללא stemming והסרנו stop word. ביצענו שליפת מסמכים רלוונטים מהאינדקס של הכותרות, עבור כל מילה בשאלתה, נשלפת מרשימת האינדקס רשימת המסמכים שבהם המילה מופיעה בכותרת. עבור כל מסמך נשמרת קבוצת מילות השאלתה שנמצאות בכותרת שלו. פונקציה הדירוג שהשתמשנו בה הינה מספר מילות השאלתה השונות שמופיעות בכותרת. המסמכים מוינו בסדר יורד לפי הציון.

לאחר שסיימנו לכתוב את הפונקציה הרצנו את השאלתא ובדקנו באופן ידני האם התוצאות של השאלתה נראות לנו הגיוניות. בהתחלה ראינו שהתוצאות לא תואמות את הציפיות שלנו, למשל בחיפוש של kanye west, הופיעו מילים שלא ציפינו כמו water ואז לאחר תיקונים, תוצאות השאלתה השתנו בהתאם למצופה. שיטת הערכה נוספת שהשתמשנו בה היא שימוש בקובץ queries_train.json כדי לבדוק עד כמה התוצאות שהוחזרו תואמות את המסמכים

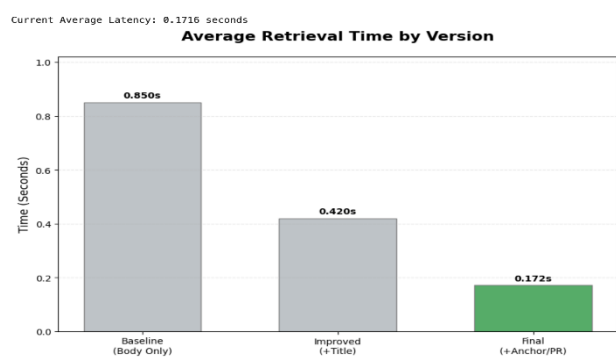
הרלוונטיים שסופקו בנתוני האימון, תוך התמקדות במיקומים הראשונים בדירוג.

לבסוף הגענו למסקנה כי חיפוש מבוסס כותרת הוא רכיב חזק ואפקטיבי במנוע חיפוש, במיוחד לשאלות קצרות, ויכול לשמש כבסיס טוב לדירוג ראשוני או כרכיב משוקלל במודל חיפוש משולב. גרף המציג את ביצועי המנוע עבור כל גרסה מרכזית במהלך פיתוח מנוע החיפוש שלנו:

harmonic mean of precision@5 and F1@30



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



עבור כל גרסה מרכזית

Engines average retrieval time(in sec)

בסעיף זה בוצעה הערכה איכותנית של תוצאות מנוע החיפוש עבור שתי שאלות: שאלתה אחת שבה המנוע הפגין ביצועים טובים במיוחד, ושאלתה נוספת שבה ביצועי המנוע היו פחות מוצלחים. עבור כל שאלתה בחנו את עשר התוצאות הראשונות שהוחזרו, בדקנו האם התוצאות רלוונטיות, את סדר הדירוג והגורמים שהשפיעו על איכות התוצאה.

שאלתה בה הוחזרו תשובות רלוונטיות : kanye west

בבחינת עשר התוצאות הראשונות שהוחזרו עבור שאלתה זו, ניתן לראות כי מנוע החיפוש סיפק תוצאות רלוונטיות ואיכותיות 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 כיסוי רחב ומדויק של נושאים הקשורים לאמן.

סדר התוצאות הגיוני ומשקף היטב את כוונת המשתמש: הערך הכללי מופיע ראשון, ואחריו תכנים פופולריים ומשמעותיים בקריירה של Kanye West. הצלחת שאלתה זו נובעת מהיותה שאלתה ישותית וברורה, שבה מונחי החיפוש מופיעים באופן עקבי בכותרות המסמכים. בנוסף, שילוב אותות פופולריות כגון PageRank ומספר צפיות מחזק את דירוגם של מסמכים רלוונטיים ומוכרים.

עשר 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 שאילתה זו, עולה כי מרבית התוצאות עוסקות בנושאים ספציפיים שאינם מייצגים את המשמעות הכללית של המונח. בין התוצאות הראשונות מופיעים ערכים 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 לטובת נושאים סדרתיים שבהם המילה "summer" מופיעה בכותרת, גם כאשר הם מייצגים תתי־נושאים צרים ואינם תואמים את כוונת המשתמש הכללית.

עשר 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 בשאילתה summer הוא ההסתמכות החזקה של מנוע החיפוש על התאמה לקסיקלית של מונחי השאילתה בכותרות המסמכים, ללא מידול של כוונת המשתמש.

מאחר שהמונח "summer" מופיע בתדירות גבוהה בכותרות של ערכים רבים ושונים, ובעיקר בערכים סדרתיים כגון אולימפיאדות קיץ ושעון קיץ אזורי, מנוע החיפוש מדרג מסמכים אלו גבוה למרות שהם אינם מייצגים את המשמעות הכללית והרחבה של השאילתה.

מה ניתן לעשות כדי לשפר:

כדי לשפר את איכות התוצאות בשאילתות כלליות ואמביוולנטיות כגון summer, ניתן להעדיף ערכים כלליים בשאילתות קצרות, להטיל ענישה על דפוסים סדרתיים החוזרים על עצמם (כגון ערכים מאותו סוג עם שנים שונות), ולשלב אותות נוספים מעבר להתאמה לכותרת בלבד, כגון התאמה לגוף הטקסט או זיהוי כוונת משתמש. שילוב שיפורים אלו צפוי לצמצם הטיות ולשפר את הרלוונטיות והגיוון של תוצאות החיפוש.