אחזור מידע-פרויקט

יהונתן קידושים 319068789 kidushim@post.bgu.ac.il

אופק כחלון 208590174 ofekkah@post.bgu.ac.il

Github Repository:

https://github.com/InformationRetrievalFinalProject/Finalproject

Google storage bucket:

https://console.cloud.google.com/storage/browser/irproject-414719bucket

בתור התחלה, עבדנו ב-Colab על קורפוס קטן (multistream1_preprocessed) וחשבנו איך נוכל לשפר את הInverted index ממטלה 3. החלטנו שפונקציית המשקל שלנו תהיה Tf-idf. בנוסף, החלטנו שפונקציית הדמיון שלנו תהיה Cosine similarity. לאחר מחשבה שביצוע החישוב של ערך ה-Tf-idf של כל term בכל מסמך ובנוסף ערך ה-nf של כל מסמך יגדיל את זמן הריצה הסופי של האחזור. נבצע את החישוב לפני השאילתה והתוצאות ישמרו בשדה Self.tfIdf שהוא יהיה מילון שהמפתח שלו זה ה-Term והערך שלו יהיה List of tuples שיהיה מהצורה שכל איבר בו הוא מזהה המסמך וערך ה-Tf-idf של הביטוי במסמך. בנוסף, יהיה שדה נוסף Self.nf שיכיל את הגודל הוקטורי של כל מסמך. ניסינו לבדוק האם פונקציית הדמיון שלנו עובדת על ידי כך שניסינו לתת את השאילתה "Barcelona" ולצפות למסמכים רלוונטיים כמה שאפשר, ואכן קיבלנו למשל את המסמך "Park Guell" שזה פארק מאוד מוכר בברצלונה. על כן הבנו כי ככל הנראה החישובים שעשינו הינם נכונים, ועל כן החלטנו לעבור לעבודה בסביבת ה-gcp.

על מנת לבחון כי אכן פונקציית הדמיון מחזירה את המסמכים הרלוונטיים הוחלט כי עלינו לבנות Inverted index קטן יותר מהקורפוס הגדול שיכיל מילים ספציפיות מהשאילתות שקיבלנו בקובץ queries_train. בחרנו במספר שאילתות באורך 1 ובנינו קורפוס מהמסמכים שמכילים אותם. מהתוצאות שקיבלנו הבנו שהאינדקס על הטקסט לא מספיק, לדוגמא השאילתה 'genetic' החזירה לנו תוצאה של precision@10 עם ערך של 0.2 ועל כן יש צורך בבניית אינדקס נוסף לכותרות. מהבנייה של האינדקס הנוסף ראינו שיפור בתוצאות, אותה שאילתה עם שילוב של האינדקס על הטקסט וכן האינדקס של הכותרות, כאשר לכל אחד מהם הבאנו משקל של 0.5, החזירה תוצאה של precision@10 עם ערך של 0.5. נציין כי הוספנו שדה של Self.tf, שיתפקד כ-Posting list של מילות הכותרת.

לאחר שראינו שיפור ניכר בתוצאות עם הקורפוס הקטן בשאילתות הנבחרות, החלטנו לעבור לקורפוס הגדול. שם התמודדנו עם קושי עיקרי שהיה ניסיון ליצור את המילון Self.tfIdf בקורפוס הגדול מפאת צריכת זיכרון גבוהה מידי. לאחר מחשבה על פתרון חלופי הוחלט למחוק את השדה Self.tfIdf ובמקומו לשנות את Self.nf כך שיהיה מילון מהצורה שהמפתח בו יהיה מזהה המסמך והערך שלו יהיה tuple של אורך המסמך והגודל הוקטורי של המסמך. המטרה העיקרית של השמירה הזאת היא להשיג את הערכים הדרושים לחישוב Cosine similarity כאשר פונקציית המשקל היא Tf-idf.

כדי לשפר את איכות האחזור שלנו הוספנו את המילונים Page views וPage ranks.

לאחר הרצת שאילתות האימון על הקורפוס הגדול, ועם ההבנה כי Cosine similarity עובד פחות טוב בפני עצמו כמו שראינו בקורפוס הקטן, ורק לאחר שילובו עם התחשבות בכותרות הביא לשינוי חיובי באחזור. לכן חשבנו כי אם נשתמש בBM25 ביחד עם הכותרות נקבל תוצאות אחזור טובות יותר.

לאחר הסתכלות על שאילתות האימון שהן שאלות, הבנו שהתשובה לשאילתה לא נמצאת בכותרת אלא בגוף המסמך. למשל, אם ניקח את השאילתה – "When was the Berlin Wall constructed?" נקבל ממוצע הרמוני של F1@30 ו-precision@5 עם ערך של 0.354, וזאת כתוצאה שנתנו משקל גדול לכותרת. על כן החלטנו לעשות אינדקס נוסף שיהיה על Anchor text. אכן נראה שיפור משמעותי בתוצאות, ואותה השאילתה "When was the Berlin Wall constructed?" החזירה ממוצע הרמוני של F1@30 ו-precision@5 עם ערך של 0.441, כאשר התייחסנו לAnchor text .

לאחר ניסיונות רבים, הבנו שלשאילתות באורך 1(לאחר stemming וסינון stopwords) ולשאילתות באורך שגדול מ1 יש התייחסות שונה לאופן האחזור.

<u>אלגוריתם סופי:</u>

עבור שאילתות <u>באורך 1</u>, מסמך מקבל ניקוד על ידי החישוב:

$$\boldsymbol{for\ each\ d\ in\ C : score[d] = n \cdot (1 + views[d] + PageRanks[d])}$$

$C -$ קורפוס , $n -$ מספר המילים מהשאילתא שהופיעו בכותרת המסמך , $d -$ מסמך

עבור שאילתות <u>שאורכן גדול מ-1</u> נסנן 110 מסמכים על ידי החישוב:

$$\boldsymbol{for\ each\ d\ in\ C : score[d] = n \cdot (1 + views[d] + PageRanks[d])}$$

ציון זה יהווה 25% מציון המסמך הסופי.

בנוסף, נסנן 140 מסמכים על פי החישוב:

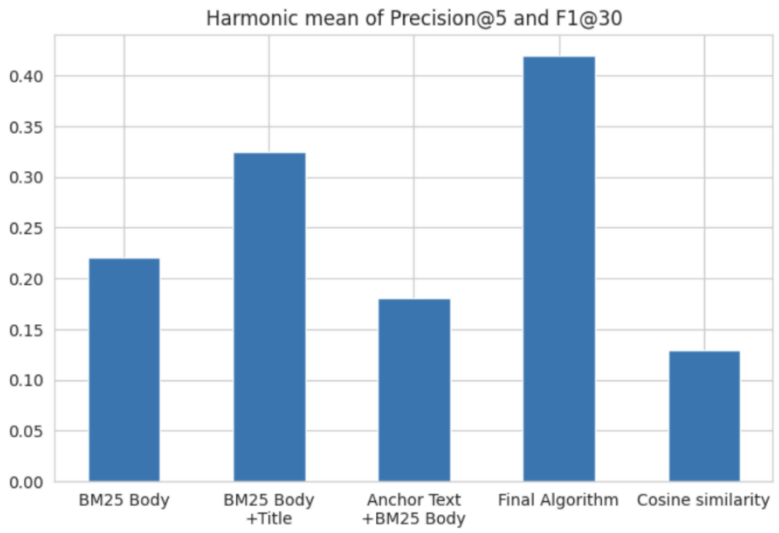
$$\boldsymbol{for\ each\ d\ in\ C : score[d] = n}$$

$n -$ מספר המילים מהשאילתא שהופיעו ב $AnchorText$

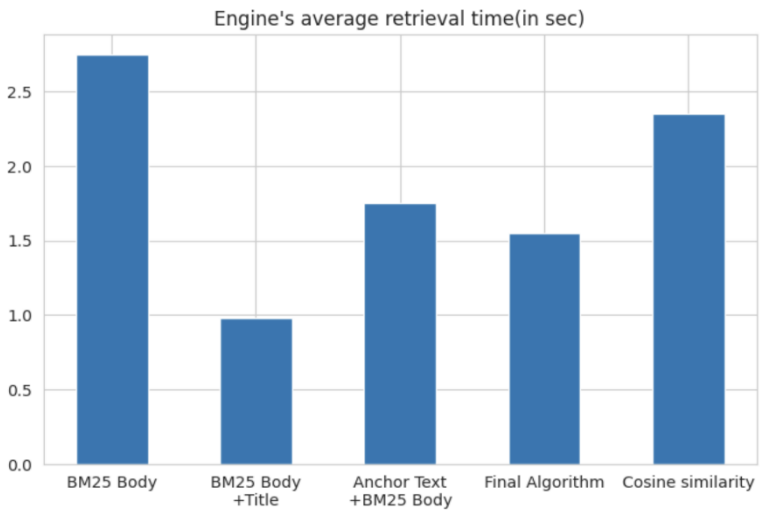
ציון זה יהווה 15% מציון המסמך הסופי.

לבסוף, נחשב BM25 לכ-250 המסמכים שסוננו, כאשר ציון זה יהווה 60% מציון המסמך הסופי.

להלן גרף המציג את ביצועי המנוע עבור כל גרסה מרכזית במהלך פיתוח מנוע החיפוש שלנו :



להלן גרף המציג את זמן האחזור הממוצע של המנוע עבור כל גרסה מרכזית :



38673321 , Construction 3D printing

1305947 , 3D printing

53292889 , Applications of 3D printing

67944537 , Reinforcement in concrete 3D printing

31260053 , Powder bed and inkjet head 3D printing

41615704 , 3D printing marketplace

53292993 , 3D printing processes

62443231 , Multi-material 3D printing

45527564 , 3D Print Canal House

56543723 , 3D makeR Technologies

11113705 , Fathers' rights movement in the United States

540802 , Fathers' rights movement in the United States

875858 , List of United States post offices

14163302 , Lists of United States Congress

17300328 , Outline of the United States

435795 , United States Board on Geographic Names

32212 , United States Armed Forces

19908980 , List of presidents of the United States

37968119 , List of Columbia University people in politics, military and law

2234851 , List of University of Pennsylvania people

עבור השאילתה - 'Who is considered the "Father of the United States"'

עבור השאילתה הנ"ל מנוע החיפוש הצליח להחזיר רק מסמך אחד רלוונטי מתוך עשרת המסמכים המצורפים. עבור השאילתה הנ"ל, מהסתכלות על המסמכים שסווגו כרלוונטיים בשאילתות האימון נראה כי המסמך הראשון מכיל את מילות השאילתה בכותרת, בעוד ששאר המסמכים מכילים בכותרת את שמות האנשים שהם האבות המייסדים של ארה"ב. על כן סיווג על פי הכותרת הוא לא אופטימלי, וסיווג על פי הטקסט אחזר לנו מסמכים שמכילים למשל את המילים - United States בטקסט, אך מסמכים אלה לא היו רלוונטיים. בנוסף, סיווג לפי Anchor text הביא לנו מסמכים כמו למשל- מסמך עם מספר מזהה 307, אברהם לינקולן שהוא אכן אב מייסד של ארה"ב. אך לבסוף נאלצנו לוותר על הסיווג הזה מהסיבה שמסמך זה לא נמצא כרלוונטי במסמכים הרלוונטיים בקובץ שאילתות האימון.

עבור השאילתה - '3D printing technology'

עבור השאילתה הנ"ל, ניתן לראות מעשרת הכותרות של המסמכים הראשונים שאוחזרו כי הן מכילות שתיים מתוך שלושת המילים בשאילתה. על כן אפשר לראות שסיווג ראשוני על פי הכותרת הוא סיווג חזק, וכן שילוב עם סיווגים נוספים מעלה את איכות האחזור.

גדלי האינדקסים :

Index Anchor Size

2.25 GiB gs://irproject-414719bucket/bucketAnchorText/

13.18 MiB gs://irproject-414719bucket/bucketAnchorText/indexAnchorText.pkl

Index Body Size

5.93 GiB gs://irproject-414719bucket/bucketBody/

76.77 MiB gs://irproject-414719bucket/bucketBody/indexBody.pkl

Index Title Size

375.61 MiB gs://irproject-414719bucket/bucketTitle/

206.74 MiB gs://irproject-414719bucket/bucketTitle/indexTitle.pkl

168.88 MiB gs://irproject-414719bucket/bucketTitle/dictIdTitle.pkl

Index Page ranks Size

145.21 MiB gs://irproject-414719bucket/page_ranks/

145.21 MiB gs://irproject-414719bucket/page_ranks/pageRanks.pickle

Index Page views Size

143.78 MiB gs://irproject-414719bucket/page_views/

143.78 MiB gs://irproject-414719bucket/page_views/pageviews.pkl

* את הנספח המלא של כל קבצי האינדקס ניתן לראות בקישור הגיט המצורף