**דו"ח פרויקט אחזור מידע**

**תיאור האינדקסים:**

***text-index*** – באינדקס זה לקחנו את כל הטקסטים של כל המסמכים בקורפוס וייצרנו אותו עם פוסטינג ליסט שמכיל עבור כל מילה את כמות המופעים שלה במסמך. כמו כן שמרנו שדות באינדקס אשר מכילים את אורך המסמכים, מספר המופעים של מילה בכל המסמכים, כמות המסמכים שמילה הופיעה בה ואת המיקום של כל פוסטינג ליסט בזיכרון.

***text-index-stem*** – באינדקס זה לאחר שביצענו tokenization עשינו stemming ואז ביצענו את כל הפונקציות כמו בtext-index ושמרנו את אותן שדות.

***title-index*** – באינדקס זה לקח את הכותרות של כל המסמכים וייצרנו פוסטינג ליסט שמכיל עבור כל מילה במסמך את כמות המופעים שלה בכותרת. שמרנו את השדות אורך המסמך, כמות המסמכים שהמילה מופיעה בהם ואת מיקום הפוסטינג ליסט בזיכרון.

***anchor-index*** *–* באינדקס זה לקחנו עבור כל מסמך את האנקורים שלו וראינו לאן הוא מצביע. את האינדקס בנינו כך שהמילים מיוחסות למסמך שאליו האנקור הצביע. כמו בכל שאר האינדקסים שמרנו את השדות אורך המסמך, כמות המסמכים שהמילה מופיעה בהם ואת מיקום הפוסטינג ליסט בזיכרון.

**קבצי pickle שנשמרו:**

***Page-rank*** – בקובץ זה שמרנו את ציון הpage rank של כל מסמך.

***Doc-norm*** – בקובץ זה שמרנו את הנורמה של כול מסמך בשביל חישוב הcosin similarity.

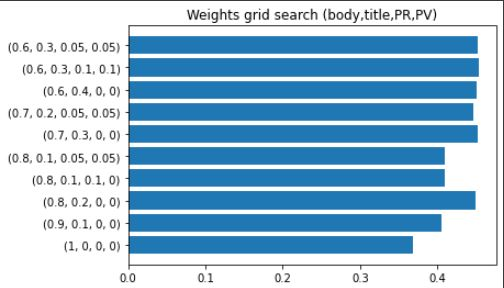
***Page-views*** - בקובץ זה שמרנו את ציון הpage view של כל מסמך עבור התאריך הנתון.

**תיאור הניסויים:**

בהרצה הראשונית של פונקציית הsearch השתמשנו index-text שדירג מסמכים לפי tf-idf וcosin similarity ובנוסף הוספנו משקול על הtitle-index והשתמשנו בו בחיפוש בינארי, ויצאו התוצאות הבאות:

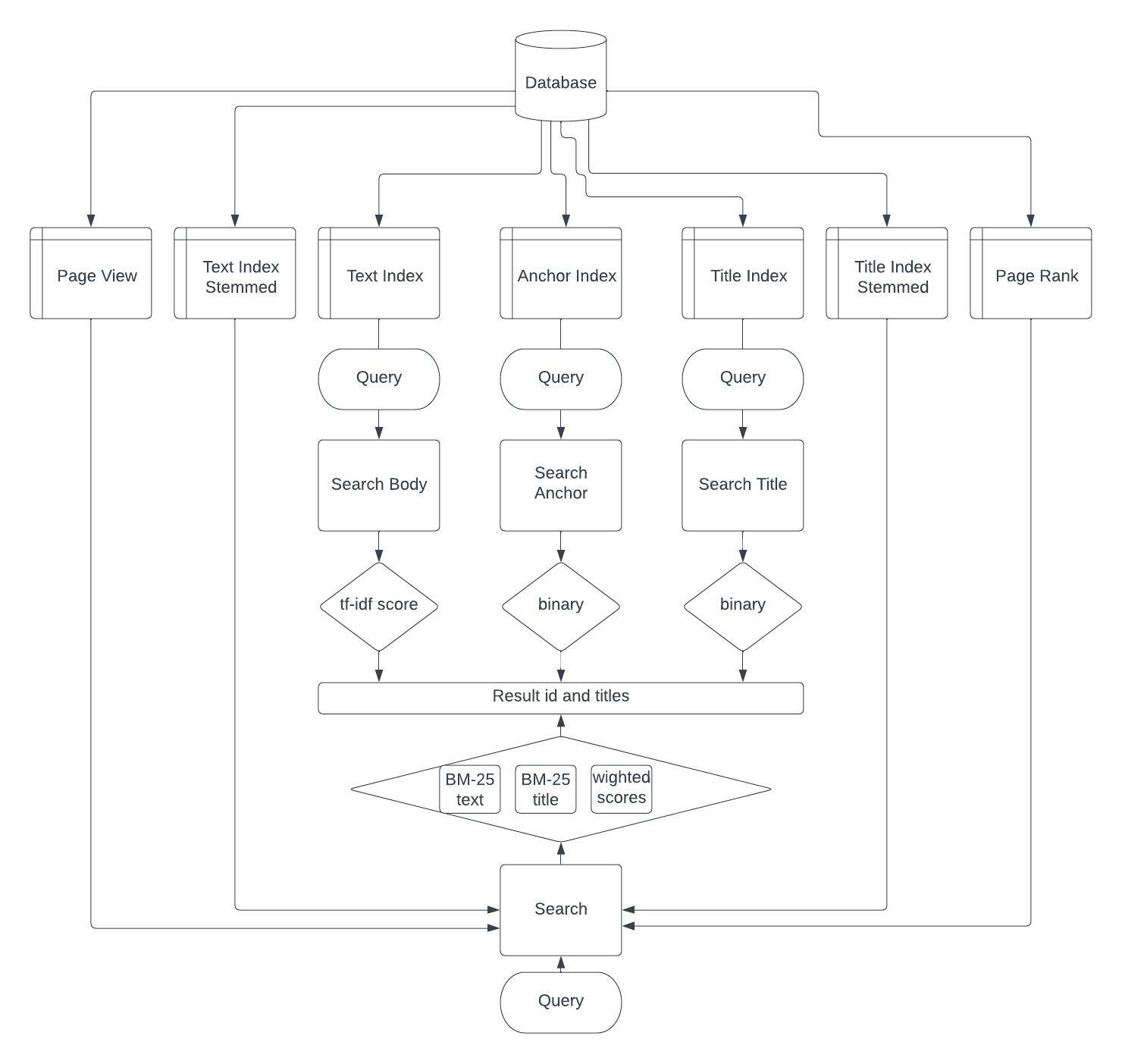
תוצאות אלו לא סיפקו אותנו ולכן בהרצה השנייה השתמשנו רק בtext-index אבל הפעם השתמשנו ב BM25כשיטת דמיון. התוצאות שקיבלנו הן:

ניתן לראות שאומנם דייקנו יותר אחר לקח לנו יותר זמן. לכן בניסוי השלישי שמרנו משתנים לפני עבור ה BM25והשתמשנו בו גם על הtext-index וגם על הtitle-index. בניסוי זה נתן משקל לכל אחד מהאינדקסים וקיבלנו את התוצאות הבאות:



בהרצה הרביעית ביצענו שקלול של הtext-index,title-index,page rank ו-page view. ביצענו אופטימיזציה על המשקולות והנה התוצאות:

בגרף זה בציר הy ניתן לראות משקול לפי (text, title, page rank, pageview) ובציר הx ניתן לראות את MAP@40 עבור כל קומבינציה. בהרצה זו כל פעם החזרנו 100 מסמכים. כמו כן ניתן לראות שכאשר אנחנו מתייחסים רק לחישוב הBM-25 על הtext-index הציון MAP@40היה הכי נמוך וכאשר שמנו משקולות על הtitle page rank ו page view אנחנו מקבלים תוצאות טובות יותר. לכן הבנו שכרגע הערכים הטובים ביותר שלנו עבור המשקולות הם ואלה התוצאות שקיבלנו:

ניתן לראות כאן תרשים של מנוע האחזור שלנו ואיך הוא עובד. ניתן לראות מה קורה כאשר נכנסת לאחת מפונקציות הsearch ואיך כל פונקציית search עובדת ובאילו אובייקטים מהזיכרון היא משתמשת. כמו כן ניתן לראות את שיטת האחזור בכל search ולבסוף נקבל את הresult.