**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.

***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.

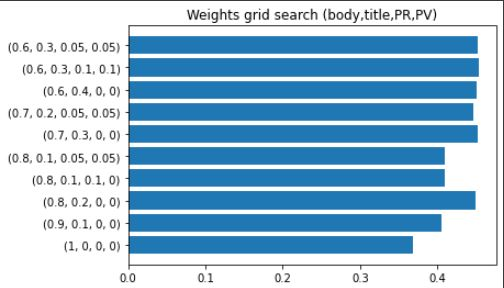
***Page-views*** - בקובץ זה שמרנו את ציון הpage view של כל מסמך עבור התאריך הנתון.

**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:

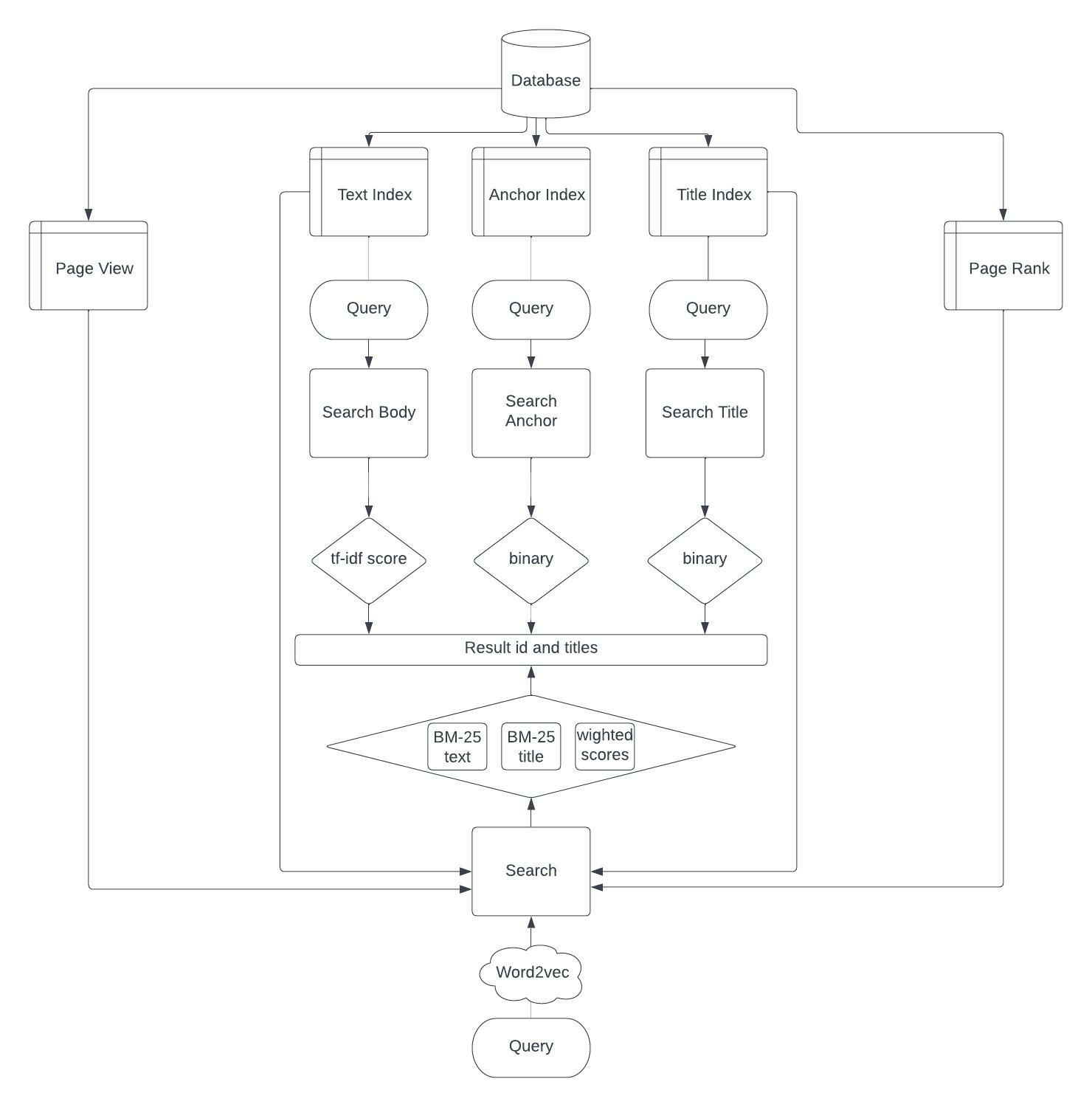
ואלה התוצאות שקיבלנו:

בהרצה החמישית לאחר שהצלחנו לדייק את המנוע של יותר ובנוסף להוריד את זמני הריצה חשבנו שאולי יהיה כדאי להוריד את כמות המסמכים שאנו מחזרים ואולי זה מה שיעזור לנו לדייק את מנוע האחזור שלנו עוד יותר. לכן בהרצה זו החזרנו 40 מסמכים ובדקנו שוב עם אותם פרמטרים מהמודל הקודם, ואילו הם התוצאות:

בהרצה השישית ניסינו לדייק את המודל שלנו עוד יותר בעזרת המודל word2vec. במודל זה השתמשנו בכדי למצוא מילים דומות למילים בשאילתה המקורית שלנו, ולאחר מכן להוסיף מילים אלו לשאילתה כדי לראות האם המסמכים שחוזרים כתוצאה מהשאילתה החדשה טובים יותר. אילו התוצאות שקיבלנו לאחר הוספת המודל word2vec:

*ניתן לראות שלאחר הוספת המודל של word2vec אכן זמני הריצה גדלו מעט אבל ניתן לראות כי הדיוק גדל בקצת ולכן זה מודל שאולי נעדיף.*

*ניתן לראות בגרף למטה את תהליך הלמידה של מודל האחזור שלנו. איזה ניסוי עשינו איזה דיוק קיבלנו ובכמה זמן לקח לנו להחזיר את התוצאות:*

ניתן לראות בתרשים למטה את מנוע האחזור שלנו ואיך הוא עובד. ניתן לראות מה קורה כאשר נכנסת שאילתה 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.

למנוע האחזור שלנו יש כמה שאילתות שהוא בדרך כלל מביא עליהם תוצאות טובות (כלומר מסמכים רלוונטיים) בחלקם מביא תוצאות לא טובות (מסמכים לא רלוונטיים). שאילתה שמקבלת אצלנו דיוק טוב LinkedIn. הסיבה לכך נובעת כנראה מכך שאורך השאילתה קצר וכמו כן יש לה קישוריות מאוד ברורה שזה לחברה LinkedIn. שאילתה שאנחנו לא מקבלים עליה תוצאות טובות היא what is the best place to live in?. הסיבות לכך שלדעתנו אנחנו לא מקבלים על שאילתה זו תוצאות טובות היא מכיוון שאין ערך ויקיפדיה שמגדיר בדיוק מה המקום הטוב ביותר לחיות בו, כמו כן שאילתה זו ארוכה ממש וגם מוצגת בצורה של שאלה כך שבדרך כלל בשאלות סדר המילים והופעתן לא בהכרח תואם את המילים במסמך. עוד סיבה שיכולה להיווצר היא שלתשובה זו יכולות להיות הרבה מאוד תשובות ואין בהכרח דרך לדעת מה רלוונטי או מה לא, כלומר האם החזרת מסמכים של מדינות תיתן תוצאה טובה? או האם החזרת ערכי מסמכים של ערים? או האם בכלל יש שיקולים אחרים בהגדרה של המקום שהכי טוב לגור בו. בכדי שנוכל איכשהו להעריך שאילתה זו לדעתנו צריך להרחיב את כמות המסמכים בקורפוס שעליהם אנחנו בונים את מנוע האחזור שלנו. בדרך כלל בויקיפדיה יש ערכים בודדים שמספרים על היסטוריה או הגדרות של כל מיני ערכים, אבל לא עונים על השאלות של מה יותר טוב או מה פחות טוב ולכן אם נוסיף עוד מידע יש מצב שנוכל לענות על שאלה זו בצורה טובה יותר.