סתיו ברק – 206477952
נעמה מימון – 207547183

דוח הגשה על הפרויקט

התחלנו את הפרויקט ביציאת 2 אינדקסים, index_body לחיפוש על גוף הטקסט ו-index_title לחיפוש על הכותרות שלו, יצרנו להם קבצי pickle.

יצרנו מילון של ה-doc_id וה- title עבור כל מסמך אותו שמרנו גם כן כקובץ pickle.

יצרנו עבור כל doc_id של מסמך וציון עמוד הדירוג שלו (page rank) קובץ pickle נוסף.

על מנת לבדוק את טיב הפרויקט, בדקנו את תוצאותיו על כל אחד מהאינדקסים שיצרנו. תחילה בנפרד על פונקציות דירוג שונות על מנת לבחור את פונקציית הדירוג הטובה עבור כל אינדקס. לאחר מכן שילבנו את החישובים ע"י משקולים בין הtitle לbody, הוספנו משקולים על תוצאות הpage rank ועשינו מספר בדיקות אם השילוב אכן משפר את התוצאות. בבדיקות הסתכלנו על 2 קטגוריות, זמן ריצה ודיוק.

<u>התהליכים השונים שהרצנו לבדיקות:</u>

1. חישבנו cosine-similarity פעמיים, כל פעם על אינדקס אחר – הזמן ריצה לא היה טוב וכך גם הדיוק, הרבה אפסים.

2. ניסינו את פונקציית הדירוג BM25 בה יש יותר חשיבות לתדירות המילה ופחות חשיבות לאורך המסמך של מופע באותו מסמך. חישבנו אותה גם כן פעמיים, כל פעם על אינדקס אחר – זמן ריצה יותר טוב משיטה 1, תוצאות טובות יותר מבחינת הדיוק אך עדיין יש מקום לשיפור.

3. חישבנו את BM25 ל-title ואת BM25 ל-body, סכמנו את תוצאותיהם ושינינו את המשקלים שהבאנו לכל חישוב עד שקיבלנו תוצאות יותר גבוהות מבחינת הדיוק ונמוכות מבחינת הזמן ריצה. אך גם זה לא הספיק.

4. חישבנו משקלים שונים כתלות באורך בשאילתה. הוספנו מספר תנאים שכל תנאי נותן משקל אחר ל-title ול-body, שינינו את המשקלים עד שהתוצאות השתפרו יותר מתוצאות סעיף 4.

5. שיקללנו את התוצאה מסעיף קודם עם התוצאה שחוזרת מחישוב הpage rank (לאחר נורמליזציה) – שינינו את המשקלים בהתאם.

6. בריצות הסופיות שלנו שינינו את אופן מימוש מיון התוצאות למיון ערימה.

סתיו ברק – 206477952
נעמה מימון – 207547183

לאחר ביצוע הבדיקות, הגענו למסקנה כי עלינו לשלב את פונקציית הדירוג BM25 על הכותרת בשילוב חיפוש על גוף הטקסט, כך שבמצב בהן השאילתות יותר ארוכות ומחפשות יותר מידע, אנחנו נרצה לגשת לגוף הטקסט ובמקרה שהן קצרות, הדיוק של התוצאות מה-title יהיה יותר גבוה מכיוון שהכותרות לרוב מכילות מילות מפתח שעליהן מדובר המסמך.

כתוצאה מכך, התחלנו לנסות למשקל את התוצאות שחוזרות לנו מכל אחת מפונקציות הדירוג.

אם נסתמך רק על תוצאות משאילתות האימון, נהיה עלולים להגיע למצב של overfit על הכותרת, וכפי שאנו יודעים זה מצב מסוכן מכיוון שכך למודל שלנו יהיה קשה יותר להצליח כאשר יגיעו נתונים חדשים.

לאחר מעבר על שאילתות האימון בקובץ JSON שמנו לב שרובן היו ארוכות, מה שהקשה על פונקציית הדירוג של ה-title להצליח באחוזים גבוהים. לכן עשינו חלוקה למקרים:

מקרה בו האורך של השאילתה קטן שווה ל-1, נחשב את BM25 ע"י מתן משקל מלא לחישוב ה-title. מקרה בו האורך של השאילתה בין 2 (כולל) ל3 (כולל) נעשה שקלול של 0.6 מהחישוב של ה-title ו0.4 מהחישוב של ה-body.

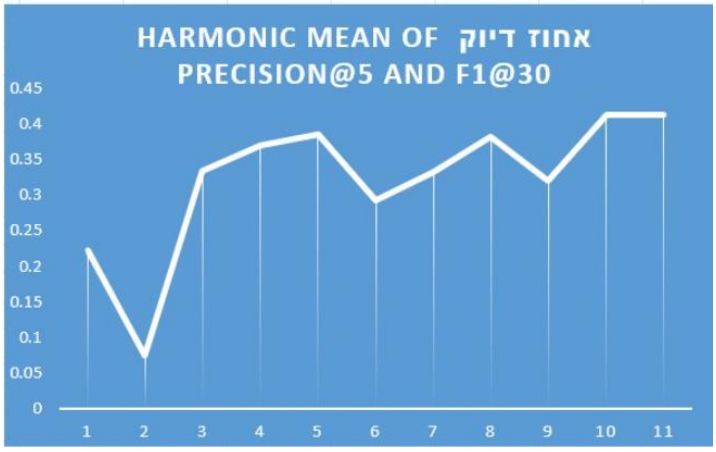
אחרת, מאורך שאילתה של 4 ומעלה, נחשב את BM25 ע"י מתן משקל מלא לחישוב ה-body.

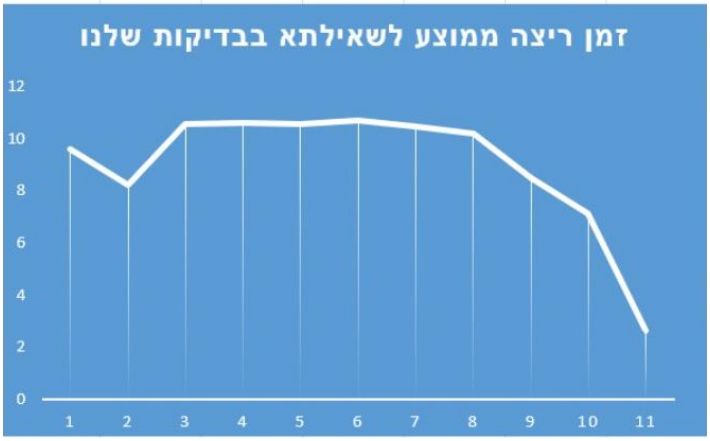
התוצאות השתפרו מבחינת הדיוק והזמן ריצה, אך לא מספיק.

הוספנו פונקציה שמחשבת page rank (אותה עשינו בעבודה 3) ונרמלנו את הscore-ים שחוזרים ממנה ע"י log, על מנת לשמור על היחסים. הוספנו שקלול של 0.6 מהערך שהתקבל מהחלק הקודם של חישוב ה-BM25 על ה-title ו/או ה-body, ו-0.4 ל- page rank.

במיון 100 התוצאות הכי טובות השתמשנו במיון ערימה, מה שגרם לתוצאות שלנו להישמר מבחינת הדיוק ושיפרנו באופן משמעותי את זמן הריצה (בהשוואה לבדיקות שעשינו לפני) ולכן בחרנו להפעיל את מנוע החיפוש שלנו לפי שיטה זו.

סתיו ברק – 206477952
נעמה מימון – 207547183





- הנקודה הנמוכה ביותר בגרף הדיוק היא הריצה שביצענו עם Cosine Similarity עם משקלים בין ה-title (0.6) לבין ה-body (0.4), שהובילה לדיוק של 0.07467 וזמן ריצה 9.5816 שניות. נציין כי גם ריצה בנפרד על כל אחד מהאינדקסים החזירה תוצאה נמוכה.
- לאחר מכן עבדנו רק עם BM25: ביצענו בדיקות נוספות עם משקולים שונים עד שמצאנו את התוצאה הכי טובה שהצלחנו להגיע אליה. שיפרנו את אחוז הדיוק שלנו ל-0.221 עם משקלים בין ה-title (0.6) לבין ה-body (0.4).
- לאחר מכן, הוספנו מדד של page-rank ששיפר לנו את אחוז הדיוק ל-0.33 (התוצאות השתפרו אך הזמנים קצת עלו). המשכנו בדיקות נוספות עם משקולים שונים. התוצאה הטובה ביותר שקיבלנו הייתה עם משקולים לפי תנאים (כפי שהסברנו לפני) ושילוב page-rank ב0.4 אחוז, בשילוב חישוב ה-BM25 ב0.6 אחוז.
- שינינו את המיון תוצאות שלנו לשיטת מיון ערימה והרצנו בGCP וקיבלנו אחוז דיוק של 0.41347 וזמן ריצה ממוצע של 2.62 שניות.

ניתן לראות שבשאילתות שמורכבות ממילה אחת (לאחר טוקניזציה) כמו למשל: snowboard, genetics תוצאות הדיוק וזמן הריצה מאוד טובות מכיוון שבמצב זה נבדוק את פונקציית הדירוג על הכותרת בלבד, מהסיבות שציינו למעלה.

שאילתות נוספות שקיבלנו עליהן אחוז דיוק יחסית גבוה:

When was the United Nations founded?

Describe the process of water erosion.

When was the Gutenberg printing press invented?

סתיו ברק – 206477952
נעמה מימון – 207547183

אלגוריתם BM25 מתפקד היטב בזיהוי מסמכים רלוונטיים כאשר קיימת כמות גבוהה של מילים משותפות בין השאילתה למסמך.

שאילתה פחות טובה שראינו היא Who is considered the "Father of the United States"?

הזמן ריצה שלה היה בכל הריצות שלנו נורא גבוה והאחוז דיוק שקיבלנו עליה הוא 0.

הסבר אפשרי: תדירות נמוכה של רצף המילים המלא (לאחר טוקניזציה)

("Considered Father United States") במסמכים רלוונטיים.

בריצה האחרונה שביצענו ב-GCP, קיבלנו את זמן החישוב המהיר ביותר ואת הדיוק הגבוה ביותר, ביחס לכל הניסיונות הקודמים.

כדי לשפר את הטיפול של המודל שלנו ברצפי מילים נדירים, אנו יכולים להרחיב שאילתות באמצעות טכניקות שונות למשל נוכל לכלול מילים נרדפות וצירופים קשורים.

<u>קישורים</u>:

קישור ל-GitHub:

https://github.com/naama2399/IR-_Naama_Stav

קישור לBucket שמכיל את אינדקס title:

https://console.cloud.google.com/storage/browser/bucket_207547183_title

קישור לBucket שמכיל את הקבצי pickle של doc_id ו-title ושל ה-page rank:

https://console.cloud.google.com/storage/browser/bucket_207547183

קישור לBucket שמכיל את אינדקס body:

https://console.cloud.google.com/storage/browser/bucket_207547183_body

קישור מלא למנוע עובד: http://34.170.59.90:8080