**דוח דרישות – פרויקט אחזור מידע**

1. **מגישות:**

רותם אמיר: 319041208 [amirrot@post.bgu.ac.il](mailto:amirrot@post.bgu.ac.il)

ליאור אפטבי: 209259993 [lioraft@post.bgu.ac.il](mailto:lioraft@post.bgu.ac.il)

1. **קישור לגיט-האב:** https://github.com/lioraft/IR\_SearchEngine
2. **קישור ל:Google Storage Bucket** https://console.cloud.google.com/storage/browser/ass3\_new
3. **תיאור ניסויים (כולל הערכות ומסקנות):**
4. ניסוי ראשון: הרצת tf-idf בשילוב cosine similarity על כותרות בלבד. בצענו pre-processing לכותרות על ידי סינון stopwords, הרצנו גם מתודה שהדפיסה לנו את המילים הכי נפוצות במילון DF והוספנו לרשימת stopwords מילים שלא מוסיפות משמעות סמנטית. בשלב זה לא המרנו את כל הקבצים למילונים עדיין, ולכן זה השפיע על זמן הריצה. הניסוי כלל פתרון יצירתי בו ניסינו להחליף את ה-tf בנוסחה של cosine similarity בציון tf-idf. בניסוי זה, מנוע החיפוש מחזיר 100 מסמכים. ביצענו ניסוי זה פעמיים באמצעות שני מילונים שונים, המילון הראשון כלל Stemming באמצעות Porter Stemmer, ואילו המילון השני לא כלל Stemming.  
   תוצאות הניסוי: ממוצע המדדים

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| stemmer | duration | rq | precision@5 | precision@10 | f1@30 | recall@5 | recall@10 |
| No | 3.235 | 0.034 | 0.080 | 0.070 | 0.034 | 0.011 | 0.021 |
| Yes | 3.312 | 0.121 | 0.187 | 0.170 | 0.114 | 0.027 | 0.048 |

מסקנות: עבור שאילתות שעברו stemming הוחזרו תוצאות יותר טובות בכל המדדים, והפרש הזמנים הינו מינימלי. עם זאת, בכשליש מן השאילתות המנוע לא החזיר מסמכים כלל. הבנו שיש צרוך בשיפור המנוע שיצליח להחזיר מסמכים לרוב השאילתות.

1. ניסוי שני: החלטנו להוסיף למנוע אינדקס על הגוף המסמך בנוסף לכותרת. ביצענו stemming על הכותרות בלבד לאור תוצאות הניסוי הקודם. את הדמיון בין המסמך לשאילתה חישבנו על ידי מספר שיטות: cosine similarity, tf-idf ו-bm25. בדקנו שילובי משקלים שונים לכותרת ולגוף, והמשקלים שיצאו הכי טובים היו כך שחישוב הכותרת היווה 30% מהציון וחישוב הגוף היווה 70% מהציון.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| method | duration | rq | precision@5 | precision@10 | f1@30 | recall@5 | recall@10 |
| Cos sim | 5.467 | 0.139 | 0.215 | 0.196 | 0.130 | 0.031 | 0.055 |
| Tf-idf | 6.432 | 0.121 | 0.186 | 0.172 | 0.114 | 0.027 | 0.049 |
| Bm25 | 5.349 | 0.110 | 0.157 | 0.146 | 0.097 | 0.022 | 0.041 |

תוצאות הניסוי: ממוצע המדדים

מסקנות: בכל השיטות זמן החישוב התארך, אך ניתן לראות שיש שיפור ניכר במדדים. כמעט לכל השאילתות הוחזרה תשובה. בשלב זה, השיטה הכי טובה בכל המדדים הינה cos sim.

1. ניסוי שלישי: בניסוי זה ניסינו לשפר את התוצאות באמצעות הוספת page rank . בשלב זה המרנו את page rank ו-tfidf עבור כותרות למילונים כבר בשלב טעינת המנוע, בכדי להפחית מזמן הריצה. בנוסף, שמנו לב שעבור שאילתות קצרות יותר המנוע מתקשה להחזיר מסמכים, כנראה כי אין לו מספיק מידע. למשל עבור genetics לא קיבלנו מסמכים רלוונטיים כלל – כאשר עשינו stem לשאילתה קיבלנו genet שזה גחן, ועבור genetics לא חזרו הרבה מסמכים רלוונטיים. לכן הוספנוquery expansion בעזרת word2vec עבור שאילתות קצרות באורך מילה (בדקנו גם עבור שאילתות ארוכות, אך זה לפעמים הגיע ל-timeout ולא היה שיפור משמעותי במדדים). שקלול הדמיון חולק באופן הבא: כותרת 30%, גוף 60%, page rank 10%. הניסוי שערכנו הינו בדיקה כמה stemming לגוף המסמך תרם לשיפור המדדים. לשם כך יצרנו df ו-posting list למסמכים שעברו stemming.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| method | stem | duration | rq | precision@5 | precision@10 | f1@30 | recall@5 | recall@10 |
| Cos sim | No | 7.37 | 0.232 | 0.44 | 0.34 | 0.169 | 0.055 | 0.085 |
| Cos sim | Yes | 10.87 | 0.237 | 0.453 | 0.346 | 0.173 | 0.056 | 0.085 |
| BM25 | No | 3.13 | 0.253 | 0.487 | 0.36 | 0.184 | 0.055 | 0.084 |
| BM25 | Yes | 4.05 | 0.261 | 0.473 | 0.353 | 0.193 | 0.056 | 0.081 |
| TF-IDF | No | 2.78 | 0.197 | 0.4 | 0.29 | 0.146 | 0.05 | 0.073 |

תוצאות הניסוי: ממוצע המדדים

מסקנות: השיטה של w2v תרמה למדדים של שאילתות קצרות, ו-page rank העלתה את כלל המדדים. עם זאת, לא היה הבדל משמעותי במדדים בין stemming לבין ללא stemming, אך זמן השאילתה עבור stemming התארך בשנייה ולכן החלטנו לא להשתמש ב-stemming עבור גוף המסמך. cos sim ו-bm25 בעלי תוצאות דומות במדדי הדמיון, החלטנו לקחת את bm25 בגלל שהתוצאות שלה היו מעט גבוהות יותר והיא מחזירה תוצאות מהר יותר באופן משמעותי.

1. ניסוי רביעי: לקחנו את מודל BM25 ללא stemming ועשינו לו אופטימיזציה. שינינו את היחס בין הניקוד על הגוף, הכותרות וה-page rank, ושינינו מעט את הפרמטרים של BM25 כדי לראות אם יש שינוי משמעותי בתוצאות.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| title | body | PR | b | K1 | K3 | duration | rq | precision@5 | f1@30 | recall@10 |
| 60% | 30% | 10% | 0.75 | 1.2 | 1 | 3.108 | 0.275 | 0.513 | 0.205 | 0.095 |
| 60% | 20% | 20% | 0.75 | 1.2 | 1 | 3.430781 | 0.305 | 0.52 | 0.229 | 0.108 |
| 60% | 10% | 30% | 0.75 | 1.2 | 1 | 3.066 | 0.280 | 0.513 | 0.213 | 0.098 |
| 80% | 0% | 20% | 0.75 | 1.2 | 1 | 0.699 | 0.201 | 0.4 | 0.147 | 0.076 |
| 60% | 20% | 20% | 0.75 | 1.2 | 0 | 3.494 | 0.305 | 0.52 | 0.229 | 0.108 |
| 60% | 20% | 20% | 0.75 | 2 | 1 | 3.504 | 0.305 | 0.526 | 0.228 | 0.102 |

תוצאות הניסוי: ממוצע המדדים

מסקנות: החלוקה של 60% לכותרת, 20% לגוף ו-20% ל-PR הניבה תוצאות הכי טובות. לא היו הבדלים גדולים בזמנים, למעט החישוב בו לא הבאנו משקל לגוף כלל, אך הוא הניב תוצאות הרבה פחות טובות. מבחינת פרמטרים ל-BM25, לא ראינו הבדל גדול בשינוי הפרמטרים. לכן, החלטנו לקחת את הפרמטרים הראשוניים שבחרנו, שהניבו תוצאות מעט טובות יותר בזמן מעט קצר יותר.

**המודל הסופי:** 60% עבור אינדקס על כותרות שעברו stemming בשיטה משולבת של TF-IDF ו-Cos Sim.

20% עבור אינדקס על הגוף (ללא Stem) באמצעות BM25, עם הפרמטרים k1=1.2, k3=1, b=0.75.

20% עבור דירוגי Page Rank. בנוסף, שאילתות בעלות מילה אחת הורחבו באמצעות word2vec בכ-3 מילים.

1. **הערכת רלוונטיות:**

השאילתה הכי טובה שהייתה לנו היא genetics עם rq= 0.619, p@5=1, p@10 = 0.9. התוצאות שקיבלנו היו להלן:

Genetics, Genome, Dominance (genetics), Population genetics, Genetic disorder, Genetic engineering, Genetic recombination, Molecular genetics, Genetic code, Genomics.

ניתן לראות שכל התוצאות הן בשדה הסמנטי של גנטיקה, ומכילות את המילים genetic או genom.

בהתחלה קיבלנו תוצאות לא טובות על השאילתה הזו, אך הן השתפרו מאחר וביצענו query expansion שהכניס את המילה genome וחיזק את השדה הסמנטי של הביטוי, כך שהשאילתות שעוסקות בגנטיקה קיבלו ציונים יותר טובים גם בחישוב על גוף המסמך. בנוסף, ה-page rank חיזק את הדפים המרכזיים בגנטיקה כיוון שהרבה דפים בתחום מצביעים אליהם, ולכן הם קיבלו דירוג סופי גבוה יותר.

השאילתה הכי פחות טובה שהייתה לנו היא Who is considered the "Father of the United States"?, שקיבלה ציון 0 בכל המדדים. התוצאות שקיבלנו היו להלן:

Honorific nicknames in popular music, Father, American Australians, List of NBA players born outside the United States, King v. Smith, Samuel C. Sample, Parental responsibility (access and custody), Elizabeth Morgan Act, Putative father, Andrew Gregg.

חקרנו את המסמכים שחזרו, וגילינו שרובם אכן עוסקים באבות ובארה"ב, אך כיוון שהמנוע לא מבין משמעות סמנטית של ביטויים ומחפש מסמכים שהמילים הללו חזקות בהם באופן בדיד, הוא לא הצליח להבין שמדובר בקשר מסויים בין האומה לבין מי שהקים אותה. למשל במסמך הראשון שחזר, מופיעים הרבה אמנים, חלקם אמריקאים, שמכונים אבות של תחום מסוים במוזיקה. המסמכים שבאים אחר כך עסקו בעיקר באבהות או בסוגיות שקשורות לארה"ב, והיו מספר דפים אחר כך שעסקו בחוקים בארה"ב בנוגע לאבהות, אך אף אחד מהעמודים לא קשור למייסדי ארה"ב. אם היינו משתמשים במנוע שמחזיק קונספטים ומקשר ביניהם, כמו למשל LSI, ייתכן שהיינו מצליחים לענות על שאילתה זו בצורה טובה יותר.

1. תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

   התיאור נוצר באופן אוטומטי**גרף תיאור ביצועי המנוע:**

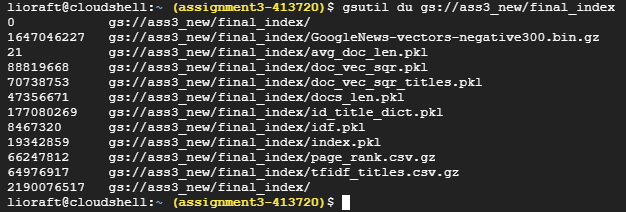
**7.גרף ממוצע זמן אחזור עבור שאילתה:**

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטי

**נספחים:**

רשימת קבצים שבהם השתמשנו באינדקס במהלך הפרויקט, איחדנו אותם תחת ה-directory הנ"ל:

****

יצרנו קבצים נוספים שבחרנו לא להשתמש בהם, כולם נמצאים ב-bucket. כמו כן, השתמשנו גם ב-posting list שכתבנו בעבודה 3 (תחת תיקיית gcp\_postings).