דוח דרישות – פרויקט אחזור מידע

1. **מגישות:**

רותם אמיר: 319041208, amirrot@post.bgu.ac.il

ליאור אפטבי: 209259993, lioraft@post.bgu.ac.il

2. **קישור לגיט-האב:** https://github.com/lioraft/IR_SearchEngine/tree/fixes1003
3. **קישור לGoogle Storage Bucket:** gs://ass3_new
4. **תיאור ניסויים (כולל הערכות ומסקנות):**

א. 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.

ב. 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. Bm25 נבדק בשלב זה באמצעות פרמטרים דיפולטיביים. בדקנו שילובי משקלים שונים לכותרת ולגוף, והמשקלים שיצאו הכי טובים היו כך שחישוב הכותרת היווה 30% מהציון וחישוב הגוף היווה 70% מהציון.

תוצאות הניסוי: ממוצע המדדים

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

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.

ג. 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 הוספנו query 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.

תוצאות הניסוי: ממוצע המדדים

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

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.

ד. <u>ניסוי רביעי:</u> לקחנו את מודל BM25 ללא stemming ועשינו לו אופטימיזציה. שינינו את היחס בין הניקוד על הגוף, הכותרות וה-page rank, ושינינו את הפרמטרים של BM25 בחלק מהווריאציות כדי לראות אם יש שינוי משמעותי בתוצאות.

<u>תוצאות הניסוי: ממוצע המדדים</u>

recall@10	f1@30	precision@5	rq	duration	K3	K1	b	PR	body	title
0.095	0.205	0.513	0.275	3.108	1	1.2	0.75	10%	30%	60%
0.108	0.229	0.52	0.305	3.430781	1	1.2	0.75	20%	20%	60%
0.098	0.213	0.513	0.280	3.066	1	1.2	0.75	30%	10%	60%
0.076	0.147	0.4	0.201	0.699	1	1.2	0.75	20%	0%	80%
0.108	0.229	0.52	0.305	3.494	0	1.2	0.75	20%	20%	60%
0.102	0.228	0.526	0.305	3.504	1	2	0.75	20%	20%	60%
0.113	0.252	0.56	0.333	3.166	1	1.2	0.5	20%	20%	60%
0.114	0.256	0.573	0.34	2.92	0	2	0.5	20%	20%	60%

<u>מסקנות:</u> 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 בחירת פרמטרים ל-BM25, בחנו ערכים שונים עבור כל פרמטר ושילובים בניהם, ולקחנו את הפרמטרים שהניבו את התוצאות הכי טובות מבחינת rq, recall, presicion.

<u>המודל הסופי:</u> 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=2, k3=0, b=0.5.
20% עבור דירוגי Page Rank. בנוסף, שאילתות בעלות מילה אחת הורחבו באמצעות word2vec בכ-3 מילים.

5. <u>הערכת רלוונטיות:</u>

השאילתה הכי טובה שהייתה לנו היא genetics עם rq, precision גבוהים מאוד. התוצאות שקיבלנו היו להלן:

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

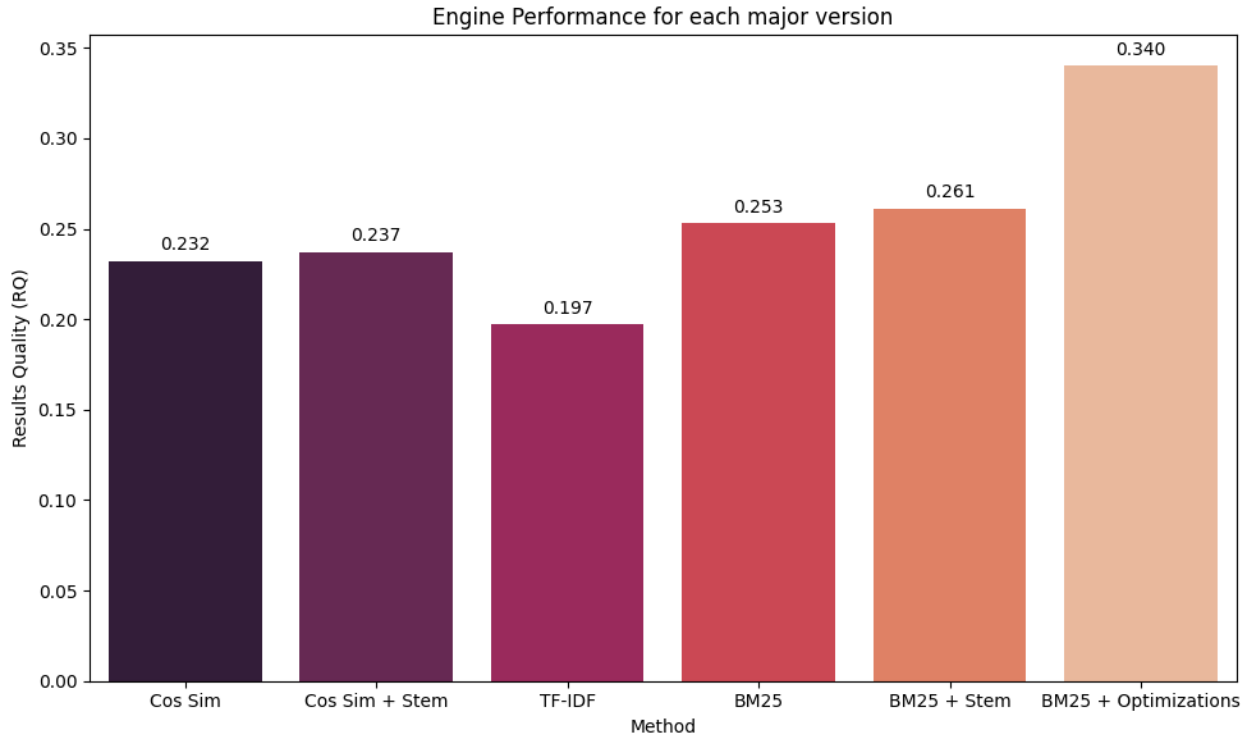
ניתן לראות שכל התוצאות הן בשדה הסמנטי של גנטיקה, ומכילות את המילים genetic או genom.
בהתחלה קיבלנו תוצאות לא טובות על השאילתה הזו, אך הן השתפרו מאחר וביצענו query expansion שהכניס את המילה genome וחיזק את השדה הסמנטי של הביטוי, כך שהשאילתות שעוסקות בגנטיקה קיבלו ציונים יותר טובים גם בחישוב על גוף המסמך.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

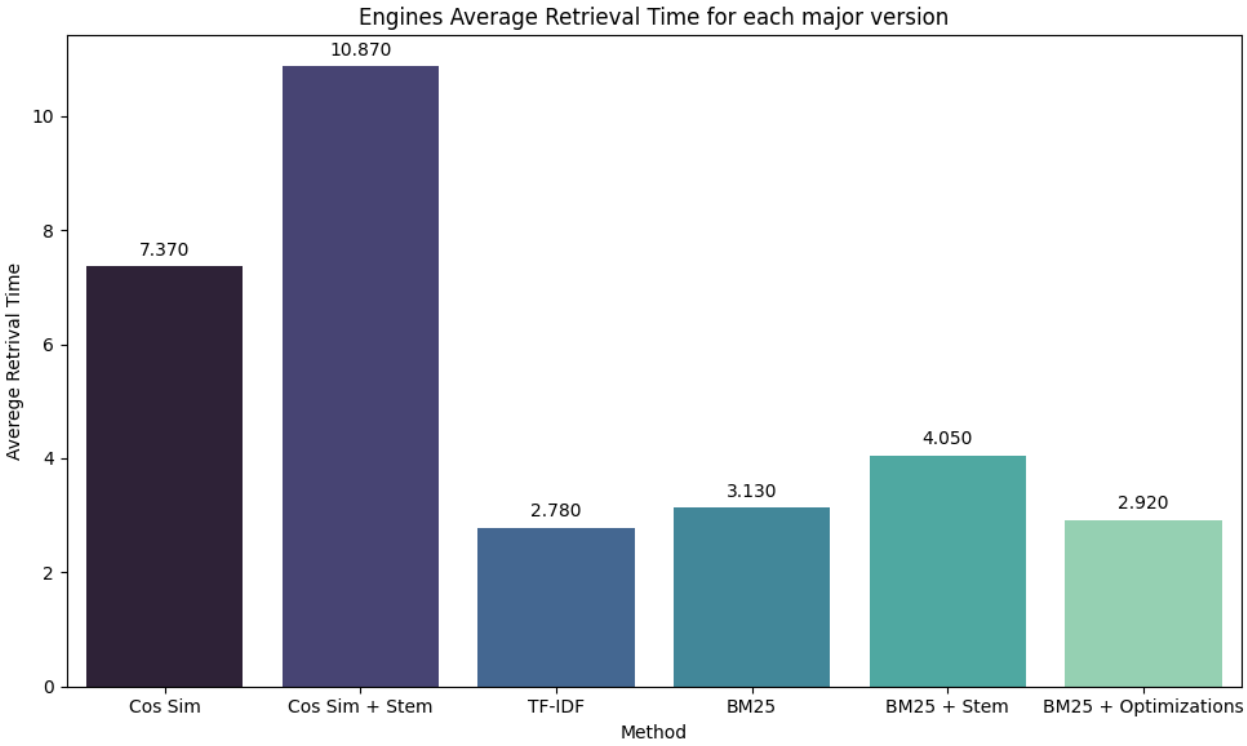
השאילתה הכי פחות טובה שהייתה לנו היא "Who is considered the "Father of the United States?, שקיבלה ציון 0 בכל המדדים. התוצאות שקיבלנו היו להלן:

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.

6. <u>גרף תיאור ביצועי המנוע וגרף ממוצע זמן אחזור עבור שאילתה:</u>





נספחים:

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

```
lioraft@cloudshell:~ (assignment3-413720)$ gsutil du gs://ass3_new/final_index
0            gs://ass3_new/final_index/
1647046227   gs://ass3_new/final_index/GoogleNews-vectors-negative300.bin.gz
21           gs://ass3_new/final_index/avg_doc_len.pkl
88819668     gs://ass3_new/final_index/doc_vec_sqr.pkl
70738753     gs://ass3_new/final_index/doc_vec_sqr_titles.pkl
47356671     gs://ass3_new/final_index/docs_len.pkl
177080269    gs://ass3_new/final_index/id_title_dict.pkl
8467320      gs://ass3_new/final_index/idf.pkl
19342859     gs://ass3_new/final_index/index.pkl
66247812     gs://ass3_new/final_index/page_rank.csv.gz
64976917     gs://ass3_new/final_index/tfidf_titles.csv.gz
2190076517   gs://ass3_new/final_index/
lioraft@cloudshell:~ (assignment3-413720)$
```

יצרנו קבצים נוספים שבחרנו לא להשתמש בהם במודל הסופי, כולם נמצאים ב-bucket. כמו כן, השתמשנו גם ב-posting list שכתבנו בעבודה 3 (תחת תיקיית gcp_postings שמופיעה בשורה האחרונה).

```
1999998      gs://ass3_new/postings_gcp/9_015.bin
1999998      gs://ass3_new/postings_gcp/9_016.bin
1999998      gs://ass3_new/postings_gcp/9_017.bin
1999998      gs://ass3_new/postings_gcp/9_018.bin
1999998      gs://ass3_new/postings_gcp/9_019.bin
1999998      gs://ass3_new/postings_gcp/9_020.bin
1999998      gs://ass3_new/postings_gcp/9_021.bin
1318284      gs://ass3_new/postings_gcp/9_022.bin
102753       gs://ass3_new/postings_gcp/9_posting_locs.pickle
6341427666   gs://ass3_new/postings_gcp/
lioraft@cloudshell:~ (assignment3-413720)$
```