**דוח דרישות – פרויקט אחזור מידע**

1. מגישות:

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

ליאור אפטבי:

1. קישור לגיט-האב:
2. קישור לGoogle Storage Bucket
3. רשימה של מסמכי אינדקס
4. תיאור ניסויים: (כולל הערכות ומסקנות)
5. גרף תיאור ביצועי המנוע (השוואת גרסאות)
6. גרף ממוצע זמן אחזור עבור שאילתה (השוואת גרסאות)
7. הערכת רלוונטיות עבור 10 המסמכים הראשונים שהוחזרו עבור שאילתה אחת שהמנוע עבדה בצורה טובה ועבור שאילתה שבה המנוע לא עבדה בצורה טובה.

**שיטות עבודה:**

אינדקס:

1. Tf-idf
2. Tf-idf ממושקל
3. Lsi
4. Page rank
5. Stemming - porter
6. Stop words

בקוד שלנו:

1. TF
2. DF
3. IDF
4. הורדה של stop words
5. Tf-idf weighted.
6. Page rank
7. בהמשך חייב להוסיף stemming

דמיון וחיפוש:

1. Cosine similarity
2. Bm25

תוכנית עבודה:

1. אינדקס
2. מנוע חיפוש
3. ניסויים על 1000 מסמכים + מדדים כדי לדעת אם האנדקס מספיק מקיף
4. אינדוס על כל הקורפוס
5. שיפור המנוע חיפוש