**איחזור מידע-פרויקט**

יהונתן קידושים 319068789 [kidushim@post.bgu.ac.il](mailto:kidushim@post.bgu.ac.il)

אופק כחלון 208590174 [ofekkah@post.bgu.ac.il](mailto:ofekkah@post.bgu.ac.il)

Github Repository:

<https://github.com/InformationRetrievalFinalProject/Finalproject>

Google storage bucket:

<https://console.cloud.google.com/storage/browser/ir-project-bucket-irproject-414719>

בתור התחלה, עבדנו בcolab וחשבנו איך נוכל לשפר את הinverted index ממטלה 3. החלטנו שפונקציית המשקל שלנו תהיה Tf-idf. בנוסף, החלטנו שפונקציית הדימיון שלנו תהיה .Cosine similliarity התחלנו מבדיקה של מה חסר לנו עבור ה Cosine similliarityו- tfidf, לאחר מחשבה שביצוע החישוב של ערך הtfidf של כל term בכל מסמך ובנוסף ערך הnf של כל מסמך יגדיל את זמן הריצה הסופי של האיחזור, אנחנו מבצעים את החישוב לפני השאילתא והתוצאות ישמרו במשתנים גלובליים. לצורך נכונות ובדיקת החישובים, עבדנו על חלק קטן מהקורפוס שנקרא- multistream1\_preprocessed.

לאחר הבנה כי החישובים נכונים, החלטנו לעבור לעבודה בסביבת הgcp, על מנת לבחון כי אכן פונקציית הדימיון מחזירה את המסמכים הרלוונטים הוחלט כי עלינו לבנות inverted index קטן יותר מהקורפוס הגדול שיכיל מילים ספציפיות מהשאילתות שקיבלנו בקובץ queries\_train. בחרנו ב-