כריית מידע – תרגיל בית מספר 1

מגישים:

* אמיר לביא 300202595
* אלירן יחזקאל 021682042

תיעוד:

* בתחתית הקובץ תוכלו למצוא 4 סקריפטים להרצת הקבצים הנדרשים train ו test
  + מחקו את הערת הסקריפט אותו תרצו להריץ
  + יש להריץ 2 סקריפטים run ו read של אותו קובץ
* הסקריפט run יקרא את ה csv וייצר את מילון החתימות
* הסקריפט read יקרא את מילון החתימות מקבצי ה pkl
  + בשני המקרים הנ"ל תורץ פונקציית ג'קארט אשר תחשב את המרחק בין וקטורי המסמכים
* בסוף הרצת סקריפט ה train יודפס הזמן הממוצע אשר לחשב כל חתימת מסמך
* הגלובלי dictionary יחזיק את מאגר החתימות
* הגלובלי jaccard\_compatibility יחזיק את המסמכים להם יש את הציון ג'קראט 0.8 ומעלה