ניהול נתונים באינטרנט – מטלה 3

מגישים :
שם: נתן בלוך , ת.ז: 316130707
שם: שי פוקס , ת.ז: 313452252

<u>תיאור הפרויקט:</u>

נתאר את הקבצים השונים בפרויקט ואת תפקידם-

- extraction.py – בקובץ זה אנו בונים את ה-inverted_index לפי הארכיטקטורה שנלמדה בכיתה(שילוב של טבלאות hash) כאשר הפלט הוא קובץ json.
  ראשית אנו מחלצים את המידע מתוך קבצי ה-xml שנמצאים בתיקיית הקלט באמצעות XPATH כאשר המידע הרלוונטי שחילצנו נמצא תחת האלמנטים : TITLE, ABSTRACT ,EXTRACT.
  לאחר מכן ביצענו טוקניזציה ,ניקוי של stop-words , stemming וניקוי של תווים מספריים.
  קובץ ה-json מכיל עבור כל מילה שחילצנו מקבצי ה-xml(לאחר טוקניזציה, stemming וכו'):
  - ציון df_i עבור המילה.
  - ציון idf_i עבור המילה.
  - רשימת מסמכים שהמילה מופיעה בהם וציון ה-tf של המילה והמסמך.

  בנוסף קובץ ה-json מכיל את מס' המסמכים בקורפוס ואת "אורכי" המסמכים לפי החישוב שנלמד בכיתה: $\left|d_j\right| = \sqrt{\sum_{i=1}^{t} w_{ij}^2}$ כאשר $w_{ij} = tf_{ij} \cdot idf_i$.
- query.py – מקבל כארגומנטים שאילתה ואת ה-inverted_index ומחשב כפלט את המסמכים הרלוונטיים ביותר לשאילתה לפי חישוב $F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} > 0.09$.
- vsm_ir.py – קובץ פייתון שמריץ את התכנית לפי ההנחיות בעבודה.