

## קובץ תיעוד לפרויקט 2 - ניהול נתונים באינטרנט

לירון כהן 207481268, יובל מור 209011543

### תיאור הקוד

#### חלק א' - בניית *Inverted Index* (*create index*)

- עבור כל קובץ *xml* בתיקייה הנתונה:
  - פרסור הקובץ ל-*Element Tree*
  - מציאת כל המסמכים (*records*) בקובץ
  - עבור כל מסמך:
    - חילוץ שדה ה-*record num*
    - חילוץ כל המילים במסמך (*get\_words\_from\_record*)
  - השדות שבחרנו לחלץ והובילו לתוצאות טובות הם:  
*TITLE, ABSTRACT, EXTRACT, TOPIC*
  - עבור כל מילה במסמך:
    - ביצוע *stemming*, הסרת סימני פיסוק, ניפוי *stopwords* והמרה ל-*lower case*.
    - שמירת המילה במילון המילים (*update\_words\_dict*).
    - העלאת ה-*counter* של המילה במילון המסמכים תחת המסמך המתאים.
    - עדכון ערך ה-*max frequency* של המסמך בעת הצורך.
    - חישוב ערכי ה-*tf* של המילים במסמך (*calc\_tf\_values*):
    - עבור כל מילה במילון המילים, אם היא נמצאת במסמך, נחלק את הערך שלה ב-*max frequency* של המסמך.
    - חישוב ערכי ה-*idf* של המילים (*calc\_idf\_values*):
      - חישוב *D*, שהוא כמות המסמכים.
      - עבור כל מילה במילון המילים נבצע את חישוב הנוסחה לקבלת ה-*idf* מתוך ה-*tf* ו-*D*.
    - חישוב ערכי המשקלים של המילים (*calc\_weight\_values*):
      - הכפלת ה-*tf* ב-*idf* שחושב לעיל לקבלת ערך ה-*tf - idf* של כל מילה.
    - שמירת מילון המסמכים ומילון המילים לקובץ *json* (*save\_index\_dict\_to\_json*).

#### חלק ב' - אחזור מידע בהינתן שאלה (*ask question*)

- טעינת מילון המסמכים ומילון המילים מקובץ ה-*json* (*load\_index\_dict\_from\_json*).
- פרסור השאלה (*parse\_query*)
- אם מדובר ב-*tf - idf*, חישוב ציונים בהתאם (*calc\_tfidf\_grades*):
  - עבור כל מסמך:
    - חישוב ערכי *cosine similarity* באמצעות הנוסחה המתאימה (*calc\_cosine\_similarity*).
    - אם התוצאה מעל *threshold* מסוים, נכניס את התוצאה למילון *relevant\_records*.
  - מיון המילון בסדר יורד לפי הערכים שהתקבלו והחזרתו.

- אם מדובר ב- $BM25$ , חישוב ציונים בהתאם  $(calc\_bm25\_grades)$ :
  - חישוב הקבועים הכלליים הנדרשים לנוסחה  $(avgdl, N)$ .
  - עבור כל מסמך:
    - חישוב ערך ה- $D$  המתאים.
    - חישוב ערך ה- $BM25$  באמצעות הנוסחה המתאימה  $(calc\_bm25\_grade\_for\_record)$ .
  - אם התוצאה מעל  $threshold$  מסוים, נכניס את התוצאה למילון  $relevant\_records$ .
- מיון המילון בסדר יורד לפי הערכים שהתקבלו והחזרתו.
- שמירת התוצאות לקובץ טקסט  $(save\_query\_result\_to\_txt)$ .