שאילתות באינטרנט

תרגיל 2 – בניית האינדקס

תאריך הגשה: ט"ז סיון (19 ביוני)

1. תיאור התרגיל

בתרגיל זה, המטרה היא לשפר את הזמן הנדרש לבניית האינדקס ולאפשר לו לגדול ולתמוך בערכות נתונים גדולות מאוד.

לצורך כך עליכם לממש את המחלקה הבאה בנוסף למחלקות הקודמות שנועדו לקריאה מהאינדקס (ReadIndex מתרגיל מספר 1)

```
class IndexWriter:
     def write(self, inputFile, dir):
          """Given product review data, creates an on
          disk index
          inputFile is the path to the file containing
          the review data
          dir is the directory in which all index files
          will be created
          if the directory does not exist, it should be
          created"""

     def removeIndex(self, dir):
          """Delete all index files by removing the given
          directory"""
```

בשלב זה אין להניח כי יש מספיק זיכרון כדי לאחסן את כל הנתונים הגולמיים בבת אחת. במקום זאת עליכם להשתמש באלגוריתמים שיילמדו בהרצאה (או לפתח אחרים משלכם) המאפשרים את יצירת האינדקס למרות גודלו של קובץ הקלט.

סביר להניח כי מבנה האינדקס יהיה זהה למבנה בתרגיל מספר 1 אולם אתם רשאים לשנותו.

שימו לב, כי כל הפעולות של IndexReader צריכות להישאר יעילות למרות הגודל של ערכת הנתונים. אם IndexReader תוכנן היטב בתרגיל מספר 1, זה יקרה מעצמו. אחרת, יהיה צורך לעשות בו שינויים.

2. ניתוח ביצועים

בנוסף להגשת הקוד, עליכם לבצע ניתוח של ביצועי התכנית שבניתם.

עליכם להריץ את התכנית שכתבתם תחילה על קובץ עם 1000 ביקורות ולאחר מכן עם מספר עולה של ביקורות, בכל פעם פי 10 ביקורות מהניסוי הקודם. עד שהצלחתם לבנות אינדקס של כל הביקורות בקובץ או עד שהתכנית שלכם לא הצליחה לבנות את האינדקס בזמן סביר (כשעתיים).

לצורך הבדיקות עליכם להשתמש בקובץ books.txt שקישור אליו מופיע באתר הקורס. הקובץ מכיל קרוב ל 9 מיליון ביקורות. עליכם לבנות מהקובץ קבצים בגודל הנדרש ועליהם לבצע את הבדיקות.

2.1. דרישות ניתוח הביצועים

ניתוח הביצועים צריך להכיל את הנתונים הבאים:

- זמן ריצה של בניית האינדקס.
- גודל האינדקס על הדיסק.
- הזמן שלוקח להריץ 100 שאילתות אקראיות של getReviewsWithToken ו 100 שאילתות אקראיות של TokenFrequency.

את האנליזה יש להראות בעזרת גרפים המראים את זמן הריצה ואת גודל האינדקס כפונקציה של מספר הביקורות באינדקס.

ביצועי התכנית תלויים כמובן במחשב עליו אתם מריצים אותה. לכן עליכם לציין בדו"ח ניתוח הביצועים את סוג המחשב עליו הרצתם את התכנית. הנתונים אותם עליכם לציין הם:

- מערכת ההפעלה
- מהירות המעבד
- גודל הזיכרון
- סוג הדיסק.

<u>לתשומת ליבכם! ביצוע בדיקות הביצועים – הכנת הקבצים והרצתם לוקח לא מעט זמן, אל תתחילו את שלב הבדיקות ברגע האחרון!</u>

3. הגשת התרגיל

- התרגיל יוגש דרך אתר המכללה בפורמט ZIP.
- עבור כל זוג יש להגיש רק הגשה אחת. שם הקובץ צריך להיות ID1_ID2.zip כאשר ID1 ו ID2 הם מספרי הזהות של הסטודנטים המגישים את הפרויקט.

- **<u>קבצי הקוד צריכים לכלול שני קבצים עם השמות IndexWriter.py ו IndexReader.py.</u>**
  <u>תכנית הבדיקה תייבא (import) קבצים אלו כך שחשוב שתקפידו על השמות הנכונים (כולל אותיות גדולות וקטנות) קבצים אלו יכילו את המחלקות שנדרשתם לפתח.</u>
- יש לוודא כי הקובץ שהועלה הוא בפורמט הנכון ומכיל את כל הקבצים הרלוונטיים כולל:
  - קבצי הקוד **<u>(יש לכלול את הקוד של IndexReader גם אם הוא לא השתנה מהתרגיל הקודם)</u>**
  - קובץ README הכולל הנחיות רלוונטיות לקומפילציה או להרצה, שמות ותעודת זהות של הסטודנטים המגישים את התרגיל.
  - קובץ בשם analysis.pdf שיכיל את ניתוח הביצועים שהתבקש בתרגיל.