תרגיל מס' 2 – שיטות אינדוקס והשוואתן

בתרגיל זה תשתמשו בנתונים מפלטפורמת גיוס כספים חברתית בשם Kiva אשר מאפשרת לאנשים להלוות כסף באמצעות רשת האינטרנט ליזמים וסטודנטים בכ-80 מדינות.

ברשותכם קובץ טקסט המאורגן כ-heap (kiva_loans.txt) שמכיל פרטים על 10,000 ההלוואות שבוצעו:

1. lid- קוד מזהה ייחודי להלוואה Loan ID.
2. loan amount- סכום ההלוואה.
3. currency- מטבע.
4. sector- ענף, לדוגמה: חקלאות, מזון, קמעונאות ועוד.

עליכם לממש את שיטות האחסון הבאות:

1. קובץ לא ממוין- קובץ heap.
2. אינדקס ממוין שעליו נבצע חיפוש בינארי (באינדקס).
3. שימוש בקובץ אינדקס המאורגן בשיטת hash table עם פונקציית hash: value % N + 1.
   כאשר N מייצג את מרחב הכתובות (כמות המגרות).
   עבור מחרוזת, פונקציית הhash תחושב על המספר המייצג את קוד הUnicode של התו הראשון במחרוזת.
   לדוגמה: עבור המחרוזת 'apple' הפונקציה תחושב עבור value = 97.

עליכם לדווח על זמני ריצה ממוצעים של הפעולות הבאות באמצעות כל אחת מהשיטות: heap (לא ממוין), אינדקס ממוין וטבלת ערבול עם שלושת האפשרויות הבאות לכמות המגרות N=10,100,1000:

1. משך אינדוקס (יצירת האינקדס) של כל קובץ הנתונים עבור כל אחת מארבעת האפשרויות.
2. הוספת רשומה לקובץ הנתונים ובמידת הצורך לקובץ האינדקס.
3. עדכון רשומה בקובץ הנתונים ובמידת הצורך בקובץ האינדקס.
4. מחיקת רשומה ע"י סימון בקובץ הנתונים ובמידת הצורך מחיקה מקובץ האינדקס.

הריצו את הפעולות הנ"ל עבור 1,000 הרשומות הראשונות בקובץ על השדות הבאים: lid, loan amount, sector.
הציגו את זמני הריצה הממוצעים עבור שיטות האחסון השונות, תעזרו במודול timeit למדידת זמני הריצה.
את פעולות ההוספה, עדכון ומחיקה (סעיפים 2-4) יש לדווח על ממוצע משך הפעולה עבור 50 רשומות.

הציגו תרשימים עבור משך הזמן הממוצע של כל פעולה (אינדקוס, הוספה, עדכון ומחיקה) של כל אפשרות. הגרפים יכילו את רווחי בר הסמך 95% של כל ממוצע על פי סטיית התקן. דונו בתוצאות ובדקו את מובהקותן.

הנחיות והערות נוספות:

- **אסור** להעלות לזיכרון את כל קובץ הheap או קובץ האינדקס.
- **אסור** לשנות את החתימות של הפונקציות הקיימות.
- הפונקציות צריכות להיות ממומשות בצורה גנרית, כך שיהיו ניתנות לריצה על קבצי נתונים בפורמט CSV עם ערכים אחרים. שורה ראשונה תכיל את שמות העמודות כמו בkiva.
- במידה והוספתם למחלקות פונקציות פרטיות, יש לתעדם היטב.

- אין לבצע הדפסות מיותרות מתוך הפונקציות.
- בסיום כל פונקציה כל הקבצים עימם עבדתם צריכים להיות סגורים.
- הניחו כי ערכי הקלט תקינים (הקובץ קיים, העמודה קיימת בקובץ וכו').
- **אסור** לייבא מודולים נוספים, כולל pandas וnumpy.
- בעבודה תמצאו מספר בדיקות שנועדו לשימושכם בלבד, בנוסף יערכו בדיקות שלמות ותקינות לקוד.

הוראות הגשה

- ההגשה בזוגות **בלבד**, גם עבור סטודנטים שלא הגישו בזוג בתרגיל הקודם.
- יש להגיש קובץ ZIP המכיל תיקייה בשם ex2 ובתוכה קובץ קוד בשם ex2.py וקובץ PDF בלבד.
  כלומר, ללא קבצי הנתונים.
- יש לקרוא לקובץ הZIP לפי ת"ז המגישים ID1_ID2.zip.
- רק אחד מבני הזוג יגיש את העבודה.
- תאריך הגשה: 24-05-2018 חצות.
- העבודה תכתב בPython 2.7.
- יש לשאול שאלות רק בפורום באתר הקורס.
- מקרים פרטניים של בקשת הארכה יתבצע דרך המייל.

בהצלחה!