שאילתות באינטרנט

תרגיל 1 – מבנה האינדקס

1. ערכת הנתונים

נתון קובץ המכיל אוסף של ביקורות על מוצרים. כל אחת מהביקורות היא מהמבנה הבא:

product/productId: B001E4KFG0

review/userId: A3SGXH7AUHU8GW

review/profileName: delmartian

review/helpfulness: 1/1

review/score: 5.0

review/time: 1303862400

review/summary: Good Quality Dog Food

review/text: I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most.

מכיוון שלביקורות אין מזהה (ID), נמספר אותן בסדר עולה. כלומר הביקורת הראשונה תקבל את המזהה 1, הביקורת השנייה תקבל את המזהה 2 וכן הלאה.

מתוך הנתונים שיש לכל אחת מהביקורות, נתעניין בשדות הבאים:

- product/productId (string containing 10 characters)
- review/helpfulness (two integers)
- review/score (integer between 1 and 5)
- review/text

עיבוד הטקסט של הביקורות למילים יתבצע באופן הבא:

- חלוקת הטקסט למילים נפרדות בכל מקום שבו יש תו שאינו אלפאנומרי (אינו אות מהאלפבית האנגלי או ספרה). התווים שאינם אלפאנומריים צריכים להיות מושלכים.
- נרמול הטקסט על ידי הפיכת כל תווי האותיות לאותיות קטנות (lowercase)

הערה: מכיוון שזהו מידע אמתי, הוא עלול להכיל תוכן משונה. למשל ייתכן כי יימצא תו של newline באמצע profileName, או מילים מאוד ארוכות. עליכם לכתוב את התכנית בצורה שתתמודד בצורה טובה עם הפתעות שכאלה.

ניתן להניח כי הטקסט הוא רק בשפה האנגלית.

לצורך הרצת התכנית, ניתן להוריד מאתר הקורס שני קבצים של נתונים גולמיים. קובץ אחד מכיל 100 ביקורות והקובץ השני 1000 ביקורות.

2. תיאור התרגיל

בהינתן קובץ הקלט עם הנתונים הגולמיים, עליכם לבנות:

1. מילון שיאפשר מענה יעיל לשאילתות המתוארות בסעיף הבא. על המילון להיות דחוס כפי שמתואר בהמשך ולהישמר על הדיסק כדי שיהיה אפשר להשתמש בו שוב מבלי שיהיה צורך לבנות אותו מחדש..
2. קובץ נוסף שיאפשר מענה לשאילתות הנוספות שהמידע עליהן אינו נמצא במילון.

לאחר בניית המילון והקובץ אין צורך בקובץ הנתונים הגולמיים. כלומר, כל המידע הנדרש לצורך מענה לשיטות צריך להימצא בקבצים שנוצרו.

בשלב זה אין צורך לדאוג לכך שתהליך יהיה יעיל. עם זאת, מבנה המילון והמימוש צריך להיות ניתן לשדרוג כך שיוכל לאפשר בנייה ואחסון של כמות נתונים גדולה מאוד כך שתוכלו להשתמש בו כמו שהוא כאשר נרצה לבנות אינדקס עבור כמות גדולה יותר של נתונים.

2.1. המילון

עליכם לדחוס את במילון בשיטה של front coding 10 in 9.

כלומר, במחרוזת כל מילה עשירית תופיע בשלמותה ותשע המילים העוקבות לה יהיו מקודדות עם front-coding.

הטבלה תכיל שורה עבור כל בלוק. בכל שורה יהיה מצביע אל המקום במחרוזת בו נמצאת המילה הראשונה בבלוק (4 bytes). ובנוסף, עבור **כל מילה** בבלוק יופיעו הנתונים הבאים:

- אורך רשימת התפוצה (frequency) (4 bytes)
- אורך המילה (למעט עבור המילה האחרונה בבלוק שאורכה יכול להיות מחושב בעזרת המצביע לבלוק הבא) (1 byte)
- גודל התחילית המשותפת עם המילה הקודמת (למעט עבור המילה הראשונה בבלוק שאורך התחילית שלה הוא 0) (1 byte)

סך הכל, גודל כל שורה בטבלה הוא: 62 bytes

המילון צריך להישאר בזיכרון בצורתו הדחוסה. כל חיפוש במילון יבצע חיפוש בינארי על הטבלה כדי למצוא את הבלוק בו נמצאת המילה ובתוך הבלוק יתבצע חיפוש סדרתי למצוא את המילה עצמה.

<u>מבנה קובץ המילון:</u>

לקובץ המילון של ה text עליכם לקרוא בשם text.dic

4 בתים ראשונים – אורך המחרוזת בבינארית.

המחרוזת: מקודדת ב ASCII

הטבלה: לכל בלוק 62 בתים המכילים את המידע כמתואר למעלה.

string length (4)	the string (variable length)	block1 (62)	block2 (62)	block3 (62)	…

2.2. נתונים כלליים

מבנה הקובץ הנוסף עם הנתונים הכלליים נתון לבחירתכם.

3. דרישות הקוד

התכנית תכיל לפחות את שתי המחלקות הבאות: (ככל הנראה התכנית תכלול מחלקות נוספות הנחוצות לצורך מימושן)

3.1. FirstIndexWriter: בהינתן נתונים גולמיים, המחלקה תיצור את הקבצים הנדרשים על הדיסק בספריה שתתקבל כקלט.

בתרגיל זה ניתן להניח כי כאשר בונים את הקבצים, כל הנתונים יכולים להיות מאוחסנים בזיכרון.

המחלקה מאפשרת גם למחוק את הקבצים מהדיסק על ידי מחיקת כל הקבצים מהספרייה.

3.2. FirstIndexReader: לאחר שנוצרו הקבצים על הדיסק ניתן להשתמש במחלקה כדי לגשת למגוון של נתונים. כלומר מבנה הקבצים צריך לתמוך במימוש יעיל של המתודות המוגדרות. ניתן להניח כי המתודות יופעלו רק לאחר שהאינדקס ייבנה על ידי FirstIndexWriter.

תיאור של הממשק שצריך להיות ממומש מתואר בעמודים הבאים.

```
class FirstIndexWriter:
     def __init__(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 path of 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
          dir is the path of the directory to be
          deleted"""
```

```
class FirstIndexReader
     def __init__(self, dir):
          """Creates a FirstIndexReader object which will
          read from the given directory
          dir is the path of the directory that contains
          the index files"""

     def getTokenFrequency(self, token):
          """Return the number of reviews containing a
          given token (i.e., word)
          Returns 0 if there are no reviews containing
          this token"""


     def getNumberOfReviews(self):
          """Return the number of product reviews
          available in the system"""

     def getTokenSizeOfReviews(self):
          """Return the number of tokens in the system
          (Tokens should be counted as many times as they
          appear)"""
```

4. בדיקת התרגיל

בבדיקת התרגיל ייעשה שימוש בערכת נתונים קטנה בסדרי גודל של הערכות הנמצאות באתר הקורס.

התרגיל ייבדק שאכן המימוש נותן מענה נכון לשאילתות וכן שפורמט קובץ המילון תואם את הדרישות.

בדיקת התרגיל נעשית בעזרת מערכת אוטומטית. כדי שבדיקת התרגיל שלכם לא תיכשל (ותגרום להורדה בציון) הקפידו היטב על ההנחיות שבסעיף 5.

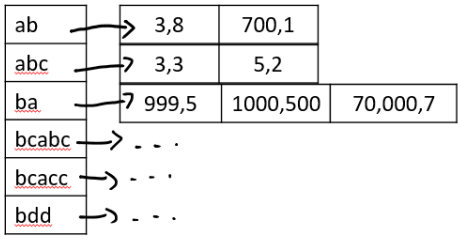
5. הגשת התרגיל

- 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.
- **קבצי הקוד צריכים לכלול שני קבצים עם השמות FirstIndexWriter.py ו FirstIndexReader.py. תכנית הבדיקה תייבא (import) קבצים אלו כך שחשוב**

<u>שתקפידו על השמות הנכונים (כולל אותיות גדולות וקטנות) קבצים אלו יכילו את המחלקות שנדרשתם לפתח.</u>

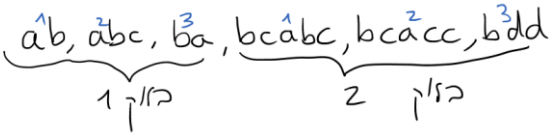
- **<u>במידה והפרויקט שלכם מכיל קבצי קוד נוספים (מחלקות נוספות), באחריותכם לייבא אותם (import) מתוך הקבצים FirstIndexWriter.py ו FirstIndexReader.py.</u>**
- יש לוודא כי הקובץ שהועלה הוא בפורמט הנכון ומכיל את כל הקבצים הרלוונטיים.

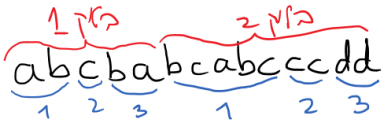
דוגמה לקידוד



קידוד המילון ב 2-in-3 front coding:

<u>קידוד המחרוזת:</u>





<u>קידוד הטבלה:</u>

	str ptr	term1		term2			term3	
		freq	length	freq	length	prefix	freq	prefix
block1	0	2	2	2	3	2	3	0
block2	5	…	5	…	5	3	…	1

סך הכל קידוד הקובץ:

string size (4bytes)	the string (14bytes)	block1 (20bytes)	block2 (20bytes)	…
14	abcbabcabcccdd	…	…	…