מטלת בית מס' 1

הנחיות:

- הגשה:
  - יש להגיש את המטלה עד לתאריך 9.4.18.
  - ההגשה תתבצע ע"י קובץ zip המכיל את הקוד שנכתב בהתאם לדרישות. פורמט קובץ ה-zip יהיה ID01_ID02.zip. שם קובץ ההרצה יהיה Ex1.py.
  - ההגשה הינה **בזוגות**.
  - חבר קבוצה אחד בלבד יעלה את הפתרון לאתר.
  - בעיות אישיות בנוגע למועד ההגשה יש להפנות לבודק התרגילים הקורס טרם מועד ההגשה.
  - **כל חריגה מנהלים אלו, ללא אישור בכתב מצוות הקורס, מהווה עילה לפסילת המטלה או להפחתת נקודות.**
  - **אין להעתיק פתרונות ואין לשתף קוד בין סטודנטים. אין להעתיק קוד מוכן באינטרנט!**
  - **לפתרון המטלה יש להשתמש בגרסת פייטון 2.7 בלבד ובחומר הנלמד במסגרת ההרצאות בקורס בלבד.**
  - להבהרות, הכוונות או כל עזרה אחרת ניתן לשלוח מייל - nahmiasd@post.bgu.ac.il
  - בדיקת המטלה תתייחס בין השאר לפרמטרים הבאים: נכונות הקוד, יעילות הקוד וזמני ריצה. יש לבדוק מקרי קצה.

בתרגיל נממש פעולות עיבוד מידע מתוך קבצים באמצעות Python.

לא ניתן להשתמש בתרגיל זה בספריות חיצוניות שתואמות לפעולות בעבודה, אלא במבנים קיימים של python בלבד (כגון set וכו'). ניתן להשתמש בספריה המבצעת קריאה מקבצי txt או קבצי csv.

בMoodle מצורף לתרגיל קובץ zip בשם Ex1Samples.zip.

הקובץ מכיל את הקבצים הבאים (קבצים אלו הינם לדוגמא בלבד):

- *Users.txt – מכיל את הפריטים שהמשתמשים דירגו במערכת*
- *Ratings.txt – מכיל את פרטי הדירוגים שנתנו המשתמשים לפריטים במערכת*
- *Items.txt – מכיל את הפריטים הקיימים במערכת*
- *Users(2).txt – מכיל פריטים נוספים שמשתמשים אחרים דירגו במערכת*
- *Ratings(2).txt – מכיל פרטי דירוגים נוספים שנתנו המשתמשים לפריטים במערכת*
- *Items(2).txt – מכיל פריטים נוספים הקיימים במערכת*
- *itemsMerged.txt – מכיל פריטים מאוחדים לאחר פעולת Union*

הפעולות אשר יש לממש במטלה זו:

א. UNION:

פעולת איחוד בין טבלאות. מקבלת כקלט נתיבים ל-2 קבצים (מסוג csv או txt), ונתיב לשמירת התוצאה.

קובץ התוצאה צריך להיות מאותו הסוג של קבצי הקלט ולכלול את איחוד כל הרשומות מ-2 קבצי הקלט.

ניתן להניח כי אותה שורה לא תהיה באותו הקובץ פעמיים.

יש לשים לב שיתכן שמספור הפריטים (או המשתמשים) בין הקבצים הוא זהה, אך הכוונה היא לפריטים **שונים**.

הטבלאות R1 (קובץ ראשון) ו-R2 (קובץ שני) צריכות לקיים מבנה זהה:

- אותו מספר עמודות
- תחום (סוג ערך) של עמודה מס' i ב- R1 זהה לתחום של עמודה מס' i ב- R2

במידה והטבלאות לא מקיימות את המבנה הזהה, יש להפיק הודעת שגיאה.

על מנת שיהיה ניתן מהקובץ המאוחד להפריד חזרה לשני הקבצים, יש להוסיף עמודה חדשה לקובץ המאוחד שערכה יהיה מספר (או ערך) ייחודי של הקובץ שממנו הגיעה שורת הנתונים. לכן, במידה ואותה שורה קיימת בשני הקבצים, היא צריכה להופיע פעמיים בקובץ התוצאה.

שימו לב כי בקובץ התוצאה סדר הרשומות צריך להיות לפי סדר קבצי הקלט (קודם הרשומות המופיעות מקובץ הקלט הראשון ולאחריהם הרשומות מקובץ הקלט השני), וכן סדר העמודות צריך להיות זהה לסדר העמודות בקבצי הקלט.

דוגמת הפעלה:

```
$> Python Ex1.py UNION Users.txt Ratings.txt OUTPUTUNION.txt
$> Error! The tables' format does not match
$> Python Ex1.py UNION Users.txt Users(2).txt OUTPUTUNION.txt
```

ב. SEPERATE:

פעולת הפרדה של קובץ מאוחד ל-2 קבצים נפרדים. מקבלת כקלט נתיב לקובץ המאוחד ונתיבים ל-2 קבצי פלט של הקבצים הנפרדים.

קבצי התוצאה צריכים להיות מאותו סוג של קובץ הקלט ולכלול את ההפרדה בין הרשומות מקובץ הקלט עפ"י הערך של העמודה האחרונה (כל ערך המפריד בין 2 קבצים). במידה ויש יותר משני ערכים שונים בכל הקובץ המאוחד בעמודה האחרונה, יש להפיק הודעת פלט מתאימה.

ניתן להניח כי אותה שורה בדיוק לא תהיה בקובץ המאוחד פעמיים.

בקבצי הפלט הטבלאות R1 (קובץ הפלט הראשון) ו-R2 (קובץ הפלט השני) צריכות לקיים מבנה זהה:

- אותו מספר עמודות
- תחום (סוג ערך) של עמודה מס' i ב- R1 זהה לתחום של עמודה מס' i ב- R2

במידה ולא ניתן להפריד את קובץ הקלט ל-2 קבצים נפרדים בדיוק, יש להפיק הודעת שגיאה.

שימו לב כי בקבצי התוצאה סדר הרשומות צריך להיות לפי סדר הרשומות בקובץ המאוחד, וכן סדר העמודות בקבצי הפלט צריך להיות זהה לסדר העמודות המופיע בקובץ המאוחד.

<u>דוגמת הפעלה:</u>

$> Python Ex1.py **SEPERATE** ItemsMerged.txt items1.txt items2.txt

ג. DISTINCT:

פעולת שליפה של ערכים ייחודיים (לפי סדר) עבור עמודה מסוימת.

מקבלת כקלט נתיב לקובץ הקלט, אינדקס של עמודה בקובץ הנ"ל (עמודה 0 היא העמודה הראשונה בקובץ) ונתיב לקובץ הפלט.

יש לבדוק האם אינדקס העמודה קיים בקובץ הקלט ובמידה והעמודה לא קיימת בטבלה יש להוציא את ההודעה:

"Error! Column does not exist in table"

הפעולה תייצר קובץ בנתיב הפלט בו יופיעו הערכים הייחודיים מהעמודה הנ"ל, כל ערך בשורה נפרדת.

יש למיין את הערכים בקובץ הפלט לפי סוגם (אם מספרים – קודם הקטן ואז הגדול ביותר, אם אלו מחרוזות – לפי סדר לקסיקוגרפי). אין משמעות למיון של מערכים ומבנים איטרטיביים ולכן יוחזרו מערכים באותו סדר שהופיעו בקובץ המקורי.

מידע נוסף על פעולת ה-DISTINCT ניתן למצוא ב: https://www.w3schools.com/sql/sql_distinct.asp

ד. LIKE:

מבצע שליפת רשומות בעלות מבנה מסוים בעמודה מסוימת.

מקבלת כקלט נתיב לטבלה, אינדקס של עמודה בטבלה, ופרמטר לשליפה. במידה ולא ישלח לפונק' ההפעלה פרמטר לשליפה, הפרמטר שיישלח הוא * (כלומר כל הרשומות של העמודה יישלפו).

יש לבדוק האם אינדקס העמודה קיים בטבלה. במידה והעמודה לא קיימת בטבלה יש להוציא את ההודעה:

"Error! Column does not exist in table"

הפורמט של הפרמטר לשליפה יינתן כביטוי רגולרי המוגדר ב: https://docs.python.org/2/howto/regex.html, במקום לפורמט הסטנדרטי של SQL.

קובץ התוצאה צריך להכיל את השורות מהטבלה (ללא חזרות) בהם הערך בעמודה שהוגדרה תואם לפרמטר השליפה.

מידע נוסף על פעולת הLIKE ניתן למצוא ב: https://www.w3schools.com/sql/sql_like.asp.

<u>בהצלחה! ☺</u>