תרגיל בית 5 – קבצים וטיפול בשגיאות

הנחיות כלליות:

- קראו **בעיון** את השאלות והקפידו שהתכניות שלכם פועלות בהתאם לנדרש.
- את התרגיל יש לפתור לבד!
- הקפידו על כללי ההגשה המפורסמים באתר. בפרט, יש להגיש את כל הפתרונות לשאלות יחד בקובץ ex5_012345678.py המצורף לתרגיל, לאחר החלפת הספרות 012345678 שבשם הקובץ במספר תעודת הזהות שלכם, כל 9 הספרות כולל ספרת ביקורת .
- אופן ביצוע התרגיל: שימו לב, בתרגיל זה עליכם להשלים את הקוד בקובץ המצורף.
- **אין לשנות את שמות הפונקציות והמשתנים שכבר מופיעים בקובץ השלד של התרגיל.**
- **אין למחוק את ההערות שמופיעות בקובץ השלד.**
- היות ובדיקת התרגילים עשויה להיות אוטומטית, **יש להקפיד על פלטים מדויקים על פי הדוגמאות (עד לרמת הרווח).**
- בדיקה עצמית: כדי לוודא את נכונותן ואת עמידותן של התוכניות לקלטים שגויים, בכל שאלה הריצו את תוכניתכם עם מגוון קלטים שונים, אלה שהופיעו כדוגמאות בתרגיל וקלטים נוספים עליהם חשבתם) וודאו כי הפלט נכון).
- **ניתן להניח שהקלט תקין בהתאם להערות המפורטות בהוראות כל שאלה.**
- מועד אחרון להגשה: כמפורסם באתר.

שאלה 1 – סכימת מספרים המופיעים בקובץ

ממשו את הפונקציה **sum_nums(file)** המקבלת מחרוזת המציינת שם של קובץ קלט (file).
הניחו שבתוך הקובץ מופיעה שורה בודדת הכוללת סדרת מספרים שלמים המופרדים על ידי רווח בודד.
על הפונקציה לקרוא את הקובץ ולהחזיר את סכום המספרים המופיעים בו.

עבור קובץ קלט המכיל את השורה הבאה:

```
4 55 3 67 10
```

הפונקציה תחזיר את הערך 139.

הערה: בשאלה זו ניתן להניח שהקלט תקין ואין צורך לטפל בשגיאות.

שאלה 2 – חישוב שכיחויות מילים המופיעות בקובץ

ממשו את הפונקציה **get_x_freqs (infile, outfile, x)** המקבלת שם של קובץ קלט (המחרוזת **infile**), שם של קובץ פלט (המחרוזת **outfile**) ומספר שלם חיובי **x**.
הפונקציה תכתוב לקובץ הפלט בשורות נפרדות וללא חזרות את רשימת x המילים השכיחות ביותר בקובץ הקלט ואת מספר המופעים של כל אחת מהן. על המילים להיכתב כשהן ממוינות בסדר יורד לפי שכיחות ההופעה שלהן בקובץ הקלט. רווח בודד יפריד בכל שורה בין המילה לבין מספר המופעים שלה.

הערות:

- קובץ הקלט infile הוא קובץ טקסט המכיל מילה אחת או יותר בכל שורה כאשר המילים מופרדות על ידי רווחים (רווח בודד בין מילים וירידת שורה בין שורות). כל המילים מורכבות מאותיות קטנות בלבד : ללא סימני פיסוק, מספרים וכו' כל הקובץ מכיל רק אותיות קטנות ורווחים.
- בשורה הראשונה בקובץ הפלט outfile תופיע המילה השכיחה ביותר בקובץ הקלט infile. במידה ויש כמה מילים המופיעות באותה שכיחות, אין חשיבות לסדר דירוגן הפנימי בקובץ הפלט.
- על התוכנית להדפיס בדיוק x מילים. במידה וישנן כמה מילים עם אותו מספר מופעים, יתכן ורק חלק מהמילים תודפסנה לקובץ (כדי לא לחרוג מ-x). במצב כזה אין חשיבות לבחירת המילים מתוך אוסף המילים בעלות אותו מספר מופעים.
- במידה והמחרוזות infile או outfile הן ריקות, יש להעלות שגיאה מסוג ValueError (יש להשתמש ב-raise) ואת הודעת השגיאה: 'Invalid file name' במדוייק . במידה והמחרוזות אינן ריקות, ניתן להניח שהקלט תקין.
- הדרכה: היעזרו במילון לחישוב שכיחויות המילים כפי שראינו בדוגמאות בכיתה.

לדוגמא, אם נפעיל את הפונקציה על קובץ הקלט ('the_wheels_on_the_bus.txt') שנמצא בין קבצי התרגיל ונגדיר ש-x=3, אז היות והמילה "round" מופיעה 8 פעמים בקובץ הקלט, המילה "the" מופיעה 5 פעמים והמילה "and" מופיעה 4 פעמים, אז בקובץ הפלט יופיע הטקסט הבא :

round 8

the 5

and 4

שימו לב שאין חשיבות לסדר הפנימי של הופעת מילים בקובץ הפלט אם יש להן אותו מספר מופעים בקובץ הקלט. לדוגמא, אם נפעיל את הפונקציה על קובץ הקלט ('the_wheels_on_the_bus.txt') שנמצא בין קבצי התרגיל ונגדיר ש-x=4,אז היות ולכל המילים הבאות:

wheels, go, bus, on

יש מספר מופעים שווה של 2 בקובץ הקלט, אז יהיו לנו מספר אפשרויות לתשובה נכונה בקובץ הפלט:

round 8

the 5

and 4

wheels 2

Or

round 8

the 5

and 4

go 2

Or

round 8

the 5

and 4

on 2

Or

round 8

the 5

and 4

bus 2

שאלה 3 - פיענוח קובץ טקסט מוצפן

ממשו את הפונקציה **decode(in_file, out_file)** הקוראת טקסט מוצפן מהקובץ in_file, מפענחת אותו על פי החוקיות שתוגדר בהמשך, וכותבת את הטקסט המפוענח לקובץ out_file.
את הפענוח יש לבצע ע"פ קידוד ASCII [לחיצה תוביל לטבלת ASCII] המגדיר לכל תו ערך מספרי כלשהו.
עליכם לפענח את הטקסט שבקובץ הקלט על ידי החלפת כל אות אנגלית באות **הקודמת** לה. כל תו שאינו אות באנגלית (רווחים וירידות שורה) יש להשאיר בדיוק כפי שהוא בקובץ הקלט.
לדוגמא, האות B בקובץ הקלט תוחלף באות A שתיכתב במקומה לקובץ הפלט, האות H תוחלף באות G שתיכתב במקומה לקובץ הפלט.
שימו לב האות a תוחלף ב-z, האות A תוחלף באות Z,

לדוגמא, הטקסט המוצפן "Qzuipo Qsphsbnnjoh gps Fohjoffst" יפוענח ל-
"Python Programming for Engineers" .

- אם אירעה שגיאת IO במהלך הקריאה או הכתיבה לקבצים יש "לתפוס" אותה ולהדפיס למסך את ההודעה :

'Can't decipher {**in_file**} due to an IO Error. '

עליכם להחליף את {in_file} בערך של שם הקובץ כאשר מחזירים את ההודעה.
במקרה זה יש לצאת מהתוכנית בצורה מסודרת ולנסות לסגור את הקבצים בטרם היציאה.

הקובץ q3.txt מכיל טקסט מוצפן. אם תבצעו את הפענוח נכון, התוצאה תהיה זהה לתוכן הקובץ q3_deciphered.txt

שימו לב:

- ניתן להניח שקובץ הקלט כולל אותיות גדולות או קטנות באנגלית, רווחים, וסימן ירידת שורה בלבד.
- סימן ירידת שורה בחלונות ובלינוקס שונה (חלונות- \n לינוקס - \r\n)
- יש לוודא שחרור משאבים על ידי סגירה מסודרת של הקבצים גם אם היתה שגיאה. רמז: השתמשו ב-finally.
- יש להשתמש בפוקנציית ord בשאלה על מנת להשוות בין אותיות.
- עבור קובץ שלא קיים – לדוגמה 'not_exist.txt' יודפס למסך –

'Cannot decipher not_exist.txt due to an IO Error'

שאלה 4 - עיבוד קובץ נתונים במבנה טבלאי

קובץ CSV הוא קובץ טקסט המכיל נתונים במבנה של טבלה מלבנית כאשר פסיקים משמשים כתו מפריד בין השדות בכל שורה (ראו מצגת תירגול 5).

ממשו את הפונקציה **process_contacts(contacts_file)** המקבלת כקלט שם של קובץ csv המכיל טבלה המתארת את פרטיהם של אוסף אנשי קשר. כל שורה תייצג איש קשר ותכלול 4 שדות המופרדות על ידי פסיק: שם פרטי, שם משפחה, כתובת, ועיר מגורים. הפונקציה תקרא את הקובץ, תעבד את הנתונים המאוחסנים בו ותחזיר מילון שממפה לכל שם עיר את רשימת שמות המשפחה של תושביה (ללא חזרות).

שימו לב שהעמודה השנייה מציינת שם משפחה של איש הקשר והעמודה הרביעית מציינת את עיר המגורים שלו.

על הפונקציה לבדוק שקובץ הקלט תקין, כלומר שבכל שורה יש אותו מספר שדות, ושאף שדה אינו ריק. במידה וקובץ הקלט זוהה כקובץ שאינו תקין יש להעלות שגיאה (באמצעות raise) מסוג ValueError ואת הודעת השגיאה 'Invalid input file' ולשים לב לסגור את הקובץ. ניתן להניח שמלבד בעיות אלו, הקובץ קיים ותקין.

במקרה של שגיאת IOError, יש לתפוס את השגיאה, להדפיס למסך הודעת שגיאה מתאימה, לסגור את הקובץ, ולהחזיר מילון ריק.

על הפונקציה להתעלם משורות הערה אשר מתחילות בסולמית ('#').

לדוגמה ,עבור הקובץ המצורף , 'good.csv' שהוא קובץ CSV תקין המכיל נתונים על אנשי קשר לפי הפורמט שקבענו

```
Avi,Levi,Kushnir 7,Jerusalem
Moshe,Yarden,Hamakabim 4,Tel Aviv
Michael,Cohen,Herzel 70,Tel Aviv
#This is a comment
Eli,Cohen,Haroe 6,Jerusalem
Moti,Cohen,shalom 5,Tel Aviv
```

הפונקציה תחזיר את המילון הבא המכיל שני מפתחות:

'Jerusalem' -> ['Levi', 'Cohen']

'Tel Aviv' -> ['Cohen', 'Yarden']

עבור הקובץ 'bad.csv' שמכיל את הטקסט הבא:

```
Avi,Levi,Kushnir 7,Jerusalem
Moshe,,Hamakabim 4,Tel Aviv
Michael,Cohen,Herzel 70 Tel Aviv
Eli,Cohen,Haroe 6,Jerusalem
Moti,Cohen,shalom 5,Tel Aviv
```

הפונקציה תעלה שגיאה כי שם המשפחה בשורה השניה חסר, וגם כי מספר השדות בשורה השלישית שונה מ-4.

בהצלחה!