תרגיל בית 4

הנחיות כלליות:

- קראו בעיון את השאלות והקפידו שהתוכניות שלכם פועלות בהתאם לנדרש.
- את התרגיל יש לפתור לבד!
- הקפידו על כללי ההגשה המפורסמים באתר. בפרט, יש להגיש את כל השאלות יחד בקובץ ex4_012345678.py המצורף לתרגיל, לאחר החלפת הספרות 012345678 במספר ת.ז. שלכם, כל 9 הספרות כולל, ספרת ביקורת.
- אופן ביצוע התרגיל: בתרגיל זה עליכם להשלים את הקוד בקובץ המצורף.
- 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).
- חלק מהשאלות נבדקות באופן אוטומטי. לכן, עליכם לרשום את הקוד שלכם אך ורק במקומות המתאימים לכך בקובץ השלד.
- ניתן להניח כי הקלט שמקבלות הפונקציות תקין (אלא אם נכתב אחרת).
- אין לשנות שמות פונקציות או משתנים שקיימים בקובץ השלד של התרגיל.
- אין למחוק את ההערות שמופיעות בשלד.
- אין להשתמש בקריאה לספריות חיצוניות (אסור לעשות **import**).
- מועד אחרון להגשה: כמפורסם באתר.

מילונים

שאלה 1

כתבו פונקציה בשם second_most_popular_character שמקבלת מחרוזת ומחזירה את האות השנייה הכי שכיחה בה. במידה ויש יותר מאות אחת כזו, יש להחזיר את האות הקטנה ביותר מהן (על פי סדר מילוני).

דוגמאות הרצה:

```
>>> second_most_popular_character('HelloWorld')
'o'
```

הסבר: האות 'l' מופיעה 3 פעמים ו 'o' פעמיים, שאר האותיות מופיעות פעם אחת בלבד. לכן 'o' היא השנייה השכיחה ביותר.

```
>>> second_most_popular_character('cccaabb')
'a'
```

הסבר: האותיות 'a' ו 'b' מופיעות פעמיים, ו 'c' מופיעה שלוש. 'a' נבחרה כי היא קטנה מ 'b'.

הדרכה: להלן דרך אפשרית לפתרון השאלה .עליכם לממש את הצעדים הבאים:

1. בנו מילון המכיל בתוכו את האותיות של המחרוזת ואת מספר המופעים של כל אות.
2. מצאו את הערך השני הכי גדול במילון.
3. מצאו את המפתח הקטן מבין הערכים הנ"ל (במידה ויש כמה).

ניתן להניח שהמחרוזת מקבלת רק אותיות (אנגליות).

הבהרות לגבי שאלה 1:
א) יש להניח שהפונקציה מקבלת מחרוזת המכילה יותר מאות אחת. כלומר, אין להזין 'a' או 'aaa' .
ב) אותיות קטנות שונות מאותיות גדולות. כלומר, 'a' ו- 'A' הן אותיות שונות.
ג) בפייתון, אותיות גדולות מופיעות לפני אותיות קטנות. כלומר, A-Z < a-z. עוד דוגמאות: ' A'<'a', 'Z'<'a'.
ד) במידה ויש כמה אותיות החוזרות אותו מספר פעמים בשכיחות הכי גבוהה, למשל 'aaabbbcc', ניתן להחזיר את אחת משתי האפשרויות הבאות:
1. את 'c', כי היא מופיעה פעמיים בעוד ש 'a' ו- 'b' מופיעות שלוש פעמים.
2. את 'a', כי היא מופיעה את כמות המופעים הכי שכיחה והשנייה הכי שכיחה (כמו 'b') והיא הקטנה מביניהן.
שתי האפשרויות תתקבלנה כתוצאות נכונות על ידיי הבודקים.

שאלה 2

בתרגול ראינו ייצוג של sparse matrix באמצעות מילון. בייצוג כזה, עבור כל תא במטריצה שאינו אפס,

יאוחסן במילון מפתח מסוג tuple המיצג את קורדינטות התא, והערך ייצג את ערכו של התא במטריצה.
ממשו את הפונקציה diff_sparse_matrices(lst) אשר מקבלת רשימה של מילונים (2 או יותר)
המייצגים sparse matrices ומחזירה מילון המייצג את מטריצת ההפרש.

lst = [M1, M2, …. , Mn].

הפרש מטריצות מחושב כהפרש בין כל האיברים במטריצה א' לכל האיברים במטריצה ב', בהתאמה.
כלומר ההפרשM1 – M2 יתבצע כ: M1(i, j) – M2(i, j) עבור כל i, j

באותו האופן אפשר לחשב הפרש של יותר משתי מטריצות:

M1 – M2 – M3 - … - Mn = M1(i, j) – M2(i, j) – M3(i, j) - … - Mn(i, j) , עבור כל i, j

לדוגמה:

```
In[2]: diff_sparse_matrices([{(1,3):2, (2,7):1} , {(1,3):6}])
Out[2]: {(1, 3): -4, (2, 7): 1}
In[3]: diff_sparse_matrices([{(1,3):2, (2,7):1} , {(1,3):2}])
Out[3]: {(2, 7): 1}
```

במקרה השני ערך מטריצת ההפרש במיקום (1,3) הוא 0, ולכן ערך זה הוסר מהמילון.

הפונקציה אמורה לקבל רק רשימה המכילה מטריצות מהתצורה הנ"ל- שתיים או יותר. מספר האיברים בכל מילון לא חייב להיות זהה (כלומר, מספר האיברים השונים מ 0 בכל מטריצה, לא חייב להיות זהה).

שאלה 3

ממשו פונקציה בשם find_substring_locations(s, k) המקבלת מחרוזת ואורך של תת-מחרוזת
כמספר שלם ומחזירה את המילון הבא:

- המפתחות הם כל תתי המחרוזות של המחרוזת s, באורך k (אוסף רצוף של תווים מתוך s).
- הערך המתאים לכל מפתח היא רשימה של כל האינדקסים בהם הוא מופיע ב s (כל מיקום מצויין על ידי אינדקס התו הראשון של התת-מחרוזת). למשל: "ge" מתוך "drge" יופיע במיקום 2.

דוגמת הרצה:

```
In[8]: find_substring_locations('TTAATTAGGGGCGC', 2)
Out[8]:
{'TT': [0, 4],
 'TA': [1, 5],
 'AA': [2],
 'AT': [3],
 'AG': [6],
 'GG': [7, 8, 9],
 'GC': [10, 12],
 'CG': [11]}
In[9]: find_substring_locations('TTAATTAGGGGCGC', 3)
Out[9]:
{'TTA': [0, 4],
 'TAA': [1],
 'AAT': [2],
 'ATT': [3],
 'TAG': [5],
 'AGG': [6],
 'GGG': [7, 8],
 'GGC': [9],
 'GCG': [10],
 'CGC': [11]}
In[10]: find_substring_locations('Hello World', 3)
Out[10]:
{'Hel': [0],
 'ell': [1],
 'llo': [2],
 'lo ': [3],
 'o W': [4],
 ' Wo': [5],
 'Wor': [6],
 'orl': [7],
 'rld': [8]}
```

כפי שניתן לראות, הפונקציה מחלקת את המחרוזת לתת-מחרוזות באורך k, בצורה רציפה, וסופרת כמה פעמים כל תת-מחרוזת הופיעה במחרוזת s.

הפונקציה צריכה לדעת להתמודד עם k בגודל:

k <= len(s) => 1

שאלה 4 (שאלות 4-6 בנושא קבצים ושגיאות)

בראיון עבודה לחברה המתמחה בניתוח טקסטים אוטומטי, אתם מתבקשים לממש את הפונקציה count_lines(in_file, out_file) המקבלת שתי מחרוזות המייצגות נתיבים של קבצים. הפונקציה תכתוב לקובץ הפלט out_file את מספר השורות בקובץ in_file.

בשאלה זו ניתן להניח שקובץ הקלט הוא קובץ טקסט תקין, ואין צורך לטפל בחריגות.

- מצורף לתרגיל q4_input_example_1.txt כבדיקה לפונקציה. מומלץ ליצור קבצי בדיקה נוספים למקרי קצה בעצמכם.

דוגמא:

בקובץ q4_input_example_1.txt מופיע הטקסט הבא -

line 1

line 2

line 3 -> thus, your code should write to the output file 3

הפעלת הפונקציה עם קלט in_file שמכיל נתיב לקובץ זה, ונתיב נוסף out_file, תכתוב קובץ חדש, בנתיב שהוזן out_file, שיכיל את הטקסט הבא -

3

שאלה 5

התקבלתם לעבודה, ובתור משימה ראשונה אתם מתבקשים לסווג את אופי תוכן המסמכים בתור שמחים או עצובים באופן אוטומטי. אתם כמובן נלהבים להשתמש בטכניקות המתקדמות ביותר של למידת מכונה, אך ראש הצוות מזכירה לכם שפרקטיקה הנדסית חשובה היא להתחיל עם פתרון פשוט, ולהשתמש בו כדי לבדוק אם ועד כמה פתרון מתוחכם נדרש ועדיף.

אם כן, ראש הצוות מבקשת מכם לממש את הפונקציה simple_sent_analysis(in_file).

פונקציה זו מקבלת הנתיב של קובץ הטקסט שצריך לסווג (משתנה בשם in_file) ומחזירה מילון ובו מספר הפעמים שהופיעה המילה happy ומספר הפעמים שהופיעה המילה sad, שיראה בצורה הבאה: {happy':num_happy,'sad':num_sad}.

ניתן להניח שהטקסט בקובץ הקלט מכיל רק אותיות באנגלית, מספרים, רווח, את התווים הבאים: !?, -;$%. וכן הוא יכול להכיל מספר שורות.

- הספירה צריכה להיות case insensitive, כלומר happy, Happy, hapPY, HAPPY, כולם נספרים כ-happy
- יש להקפיד לא לספור מילים שמכילות happy או sad. למשל המילה saddle לא תספר כ-sad.

- יש להתעלם מהתווים בתוך הסוגריים (!?, -;$%.). כלומר, הקפידו לספור את happy או sad גם אם מופיע אחד התווים לפניהם או אחריהם. למשל happy? או happy% נספרים כ-happy, אבל hap?py לא נספר כ-happy.

- אם אירעה שגיאת IO יש "לתפוס" אותה, ולסיים את הריצה בצורה מסודרת (ללא קריסה). יש לוודא סגירה של כל הקבצים שנפתחו. יש להחזיר מילון ריק, וכן להדפיס את ההודעה: "Cannot encode $in_file due to IO error".

כאשר $in_file מציין את שם קובץ הקלט.

- מצורף לתרגיל q5_input_example_1.txt כבדיקה לפונקציה. מומלץ ליצור קבצי בדיקה למקרים אחרים, לרבות מקרי קצה
- הצעה, שימוש במתודות של מחרוזות יכול להקל מאוד על הפתרון. למשל המתודה str.replace(old,new) מאפשרת להחליף כל מחרוזת old המופיעה בstr, במחרוזת new

- רשות, להעשרה, למי שמתעניין בתחום עיבוד שפה (מחוץ לגבולות הקורס), ניתן אפשר לקרוא עוד על בעיות מסוג ניתוח סנטימנט

דוגמא:

בקובץ q5_input_example_1.txt מופיע הטקסט הבא (הצבעים לצורך הסבר ולא מופיעות בקובץ עצמו) –

happy hap saddle,

sad bla sad s!ad

!shimi happy

עבור הפעלת הפונקציה עם נתיב לקובץ הנ"ל, יתקבל כפלט המילון הבא -

{happy': 2, 'sad': 2'}

שאלה 6

סיווג המסמכים שמימשתם עובד מצוין, ובזכותו החברה קיבלה עבודה מחברת הפקות של סדרות טלוויזיה. חברת ההפקות מעוניינת לדעת האם רווחי יותר להפיק סדרות שמחות, עצובות או ניטרליות. החברה סיפקה תסריטים לכל הסדרות שלהם, וכן את הרווחים מהם. צוות אפליקציה כבר הריץ את האלגוריתם שלכם על כל התסריטים, וסיפק לכם קובץ csv שמכיל את שם הסדרה, כמה הרוויחה, סיווגה (happy, sad, או neutral).

ממשו את הפונקציה calc_profit_per_group(in_file), המקבלת כקלט את הקובץ csv האמור, ומחזירה מילון בו מופיע הרווח הממוצע עבור כל קטגוריה.

במידה ולא מופיעה בכלל אחת מהקבוצות, יופיע במקום הרווח הממוצע 'NA'. זה נחשב מצב תקין ואין להקפיץ שגיאה. ראו דוגמא שניה.

- אם אירעה שגיאת IO יש "לתפוס" אותה, ולסיים את הריצה בצורה מסודרת (ללא קריסה), לא לכתוב לקובץ הפלט, ולהדפיס את ההודעה:

  ".Cannot use $in_file due to IO error"

  כאשר $in_file מציין את שם קובץ הקלט.

- במידה וסדרה מופיעה יותר מפעם אחת תתקבל שגיאה מסוג ValueError עם המחרוזת 'The series $series_name appears more than once.'

כאשר במקום $series_name יודפס שם הסדרה, (איך להחזיר ערך). במידה ויש יותר מסדרה אחת שחוזרת על עצמה, מספיק להחזיר את שם הסדרה הראשונה שמופיעה פעמיים. ראו דוגמא שלישית.

- יש לבצע את בדיקות הקלט הבאות:
  - לוודא שישנן 3 עמודות
  - עמודה שניה מכילה מספרים בלבד
  - עמודה שלישית מכילה רק את הערכים sad/happy/neutral בlowercase

במידה ונמצא הפרה יש לתת שגיאה מסוג ValueError עם המחרוזת: 'Invalid input.'

ראו דוגמא רביעית.

מצורף לתרגיל q6_input_example_1.txt כבדיקה לפונקציה. מומלץ ליצור קבצי בדיקה למקרים אחרים, לרבות מקרי קצה בעצמכם.

דוגמא ראשונה לטקסט הנמצא בקובץ בנתיב in_file:

Descendant Without A Conscience,505.4,happy

Wolf Of The Solstice,30000,sad

Women Of Hope,-4000,neutral

Pirates Of Perfection,65467,neutral

Warriors And Soldiers,-5435,sad

Butchers And Soldiers,76542,sad

World Of The Mountain,6536543,sad

Ruination Of Dusk,-2000,happy

Destroying The Stars,5435,happy

Blinded In My Enemies,765745.5,happy

פלט עבור הקלט הנ"ל:

{'happy': 192421.475, 'sad': 1659412.5, 'neutral': 30733.5'}

דוגמא שניה לטקסט הנמצא בקובץ שבנתיב in_file (כמו הדוגמא הקודמת, רק במחיקת השורות עם neutral) :

Descendant Without A Conscience,505.4,happy

Wolf Of The Solstice,30000,sad

Warriors And Soldiers,-5435,sad

Butchers And Soldiers,76542,sad

World Of The Mountain,6536543,sad

Ruination Of Dusk,-2000,happy

Destroying The Stars,5435,happy

Blinded In My Enemies,765745.5,happy

פלט עבור הקלט הנ"ל:

{'happy': 192421.475, 'sad': 1659412.5, 'neutral': 'NA'}

דוגמא שלישית לטקסט הנמצא בקובץ בנתיב in_file (ההדגשה לצורך הסבר ולא מופיעה כך בטקסט) :

Descendant Without A Conscience,505.4,happy

Wolf Of The Solstice,30000,sad

Women Of Hope,-4000,neutral

Pirates Of Perfection,65467,neutral

Warriors And Soldiers,-5435,sad

Butchers And Soldiers,76542,sad

World Of The Mountain,6536543,sad

Ruination Of Dusk,-2000,happy

Destroying The Stars,5435,happy

Blinded In My Enemies,765745.5,happy

Women Of Hope,-3000,neutral

Wolf Of The Solstice,30000,sad

במקרה זה תתקבל הודעת שגיאה:

ValueError: The series Women Of Hope appears more than once.

(ולא יוחזר כלום). שימו לב שהודפסה הסדרה שהמופע השני שלה הופיע ראשון.

דוגמא רביעית לטקסט הנמצא בקובץ בנתיב in_file (הצבעים לצורך הסבר ולא מופיעים כך בטקסט) :

Descendant Without A Conscience,505.4,Happy

Wolf Of The Solstice,30000,glad

,Women Of Hope,-4000

Pirates Of Perfection,a lot money,neutral

Warriors And Soldiers,-5435,sad

Butchers And Soldiers,76542,sad

World Of The Mountain,6536543,sad

Ruination Of Dusk,-2000,happy

Destroying The Stars,5435,happy

Blinded In My Enemies,765745.5,happy

בדוגמא זו הקלט לא תקין, ולכן תתקבל הודעת שגיאה:

.ValueError: Invalid input