תכנות בפייתון למדעים דיגיטליים להייטק , סמסטר א׳ תשפ״א 2020-2021

תרגיל בית 9
Numpy

הנחיות כלליות:

- קראו **היטב** את השאלות והקפידו שהתכניות שלכם פועלות בהתאם לנדרש.
- את התרגיל יש לפתור לבד!
- הקפידו על כללי ההגשה המפורסמים באתר. בפרט, יש להגיש את כל השאלות יחד בקובץ ex9_012345678.py המצורף לתרגיל, לאחר החלפת הספרות 012345678 במספר ת.ז שלכם, כל 9 הספרות כולל ספרת הביקורת.
- מועד אחרון להגשה: כמפורסם באתר.
- 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 וכי התוכנית אינה קורסת).
- היות ובדיקת התרגילים עשויה להיות אוטומטית, **יש להקפיד על פלטים מדויקים על פי הדוגמאות (עד לרמת הרווח).**
- אופן ביצוע התרגיל: בתרגיל זה עליכם להשלים את הקוד בקובץ המצורף.
- **אין לשנות את שמות המשתנים שכבר מופיעים בקובץ השלד של התרגיל.**
יש לעבוד עם המשתנים שמופיעים בשלד התרגיל. על הקוד של כל שאלה לעבוד ולספק את התוצאה הדרושה עבור קלט שיוזן במשתנים שמופיעים בשלד (המשתנים שלידם סימני שאלה ומחכים לקלט כפי שראינו בדוגמא מהתרגול). יחד עם זאת, אתם רשאים להוסיף משתנים נוספים כראותם עינכם.
- **שימו לב מתי מבקשים מכם להחזיר ערך מפונקציה (return) ומתי מבקשים להדפיס למסך (print).**
- אין למחוק את ההערות שמופיעות בשלד.

חלק א' – ניתוח נתוני צפיה

שאלה 1:

לרגל אירוע חגיגות העשור השלישי לסדרת הטלוויזיה האגדית "מסע בין כוכבים – הדור הבא" החלו מארגני האירוע בניתוח נתוני הצפייה העולמיים בסדרה לאורך עונותיה השונות. ברשות המארגנים מספר קבצי טקסט בפורמט CSV, כשכל קובץ מתאר את נתוני הצפייה במדינה מסוימת.

בכל קובץ מאוחסנת טבלת מספרים מלבנית המייצגת את מספר הצופים (בעשרות אלפים) בכל פרק, כאשר כל שורה מייצגת עונה (Season) וכל עמודה מייצגת פרק (Episode). ניתן להניח שבכל העונות ישנו אותו מספר פרקים, ושבכל הקבצים יש את אותו מספר שורות ואת אותו מספר עמודות.

דוגמא – קבצי נתוני הצפייה עבור ישראל וספרד:

Israel.csv

```
8,9,9,10,11,10,9,10,10,11
10,10,9,10,9,10,10,9,9,10
9,9,9,8,7,9,9,11,8,10
8,8,8,9,9,9,7,10,10,12
9,9,9,8,7,6,22,23,21,20
15,15,14,14,14,12,12,11,11,10
11,11,10,9,9,9,8,7,7,9
```

Spain.csv

```
67,68,74,76,77,78,74,77,80,87
74,76,75,77,68,76,77,70,74,80
72,70,70,70,70,70,69,71,72,78
66,67,65,69,69,70,56,80,80,94
70,68,72,66,55,47,134,136,138,160
120,116,112,110,111,97,94,86,84,90
88,84,78,70,72,68,66,58,55,68
```

השלימו את מימוש הפונקציה **analyze_rating_data (filename)** המקבלת שם של קובץ נתוני צפייה בודד של מדינה מסוימת, קוראת אותו בעזרת הפקודה loadtxt ומדפיסה למסך חיתוכים שונים על הנתונים.

הפקודה הראשונה בפונקציה, קוראת את הטבלה שבקובץ filename ומחזירה מערך דו מימדי שמחזיק את נתוני הטבלה:

```
rating = np.loadtxt(filename,delimiter=',')
```

החליפו את סימני השאלה בתוך פקודות ה-print שבשלד תרגיל 2 כך שיודפסו חיתוכי המידע הבאים עבור המערך rating:

1. מספר העונות שמיוצגות במערך rating (מספר השורות במערך).
2. נתון הצפייה הגבוה ביותר שנמדד עבור פרק בודד (מקסימום על איברי המערך).
3. מהו ממוצע נתוני הצפייה עבור הפרקים הראשונים בכל העונות (ממוצע על העמודה הראשון משמאל במערך)?
4. עבור כמה פרקים בטבלה יש נתון צפייה הקטן מ-8 (מספר איברי המערך שקטנים מ-8)?
5. האם יש פרק כלשהו בטבלה שנתון הצפייה שלו שווה בדיוק ל-15? יש להדפיס True אם קיים במערך לפחות איבר אחד השווה ל-15, ו-False אחרת.
6. מהו נתון הצפייה העונתי הגבוה ביותר ? (מקסימום על סכומי שורות המערך)?
7. מהם נתוני הצפייה הנמוכים ביותר עבור כל פרק מבין כל העונות? (יש להדפיס וקטור המכיל את המינימום של כל עמודה).

- יש לכתוב מימוש כללי שיעבוד עבור כל קובץ עם מספר כלשהו של עונות ועם מספר כלשהו של פרקים ולא להניח שקבצי הקלט יראו כמו בדוגמאות הנ"ל. עם זאת, ניתן להניח שישנן לפחות 2 עונות בסדרה, ושבכל העונות יש מספר קבוע של פרקים שהוא לפחות 2.
- בשאלה זו אין צורך לטפל בשגיאות IO.

דוגמאות הרצאה:

```
>>> analyze_rating_data('Israel.csv')

The number of seasons:
7
The highest rating ever recorded for an episode:
23.0
Average rating for the first episode over all seasons:
10.0
Number of episodes which had a rating lower than 8:
6
Is there at least one episode with a rating of 15:
True
The maximal total season rating:
134.0
Minimal rating for each episode:
[8. 8. 8. 8. 7. 6. 7. 7. 7. 9.]

>>>analyze_rating_data('Spain.csv')

The number of seasons:
7
The highest rating ever recorded for an episode:
160.0
Average rating for the first episode over all seasons:
79.57142857142857
Number of episodes which had a rating lower than 8:
0
Is there at least one episode with a rating of 15:
False
The maximal total season rating:
1020.0
Minimal rating for each episode:
[66. 67. 65. 66. 55. 47. 56. 58. 55. 68.]
```

חלק ב' – ניתוח נתוני תחלואה עולמיים:

בשאלה זו ננתח בעזרת חבילת Numpy את נתוני התחלואה העולמיים של וירוס הקורונה כפי שהורדו מהאתר

https://covid.ourworldindata.org/data/owid-covid-data.csv?v=2021-01-02

שאלה 1 – טעינת קובץ הנתונים

ממשו את הפונקציה **load_covid_world_matrix(filename, fieldname)** אשר מקבלת מחרוזת המציינת שם של קובץ נתונים filename ומחרוזת המציינת שם של שדה fieldname המופיעה בשורה הראשונה בקובץ הנתונים.

כל שורה בקובץ הנתונים החל מהשורה השנייה, מכילה נתוני תחלואה בווירוס הקורונה לגבי מיקום מסוים בעולם בתאריך מסוים.

הפונקציה תקרא את נתוני הקובץ ותחזיר 3 מערכים של Numpy לפי הפירוט הבא:

countries – מערך חד-ממדי של מחרוזות המכיל את שמות כל המדינות (ללא World או International) שהופיעו בעמודה השלישית בקובץ הנתונים, ללא חזרות, וכשהן ממוינות אלפביתית בסדר עולה.

dates – מערך חד-ממדי של מחרוזות המכיל את כל התאריכים שהופיעו בעמודה הרביעית בקובץ הנתונים, ללא חזרות, וכשהם ממוינים בסדר עולה (מיון אלפביתי של המחרוזות ישיג זאת).

matrix – מערך דו-ממדי של מספרים (float) המכיל את כל הנתונים שהופיעו בשדה שכותרתו fieldname עבור כל מדינה שמופיעה במערך countries (שורות המטריצה) ועבור כל תאריך שמופיע במערך dates (עמודות המטריצה). מספר השורות במטריצה יהיה כאורך המערך countries, ומספר העמודות במטריצה יהיה כאורך המערך dates. המטריצה תכיל אפסים בתאים שאין עבורם מידע בקובץ הקלט.

דוגמא להרצת הפונקציה:

```
countries, dates, matrix = load_covid_world_matrix('owid-covid-data_2-1-2021.csv', 'new_cases')

print(countries.shape)
print(dates.shape)
print(matrix.shape)
print(matrix.sum())

np.savetxt("matrix_new_cases.csv", matrix, delimiter=",", fmt='%f')
np.savetxt("dates_new_cases.csv", dates, delimiter=",", fmt='%s')
np.savetxt("countries_new_cases.csv", countries, delimiter=",", fmt='%s')
```

פלט שיודפס למסך:

```
(190,)
(367,)
(190, 367)
83423170.0
```

שימו לב

- יש להתעלם משורות בקובץ הקלט שבשדה המדינה מופיעה המחרוזת World או International
- ניתן להניח שמבנה הקובץ תקין אך בכל מקרה של במקרה של שגיאת IO יש להדפיס למסך את המחרוזת 'IOError encountered' ולצאת בצורה מסודרת מהתוכנית.

- דוגמת ההרצה לעיל מדפיסה למסך את ממדי המערכים שנוצרו ואת סכום האיברים במטריצה. הדוגמא גם שומרת לקבצים את תוכן 3 המערכים. השתמשו בפלטים אלו כדי לבדוק את הקוד שלכם.

שאלה 2 – ניתוח קובץ הנתונים

בשאלה זו נענה על שאלות על בסיס הנתונים שטענו למערכים, בעזרת חבילת ההרחבה Numpy.

ממשו את הפונקציה **analyze_covid_data(countries, dates, matrix)** אשר מקבלת את שלושת המערכים שהוחזרו על ידי הפונקציה מהסעיף הקודם, ומדפיסה למסך את הנתונים הבאים:

1. האם יש לפחות נתון שלילי אחד במטריצה matrix? יש להחזיר True או False .
2. בכמה ימים התגלו בישראל למעלה מ-8000 מקרים ? (הפקודה צריכה למצוא את השורה בטבלה שמתאימה ל-'Israel').
3. בכמה מדינות התגלו למעלה ממיליון מקרים במצטבר?
4. מהו שמה של המדינה בעלת מספר המקרים המצטבר הגבוה יותר ב-30 הימים הראשונים שמופיעים בטבלת הנתונים?
5. מהו התאריך בו מספר המקרים היומי בכל המדינות ביחד היה הגבוה ביותר?

דוגמת הרצה:

```
analyze_covid_data(countries, dates, matrix)
```

```
Are there any negative values in the table?
True
In how many days more than 8000 new cases were identified in Israel?
3
Number of countries with more than 1 million total cases:
18
Name of country with the highest total number of daily cases in the first
30 days appearing in the table:
China
Date with maximal number of new cases in all countries together:
2020-12-10
```

<u>שימו לב:</u>

בסעיף זה יש להשתמש בפקודות numpy בלבד. אין להשתמש במשפט IF, בלולאות, או ברקורסיות.

שאלה 3 – Covid-19 Data Visualization

בשאלה זו נבצע הדמיית נתונים בעזרת חבילת ההרחבה **matplotlib**.

שימו לב: היעזרו בתיעוד של חבילת matplotlib (או באתר אחר לבחירתכם) כדי להבין איך משתמשים בכל פקודה.

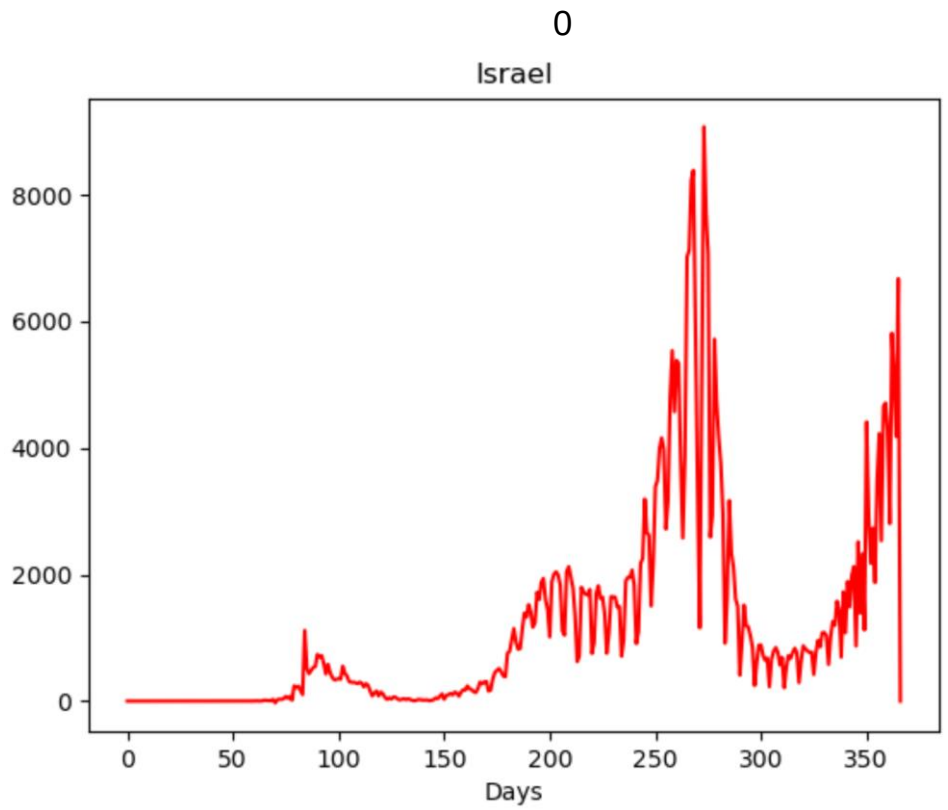
1. ממשו את הפונקציה **plot_country_data(matrix, countries, country)** אשר מקבלת את המערכים matrix ו-countries שהוחזרו על ידי הפונקציה של שאלה 1, בנוסף למחרוזת בשם country המכילה שם של מדינה (אפשר להניח שהמדינה קיימת במערך countries).
הפונקציה תדפיס למסך גרף המתאר את מספר המקרים היומי במדינה המבוקשת כפונקציה של הימים.

הערות:

- השתמשו בפקודות plot, title ו-xlabel כדי לייצר גרף על פי הדוגמא.
- הפקודה plot מצפה לקבל את סדרות הנתונים שעליה לשרטט כעמודות ולכן יש לבצע transpose למטריצה שניתנת לפקודה.

דוגמת הרצה:

```
plot_country_data(matrix, countries, 'Israel')
```



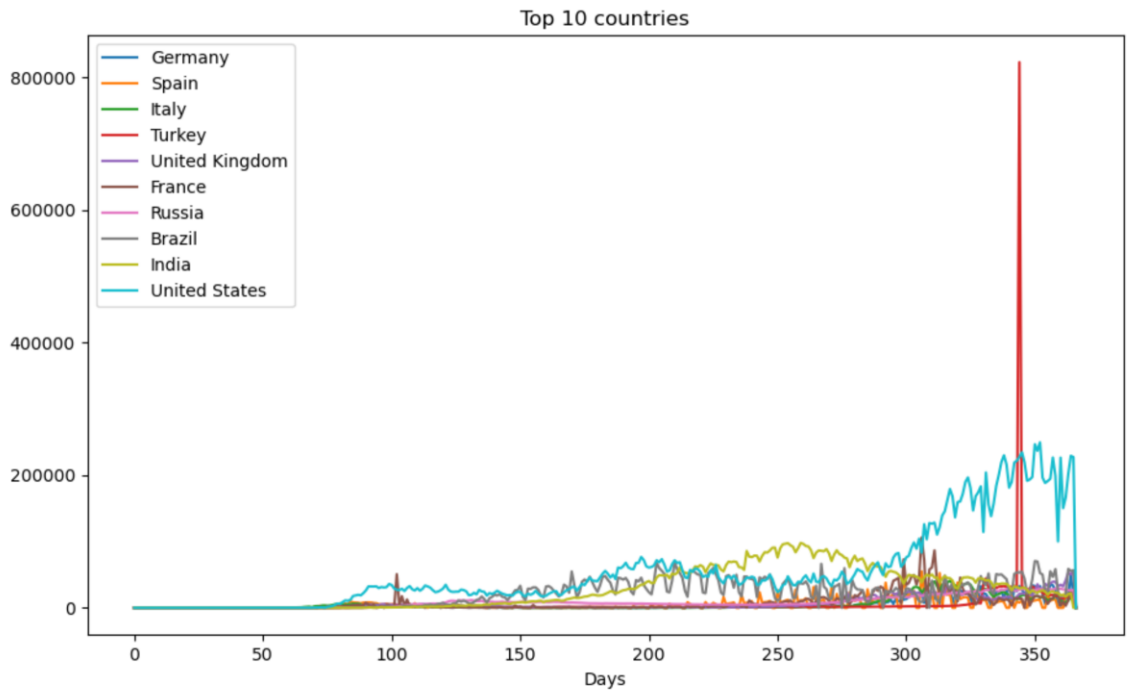
2. ממשו את הפונקציה **plot_top_countries(matrix, countries)** אשר מקבלת את המערכים matrix ו-countries שהוחזרו על ידי הפונקציה של שאלה 1, ומשרטטת גרף של מספר המקרים היומי כפונקציה של הימים, עבור 10 המדינות עם מספר המקרים המצטבר הגבוה ביותר .

הערות :

- השתמשו בפקודות plot, title, xlabel ו-legend כדי לייצר גרף על פי הדוגמא.
- הפקודה plot מצפה לקבל את סדרות הנתונים שעליה לשרטט כעמודות ולכן יש לבצע transpose למטריצה שניתנת לפקודה.

דוגמת הרצה :

```
plot_top_countries(matrix, countries)
```



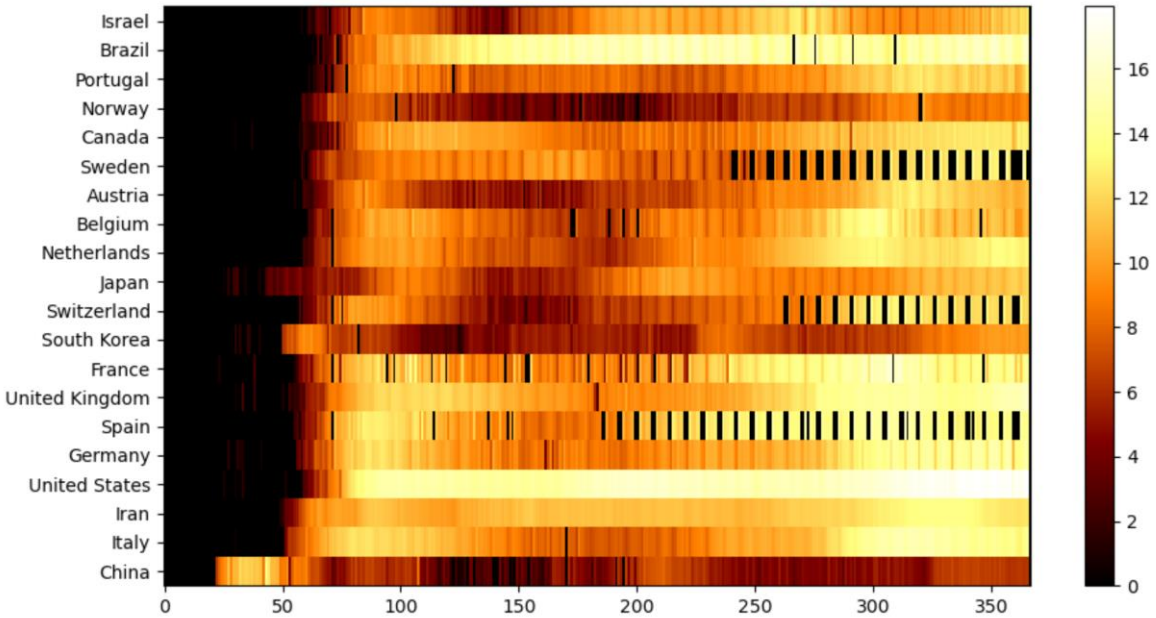
3. ממשו את המתודה **draw_covid_heatmap(matrix, countries)** אשר מקבלת את המערכים matrix ו-countries שהוחזרו על ידי הפונקציה של שאלה 1, מעבדת את נתוני המטריצה לפי הפירוט שבהמשך ומשרטטת גרף מפת-חום של 20 המדינות בהן מספר המקרים המצטבר היה הגבוה ביותר ב-100 הימים הראשונים בטבלה. המדינות תופענה כשהן ממוינות לפי קריטריון זה. יש להציג את הנתונים לגבי כל הימים המופיעים בטבלה.

שימו לב:

- היעזרו בפקודות imshow, yticks, colorbar כדי לייצר גרף על פי הדוגמא.
- להשגת המראה המוצג בדוגמא, הפקודה imshow צריכה לקבל מלבד המטריצה לשרטוט גם את הפרמטרים האופציונאליים האלו: cmap='afmhot', interpolation='none', aspect='auto'
- לפני שירטוט המטריצה יש לבצע עליה את העיבודים הבאים:
  - יש לשנות ל-1 כל ערך שקטן מ-1
  - יש להפעיל על המטריצה את הפונקציה np.log2 (ממתן ערכים גבוהים מדי כדי שלא יתפסו את כל סקאלת הצבעים. נסו לבדוק מה קורה ללא הפעלת הלוג על המטריצה לפני השירטוט).

דוגמת הרצה:

```
draw_covid_heatmap(matrix, countries)
```



בהצלחה !