מטלת מבחן ליום אחד - מסלול דאטה

## **מיני - פרויקט הכנה לניתוח נתוני סיווג אנטישמיות בטוויטר**

## **הנחיות:**

### **אופן עבודה:**

* יש לעבוד ללא עזרי AI לסוגיהם, **תלמיד שייתפס משתמש בכלי AI - מבחנו ייפסל ללא ערעור**.
* ניתן להשתמש באתרים הבאים בלבד (**אין לבצע חיפוש באינטרנט**):
  + [https://www.w3schools.com](https://www.w3schools.com/)
  + <https://stackoverflow.com/questions>
  + <https://pandas.pydata.org/>
  + <https://realpython.com/>
* ניתן להיעזר בעבודות קודמות שלכם בקורס הנמצאות ב-GitHub.
* **אין להעתיק קוד מהאינטרנט או מחברים.**

### **הגשה:**

* יצירת repository ציבורי מסודר ב-GitHub
* שיתוף הקוד ע"י הוספת קישור ל-repository בטופס הבא: [הגשות מטלת יום](https://docs.google.com/spreadsheets/d/1Z3SGFpStXfLEPf-DVhC_zrK5UJHH3z3QSRnbc77p-4c/edit?gid=0#gid=0)

### **זמן לביצוע:**

* משעה 09:00 ועד 16:00 חובה עליכם להשתתף בשיעור תורה בשעה 1300, ניתן לצאת להפסקות על פי לוז שיוגדר בכיתה.
* **שימו לב** - מבחינת בדיקת המיני פרויקט - COMMITS לאחר השעה 16:00 לא יחושבו בבדיקה, יש לעמוד במסגרת הזמנים לציון מלא.

#### **מטרה:** לבצע ניתוח נתונים וניקיון נתונים בסיסיים למידע טקסטואלי (ציוצים אנטישמיים ולא אנטישמיים)

### **מה תעשו:**

* תבצעו ניתוח טקסטואלי בסיסי.
* תנקו את הנתונים (למשל להסיר סימני פיסוק ולהמיר אותיות גדולות לקטנות).

### **מבנה תיקיות:**

project /

├── README.md ← הסבר על הרעיון, שלבים, ומה הצליח / לא הצליח

├── data/

│ ─── tweets\_dataset.csv ← הקובץ שקיבלתם (לא לשנות את הנתונים המקוריים)

├── src/

│ ─── \*.py ← תיקיה לקבצי הקוד

└── results/

│ ─ tweets\_dataset\_cleaned.csv

└── results.json ← קובץ פלט

#### 

#### 

#### **דגשים לפיתוח:**

* תיעוד וקריאות:
  + הערות קוד רלוונטיות - לדוגמא: אם יש לכם פונקציה בשם clean - אין צורך לתעד שמדובר במנקה, יש להסביר את תפקידה וכו'
  + שמות ברורים ובעלי משמעות (משתנים, פונקציות, מחלקות) ובהתאם למוסכמה הפייתונית
* כתיבה מודולרית וברת תחזוקה
  + לשים לב לפצל את הלוגיקה למחלקות וקבצים - בהתאם לעקרונות OOP ו SOLID   
    (לדוגמא: DataAnalyzer, ReportBuilder)

### **דאטאסט העבודה:**

מצורף למטלה קובץ בשם: tweets\_dataset.csv

* כל שורה היא ציוץ מטוויטר.
* שימו לב ל-2 עמודות רלוונטיות:
  + text - תוכן הציוץ
  + biased - סיווג/קטגוריה:
    - 1 - אנטישמי
    - 0 - לא אנטישמי

### **משימות:**

### **1.חקירת הדאטה:**

#### כמה ציוצים יש מכל קטגוריה (לפי קטגוריה, לא מוגדר, וסה"כ)

* מה האורך הממוצע (במילים) של ציוצים (לפי קטגוריה וסה"כ)
* מהם 3 הציוצים עם כמות המלל הגדולה ביותר (לפי קטגוריה)
* מהם 10 המילים הנפוצות ביותר בכל הציוצים (מכל הקטגוריות)
* כמה מילים באותיות גדולות (לפי קטגוריה וסה"כ) [הארה: מאפיין צעקות]

### **2. ניקוי בסיסי של הטקסטים**

* שמירת העמודות הרלוונטיות מתוך קובץ הנתונים (שימו לב לדגשים)
* הסרת סימני פיסוק
* המרה לאותיות קטנות
* הסרת ציוצים ללא סיווג

#### 

#### **3. ייצוא תוצאות החקירה והניקוי:**

* יש לשמור את הדאטהסט הנקי בתור קובץ בפורמט CSV בשם tweets\_dataset\_cleaned.csv .
* עליכם לייצא את תוצאות חקירת הדאטה והניקוי לקובץ בפורמט JSON בשם results.json .
  + על הקובץ לייצג את התוצאות בצורה הבאה (שימו לב - הערכים הם לצורך הדוגמא ולא מייצגים ערכים אמיתיים):

{

"total\_tweets": {

"antisemitic": 512,

"non\_antisemitic": 741,

"total": 12800,

<unspecified>: 27

},

"average\_length": {

"antisemitic": 18.3,

"non\_antisemitic": 20.7,

"total": 24.2

},

"common\_words": {

"total": ["the","and","is","to","jews","support","zionists","peace","fight","truth"]

},

"longest\_3\_tweets": {

"antisemitic": [

"Zionists control everything and lie about the holocaust for sympathy and power. Wake up, world.",

"Jews are manipulating global politics again. Open your eyes, they’ve done this before.",

"The media is owned by Zionists, they spread lies daily. Don’t believe anything they say!"

],

"non\_antisemitic": [

"Today we honor all communities and stand against hate. Peace and unity are stronger than division.",

"I’m proud to support equal rights for all people, regardless of religion, race or belief.",

"Education is the best way to fight ignorance and bigotry. We must keep learning and listening."

]

},

"uppercase\_words": {

"antisemitic": 139,

"non\_antisemitic": 86,

"total": 421

}

}

* + שימו לב - ב-unspecified הכוונה לכל קטגוריה שהיא לא הקטגוריות לפיהם המידע מסווג (לצורך העניין - אם המידע הוגדר כאנטישמי ולא אנטישמי - אז קטגוריה/ות נוספות שנמצאות מלבדיהן במידע אינן רלוונטית).