אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

Working with Real Data – 5 תרגיל

הוראות הגשה:

1. בתרגיל הבא יש לענות על השאלות באמצעות שימוש בקוד **פייתון ושימוש ב-** **Regular Expressions**.
2. יש להגיש את העבודה **בזוגות בלבד**.
3. התרגיל יוגש כמחברת colab, כאשר לתיבת הגשה יש להגיש קובץ המכיל לינק למחברת עם **הרשאות קריאה**
4. שם הקובץ יהיה מספרי הזהות של המגישים בצורה הבאה: זהות1_זהות2

במחברת הפתרון, יש לציין את מספר השאלה עליה עניתם עבור כל חלק בפתרון

ענו על השאלות הבאות באמצעות הנתונים של **Blog Authorship Corpus**:

1. החזירו את המזלות (sign) בעלי 8 אותיות או מספרים ומעלה.
2. חשבו כמה פוסטים פרסמו בכל שנה .
3. חשבו כמה כתובות מייל חוקיות התפרסמו בכל קטגוריה של בלוגים (מייל חוקי _@_ לא משנה כמה יש לאחר ה-@).
4. הציגו את הבלוגר שבבלוג שלו התפרסמו הכי הרבה כתובות מייל?
5. מצאו את הבלוגרים אשר בטקסט שלהם מופיע המספר הכי ארוך (שימו לב, מספר יכול להופיע עם פסיקים). רמז: היעזרו בתבנית +[,\d]

ענו על השאלות הבאות באמצעות הנתונים של UFO Sightings:

1. כתבו פונקציה שמקבלת כקלט דאטה, מדינה ושנה ומחזירה את מספר התצפיות של UFO בה .
def get_numbers_of_UFO(df, country,year)
החזירו דוגמה עבור המדינה us בשנת 1949
2. מצאו את החודש בשנה שבה נצפו הכי הרבה UFO.
3. הביאו את המחוזות **בלבד** בסדר יורד על פי כמות UFO.
4. מצאו את הערה שיש בה הכי הרבה זוגות מילים, כאשר שני המילים באורך גדול או שווה ל-7.
5. מצאו את המדינה שבה בהערות הופיעו הכי הרבה מספרים בני 6 ספרות .
6. החזירו את כל העדויות שבהערות שלהם הופיעה התבנית xxXx כאשר x היא אות a-zA-Z או מספר X היא סיפרה