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 למייל** + tamarin@post.bgu.ac.il yaelhoc@post.bgu.ac.il
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. איזה מזל (sign) משתמש הכי הרבה במילים ארוכות בעלי 8 אותיות או מספרים ומעלה.
2. חשבו כמה פוסטים פרסמו בכל יום בשבוע.
3. חשבו כמה כתובות מייל חוקיות התפרסמו בכל קטגוריה של בלוגים.
4. מיהם הבלוגרים שבבלוגים שלהם התפרסמו הכי הרבה כתובות URL?
5. מצאו את הנושא של הבלוגים שמופיע בו 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 מדינה ושנה ומחזירה את מספר התצפיות של UFO בה.
2. מצאו את השנה שבה נצפו הכי הרבה UFO.
3. מצאו את השעה ביום שבה נצפו הכי הרבה UFO.
4. מצאו את הערה שיש בה הכי הרבה זוגות מילים, כאשר שני המילים באורך גדול או שווה ל-6.
5. מצאו את המדינה שבה בהערות הופיעו הכי הרבה מספרים בני 4 ספרות.
6. החזירו את כל העדויות שבהערות שלהם הופיעה התבנית xxXx כאשר x היא אות a-zA-Z או מספר וX היא סיפרה