# למידת מכונה – ניתוח טקסט בעברית

## הצגת בעיית הלמידה

נתון corpus (מאגר) מתויג של סיפורים (labeled corpus, כלומר כולל target values).
לכל סיפור יש תגית המציינת את המגדר של כותב הסיפור.

המגדר מיוצג בתור מחרוזת עם 2 אפשרויות:

- 'm' – עבור כותב (male)
- 'f' - עבור כותבת (female)

**עליכם לבנות מודל סיווג שתפקידו לסווג את המגדר של פסקה סיפורית כנ"ל.**

## החומרים בהם יהיה מותר להשתמש

מותר להשתמש בכל חומר אותו למדנו הכולל python בסיסי המודולים , numpy, pandas, sickit learn (sklearn),
ובמודול re עבור regular expressions.

בנוסף מותר להשתמש במודולים ובכלי ניתוח טקסט המופיעים במודל (מלבד stop words, שאסור בשימוש)

## החומרים בהם אסור להשתמש

אסור להשתמש בשום מודול נוסף מלבד אלו המוזכרים לעיל

אסור להשתמש בשום קובץ חיצוני, כולל stop words
- שימוש ברשמית stop words אם יותר, יוסבר בהמשך.

אסור לצרף רשימות של מילים (כולל stop words) ולהשתמש בהם לסיווג.

## הקבצים המצורפים למטלה (3 קבצים):

### קובץ 1. Corpus מתויג – עבור ה-training

**שם הקובץ**: annotated_corpus_for_train.csv
קובץ csv של ה-corpus המתויג. מדובר בקובץ שמכיל train data, בצורה גולמית ושיש להפוך אותו ל- feature vectors כפי שלמדנו. הקובץ מכיל 2 עמודות:

- 'story' – עמודה המכילה פסקה סיפורית
- 'gender' – עמודה המכילה את המגדר של כותב/ת הפסקה

### קובץ 2. Corpus לא מתויג -עבור סיווג ה-test

**שם הקובץ**: corpus_for_test.csv
קובץ csv נוסף, המכיל פסקאות סיפוריות. הקובץ מכיל דוגמאות חדשות אותם יש לסווג.
הוא מכיל את העמודות הבאות:

- 'story' – עמודה המכילה פסקה סיפורית
- 'test_example_id' –המסמן את דוגמת ה-test

### קובץ 3. מחברת הגשה ריקה להגשת התרגיל

**שם הקובץ**: Assignment4-text-analysis.ipynb
המחברת שתריצו בה את הקוד

המחברת אינה מכילה כל קוד מחייב, מלבד המלצות לטעינה וכתיבת הפלט (אותו יש להגיש גם כן כקובץ נפרד).
את הקוד שלכם יש להגיש בקובץ זה

## בדיקת איכות המודל

*מה נחשב מודל איכותי?*

המדד הנבחר להערכת איכות המודל הוא מדד f1.

*תזכורת מדד f1*

f1=2*(precision * recall)/(precision + recall)

*המדד שנבחר להעריך את איכות המודל - Average-f1*

f1_male – מחשבים את f1 (כפי שמוזכר לעיל), כאשר מחשיבים את הכותבים כמחלקה החיובית ואת הכותבות כמחלקה השלילית.

f1_female – מחשבים את f1 (כפי שמוזכר לעיל), כאשר מחשיבים את הכותבות כמחלקה החיובית ואת הכותבים כמחלקה השלילית.

**<u>Average_f1</u> –** יחושב כך: **Average_f1=(f1_male + f1_female)/2**

**שימו לב – מדובר בעצם ב- macro average של מדד f1 (כפי שנלמד בהרצאה של שיעור 9)**

## תאריך הגשת המטלה

את המטלה יש להגיש עד יום ראשון בערב ה-9/1.

## קבצי הגשת המטלה (2 קבצים):

1. Assignment4-text-analysis.ipynb - קובץ ה-jupyter notebook המצורף למטלה (ללא שינוי שמו), המכיל את כל הקוד בו השתמשתם לצורך אימון המודל, וסיווג הדוגמאות החדשות.
   - הקוד אמור להריץ את כל השלבים שישמשו לבניית המודל ולסיווג הדוגמאות בהtest. עליכם להשתמש בקובץ ה-csv המייצג corpus המתויג לאימון ולסווג את דוגמאות האימון המופיעים ב-corpus הלא מתוייג, כפי שמפורט בפסק הבאה.
   - שימו אתם צריכים קוד עובד, מקורי שלכם, שיעבוד גם בסביבה שלנו ויפיק את אותם תוצאות שאתם מצרפים בקובץ ה-csv (המתואר בסעיף הבא).
   - יש ללוות את הקוד שלכם בהערות הסבר בגוף הקוד.
2. classification_results.csv - קובץ csv המכיל את הסיווגים שלכם עבור כל דו' ב-test. הקובץ צריך להכיל 2 עמודות:
   - 'test_example_id' – מזהה המסמן את דוגמת ה-test (לפי הסדר המקורי)
   - 'predicted_category' – עמודה המכילה את סיווג דוגמאות ה-test. התאים בעמודה, יכולים להכיל שני ערכים אפשריים (המייצגים את שני הסיווגים האפשריים):
     - 'm' – עבור הזכרים
     - 'f' - עבור הנקבות

## הסברים על ניקוד המטלה

סך הנקודות שניתן לצבור: 20 נקודות.

*על מה אתם נמדדים?*

עיקר הדגש בניקוד המטלה ינתן על איכות המודל אותו אתם בונים עבור מודל סיווג המגדר.

כדי להעריך את איכות המודל אותו אתם בונים יחושב כשהשוואה בין התוצאות שאתם מגישים בקובץ *classification_results.csv* לבין התוצאות המצופות (שלא חשופים לסטודנטים).

### *אז כיצד נשתמש ב- Average-f1 וכיצד ננקד את המטלות?*

- ע"מ לקבל 3 נקודות – יש להגיש את הקבצים: את קובץ התוצאות *classification_results.csv* ***ואת קובץ*** מחברת המטלה עם הקוד שלכם - Assignment4-text-analysis.ipynb, כאשר יש קוד עובד, אישי עם הערות המסבירות אותו וכאשר התוצאות ב-*classification_results.csv* תואמות את אלה שנריץ בקוד שלכם ב-Assignment4-text-analysis.ipynb
- את תוצאות איכות המודל, מודדים על הסיווגים על ה-test.
  - Average-f1 של 0.425 יזכה אתכם ב-4 נקודות על המטלה (תוצאה זו צפויה להתקבל עם מאמץ מינימלי ללא שום ניסיון לשפר)
  - Average-f1 של 0.71 יזכה אתכם ב18 נקודות
  - על כל תוצאה בין 0.425 ל-0.71 תקבלו את התוצאה היחסית
    - כל שיפור של כ-0.02 בערך הAverage-f1 על ה-test, תזכה בעוד כנקודה (עד תוצאה של 0.71)
  - כדי לקבל יותר מ-18 נקודות:
    - כל שיפור של כ-0.015 מעל 0.71, תזכה בעוד נקודה.
      - תוצאה של 0.725, תזכה ב19 נקודות
      - תוצאה של 0.74 תזכה ב20 נקודות
    - ניתן לקבל בונוס של עד 5 נקודות:
      - כל שיפור של כ-0.015 תזכה בעוד נקודת בונוס (עד ל5 נקודות לכל היותר)

## הסברים נוספים לגבי המטלה

יתווספו, אם יהיה בכך צורך

נא לעקוב אחר רכיב ה"שאלות ותשובות"

בהצלחה לכולם :-)