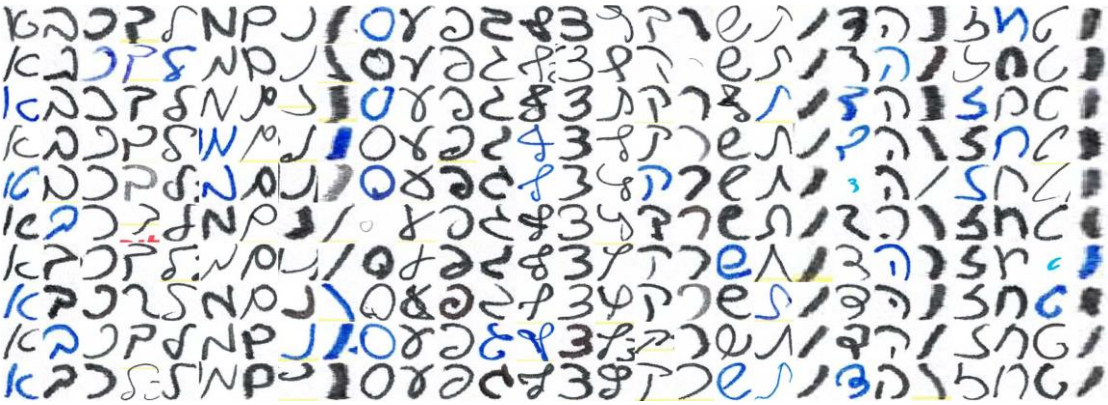
# Classifier for Handwritten Hebrew Letters

**תאריך ההגשה: 7.04.2022, שעה 23:59**

בתרגיל זה תשתמשו באלגוריתם k-Nearest Neighbor כדי לסווג תמונות של אותיות ממאגר HHD_0, שמורכב מאותיות בכתב יד בעברית. המאגר HHD_v0 מכיל בסביבות 5000 תמונות של אותיות בודדות. תמונות אלו מחולקות ל-27 תתי-קבוצות (תתי-תיקיות). כל תיקייה מכילה תמונות של אות מסוימת מתוך האלפבית העברי. פרטים אודות המאגר HHD_v0 ניתן למצוא ב-[1].



**איור 1**: דגימה ממאגר HHD_v0 של אותיות בכתב יד

מטרת התרגיל היא לאמן מסווג k-NN לסווג אותיות.

העבודה תחולק למספר צעדים:

**1.** עיבוד מקדים (pre-processing)

בשלב זה עליכם להעביר את כל האותיות לגודל אחיד.

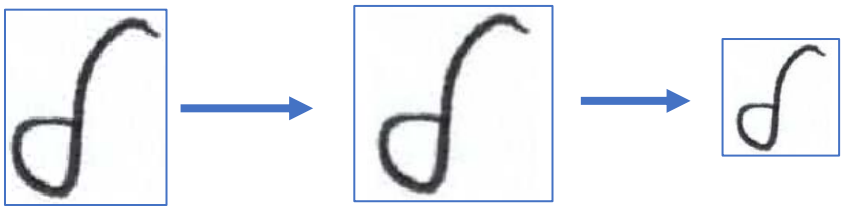
a. המירו את התמונה לגווני אפור (greyscale)

b. הוסיפו לתמונה ריפוד לבן (padding) כדי שגודלה יהיה מרובע

- אם רוחב התמונה קטן מגובה, יש להוסיף Padding מימין ומשמאל
- אחרת, אם רוחב גדול מגובה, יש להוסיף Padding מלמעלה ולמטה

אפשר להיעזר בפונקציית cv2.copyMakeBorder של OpenCV.

c. העבירו את התמונה לגודל אחיד (32,32) בעזרת פונקציית cv2.resize



input

Convert to greyscale
Pad boarders

resize

1. חילקו את המאמגר באופן אקראי לשלוש קבוצות training, validation, and testing sets.
החלוקה תהיה ביחס 80% ל-training , 10% ל-validation ו- 10% ל-testing.
חידוד: תמונות של כל אות צריכות להופיע בכל אחת מהקבוצות ביחס 80%:10%:10%.

בשלב 2, אתם תשתמשו ב-training set כדי לאמן את k-NN, וב-validation set כדי למצוא את הערך הטוב ביותר של k (ערך שנותן דיוק הגבוה ביותר).
לאחר שתמצאו את הפרמטר הטוב ביותר של k, בשלב 3 תריצו את ה-k-NN על ה-testing set כדי לחשב את הדיוק על קבוצת הנתונים אותה המודל לא ראה במהלך האימון .

2. אימון (training).
בשלב זה יש לאמן את המסווג k-NN על ערכים שונים של k, להעריך את התוצאות על validation set עבור כל ערך k, ולבחור את ערך ה-k הטוב ביותר (שנותן הדיוק הגבוה ביותר על validation set).

- יש לאמן את המסווג על הערכים של k בין 1 ל-15 בצעדים של 2.
- בתור פונקציית מרחק יש להשתמש ב- Euclidean distance.

אתם יכולים לממש את k-NN באופן עצמאי (הוא מאוד פשוט) או להשתמש ב-k-NN מתוך הספרייה sklearn (תצטרכו להתקין ספרייה זו כמובן).

3. הערכת ה-k-NN על testing set.
ברגע שמצאתם את הערך האופטימאלי של k, יש להעריך את התוצאות של k-NN על testing set ולדווח את התוצאות.

**פלט התוכנית יכלול**

1. קובץ טקסט בשם "results.txt" שיכיל:
   a. ערך k שנותן דיוק הכי גובה בפורמט
      k = …
   b. דיוק אליו הגיע המסווג עבור כל אחת מהאותיות (27 אותיות שונות) בפורמט

```
Letter     Accuracy
0               …
1               …
…
26              …
```

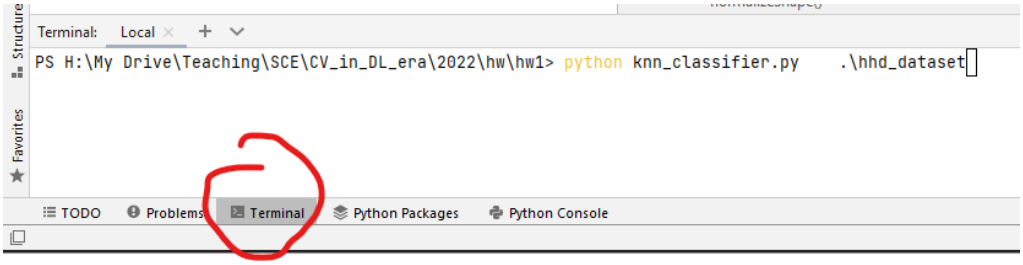
2. Confusion matrix עבור התוצאות בקובץ excel/scv בשם "confusion_matrix.csv". בלינק המצורף של ויקיפדיה נמצא הסבר מהי Confusion matrix

**הרצת התוכנית תתבצע משורת הפקודה בפורמט**

```
> python knn_classifier.py     path
```

כאשר knn_classifier.py הוא שם התוכנית ו-path הו מסלול לתיקייה עם המאגר.

ניתן לעשות זאת מתוך PyCharm באופן הבא




**שימו לב:** על מנת לייעל זמן ריצה, השתמשו ב-vectorization והימנעו מהלולאות במידת הניתן. לדוגמה, במקום לבצע פעולה מסויימת על כל איבר של המערך באמצעות לולאה, ניתן לבצע פעולה זו בו זמנית על כל הערכים כפקודה אחת – טכניקה זו נקראת vectorization.

- המימוש של sklearn.neighbors.KNeighborsClassifier משתמש ב-vectorization, אך מי שיממש kNN בצורה עצמאית, צריך לדאוג לכך, אחרת זמני ריצה יהיו מאוד ארוכים.

**הגשה:**

יש להגיש קובץ zip שמכיל את הקבצים הבאים:

1. קובץ קוד עם התוכנית. הקוד צריך לכלול את כל השלבים שתוארו בתרגיל (חלוקה ל-training/validation/testing, אימון ובחירת k, והרצת התוכנית על testing set)
2. קובץ readme.txt

The readme.txt should include the following information:

**The authors' contact information**

**Description**

A brief description of your program.

**Environment**

Describe the OS and compilation requirements needed to compile and run the program

**How to Run Your Program**

Provide instructions and examples so users how to run the program.
Describe if there are any assumptions on the input.
You can also include screenshots to show examples.

3. קבצים "results.txt" ו- "confusion_matrix.csv" בפורמט שמתואר למעלה

**אופן הבדיקה:**
הבדיקה תתבצע בצורה פרונטלית (או מקוונת אם בעקבות הגבלות הקורונה לא ניתן יהיה לבצע בדיקה פרונטלית). מועדי הבדיקה ייקבעו בהמשך.

**בכל שימוש המאגר HHD_v0, יש לתת הפנייה ל-**[1]

# עבודה נעימה!


**References**
[1] I. Rabaev, B. Kurar Barakat, A. Churkin and J. El-Sana. The HHD Dataset. The *17th International Conference on Frontiers in Handwriting Recognition*, pp. 228-233, 2020.