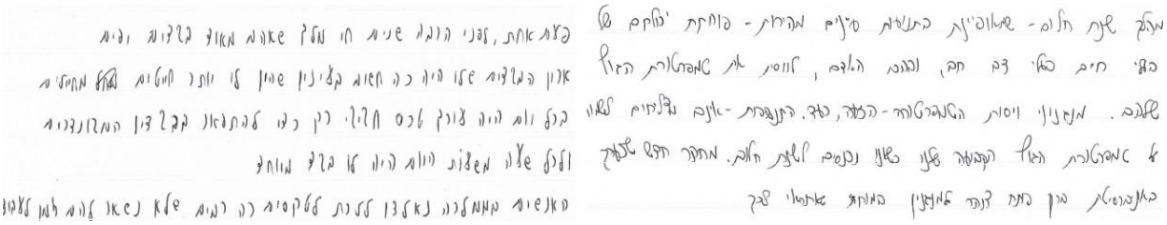
Gender Classification from Handwriting

תאריך ההגשה: 6.02.2022, שעה 23:59

תאריך ההגנה: 8.02.2022 , 9:00-12:00

בפרויקט הסופי של הקורס, תבצעו סיווג של כתב יד לפי מגדר הכותב.
סיווג אוטומטי של מגדר הכותב הוא משימה חיונית במגוון רחב של תחומים, למשל, פסיכולוגיה, סיווג מסמכים היסטוריים או ניתוח קרימינולוגי. מחקרים פסיכולוגיים של ניתוח כתב יד אישרו שניתן לבצע סיווג מגדר לפי מספר הבדלים משמעותיים בכתב יד. באופן כללי, בעוד שכתב ידה של אישה נוטה להיות אחיד יותר, מסודר ובעל מעגליות רבה יותר, כתב ידו של גבר נוטה להיות מחודד יותר, מבולגן ומלוכסן.

מטרת הפרויקט היא לאמן מודל SVM לסיווג אוטומטי של תמונת כתב יד לפי מגדר הכותב.
לאימון המודל תשתמשו במאגר HHD_gender[1]. מאגר זה מכיל בסביבות 850 דגימות של כתב יד בעברית ביחד עם ה-labels שלהם. איור 1 מציג שתי דגימות מהמאגר.



איור 1: דגימות ממאגר HHD_gender (שמאל – כתבת יד של גבר, ימין – כתב יד של אישה)

מטרת הפרויקט היא לסווג את תמונות לפי מגדר (כתב יד של גבר או כתב יד של אישה)
לשם כך תדרשו לאמן מודל SVM, לבצע ניסויים עם פרמטרים ו-kernels שונים, ולדווח איזה שילוב של פרמטרים משיג דיוק הגבוה ביותר.

העבודה תחולק למספר צעדים:

1. טעינת המאגר – המאגר כבר מחולק לשלוש תיקיות train, valid, test עבור אימון, תיקוף והערכת המודל. בכל אחת של התיקיות נמצאות שתי תיקיות female ו-male המכילות דגימות של כתב יד של אישה וכתב יד של גבר.

[1] Rabaev I., Litvak M., Asulin S., Tabibi O.H. (2021) Automatic Gender Classification from Handwritten Images: A Case Study. In: Computer Analysis of Images and Patterns. CAIP 2021. Lecture Notes in Computer Science, vol 13053. Springer, Cham. https://doi.org/10.1007/978-3-030-89131-2_30

2. Feature extraction – בשלב זה תחלצו LBP features כדי לייצג כל תמונה.
במהלך ההרצות אתם נדרשים לבצע ניסוים עם הפרמטרים הבאים ולבחור את השילוב שנותן את הדיוק הגבוה ביותר (ביחד עם בחירת ה-kernel של SVM).

radius	1	3
Number of points	8	24

תזכורת: את ה-LPB יש לחלץ מתמונות בגווני אפור.

3. אימון (training).
בשלב זה יש לאמן את ה-SVM עם ערכים שונים של radius ו-number of points, ו-kernels שונים, להעריך את התוצאות על validation set עבור כל שילוב של פרמטרים ו-kernel, ולבחור את שילוב הטוב ביותר (שילוב שנותן דיוק (accuracy) הגבוה ביותר על validation set).

- יש לאמן את המודל עם linear kernel
- יש לאמן את המודל עם RFB kernel עם הפרמטרים הבאים

```
param_grid = {'C': [0.1, 1, 10, 100],
              'gamma': [1, 0.1, 0.01, 0.001, 0.00001, 10]}
```

מצורפים לינקים לשני tutorials מצויינים עם הסברים על תיאוריה ודוגמאות הרצה:

1. Support Vector Machines (SVM)
2. SVM using Scikit-Learn in Python

4. הערכת ה-SVM על test set.
ברגע שמצאתם את השילוב האופטימאלי של של הפרמטרים ו-kernel, יש להעריך את התוצאות של המודל על test set ולדווח את התוצאות.

כפלט, על התוכנית ליצור קובץ טקסט בשם "results.txt" שיכיל

1. ערכים של הפרמטרים שנותנים דיוק הכי גובה
2. דיוק אליו הגיע המודל, למשל,

Accuracy: 75.7%

- מספיק שתי ספרות אחרי הנקודה העשרונית

3. Confusion matrix עבור התוצאות בצורה

	male	female
male		
female		

- בלינק המצורף של ויקיפדיה נמצא הסבר מהי Confusion matrix

הרצת התוכנית תתבצע משורת הפקודה בפורמט

```
> python classifier.py    path_train path_val path_test
```

כאשר classifier.py הוא שם התוכנית ו-path_train, path_val, path_test הם שמות התיקיות בהם נמצאים תמונות של train, validation and test sets.

הגשה:
יש להגיש

- קובץ/קבצי קוד עם התוכנית
- קובץ readme.txt
- קובץ "results.txt"

אופן הבדיקה:
הבדיקה תתבצע בצורה פרונטלית (או מקוונת במידה ולא ניתן יהיה לבצע בדיקה פרונטלית מסיבות שקשורות לנגיף הקורונה).

עבודה נעימה!

הערה: אם בעתיד תרצו להשתמש במאגר HHD_gender, יש לתת רפרנס למאמר הבא
https://doi.org/10.1007/978-3-030-89131-2_30