**המחלקה להנדסת תוכנה**

**פרויקט גמר – תשפ"א**

**מציאת משפטים, מילים ואותיות בכתב יד של השפה הערבית**

**Finding handwritten sentences, words and letters of the Arabic language**

**מאת**

**אביגייל הילה שרבף - 318631488**

**ליאל לוי - 207045741**

מנחה אקדמי: דר' יהודה חסין אישור: תאריך:

רכז הפרויקטים: דר' אסף שפנייר אישור: תאריך:

מערכות ניהול הפרויקט:

|  |  |  |
| --- | --- | --- |
| # | מערכת | מיקום |
| 1 | מאגר קוד | <https://github.com/liellevy88/Author-verification-by-handwriting-samples-arabic> |
| 2 | יומן | <https://trello.com/b/eWo8hygi/author-verification-by-handwriting-samples-arabic> |

מידע נוסף:

|  |  |
| --- | --- |
| סוג הפרויקט | מחקרי ממרצה במכללה |
| פרויקט מח"ר | לא |
| פרויקט ממשיך | זה פרויקט חדש שמתבסס על כלים קיימים. |
| פרויקט זוגי: | כן.  היקף הפרויקט גדול מאוד ויש הרבה דברים שצריך לחקור.  בנוסף ניתן לפתח ולהגדיל את נפח הפרויקט כך שיתאים לעבודה של שלושה אנשים ויותר. |

מבוא

בחינת הכניסה הפסיכומטרית לאוניברסיטאות משמשת ככלי מיון לכניסה לאוניברסיטאות ולמכללות השונות.

הבחינה נבנית על ידי "המרכז הארצי לבחינות ולהערכה" (מאל"ו)ומתקיימת בשפות: עברית, ערבית, רוסית, צרפתית ובנוסח משולב של אנגלית ועברית .

בבחינה הפסיכומטרית ישנה מטלת כתיבה (חיבור) המהווה 10% מציון כלל הבחינה ובה הפרויקט שלנו מתמקד.

במטלת הכתיבה נדרש הנבחן לכתוב חיבור בכתב ידו באורך של 50-25 שורות.

המרכז הארצי לבחינות ולהערכה נוקט באמצעים ומאמצים רבים על מנת להבטיח את טוהר הבחינה ולמנוע רמאות מכל סוג ככלל, והעתקות וזיופים בפרט.

למרות האמצעים אשר ננקטים כדי למנוע זיופים, עדיין ישנם מקרים של התחזות, בהם אדם אחר מבצע את הבחינה במקום המועמד.

בעקבות כך, המרכז הארצי לבחינות והערכה מעסיק מומחים לזיהוי כתבי יד, בכדי לנסות לבצע השוואה בין שני מועדי בחינה של נבחנים מסוימים אשר מוגדרים כחשודים, על ידי בדיקה של כתב ידו של הנבחן במטלת החיבור.

השוואת כתב ידו של נבחן בין שני המועדים מתבססת על ההנחה שכתב ידו של כל אדם הוא ייחודי ושניתן על פיו לזהות את כותבו.

בנוסף יש הסכמה כי אין לשני בני אדם כתב יד זהה לחלוטין ואף האדם אינו יכול לכתוב

בשנית באופן טבעי דברים שכתב בעבר בצורה זהה לגמרי.

פרויקט מחקרי זה הינו ביוזמה של המרכז הארצי לבחינות והערכה, בהנחייתו של דר' יהודה חסין. מתוך ההנחה כי קשה להכריע ששני כתבי יד זהים או שונים, נרצה לבנות מערכת שמקבלת כקלט שני חיבורים בערבית סרוקים של אותו נבחן (ממועדים שונים).

מערכת זו תקבע את הסיכוי שהחיבורים שייכים לשני אנשים שונים ככלי עזר לזיהוי רמאות בבחינה .

תיאור הבעיה

מידי שנה ניגשים כ- 70,000 נבחנים לבחינה הפסיכומטרית, כשליש מתוכם ניגשים לבחינה בשנית. על מנת להבטיח את טוהר הבחינה, מאל"ו מעסיק מומחים לזיהוי כתב יד אשר מבצעים בדיקה ידנית של מאות ואלפי בחינות.

בשל הקושי הרב במציאת מומחים לזיהוי כתב יד בערבית, מאל"ו זקוק למערכת שתשלח בחינות לבדיקה ידנית (ע"י מומחים) רק לאחר שזוהו על ידה כחשודים, וכך נצמצם את כמות הבדיקות הידניות.

הבעיה מבחינת הנדסת תוכנה

על מנת לבצע השוואה בין שני כתבי יד ואימות המחבר, נדרש תהליך המורכב משלושה חלקים מרכזיים איתם נתמודד בפרויקט: גילוי, זיהוי ואימות.

בניגוד לזיהוי כתב של תמונה עם טקסט מודפס, הנחשב קל יחסית לביצוע, התוצאות לגילוי וזיהוי כתב יד הן חלקיות בלבד.

האתגר הקשה ביותר עמו נאלץ להתמודד הוא היכולת לקבוע בסבירות גבוהה האם שני כתבי יד שייכים לאותו אדם או לא וזאת מכיוון שהכתב בשפה הערבית מחובר וקשה מאוד למצוא אותיות שמופרדות אחת מהשנייה כדי לגלות אותן בעזרת עיבוד תמונה ולאחר מכן לבצע השוואה בין אותן אותיות.

תיאור הפתרון

הפתרון הוא יצירת מערכת ממוחשבת שתקבל כקלט 2 חיבורים בשפה הערבית ותקבע מה הסיכויים שהחיבורים נכתבו על ידי אותו אדם.

המערכת תעשה זאת על ידי השלבים הבאים:

1.שלב הגילוי – בשלב זה המערכת תגלה את השורות בחיבור, לאחר מכן את המילים ולבסוף תמצא את האותיות המופרדות.

2. שלב הזיהוי – בשלב זה ניקח את אוסף האותיות משלב הגילוי ונרצה לזהות אילו אותיות גילינו.

3. שלב האימות – בשלב זה נקבל את ההחלטה האם החיבורים נכתבו על ידי אותו אדם או לא ונעשה זאת באמצעות השוואת אותיות בין שני החיבורים.

סקירת ספרות

Hybrid Feature Learning for Handwriting Verification

<https://arxiv.org/pdf/1812.02621.pdf>

המאמר העיקרי עליו אנו מבססים את השוואת המילים והאותיות (שלב האימות), מתמקד בזיהוי כותב החיבור על ידי מילה בודדת בשפה האנגלית- and.

המאמר מתבסס על כך שהמילה and היא המילה הרביעית השכיחה ביותר בשפה האנגלית ומכך שלמילה זו קיים מאגר מידע גדול שיוכל לשמש לאימון אלגוריתם הלמידה.

ההשוואה נעשית ע"י שילוב של רשת נוירונים, Auto-Encoder ואלגוריתם SIFT לקביעת אחוזי ההתאמה בין 2 מילים.

Off-line Arabic Handwriting Recognition: A Survey

<https://www.researchgate.net/profile/Venu_Govindaraju/publication/3194075_Offline_Arabic_Handwriting_Recognition_A_Survey/links/09e41512f527c59630000000/Offline-Arabic-Handwriting-Recognition-A-Survey.pdf>

* מצורף הקובץ של המאמר מכיוון שנתקלנו בבעיות לפתוח אותו דרך הקישור.

המאמר מדבר על הסוגייה והמורכבות הטכנית של זיהוי כתב יד בשפה הערבית וסוקר את השיטות שהוצעו ומומשו עד היום בנושא זה.

בנוסף, מדובר באופן ממוקד על זיהוי אותיות בתוך מסמך סרוק הרשום בכתב יד בשפה הערבית.

המאמר מתאר את תהליך הזיהוי על ידי 5 שלבים אשר בכל שלב מוצעים פתרונות שונים שמממשים את מטרת השלב.

# Recognition of cursive Arabic handwritten text using embedded

# training based on HMMs

<https://www.sciencedirect.com/science/article/pii/S2314717217300156#fig0020>

המאמר מציג מערכת לזיהוי מילים ואותיות של כתב יד מחובר בשפה הערבית.

המערכת מתבססת על מודלים נסתרים של מרקוב (HMM).

מהתמונה מחולצות תכונות המבוססות על צפיפות פיקסלים, קעירות ותכונות נגזרות באמצעות חלון הזזה.

המודל המוצע שיפר את הזיהוי והראה תוצאות טובות.

נספחים

## **תכנון הפרויקט**

|  |  |
| --- | --- |
| 25.10.2020 | קריאת מאמרים הקשורים לזיהוי כתב יד. |
| 1.11.2020 | חקירת הנושא של זיהוי כתב יד בשפה הערבית. |
| 8.11.2020 | חיפוש/ בניית DB למילים ואותיות בשפה הערבית. |
| 22.11.2020 | התאמת או כתיבת קוד לגילוי השורות והמילים בחיבור. |
| 20.12.2020 | פירוק מילים מהחיבור לאותיות בודדות.  במקרה של חוסר הצלחה מעבר לפירוק מילים. |
| 20.02.2021 | בניית רשת נוירונים שמזהה אותיות בשפה הערבית. |
| 14.03.2020 | שימוש באלגוריתמים להשוואת האותיות וקביעת אחוזי התאמה. |
| 04.04.2021 | בדיקת אחוזי הצלחה של זיהוי המחבר. |
| 18.04.2021 | שיפור וייעול לפרויקט. |