**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> |
| 3 | סרטון גרסת אלפא | <https://drive.google.com/file/d/1qJ2mwm82YsDd_yXfFqL0J9Y2iNXCxiqg/view?usp=sharing> |

נאום המעלית

בפרויקט זה נבצע מחקר למציאת אלגוריתם שבעזרתו נפתח מערכת ממוחשבת שתקבל שני טקסטים סרוקים הכתובים בכתב יד בערבית, ותחזיר כפלט את הסיכוי (באחוזים) שהטקסטים השונים נכתבו על ידי אותו אדם.

המוצר הסופי מיועד לשימוש ע"י המרכז הארצי לבחינות והערכה כדי לזהות רמאות בבחינות, אך הוא יכול להוות פתרון במגוון רחב של תחומים (לדוגמה: עבור מחלקת הזיהוי הפלילי במשטרה).

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. הבעיה הנ״ל נפתרה לשפה העברית בפרויקט גמר בשנה שעברה. בפרויקט זה נתמודד עם קשיים נוספים הקשורים לשפה הערבית והם:

1. בניגוד לשפה העברית, הכתב בשפה הערבית מחובר וקשה מאוד למצוא אותיות שמופרדות אחת מהשנייה כדי לגלות אותן בעזרת עיבוד תמונה ולאחר מכן לבצע השוואה בין אותן אותיות.
2. ישנו קושי בגילוי המילים והאותיות בכתב יד שבניגוד לכתב מודפס המרווחים בין האותיות, המילים והשורות אינם קבועים.
3. מכיוון שהכתב בערבית מחובר וגם כי כתב יד משתנה מאדם לאדם כל בדיקה של שני חיבורים תצליח למצוא כמות שונה של אותיות מופרדות דבר זה יכול להשפיע על תוצאות הבדיקה.
4. בניגוד לשפה העברית שבה כל אות נכתבת באותה צורה ללא תלות במיקומה במילה, בשפה הערבית כל אות יכולה להיכתב בשלוש צורות בהתאם למיקומה במילה (התחלה, אמצע וסוף).
5. בניגוד לשפה העברית שבה הניקוד אינו חלק מהאות ולכן אין לו חשיבות בזיהוי, בשפה הערבית ישנן מספר אותיות שנכתבות באותה צורה בדיוק ומה שמבדיל ביניהן זהו רק הניקוד. דבר זה עלול להקשות על רשת הנוירונים לזהות את האות הנכונה.

תיאור הפתרון

בפרויקט זה נשתמש בארכיטקטורת המערכת שמומשה בפרויקט של שנה שעברה (אימות מחבר המבוסס על ניתוח כתב יד בעברית).

הפתרון הוא יצירת מערכת ממוחשבת, שתקבל כקלט מהמשתמש שני חיבורים של כתבי יד בשפה הערבית. עם קבלת פקודה, התוכנה תבצע ניתוח והשוואה של כתבי היד, ותחזיר כפלט את הסיכוי שהחיבורים נכתבו על ידי אותו אדם.

בפרויקט זה נשתמש בחלק מהאלגוריתמים שמומשו ע"י הסטודנטים דניאל גבאי ושחר ישראלי בפרויקט גמר משנה שעברה שבוצע בהנחייתו של דר' יהודה חסין, אשר בנו מערכת להשוואת שני חיבורים בכתב יד בשפה העברית. מטרת הפרויקט הייתה לתת מענה לבעיה של גילוי רמאות בבחינת הפסיכומטרית ללא התערבות ובדיקה של בני אדם. בפרויקט זה הצליחו לקבוע באחוזי דיוק גבוהים האם שני חיבורים נכתבו על ידי אותו אדם או לא. הפרויקט שלנו מתעסק בפתירת הבעיה הנ״ל לשפה הערבית.

בכדי לבצע את תהליך הזיהוי של כתב היד מהתמונה, הבעיה הופרדה לשני שלבים מרכזיים: גילוי (detection) וזיהוי (recognition). שלבים אלו יהיו גם חלק מארכיטקטורת המערכת שלנו עם התאמות ושינויים לצרכינו. יתר על כן, הפרויקט שלנו מצריך שלב נוסף: שלב האימות. שלב זה הוא החלק המרכזי והמורכב בפרויקט שהוא ביצוע ההשוואה בין שני החיבורים ואימות המחבר. בכדי שנוכל לבצע את ההשוואה, נרצה משלב הגילוי והזיהוי לחלץ אותיות שישמשו אותנו כגורמי השוואה.

Start

End

Load essay one

Load essay two

Run program

Get score

Finish all essays

Yes

No

בתמונה - High level design

**פירוט השלבים:**

1. **גילוי detection))** – שלב זה מורכב מניתוח התמונה ומציאה של החלקים המכילים את הכתב אותו יש לפענח. תהליך הגילוי מורכב ממספר תתי שלבים עיקריים:
2. עיבוד מקדים- את החיבורים אנו מקבלים בפורמט tiff מהמרכז הארצי לבחינות והערכה. כל חיבור מכיל שני עמודים. בחרנו לאחד את שני העמודים לתמונה אחת על מנת שיהיה לנו עמוד אחד שיכלול את כל החיבור. לאחר מכן, נבצע חיתוך של קצוות התמונה במטרה ללכוד רק את הטקסט הנכתב על ידי הנבחן.
3. גילוי שורות – מציאת השורות בטקסט מתוך כתב היד בתמונה. גילוי השורות נעשה ע"י מציאת הרווחים בין השורות באמצעות אלגוריתם שעובר בצורה רוחבית על התמונה וסוכם את הפיקסלים של כל שורה. מבחינה גרפית, מתקבלת פונקציה בעלת הרבה 'פיקים' כך שנקודות המינימום הקרובות לכל פיק משמאל ומימין, הן הנקודות שהאלגוריתם זיהה לתחילת השורה וסוף השורה (כלומר הרווח שמעל ומתחת לשורה).
4. גילוי אותיות – הפרדת כל שורה לאותיות.

בדומה לגילוי השורות, גילוי האותיות מתבצע ע"י מציאת הרווחים בין אות לאות.

Example of detection lines and letters

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

Detect lines



Detect letters



1. **שלב הזיהוי (recognition)** –נרצה לזהות אילו אותיות גילינו בשלב הגילוי.

לאחר שמצאנו את מיקומי השורות וגילינו את מיקום האותיות בטקסט, נפעיל מודל (רשת נוירונים שאותה אנחנו בונות כרגע) שאומן לזהות את 28 האותיות בשפה הערבית. נרצה לשמור לשלב הבא רק אותיות שהאלגוריתם זיהה בהסתברות גבוהה. רמת הביטחון בזיהוי האות חשובה לנו כיוון שבשלב ההשוואה נרצה להשוות בין אותיות **זהות בלבד**.

נתאר את הרעיון הכללי של שלב האימות.

1. **שלב האימות (verification)** – שלב זה הוא השלב המרכזי בו מתקבלת ההחלטה האם שני החיבורים נכתבו על ידי אותו אדם או לא. נכון להיום, החלטנו לפעול בשתי דרכים עיקריות: השוואת אותיות ואלגוריתם 'קוף' (שלנו).
2. השוואת אותיות – משלב הגילוי והזיהוי, נשארנו עם גורמי ההשוואה משני החיבורים. כעת נרצה לבצע השוואה בין אותן אותיות משני החיבורים (אליף מול אליף וכו'). כלומר, מכל חיבור נשאיר רק את האותיות שהצלחנו לזהות בשני החיבורים. את ההשוואה נבצע באופן הבא:

נשתמש באלגוריתם Auto-Encoder (AE). באמצעות האלגוריתם נוכל לחלץ מכל דגימה (=אות) את הפיצ'רים הכי חשובים שמאפיינים אותה. מתוך הפיצ'רים שחילצנו האלגוריתם יבחר את הפיצ'רים החשובים ביותר (יבדוק מה יניב תוצאה יותר טובה), ולאחר שנחלץ משתי הדגימות את הפיצ'רים, נבצע ביניהם פונקציית חיסור ואת התוצאה נעביר לאלגוריתם סיווג נוסף (Logistic Regression/CNN) שאומן מראש לזהות האם תוצאת החיסור מאפיינת דגימות של אותו מחבר או מחברים שונים.

1. אלגוריתם קוף – עבור כל חיבור, נגדיר וקטור בגודל 28 (כמספר האותיות בשפה הערבית). בכל תא בוקטור נשמור את כמות המופעים (באחוזים) מכל אות שהמודל הצליח לזהות (תא 0 האות אליף וכו'). לאחר מכן, נבצע חיסור בערך מוחלט בין שני הווקטורים משני החיבורים. את וקטור החיסור נעביר למודל סיווג נוסף (Logistic Regression). המודל אומן לזהות וקטורי חיסור של חיבורים שנכתבו על ידי אותו אדם, ווקטורי חיסור של חיבורים שנכתבו על ידי אנשים שונים.

תיאור סכמתי של פעולת האלגוריתם:

Logistic Regression

Vector 2

Vector 1

## ארכיטקטורת המערכת

התרשים הבא מתאר את ארכיטקטורת המערכת על כל שלביה:

Detection Phase

Verification Phase

Recognition Phase

Load document

Identify letters

Load & compile trained recognition models

Store in Data Structure

Detect letters

Detect lines

Prepare document

Run 'monkey' algorithm

Store in Data Structure

Processed Both Documents?

Load & compile trained verification models

Compare by letters

Analyze results

Return score

**No**

**Yes**

## טכנולוגיות

את הפרויקט בחרנו לממש בשפת python. בחירה זו נעשתה מכיוון שאנו משתמשים בעיבוד תמונה, רשתות נוירונים (ומודלים נוספים של machine learning), חישובים מתמטיים וסטטיסטים, ואכן לכל אלו קיימות ספריות רבות, חזקות ונוחות לשימוש בשפה (numpy, sklearn, pandas, matplotlib, keras, PIL tensorflow ועוד) מה שהפך את הבחירה למובנת מאליה.

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.

תכנית בדיקות

קיבלנו מהמרכז הארצי לבחינות 39 זוגות של חיבורים, מתוכם 10 זוגות של חיבורים השארנו בצד לצורך הבדיקה הסופית של הקוד. על ה-29 זוגות החיבורים שנותרו רצינו לאמן את המודל ולשם כך הרצנו בדיקות על מנת לראות שנקבל כמות מספקת של תוצאות מספיק טובות כדי שנוכל להסתמך על התוצאה הסופית שנקבל ולהגיד שהיא מהימנה.

לאחר הרצת הקוד שמגלה את השורות בחיבורים בדקנו ומצאנו שכל השורות בכל 29 זוגות החיבורים נמצאו. תוצאה זאת טובה מאוד שכן לא איבדנו אף שורה כלומר, אף דאטה לא הלך לאיבוד וכך יכולנו להתקדם לשלב הבא, גילוי האותיות. לאחר הרצת הקוד שמגלה אותיות על החיבורים ראינו שהקוד מצא כ-64 אלף פריטים. לאחר סינון חלקי גילינו שהקוד מצא בחיבורים הנ״ל 1712 אותיות מופרדות ומתוך מספר זה 15 אותיות שונות. על פי תוצאה זו נוכל להגיד כי לאחר שנשלים את הסינון על כל הפריטים יהיו בידינו מספר רב של אותיות איתן נוכל להמשיך הלאה ולשלוח אותן לזיהוי על ידי רשת הנוירונים.

סיכום / מסקנות

בפגישה הראשונה עם דר' יהודה חסין, הוצגו לנו מספר פרויקטים אפשריים. בסופו של דבר, החלטנו לבחור בפרויקט זה מכיוון שהוא גרם לנו לעניין רב מבחינת הנושא של עיבוד תמונה ולמידה חישובית שבהם לא התעסקנו בתואר עד כה ובנוסף זה אתגר גדול עבורנו. ממחקר ראשוני שעשינו בספרות, גילינו כי הבעיה איתה אנו מתמודדים מורכבת מאוד, יש הרבה דרכים שונות לתקוף אותה ואין תשובה חד משמעית כיצד לפעול.

בחלקו הראשון של הפרויקט חקרנו מאמרים ולמדנו את הטכנולוגיות איתן נעבוד (python, עיבוד תמונה, machine learning). לאחר שהתחלנו לחשוב על דרכים אפשריות לפתור את הבעיה, התחלנו לאסוף data של האותיות בערבית וביצענו התאמות מהפרויקט שבוצע בשנה שעברה ע"י דניאל גבאי ושחר ישראלי לפרויקט שלנו.

נכון להיום סיימנו עם שלב הגילוי ואנחנו לקראת סיום שלב הזיהוי.

בסמסטר הבא נותר לנו לסיים את שלב הזיהוי ולבצע את לב האלגוריתם שלנו שהוא שלב האימות, כלומר ההשוואה בין האותיות.

לבסוף, נרצה לשקלל את כל דרכי הפתרון (השוואה בין אותיות ואלגוריתם קוף) לכדי תוצאה אחת שתקבע את רמת הביטחון שהחיבורים נכתבו על ידי אותו אדם או לא.

יש לציין כי במידה ולא נגיע לאחוזי דיוק גבוהים עבור זיהוי האותיות בגלל הבעיות הייחודיות בשפה הערבית שציינו בתיאור הבעיה, נבדוק את האופציה שהמערכת תזהה את המחבר על ידי זיהוי והשוואת מילים.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | **PRIORITIZE** | | | **PLAN TO REDUCE IMPACT** |
| **Sort by priority** | **Description of risk** | **Owner** | **Possible Impact** | **Prob (%)** | **Impact (L, M, H, VH)** | **Risk Code (color per table above)** | **Mitigation Plan / Contingency Plan** |
| 5 | אי עמידה בזמני ההגשות של פרויקט הגמר | אביגייל וליאל | אי הגשה של המשימות, הורדה בציון. | 15% | VH |  | בניית לוח זמנים עבור משימות הפרויקט ומעקב אחר סיום המשימות בזמן. |
| 4 | קושי בהתמודדות עם השפה הערבית (לא שפת האם שלנו) | אביגייל וליאל | עיכוב בהתקדמות הפרויקט בגלל תלות באנשים אחרים | 60% | VH |  | קבלת עזרה מאנשים הדוברים את השפה הערבית. |
| 2 | שילוב טכנולוגיות חדשות בפרויקט הגמר שחברי הפרויקט לא עבדו איתם בעבר | אביגייל וליאל | עיכוב בהתקדמות הפרויקט | 40% | H |  | הקצאה של חלק מהזמן ללמוד את הטכנולוגיות החדשות. |
| 3 | קושי במציאת כמות מספיקה של אותיות בודדות | אביגייל וליאל | אחוזי הצלחה נמוכים בקביעת זיהוי המחבר | 60% | H |  | שיפור האלגוריתם של זיהוי האותיות או מעבר לזיהוי מילים. |
| 1 | תקופת מבחנים + פרויקטים אחרים | אביגייל וליאל | פחות זמן עבודה על הפרויקט והתמקדות במבחנים/פרויקטים | 60% | H |  | חלוקת משימות ועבודה יותר אינטנסיבית בתקופה שלפני. |
| 7 | לחלות בקורונה | אביגייל וליאל | פחות זמן עבודה על הפרויקט בזמן המחלה | 30% | M |  | השתדלות לעמוד בלוח הזמנים שהגדרנו ואפילו להתקדם מעבר לו. |
| 6 | כתב יד לא מובן של הנבחן ,קשקושים על מחברת הבחינה אשר לא יאפשרו לבצע את הקריטריונים ההשוואה. | אביגייל וליאל | קושי בניתוח החיבור וקביעה חדש משמעית לגבי אימות הנבחן | 50% | M |  | אלגוריתם 'קוף' כפי שתואר יוכל אולי לטפל בבעיה זו. |

נספחים

\*באישור המנחה, לא בוצעו החלקים שאינם נדרשים מפרויקטים מחקריים.

## **טבלת סיכונים -- Risk assessment**

## **תכנון הפרויקט**

ניתן לגשת ליומן בקישור שבעמוד הראשון.

## **ביבליוגרפיה**

דניאל גבאי ושחר ישראלי (2020). Author verification based on handwritten text analysis

<https://github.com/DanielGabay/Author-verification-by-handwriting-samples>

**Hybrid Feature Learning for Handwriting Verification**

By:Mohammad Abuzar Shaikh, Mihir Chauhan, Jun Chu and Sargur Srihari

Publication date: 19 November 2018

**Off-line Arabic Handwriting Recognition: A Survey**

By: Liana M. Lorigo, *Member, IEEE* and Venu Govindaraju, *Senior Member, IEEE*

Publication Date: June 2006

# **Recognition of cursive Arabic handwritten text using embedded training based on HMMs**

By: Rabi Mouhcinea Amrouch Mustaphaa and Mahani Zouhirb

Publication Date: September 2018