**המחלקה להנדסת תוכנה**

**פרויקט גמר – תש"פ**

ניתוח כתב יד על מנת לזהות זיופים בבחינה הפסיכומטרית

Handwriting analysis to detect forgeries in the psychometric exams

**מאת**

**דניאל גבאי**

**שחר ישראלי**

**מנחה אקדמי: דר' יהודה חסין אישור: תאריך:**

**רכז הפרויקטים: דר' אסף שפיינר אישור: תאריך:**

מערכות ניהול הפרויקט:

|  |  |  |
| --- | --- | --- |
| # | מערכת | מיקום |
| 1 | מאגר קוד | <https://github.com/DanielGabay/Author-verification-by-handwriting-samples> |
| 2 | יומן | <https://trello.com/b/7jKnxmLL/author-verification-by-handwriting-samples> |

נאום מעלית

בפרויקט זה נרצה ליצור אלגוריתם שעבור שני טקסטים סרוקים הכתובים בכתב יד, מחזיר כפלט את הסיכוי באחוזים שהטקסטים השונים נכתבו על ידי אותו אדם. המוצר מיועד לשימוש ע"י המרכז הארצי לבחינות והערכה כדי לזהות רמאות בבחינות, אך הוא יכול להוות פתרון במגוון רחב של תחומים (לדוגמא: עבור המז"פ של המשטרה). בשלב הראשוני הפרויקט יותאם לשפה העברית אך יהיה ניתן להרחיבו לשפות נוספות.

מבוא

מבחן הכניסה לאוניברסיטאות בישראל הינו בחינה פסיכומטרית הנערכת החל משנת 1981. הבחינה משמשת ככלי מיון לכניסה לאוניברסיטאות ולמכללות השונות. הבחינה נבנית על ידי "המרכז הארצי לבחינות ולהערכה" ומתקיימת בשפות: עברית, ערבית, רוסית, צרפתית, ספרדית ובנוסח משולב של אנגלית ועברית.

הבחינה הפסיכומטרית בודקת את יכולות המועמד בשלושה תחומים: חשיבה מילולית, חשיבה כמותית ובאנגלית. בנוסף, ישנה מטלת כתיבה (חיבור) המהווה 10% מציון כלל הבחינה. ובה הפרויקט שלנו מתמקד. מטלת הכתיבה היא החלק הראשון בבחינה, בה נדרש הנבחן לכתוב חיבור באורך של 25-50 שורות. המטלה נכתבת בעיפרון, על דף בן 50 שורות המיועד לכך .מטלת הכתיבה הינה המטלה היחידה במבחן בה נדרש הנבחן לכתוב בכתב ידו.

המרכז הארצי לבחינות ולהערכה נוקט באמצעים ומאמצים רבים על מנת להבטיח את טוהר הבחינה ולמנוע רמאות מכל סוג, בפרט העתקות וזיופים. למרות האמצעים אשר ננקטים כדי למנוע זיופים, עדיין ישנם אנשים אשר מצליחים לקבל ציון גבוה בבחינה באמצעות תשלום או בקשה מאדם אחר שייגש למבחן במקומם.

בעקבות כך, המרכז הארצי לבחינות ולהערכה 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. ההשוואה אמורה לאמת/להפריך חשד עבור נבחן שרימה בבחינה.

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 שכת בעבר בצורה זהה לגמרי. עם זאת, סגנון הכתיבה של האדם מאופיין, מוגדר, אישי וייחודי לכל אדם. פעולת הכתיבה היא תוצאה של שיתוף פעולה בין מערכות אחדות בגוף, כגון הראייה, התפיסה ועיבוד הנתונים, וכן פעולות מוטוריות של שרירים רבים. שיתוף כל אלה לפעולות שגרתיות המתבצעות מדי יום, מפתח מיומנות שהיא הבסיס לסגנון כתב היד על תכונותיו הייחודיות.

פרויקט זה הינו יוזמה של המרכז הארצי לבחינות והערכה, בהנחייתו של דר' יהודה חסין.

מתוך ההנחה כי קשה למצוא שני כתבי יד זהים, נבנה תוכנית שמטרתה לבדוק האם יתכן ונבחן ביצע את מטלת החיבור במועד מסוים ובמועד אחר ביצע את מטלת הכתיבה תחת זהות שונה.

כלומר, נרצה לבדוק האם החיבור בשני מועדים שונים נכתבו על ידי אותו אדם.

תיאור הבעיה

## דרישות ואפיון הבעיה

מדי שנה נגשים כ- 70,000 נבחנים לבחינה הפסיכומטרית, כאשר שליש מתוכם נגשים לבחינה בשנית. על מנת להבטיח את טוהר הבחינה, המרכז הארצי לבחינות והערכה מעסיק מומחים לזיהוי כתב יד אשר מבצעים בדיקה ידנית של מאות ואלפי בחינות.

במטרה לייעל את תהליך אימות הנבחן ולצמצם את מספר הבדיקות הנעשות ע"י המומחים, המרכז הארצי לבחינות והערכה זקוק למערכת ממוחשבת ואוטומטית, אשר באמצעותה יהיה ניתן לקבוע בהסתברות גבוהה, האם בשני מועדי בחינה נתונים מדובר באותו הנבחן.

באמצעות המערכת ניתן יהיה לצמצם את כמות הבדיקות אשר מגיעות לידי המומחים כך שרק הנבחנים שזוהו ע"י המערכת כחשודים יגיעו לידיהם לבדיקה ידנית מעמיקה. חשוב לציין כי בכל מקרה המערכת תוכל לעזור רק בצמצום כמות הבדיקות אך לא תקבע באופן חד משמעי אם נבחן פעל שלא בתום בכתיבת הבחינה, זאת ייקבע רק ע"י המומחים.

## הבעיה מבחינת הנדסת תוכנה

על מנת לבצע השוואה בין שני כתבי יד (באופן לא מקוון) ואימות המחבר, נדרש תהליך המורכב משלושה חלקים מרכזיים איתן נתמודד בפרויקט: גילוי, זיהוי ואימות, עליהן נפרט בהמשך.

בניגוד לזיהוי כתב של תמונה עם טקסט מודפס, הנחשב קל יחסית לביצוע, התוצאות לגילוי וזיהוי

כתב יד הן חלקיות בלבד. בנוסף, אין מסד נתונים בשפה העברית של כתבי יד שבו ניתן להשתמש כדי לאמן מכונה, לעומת השפה האנגלית בה המידע נגיש ונפוץ יותר ברחבי האינטרנט.

האתגר הקשה ביותר איתו נאלץ להתמודד בפרויקט הוא היכולת לקבוע בסבירות גבוהה האם שני כתבי יד שייכים לאותו אדם או לא. אתגר זה נובע מכמה סיבות:

* בכתב יד, בניגוד לכתב מודפס, המרווחים בין האותיות, המילים והשורות אינם קובעים, דבר המקשה בגילוי המילים אשר נשתמש בהם להשוואה בין כתבי היד.
* "המרכז הארצי לבחינות והערכה" דורש כי הפרויקט יקבל חיבורים בכתב יד בשפה העברית אך שיהיה ניתן לבצע התאמה גם לחיבורים בכתבי יד בשפות נוספות כמו ערבית. כלומר הפרויקט צריך להתבצע בצורה גנרית כמה שיותר.
* כל בדיקה של שני טקסטים היא בדיקה אינדיבידואלית. כמות המידע (הטקסט) שיש ברשותנו על כל נבחן מוגבלת ויכולה להשפיע על תוצאות הבדיקה.
* ממחקר ראשוני שביצענו עולה כי הבעיה איתה אנו מתמודדים מורכבת והניסיונות לפתור אותה הניבו תוצאות חלקיות בלבד.

תיאור הפתרון

אנו חוקרים מספר אפשרויות לפתרון. אנו יוצאים מנקודת הנחה כי שילוב כל הדרכים יחד, יניב את התוצאה הנכונה ביותר לאימות זהותו של כותב המבחן.

הדרכים לפתרון אותן נחקור הם:

* ניסוח לאלגוריתם קוף
* השוואה בין אותיות זהות בין שני מועדי הבחינה.
* השוואה בין מילים זהות בין שני מועדי בחינה.

כפי שציינו, ארכיטקטורת המערכת מורכבת משלושה שלבים מרכזיים (התיאור הבא הוא עבור השוואת מילים):

1. **שלב הגילוי** – בחלק זה ננתח את התמונה (של החיבור הסרוק), במטרה לזהות את המילים שכתב הנבחן. בשלב הגילוי נעזר בחלקים מפרויקט גמר נוסף אשר נעשה ע"י איתי חפץ ומורן זרגרי אשר חקרו את נושא זיהוי כתב יד והמרתו לטקסט ממוחשב. תהליך הגילוי מורכב משלושה תתי שלבים עיקריים: אולי להוסיף פה שגם ממירים את התמונה לקובץ נכון ואיחוד של הקבצים

* חיתוך התמונה - חיתוך קצוות התמונה במטרה ללכוד רק את הטקסט הנכתב על ידי הנבחן.
* גילוי שורות - מציאת השורות בהן מופיע הטקסט הכתוב של הנבחן
* גילוי המילים - הפרדת השורות למילים.

1. **שלב הזיהוי** – לאחר שגילינו היכן בתמונה נמצאות המילים, נסתפק בזיהוי רק של חלק מהמילים בחיבור ולא כולן. כלומר, בשלב זה נרצה לסנן את המילים אשר באמצעותן נערוך את תהליך האימות. בחרנו לפעול בדרך זאת על מנת שנוכל "לאמן מכונה" במספר מצומצם של מילים נבחרות, אשר בדקנו כי חוזרות בכל החיבורים. בכך נייעל את תהליך איסוף ה- data שישמש בלמידת המכונה.

בביצוע שלב זה נשתמש באלגוריתם למידה אשר יקבע האם המילה הנתונה שייכת לקבוצת המילים הנפוצות שהוגדרו או לא.

1. **שלב האימות** – אחרי שבררנו קבוצת מילים להשוואה מכל חיבור, נבצע השוואה בין המילים הזהות משני החיבורים (הקריטריונים להשוואה יקבעו בהמשך). כלומר, בשלב זה נפתח אלגוריתם אשר ייקח בחשבון את כלל ההשוואות שבוצעו בין המילים ויחזיר כפלט את רמת הביטחון ששני החיבורים נכתבו על ידי אותו אדם (באחוזים).



בחרנו בפרויקט זה לכתוב בשפת python הכוללת ספריות רבות של עיבוד תמונה ולמידת מכונה (כגון: OpenCV, sklearn, numpy ועוד).

עריכה של שחר: onTheFace

אנו חוקרים מספר אפשרויות לפתרון. אנו יוצאים מנקודת הנחה כי שילוב כל הדרכים יחד, יניב את התוצאה הנכונה ביותר לאימות זהותו של כותב המבחן.

הדרכים לפתרון אותן נחקור הם:

* ניסוח לאלגוריתם קוף
* השוואה בין אותיות זהות בין שני מועדי הבחינה.
* השוואה בין מילים זהות בין שני מועדי בחינה.

לכל אחת מדרכי הפתרון יש ארכיטקטורה דומה המורכבת משלושה שלבים מרכזיים):

**שלב הגילוי** – בחלק זה ננתח את התמונה (של החיבור הסרוק), במטרה לזהות את גורמי ההשוואה. בשלב הגילוי נעזר בחלקים מפרויקט גמר נוסף אשר נעשה ע"י איתי חפץ ומורן זרגרי אשר חקרו את נושא זיהוי כתב יד והמרתו לטקסט ממוחשב. תהליך הגילוי מורכב משלושה תתי שלבים עיקריים:

* עיבוד התמונה – איחוד קבצי הטקסט של החיבור לתמונה אחת. לאחר מכן, נבצע חיתוך של קצוות התמונה במטרה ללכוד רק את הטקסט הנכתב על ידי הנבחן.
* גילוי שורות - מציאת השורות בהן מופיע הטקסט הכתוב של הנבחן.
* גילוי המילים - הפרדת השורות למילים.(כדי לממש את דרך הפתרון של השוואת מילים זהות)

**שלב הזיהוי** – לאחר שגילינו היכן בתמונה נמצאים גורמי ההשוואה, ננסה לזהות עבור כל אלגוריתם פתרון את גורמי ההשוואה שלו: (על הפנים)

1. עבור אלגוריתם קוף – נבצע "מערך מונים" כך שעבור כל אות נבצע ספירה של כמות הזיהויים שלה זאת במטרה לבדוק את התפלגות האחוזים של האותיות שהאלגוריתם למידה שלנו מצליח לזהות עבור מחבר המאמר.
2. עבור השוואה בין אותיות זהות: ננסה לזהות כמה שיותר אותיות בכל אחד מהמאמרים.
3. השוואה בין מילים זהות : נסתפק בזיהוי רק של חלק מהמילים בחיבור ולא כולן. כלומר, בשלב זה נרצה לסנן את המילים אשר באמצעותן נערוך את תהליך האימות. בחרנו לפעול בדרך זאת על מנת שנוכל "לאמן מכונה" במספר מצומצם של מילים נבחרות, אשר בדקנו כי חוזרות בכל החיבורים. בכך נייעל את תהליך איסוף ה- data שישמש בלמידת המכונה.

בביצוע שלבים אלו נשתמש באלגוריתמי למידה אשר מטרתם היא לזהות את גורמי ההשוואה. למשל עבור השוואה בין מילים זהות, מטרת אלגוריתם הלמידה היא לקבוע האם המילה הנתונה שייכת לקבוצת המילים הנפוצות שהוגדרו או לא.

**שלב האימות** – אחרי שבררנו את גורמי ההשוואה בכל אחד מדרכי הפתרון, נבצע השוואה בין הגורמים משני החיבורים:

1. עבור אלגוריתם קוף – נבצע הפרש בין מערכי המונים של המאמרים השונים ואת התוצאה נעביר לאלגוריתם למידה במטרה לקבוע האם התפלגות ההפרשים תקינה עבור מחבר זהה.

2+3עבור השוואה בין מילים ואותיות.אחרי שבררנו קבוצת מילים/אותיות להשוואה מכל חיבור, נבצע השוואה בין המילים/האותיות הזהות משני החיבורים (הקריטריונים להשוואה יקבעו בהמשך). כלומר, בשלב זה נפתח אלגוריתם אשר ייקח בחשבון את כלל ההשוואות שבוצעו ויחזיר כפלט את רמת הביטחון ששני החיבורים נכתבו על ידי אותו אדם (באחוזים).

בשלב האימות נפתח אלגוריתם כללי אשר ייקח בחשבון את כלל דרכי הפתרון שבוצעו ויחזיר כפלט את רמת הביטחון ששני החיבורים נכתבו על ידי אותו אדם (באחוזים).



בחרנו בפרויקט זה לכתוב בשפת python הכוללת ספריות רבות של עיבוד תמונה ולמידת מכונה (כגון: OpenCV, sklearn, numpy ועוד).

סקירת עבודות דומות \ בספרות והשוואה \ סקר שוק

Writer verification based on a single handwriting word samples

המאמר עוסק בבעיית זיהוי של מחבר של טקסט בכתב יד. המאמר מתמקד ב**אימות** המחבר ומציע גישה חדשה לאימות בהתבסס על מילה אחת בלבד (ללא צורך בהרבה נתונים).

המאמר מתאר את הדמיון בין הבעיה הנתונה לבין זיהוי חתימה המשתמש ב- Levenshtein edit distance, וכן בפתרון המוצע יש שימוש ב- Wagner-Fisher algorithm.

אלגוריתם זה נותן הערכה לעלות השינוי של הפיכת תמונה של מילה מסוימת לתמונה של מילה נוספת (יש לציין כי שתי התמונות של מייצגות אותה המילה) באמצעות הערכת העלות השינוי בין האלמנטים הבסיסיים.

<https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-016-0139-0>

Matching Handwritten Document Images

המאמר עוסק בסכמה של פעולות כדי לבדוק התאמה בין שני כתבי יד מקבצים שונים.

ע"פ המאמר, הבדיקה מבוצעת בצורה המיטבית כאשר במסמכים יש מספר גדול ביותר של התאמות מילים. הזיהוי מילים דומות בעזרת רשת נוירונים אשר השתמשו בכמה מאגרי מידע גדולים כדי לאמן אותה. מתוארים בו הקריטריונים אשר שימשו לצורך החלטה האם מדובר באותו הכתב יד: צורת כתב היד, שכיחות המילים, סדר המילים ואוצר המילים.

במאמר מדגישים את הצורך במאגר נתונים גדול כדי לבצע את ההשוואה בצורה הטובה ביותר.

<https://arxiv.org/pdf/1605.05923.pdf>

Hybrid Feature Learning for Handwriting Verification

המאמר מתמקד בזיהוי מחבר על ידי מילה בודדת בשפה האנגלית – and .

במאמר מוסבר כי הבחירה במילה זאת נובעת מכך שהיא המילה הרביעית השכיחה ביותר באנגלית ומכך שלמילה זאת קיים מאגר מידע גדול אשר אפשר להשתמש בו כדי לאמן את האלגוריתם למידה. ההשוואה נעשית ע"י שילוב של רשת נוירונים ואלגוריתם SIFT לקביעת אחוזי ההתאמה בין 2 מילים.

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

נספחים

## **תכנון הפרויקט**

|  |  |
| --- | --- |
| 8.9.19 | פגישה ראשונה עם המנחה והצגת רעיון הפרויקט |
| 11.9.19 | פגישה עם המנחה והנציגים של המרכז הארצי לבחינות והערכה, חידוד הדרישות של הפרויקט. |
| 6.10.19 | פגישה עם המנחה, הצגת התקדמות והצבת מטרות נוספות לפגישה הבאה. |
| 6.11.19 | פגישה עם המנחה, הצגת התקדמות והצבת מטרות נוספות לפגישה הבאה. |
| 8.12.19 | הגשת דוח הצעה |
| 21.3.20 | הגשת לוז עבודה לסמסטר ב' |
| 26.4.20 | הגשת דו"ח בטא |
| 26.6.20 | הגשת דו"ח סופי |
| 5-7.7.20 | הצגת הפרויקט |
| 16.7.20 | הצגת פוסטר בכנס פוסטרים |
| 20.7.20 | העברה של הפרויקט |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **הסיכון** | **חומרה** | **מענה אפשרי** |
|  | אי עמידה בזמני ההגשות של פרויקט הגמר | בינוני | בניית לוח זמנים עבור משימות הפרויקט ומעקב אחר סיום המשימות בזמן |
|  | המרכז הארצי לא יעביר בזמן הדרוש את הנתונים הדרושים להתקדמות בפרויקט | בינוני | להקדים את בקשות הנתונים ככל שניתן. |
|  | השוואה בין שני מבחנים עם מילים שלא חוזרות על עצמם – חוסר נתונים להשוואה | גבוהה | הכנת קבוצת מילים גדולה להשוואה.  מציאת גורם השוואה נוסף.  בדיקת כיוון נוסף של השוואת אותיות. |
|  | שילוב טכנולוגיות חדשות בפרויקט הגמר שחברי הפרויקט לא עבדו איתם בעבר | בינונית | חברי הפרויקט יקצו חלק מהזמן ללמוד את הטכנולוגיות החדשות |
|  | קריאה לא נכונה או ניתוח שגוי של נתונים | בינונית | בדיקה נוספת של הנתונים לפני עריכת ההשוואה. |
|  | קושי בקביעת הקריטריונים להשוואה בין שתי מילים במבחנים שונים | גבוהה | קביעת פגישה עם גרפולוג מקצועי כדי לקבל קריטריונים ואפשרויות השוואה בין שתי מילים. |
|  | סריקת המבחנים תעשה באיכות ירודה או בפורמט לא תקין אשר לא יאפשר קריאה של המילים | בינונית | קביעת פורמט אחיד לקבצים הסרוקים. לעדכן את נציגת המכון הארצי בחשיבות איכות הסריקה |
|  | כתב יד לא מובן של הנבחן ,קשקושים על מחברת הבחינה אשר לא יאפשרו לבצע את הקריטריונים ההשוואה. | גבוהה | בדיקה של קבוצת מילים גדולה להשוואה.  השוואה לפי אותיות |
|  | מאגר מילים קטן מידי אשר לא יאפשר למכונה ללמוד את המילים בצורה טובה | בינונית | להתחיל לאסוף את המילים מבעוד מועד. בנוסף, לבצע שכפול מילים בזוויות שונות עם אלגוריתם מתאים. |

**טבלת דרישות (User Requirement Document)**

מכיוון שאנו מבצעים פרויקט מחקרי, כרגע עוד לא הוחלט על אופן הגשת המוצר (במידה ותהיה הצלחה) ואנו מתמקדים בפתרון הבעיה.

עם זאת הדרישות היחידות כרגע מהלקוח הן:

* שתהיה הצלחה ברמה סבירה בקביעה האם 2 מבחנים נכתבו ע"י אותו אדם או לא.
* תכנון הפרויקט כך שיהיה בצורה גנרית ככל שניתן כך שיהיה ניתן בהמשך לבצע התאמות לשפות נוספות כגון ערבית.