**המחלקה להנדסת תוכנה**

**פרויקט גמר – תשע"ט**

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> |

מבוא

\*\*האם המבוא צריך להיות על הארגון או על הפרויקט שאנו מבצעים?

מבחן הכניסה הפסיכומטרי לאוניברסיטאות הוא בחינה פסיכומטרית הנערכת בישראל החל משנת 1981 ומשמשת ככלי מיון לכניסה לאוניברסיטאות ולמכללות השונות. את הבחינה עורך המרכז הארצי לבחינות ולהערכה. ציון משוקלל המשלב את ציון הבחינה הפסיכומטרית של המועמד עם ציוני הבגרות שלו הוא המנבא הטוב ביותר שנמצא עד כה להצלחת המועמדת בשנת הלימודים הראשונה שלו באקדמיה, כלומר, בממוצע, ככל שציוני הפסיכומטרי והבגרויות של המועמד גבוהים יותר, כך הוא ייטה לקבל ציונים גבוהים יותר גם בלימודיו באוניברסיטה.

הבחינה מתקיימת בשפות עברית, ערבית, רוסית, צרפתית, ספרדית ובנוסח משולב של אנגלית ועברית.

הבחינה מתקיימת ארבע פעמים בשנה בשפות עברית וערבית.

\*ישנם כ-90,000 נבחנים בשנה ומתוכם כ 30,000 נבחנים אשר חוזרים על הבחינה. –לבדוק נתון זה.

הבחינה הפסיכומטרית בודקת את יכולות המועמד בשלושה תחומים: חשיבה מילולית, חשיבה כמותית ואנגלית. 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 שלו, ולכן נעזר בחלק זה בפרויקט כדי לנתח את כתב ידו של הנבחן במטרה לזהות זיוף במבחן הפסיכומטרי

תיאור הבעיה

המרכז הארצי לבחינות ולהערכה נוקט באמצעים ומאמצים רבים על מנת להבטיח את טוהר הבחינה ולמנוע רמאות מכל סוג ובפרט העתקות וזיופים. בנוסף אם נבחן ניגש לבחינה חוזרת ומשפר בצורה משמעותית את ציונו, הוא מוזמן לבחינה (חלקית) נוספת כדי להבטיח שהוא אכן ראוי לציון שקיבל.

למרות האמצעים אשר ננקטים כדי למנוע זיופים, עדיין ישנם אנשים אשר מצליחים להעלות את ציונם בבחינה באמצעות תשלום או בקשה מאדם זר שייגש למבחן במקומם בכדי לקבל ציון גבוה יותר.

בעקבות כך, המרכז הארצי לבחינות ולהערכה מקצה אנשים אשר מומחים לזיהוי כתב יד בכדי לנסות לבצע השוואה בין שני מועדי בחינה של נבחנים מסוימים אשר מוגדרים חשודים, על ידי בדיקה של חלק החיבור בבחינה. ההשוואה אמורה לאמת/להפריך חשד על נבחן שרימה בבחינה.

## הבעיה מבחינת הנדסת תוכנה

האתגר העיקרי 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.
* יש צורך באיסוף data (הרבה מילים כתובות בכתב יד בעברית) מקדים על מנת לבצע אימון למחשב לזהות מילים מסוימות אשר נפוצות במבחן הפסיכומטרי מכיוון שאין data זמין ברשת.
* אחת הדרישות של הלקוח היא שהפרויקט יעבוד בשפה העברית אך שיהיה ניתן לבצע התאמה גם לשפות נוספות כמו ערבית, כלומר הפרויקט צריך להתבצע בצורה גנרית כמה שיותר.
* כל בדיקה של שני טקסטים היא בדיקה אינדיבידואלית, כלומר, כמות המידע (הטקסט) שיש ברשותנו על כל נבחן מוגבלת ויכולה להשפיע את תוצאות הבדיקה.

תיאור הפתרון

לצורך מציאת הזיוף במבחן הפסיכומטרי, אנו נשווה בין 2 מבחנים שונים של אותו אדם במטרה לבדוק האם בשני המבחנים קיים כתב יד שבוצע על ידי אותו בנאדם.

הארכיטקטורה מתארת את מבנה הבדיקה:

התוכנית תקבל כקלט 2 מבחנים ותחזיר את הסיכוי ששני המבחנים נכתבו ע"י אותו אדם.

אופן הבדיקה: ( נראה לי נוריד את החלק הזה)

* עבור כל מבחן, נבצע חיתוך של כל השורות.
* נבצע חיתוך של כל המילים הנמצאות בכל שורה
* עבור כל מילה, ננסה לזהות האם מדובר במילה אשר נמצאת במאגר המילים להשוואה.
* עבור כל המילים הנמצאות במאגר ההשוואה, נערוך השוואה בין שני מאגרי המילים ע"י זיהוי אופן כתיבת כל מילה ונחזיר את תוצאת ההשוואה. התוכנית תחזיר את ממוצע התוצאות של כלל ההשוואות שנעשו.

על מנת לבצע זיהוי של כתב יד שונה בין 2 מועדים של אותו נבחן, נדרש תהליך המורכב משלושה חלקים מרכזיים שאיתם אנו נתמודד בפרויקט זה.

1.שלב הגילוי – בחלק זה ננתח את התמונה המכילה את החיבור, 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:

* 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.

2.שלב הזיהוי – לאחר שגילינו היכן בתמונה נמצאים המילים אשר נכתבו על ידי הנבחן, נרצה לברור רק את המילים הנפוצות ביותר בקטעי החיבור. כלומר, בשלב זה נרצה לסנן את המילים אשר באמצעותם נערוך את תהליך האימות. על מנת לבצע את שלב זה, נשתמש בעזרת אלגוריתם למידה אשר תחליט האם המילה שגילינו שייכת לקבוצת המילים הנפוצות או לא.

3.שלב האימות – אחרי שבררנו קבוצת מילים להשוואה מכל חיבור, נערוך השוואה בין שני מאגרי המילים על ידי זיהוי אופן כתיבת כל מילה ונחזיר את תוצאות ההשוואה. כלומר, בעזרת אלגוריתם למידה ועיבוד תמונה, נערוך השוואה בין המילים הדומות במטרה לזהות האם כל מילה מהמבחן הראשון נכתבת באופן זהה למילה הנכתבת במבחן השני.

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 הדימיון 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. מתואר הקריטריונים אשר שומשו 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 דומה על ידי מילה בודדת בשפה האנגלית – and .

במאמר מוסבר כי הבחירה במילה זאת נובעת מכך שהיא המילה ה4 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** |
| 1 | יציאה של אחד מחברי הפרויקט למילואים לתקופה ארוכה | בינונית | הגשת בקשה למתן פטור מהמילואים עקב פרויקט הגמר. |
| 2 | אחד מחברי הפרויקט מתחיל לעבוד במשרה מלאה | גבוהה | חלוקה מחדש של משימות הפרויקט וייעול זמנים. |
| 3 | אי עמידה בזמני ההגשות של פרויקט הגמר | בינוני | בניית לוח זמנים עבור משימות הפרויקט ומעקב אחר סיום המשימות בזמן |
| 4 | המרכז הארצי לא יעביר בזמן הדרוש את הנתונים הדרושים להתקדמות בפרויקט | גבוהה | להקדים את בקשות הנתונים ככל שניתן. |
| 5 | השוואה בין שני מבחנים עם מילים שלא חוזרות על עצמם – חוסר נתונים להשוואה | גבוהה | הכנת קבוצת מילים גדולה להשוואה.  מציאת גורם השוואה נוסף. |
| 6 | שילוב טכנולוגיות חדשות בפרויקט הגמר שחברי הפרויקט לא עבדו איתם בעבר | בינונית | חברי הפרויקט יקצו חלק מהזמן ללמוד את הטכנולוגיות החדשות |
| 7 | קריאה לא נכונה או ניתוח שגוי של נתונים | בינונית | בדיקה נוספת של הנתונים לפני עריכת ההשוואה. |
| 8 | קושי בקביעת הקריטריונים להשוואה בין שתי מילים במבחנים שונים | גבוהה | קביעת פגישה עם גרפולוג מקצועי כדי לקבל קריטריונים ואפשרויות השוואה בין שתי מילים. |
| 9 | 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 |
| 10 | כתב יד לא מובן של הנבחן ,קשקושים על מחברת הבחינה אשר לא יאפשרו לבצע את קריטריוני ההשוואה. | בינונית | ביצוע חיתוך השוליים לפני בדיקת המילים, בדיקה של קבוצת מילים גדולה להשוואה. |
| 11 | 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. |

## **רשימת\טבלת דרישות**

## Use-case

## UML

## **רשימת\טבלת דרישות**

**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. הדרישה היחידה כרגע היא שתהיה הצלחה בקביעה האם 2 מבחנים נכתבו ע"י אותו אדם או לא. דרישות נוספות ותכנון המוצר יתווספו בהמשך.