**המחלקה להנדסת תוכנה**

**פרויקט גמר - תש"פ**

אימות מחבר המבוסס על ניתוח כתב יד

Author verification based on handwritten text analysis

**מאת**

**דניאל גבאי - 311223598**

**שחר ישראלי - 203713094**

**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> |
| 3 | סרטון גרסת אלפא | <https://drive.google.com/file/d/1gAy8cAt62CRi6B0vDs2cZ9trAWreiXq9/view> |

נאום המעלית

בפרויקט זה נבצע מחקר למציאת אלגוריתם שבעזרתו, נפתח מערכת ממוחשבת שתקבל שני טקסטים סרוקים הכתובים בכתב יד, 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).

בשלב הראשון הפרויקט יותאם לשפה העברית אך ניתן יהיה 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 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, והיא 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 בכלל, והעתקות וזיופים בפרט. למרות האמצעים אשר ננקטים כדי למנוע זיופים, עדיין ישנם מקרים של התחזות, בהם מצליחים מועמדים לקבל ציון גבוה בבחינה באמצעות אדם אחר אשר מבצע את הבחינה במקומם.

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. 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 שכתב 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 מחקרי זה הינו ביוזמה של המרכז הארצי לבחינות והערכה, בהנחייתו של דר' יהודה חסין. מתוך ההנחה כי קשה להכריע ששני כתבי יד זהים, נרצה לבנות תוכנית שמקבלת כקלט שני חיבורים סרוקים של אותו נבחן (במועדים שונים) ומטרתה לאמוד את הסיכוי שהחיבורים נכתבו על ידי שני אנשים שונים.

תיאור הבעיה

מידי שנה ניגשים כ- 70,000 נבחנים לבחינה הפסיכומטרית, כשליש מתוכם ניגשים 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 שתייעל את תהליך הבדיקה.

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.

## הבעיה מבחינת הנדסת תוכנה

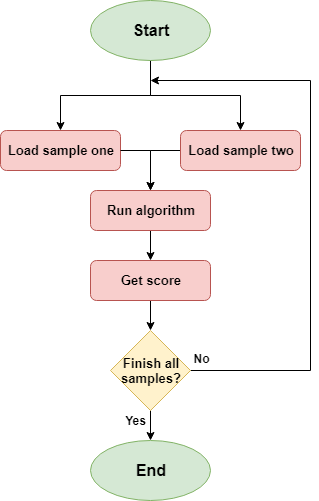
על מנת לבצע השוואה בין שני כתבי יד ואימות המחבר, נדרש תהליך המורכב משלושה חלקים מרכזיים איתם נתמודד בפרויקט: **גילוי, זיהוי ואימות**, עליהם נפרט בתיאור הפתרון.

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 עמו 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 קבועים, דבר המקשה בגילוי המילים והאותיות אשר נשתמש בהם להשוואה בין כתבי היד.
* כל בדיקה של שני טקסטים היא בדיקה אינדיבידואלית. כמות המידע (הטקסט) שיש ברשותנו על כל נבחן מוגבלת ויכולה להשפיע על תוצאות הבדיקה.
* "המרכז הארצי לבחינות והערכה" דורש כי הפרויקט יתמקד תחילה בחיבורים הכתובים בשפה העברית, אך שיהיה ניתן לבצע הרחבה לחיבורים בשפות נוספות כמו השפה הערבית. כלומר עלינו לבצע את הפרויקט בצורה גנרית ככל שניתן כדי להתאימו לשפות נוספות כך שעיקר העבודה תהיה באיסוף ה- data של השפה הדרושה.
* מבדיקה ראשונית שביצענו בספרות עולה כי הבעיה איתה אנו מתמודדים מורכבת והניסיונות לפתור אותה הניבו תוצאות חלקיות בלבד.

תיאור הפתרון

בתמונה - High level design

הפתרון הוא יצירת תוכנת מחשב, שתקבל כקלט מהמשתמש שני קבצים של כתבי יד בעברית. עם קבלת פקודה, התוכנה תבצע ניתוח והשוואה של כתבי היד, ותחזיר כפלט את הסיכוי שהחיבורים נכתבו על ידי אותו אדם.

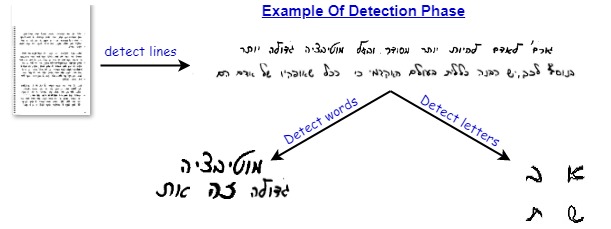
בפרויקט זה נשתמש בחלק מהאלגוריתמים שמומשו ע"י הסטודנטים מורן זרגרי ואיתי חפץ בפרויקט גמר משנה שעברה שבוצע בהנחייתו של דר' יהודה חסין, אשר חקרו את נושא זיהוי כתב יד והמרתו לטקסט ממוחשב. מטרת הפרויקט הייתה לתת למשתמש כלי להמרה של טקסטים בכתב יד לטקסט ממוחשב, על מנת לאפשר לו את היכולת לסדר ולשתף את המסמכים השונים וכן לבצע שינויים ומניפולציות על הטקסט. בפרויקט זה תוצאות זיהוי הטקסט היו חלקיות בלבד ולא מספיק בכדי להמיר את כל הטקסט מכתב יד לכדי כתב ממוחשב.

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 מילים ואותיות שישמשו אותנו כגורמי השוואה.

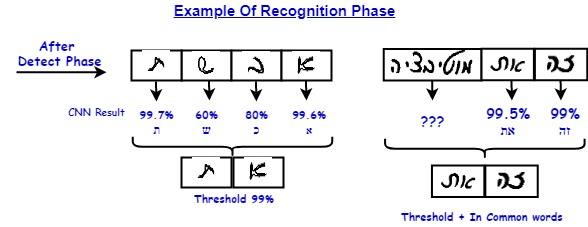
פירוט השלבים:

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 עמוד אחד או שניים, לכן בחרנו לאחד את העמודים לתמונה אחת. לאחר מכן, נבצע חיתוך של קצוות התמונה במטרה ללכוד רק את הטקסט הנכתב על ידי הנבחן. בנוסף, יש צורך ביישור התמונה והפחתת "רעש" המקשה על זיהוי הטקסט.
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 מילים – הפרדת השורות למילים.
5. גילוי אותיות – הפרדת המילים לאותיות.

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



1. **שלב הזיהוי (recognition)** – לאחר גילוי מיקום המילים והאותיות בטקסט, נרצה לזהות אילו מילים ואותיות גילינו:
2. זיהוי אותיות: לאחר שמצאנו את מיקומי השורות, נפעיל מודל (רשת נוירונים) שאומן לזהות את 27 האותיות בשפה העברית (מהפרויקט של שנה שעברה). נרצה לשמור לשלב הבא רק אותיות שהאלגוריתם זיהה בהסתברות גבוהה (כ- 99%). רמת הבטחון בזיהוי האות חשובה לנו כיוון שבשלב ההשוואה נרצה להשוות בין אותיות **זהות בלבד**.
3. זיהוי מילים: אנו נסתפק בזיהוי 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" לזהות מספר מצומצם של מילים נבחרות, אשר בדקנו כי חוזרות על עצמן בסבירות הכי גבוהה בקרב החיבורים. בכך נייעל את תהליך איסוף ה- data שישמש אותנו באימון רשת נוירונים. [ראה הסבר על אופן בחירת המילים בנספחים]



לפני שנתאר את שלב האימות, חשוב לציין מה הסיבות שהקשו על השגת מלוא המטרות של הפרויקט משנה שעברה:

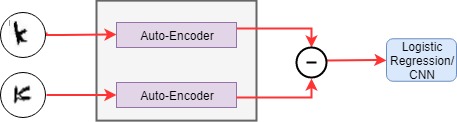
1. האלגוריתם לגילוי השורות התקשה לבצע הפרדה מדויקת של טקסט בעל שורות עם נטייה אלכסונית חדה ולא אחידה.
2. אלגוריתמי הפרדת השורות למילים ולאותיות התקשו להפריד כתבי יד מחוברים, צפופים ומקושקשים ולכן חיתוך האותיות והמילים לא תמיד היה במיקום מדויק.
3. אחוז הדיוק של רשת הנוירונים היה יחסית גבוה. עם זאת, ישנן אותיות מסוימות שהרשת לא הצליחה לזהות באחוזים גבוהים ולכן זה לא היה מספיק לצרכי המערכת כדי לספק כלי אמין.

כעת נשאלת השאלה מדוע בכל זאת החלטנו להשתמש בחלקים מהפרויקט הקודם לטובת הפרויקט שלנו? ובכן, הסיבה המרכזית היא שמטרת הפרויקט שלנו שונה ממטרת הפרויקט הקודם. למערכת שלנו אין צורך לזהות ב- 100% את כל האותיות והמילים על מנת לבצע את ההשוואה בין שני הטקסטים. מספיק שנזהה בביטחון גבוה פחות מ- 10% מהאותיות והמילים שבטקסט (ואת השאר "נזרוק") כדי להיות מסוגלים לבצע השוואה (ואכן המערכת שלהם מספקת לנו יכולת אפילו טובה מזו). את הנחה זו ביססנו על המאמר Hybrid Feature Learning for Handwriting Verification [ראה סקר שוק] שהציג תוצאות גבוהות באימות המחבר ע"י **2 דגימות בלבד** של המילה 'and' בשפה האנגלית ואנו נממש חלק מהארכיטקטורה שהוצגה בו.

3. **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 בשלוש דרכים עיקריות: השוואת מילים, השוואת אותיות ואלגוריתם 'קוף' (שלנו).

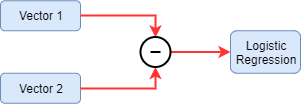
1. השוואת מילים ואותיות – משלב הגילוי והזיהוי, נשארנו עם גורמי ההשוואה משני החיבורים. כעת נרצה לבצע השוואה בין אותן אותיות ומילים משני החיבורים ('א' מול 'א', 'של' מול 'של' וכו'). כלומר, מכל חיבור נשאיר רק את האותיות והמילים שהצלחנו לזהות בשני החיבורים. את ההשוואה נבצע באופן הבא:

נשתמש באלגוריתם Auto-Encoder (AE) שהוא אלגוריתם שעושה compress ו- decompress ל- data. באמצעות האלגוריתם ניתן לחלץ מכל דגימה (=מילה/אות) את הפיצ'רים הכי חשובים שמאפיינים אותה. לדוגמא, עבור אות בודדת אנחנו שומרים תמונה בגודל 28\*28 פיקסלים (סה"כ 784 פיצ'רים). מתוך 784 הפיצ'רים האלגוריתם מחלץ את 32 או 64 הפיצ'רים החשובים ביותר(נבדוק מה יניב תוצאה יותר טובה), ולאחר שנחלץ משני הדגימות את הפיצ'רים, נבצע ביניהם בפונקציית חיסור ואת התוצאה נעביר לאלגוריתם סיווג נוסף (Logistic Regression/CNN) שאומן מראש לזהות האם תוצאת החיסור מאפיינת דגימות של אותו מחבר או מחברים שונים. יש לציין כי עבור כל אות וכל מילה נצטרך לאמן AE שונה לחלץ את הפיצ'רים החשובים (זאת נעשה באמצעות ה-data שיש ברשותנו), וכן נצטרך לאמן את אלגוריתם הסיווג לזהות האם וקטור הפיצ'רים שנותר לאחר פעולת החיסור מתאר דגימות שנכתבו על ידי אותו אדם או לא.



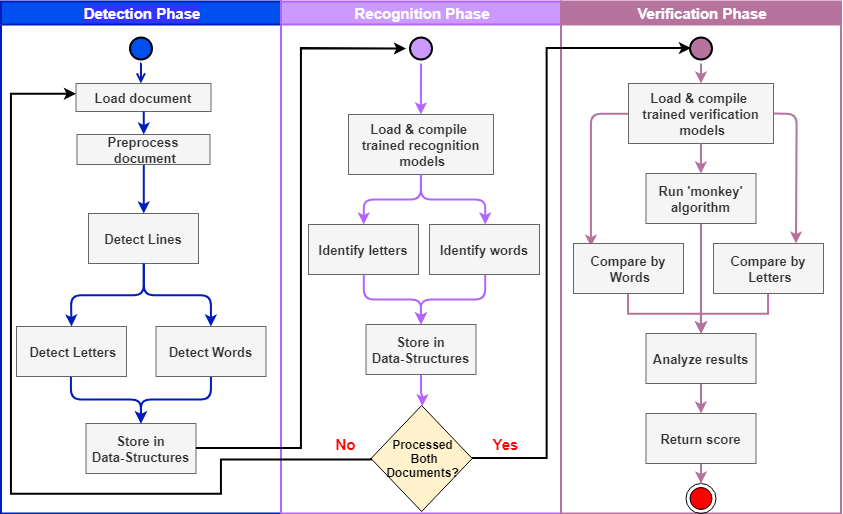
1. אלגוריתם קוף – במהלך המחקר, לאחר שהרצנו מספר בדיקות של זיהוי האותיות, גילינו כי המודל שבו אנו משתמשים מצליח לזהות בהסתברות גבוהה אותיות שונות עבור אנשים שונים. מה הכוונה? שמנו לב כי למשל עבור נבחן א' נזהה בדיוק רב את האותיות א', ח', ע', ש', ועבור נבחן ב' נזהה בדיוק רב אותיות אחרות כמו ב', ג', ח', ה', ק'. מתוך תוצאות אלו, הסקנו כי ניתן לנצל את 'חוסר הדיוק' של המודל. כלומר, אם עבור כתב יד של נבחן א' זיהינו בדיוק רב סט כלשהו של אותיות (ואותיות אחרות לא זיהינו), הרי שהאלגוריתם שלנו ידע לזהות את סט אותיות אלו גם בחיבור אחר שנכתב על ידי אותו אדם. אם בחיבור הנוסף האותיות שזוהו אינן תאומות, או תואמות חלקית, הרי שזה יכול להוות אינדיקציה על כך שהחיבורים לא נכתבו על ידי אותו אדם. קראנו לאלגוריתם זה אלגוריתם 'קוף' מכיוון שאנו מנצלים את חוסר היכולת של המודל לזהות בדיוק של 100% את כל האותיות. כלומר, לא נזדקק 'לעבוד קשה' ולבצע השוואה אמיתית בין הגורמים השונים בכדי לקבל אינדיקציה מקדימה האם מדובר באותו אדם או לא.

תיאור האלגוריתם: עבור כל חיבור, נגדיר וקטור בגודל 27 (כמספר האותיות בשפה העברית). בכל תא בוקטור נשמור את כמות המופעים (באחוזים) מכל אות שהמודל הצליח לזהות (תא 0 האות א' 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 וחיבורים שנכתבו על ידי אנשים שונים. תיאור סכמתי של פעולת האלגוריתם:



## ארכיטקטורת המערכת

התרשים הבא מתאר את ארכיטקטורת המערכת על כל שלביה:

****

## 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 למובנית מאליה.

סקירת ספרות

Hybrid Feature Learning for Handwriting Verification

המאמר העיקרי עליו אנו מבססים את ארכיטקטורת השוואת המילים והאותיות (שלב האימות). 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, Auto-Encoder ואלגוריתם SIFT (שנכון להיום בחרנו עוד לא לממש לצרכים שלנו) לקביעת אחוזי ההתאמה בין 2 מילים.

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

Writer verification based on a single handwriting word samples

המאמר עוסק בבעיית זיהוי של מחבר של טקסט בכתב יד. המאמר מתמקד באימות 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. זיהוי המילים הדומות נעשה בעזרת רשת נוירונים וכמה 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>

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 מצד אחד ואתגר גדול מצד שני. 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, קראנו מידע על זיופי חתימות, למדנו את הטכנולוגיות איתן נעבוד (python, עיבוד תמונה, machine learning) ובמקביל ביצענו התאמות מהפרויקט שבוצע בשנה שעברה ע"י מורן זרגרי ואיתי חפץ לפרויקט שלנו. לאחר שהתחלנו לחשוב על דרכים אפשריות לפתור את הבעיה, התחלנו לאסוף data של המילים הנפוצות (לאחר שבחרנו אותן כפי שמתואר בנספחים) ובנוסף בדקנו מספר אלגוריתמים נוספים עבור שלב הגילוי.

מכיוון שעלתה הדרישה מהמרכז הארצי לבחינות והערכה שהפרויקט יבוצע בצורה גנרית כדי להתאימו לשפות נוספות, ביצענו חלוקה של הקוד למודולים שונים בעלי תחומי אחריות מוגדרת, השתדלנו לבצע refactor לאורך הזמן ולשמור על הקוד נקי ויציב.

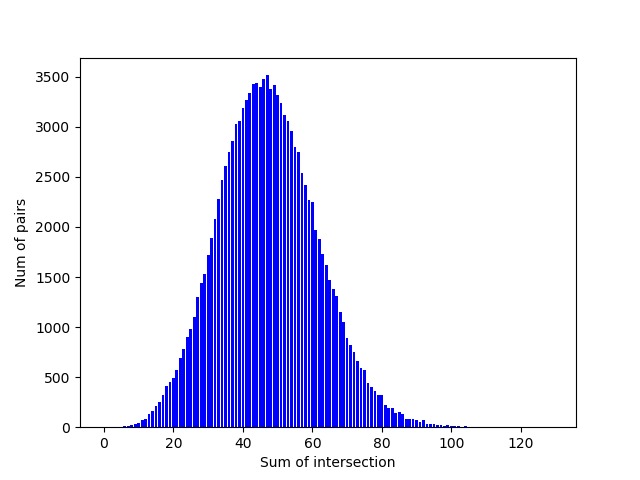
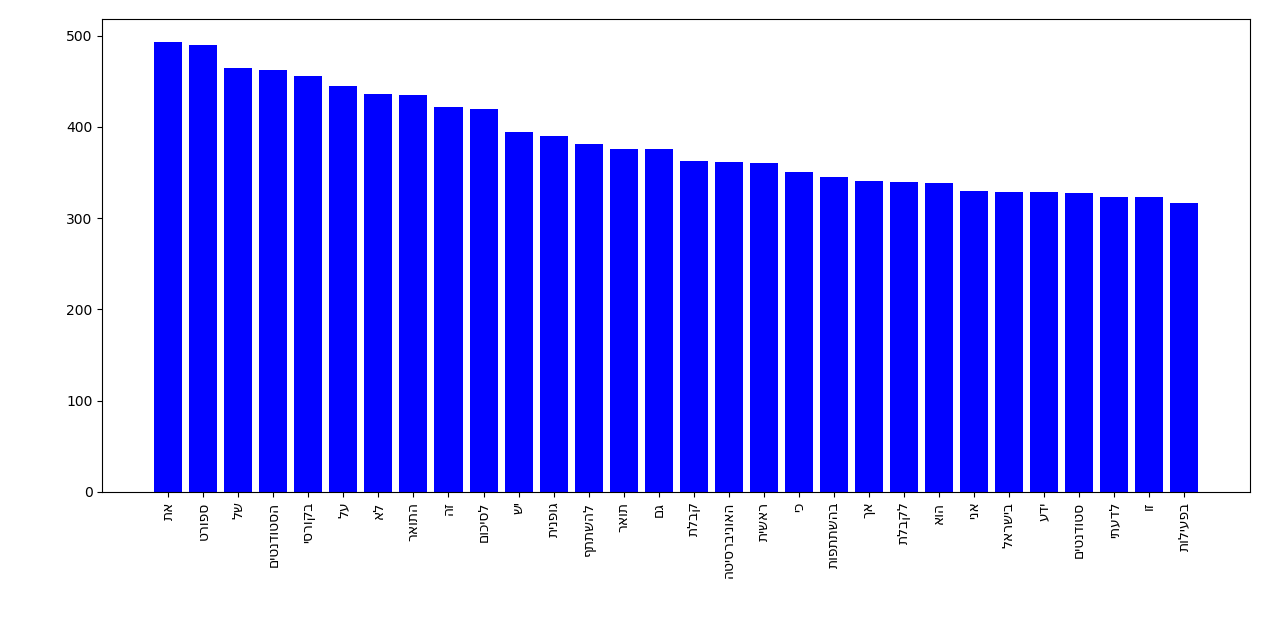
נכון להיום סיימנו עם שלב הגילוי ושלב הזיהוי עבור אותיות (נשאר עוד לבצע עבור המילים). בסמסטר הבא נותר לנו להמשיך את לב האלגוריתם שלנו שהוא שלב האימות, כלומר ההשוואה בין המילים/האותיות. בנוסף, ברשותנו 500 חיבורים של נבחנים אשר הועברו אלינו ע"י המרכז הארצי לבחינות והערכה. בעזרת החיבורים ביצענו את המימוש של אלגוריתם קוף. לצורך הלמידה חילקנו את החיבורים לשני חלקים ובכך קיבלנו "שני" חיבורים של אותו נבחן (כל 500 החיבורים שייכים לנבחנים שונים). בסה"כ ביצענו למידה על 500 זוגות של חיבורים של אותו נבחן ו- 500 זוגות של חיבורים בעלי מחבר שונה (סה"כ 1000 דגימות). הלמידה נעשתה על כ- 80% מסך החיבורים כאשר 20% השארנו לצורך בדיקה. כרגע, האלגוריתם מצליח לזהות האם החיבורים שייכים לאותו אדם או לא ב- 85% דיוק. אנו עובדים על שיפור אחוזי הדיוק של אלגוריתם קוף ועלינו להמשיך ולבדוק אילו תוצאות הוא עשוי להניב בהמשך הפרויקט.

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

נספחים

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

**המשך תיאור הפתרון – אופן בחירת המילים הנפוצות**

בכדי לבחור את המילים הנפוצות, קיבלנו מהמרכז הארצי לבחינות והערכה כ-500 חיבורים מוקלדים, ובעזרת תוכנית שכתבנו חילצנו את המילים שחוזרות הכי הרבה בכל החיבורים. יש לציין כי החיבורים שקיבלנו הם מאותו נושא בחינה, לכן מילים כמו 'אוניברסיטה', 'ספורט', 'סטודנטים' הן מילים ששייכות לנושא החיבור ולכן לא בחרנו אותן כמילים הנפוצות שנבחר לעבוד איתן. בהתבסס על התוצאות, המילים שנבחרו הן: של, לא, את, גם, לסיכום, כי ,זה, זו ,יש ,לדעתי, אני.

לאחר מכן, כדי לוודא שאכן יש 'חיתוך' בין טקסטים שונים לבין המילים המופיעות בטקסט (כלומר לוודא שנוכל לבצע השוואה בין מילים שוות בין שני טקסטים), כתבנו תוכנית נוספת שמבצעת את הבדיקה הבאה: מכל טקסט חילצנו את המילים הנפוצות שמופיעות בו (המילים שבחרנו מהפסקה הקודמת), ביצענו חיתוך עם כל שאר הטקסטים וספרנו את מספר המילים השוות. קיבלנו את הגרף הבא:

ביצענו כ-  *השוואות (כל הזוגות האפשריים עבור 500 החיבורים) וקיבלנו כי בממוצע יש חיתוך של כ- 45 מילים.*

בכדי לאמן מכונה לזהות את המילים שבחרנו, אספנו כ-300 דפים של מילים בכתב יד (3600 מופעים מכל מילה). לאחר שנבצע אימון לרשת נוירונים, כל מילה שהפרדנו בשלב הגילוי נעביר למודל בכדי לבדוק האם היא חלק מהמילים הנפוצות. את המילים שזיהינו נשמור במבנה נתונים לשלב האימות.

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | **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** |
| 7 | אי עמידה בזמני ההגשות של פרויקט הגמר | שחר ודניאל | אי הגשה של המשימות, הורדה בציון. | 15% | VH |  | בניית לוח זמנים עבור משימות הפרויקט ומעקב אחר סיום המשימות בזמן. |
| 8 | המרכז הארצי לא יעביר בזמן הדרוש את הנתונים הדרושים להתקדמות בפרויקט | שחר ודנאל | עיכוב בהתקדמות הפרויקט | 15% | H |  | להקדים את בקשות הנתונים ככל שניתן. |
| 4 | שילוב טכנולוגיות חדשות בפרויקט הגמר שחברי הפרויקט לא עבדו איתם בעבר | שחר ודניאל | עיכוב בהתקדמות הפרויקט | 40% | H |  | הקצאה של חלק מהזמן ללמוד את הטכנולוגיות החדשות. |
| 3 | חתונה של דניאל בפברואר | דניאל | פחות זמן עבודה על הפרויקט בזמן הארגונים לחתונה | 60% | M |  | חלוקת משימות ועבודה יותר חזקה בתקופה שלפני. |
| 1 | תקופת מבחנים + פרויקטים אחרים | שחר ודניאל | פחות זמן עבודה על הפרויקט והתמקדות במבחנים/פרויקטים | 60% | H |  | חלוקת משימות ועבודה יותר אינטנסיבית בתקופה שלפני. |
| 6 | התחלת עבודה חדשה | שחר | קושי בחלוקת הזמנים בין הפרויקט, העבודה והקורסים בסמסטר. | 30% | M |  | קביעת לוח זמנים מסודר, בנוסף יש לשחר קורס אחד פחות בסמסטר ב' ויוכל להקדיש יותר זמן לפרויקט |
| 5 | כתב יד לא מובן של הנבחן ,קשקושים על מחברת הבחינה אשר לא יאפשרו לבצע את הקריטריונים ההשוואה. | שחר ודניאל | קושי בניתוח החיבור וקביעה חדש משמעית לגבי אימות הנבחן | 50% | M |  | אלגוריתם 'קוף' כפי שתואר יוכל אולי לטפל בבעיה זו. |
| 2 | קושי בקביעת קריטריונים להשוואה בין החיבורים | שחר ודניאל | עיכוב בפרויקט, חוסר הצלחה באימות הנבחן | 50% | VH |  | נקיטה בכמה דרכים לפתרון ששילובם יניב את התוצאה הטובה ביותר |

**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:

* המערכת תקבל החלטה בדיוק רב ככל שניתן אם המבחנים נכתבו ע"י אותו אדם או לא.
* תכנון הפרויקט בצורה גנרית ככל שניתן כך שיהיה ניתן בהמשך לבצע התאמות לשפות נוספות כגון ערבית.

**תכנון הפרויקט**

ניתן לגשת ליומן בקישור שבעמוד הראשון

**ביבליוגרפיה**

זרגרי מורן, איתי חפץ (2019). Handwriting detection recognition.

<https://github.com/moranzargari/Handwriting-detection-recognition>

**Building Autoencoders in Keras**

By: Francois Chollet

Publication Date: 14 May 2016

**Hybrid Feature Learning for Handwriting Verification**

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

Publication Date: 19 November, 2018

**National institute for testing & evaluation**

<https://www.nite.org.il/>

**Matching Handwritten Document Images**

By: Praveen Krishnan, C.V Jawahar

Publication Date: 19 May, 2016

**Writer verification based on a single handwriting word samples**

By: Ameur Bensefia, Thierry Paquet

Publication Date: 2016