בס"ד

Azrieli College of Engineering Jerusalem

עזריאלי מכללה אקדמית להנדסה ירושלים

המחלקה להנדסת תוכנה

מציאת משפטים, מילים ואותיות בכתב יד של השפה הערבית

Finding handwritten sentences, words and letters of the Arabic language

מאת

ליאל לוי - 207045741

אביגייל הילה שרבף - 318631488

תמוז, תשפ"א **יולי, 2021**

בס"ד

Azrieli College of Engineering Jerusalem

עזריאלי מכללה אקדמית להנדסה ירושלים

המחלקה להנדסת תוכנה

מציאת משפטים, מילים ואותיות בכתב יד של השפה הערבית

Finding handwritten sentences, words and letters of the Arabic language

חיבור זה מהווה חלק מהדרישות לקבלת

תואר ראשון בהנדסה

מאת

ליאל לוי - 207045741

אביגייל הילה שרבף - 318631488

מנחה אקדמי: דר' יהודה חסין אישור: תאריך:

רכז הפרויקטים: דר' אסף שפנייר אישור: תאריך:

Azrieli College of Engineering Jerusalem
עזריאלי מכללה אקדמית להנדסה ירושלים

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
4	סרטון גרסה סופית	https://youtu.be/m17Y1jtMpsQ

מידע נוסף

סוג הפרויקט	מחקרי ממרצה במכללה
פרויקט מח"ר	לא
פרויקט ממשיך	פרויקט חדש שבחלקו מתבסס 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.

הצהרה:
העבודה נעשתה בהנחיית ד"ר יהודה חסין, עזריאלי המכללה האקדמית

להנדסה ירושלים - המחלקה להנדסת תוכנה,
החיבור מציג את עבודתנו האישית ומהווה חלק מהדרישות לקבל

תואר ראשן בהנדסה

[redacted]

תוכן עניינים

Azrieli
College of Engineering
Jerusalem
עזריאלי
מכללה אקדמית להנדסה
ירושלים

נאום המעלית

בפרויקט זה ביצענו מחקר למציאת אלגוריתם שבעזרתו נפתח מערכת ממוחשבת שתקבל שני חיבורים סרוקים הכתובים בכתב יד בערבית, ותחזיר כפלט את הסיכוי (באחוזים) שהחיבורים 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).

תקציר

הבחינה הפסיכומטרית 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) בה נדרש הנבחן לכתוב חיבור באורך של 25-50 שורות הנכתבת בעיפרון והיא המטלה היחידה במבחן בה נדרש הנבחן לכתוב בכתב ידו.
המטלה מהווה 10% מציון כלל הבחינה ובה הפרויקט שלנו מתמקד.
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.

jce.ac.il | *9087
עזריאלי מכללה אקדמית להנדסה, יעקב שרייבום 26, רמת בית הכרם, ירושלים

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.

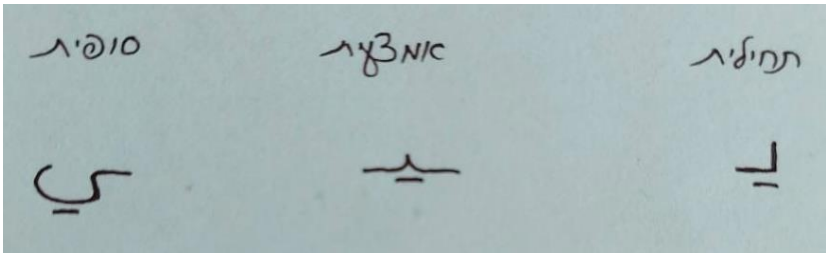
הבעיה מבחינת הנדסת תוכנה

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 לגלות אותיות בעזרת עיבוד תמונה ולאחר מכן לבצע השוואה בין אותן אותיות.
2. בניגוד לשפה העברית בה כל 26 האותיות נכתבות בנפרד אחת מהשנייה, בשפה הערבית ישנן רק 8 אותיות מתוך ה-28 שנכתבות לא במחובר, בנוסף לכך גם האותיות שנכתבות בנפרד נכתבות כך רק במיקום ספציפי במילה (למשל רק בתחילת המילה או בסופה).
3. ישנו קושי בגילוי המילים והאותיות בכתב יד שבניגוד לכתב מודפס המרווחים בין האותיות, המילים והשורות אינם קבועים.
4. הכתב בערבית מחובר ובנוסף כתב יד משתנה מאדם לאדם ולכן כל בדיקה של שני חיבורים תצליח למצוא כמות שונה של אותיות מופרדות.
   דבר זה יכול להשפיע על תוצאות הבדיקה, על כמה וכמה בשפה הערבית שבה יש מספר מאוד מוגבל של אותיות שניתן להשתמש בהן לצורך השוואה.

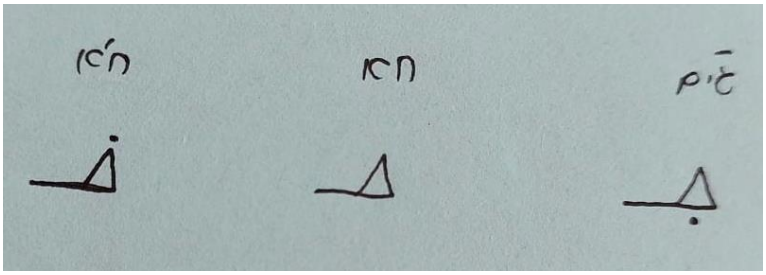
5. בניגוד לשפה העברית שבה כל אות נכתבת באותה צורה ללא תלות במיקומה במילה, בשפה הערבית כל אות יכולה להיכתב בשלוש צורות בהתאם למיקומה במילה (התחלה, אמצע וסוף). [דוגמא באיור 1]

איור 1 – האות יא בצורותיה השונות בהתאם למיקומה במילה



6. בניגוד לשפה העברית שבה הניקוד אינו חלק מהאות ולכן אין לו חשיבות בזיהוי האות, 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. [דוגמא באיור 2]

איור 2 – 3 אותיות שונות הנכתבות בצורה זהה מלבד ניקודן

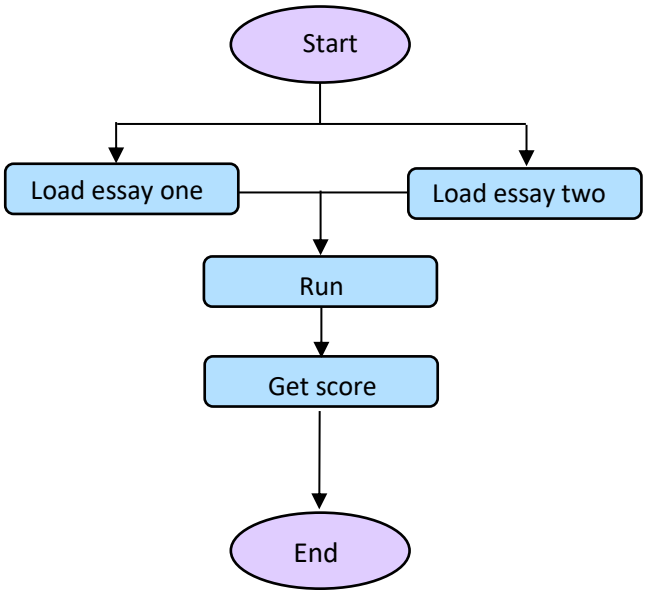


תיאור פתרון

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.

ביצוע תהליך הזיהוי של כתב היד מהתמונה הופרד לשני שלבים מרכזיים: גילוי (detection) וזיהוי (recognition) שלב זה הוא החלק המרכזי והמורכב בפרויקט. יתר על כן, הפרויקט שלנו מצריך שלב נוסף: שלב האימות - ביצוע ההשוואה בין שני החיבורים ואימות המחבר. בכדי שנוכל לבצע את ההשוואה, נרצה משלב הגילוי והזיהוי לחלץ אותיות שישמשו אותנו כגורמי השוואה.

תרשים 3 - High level design



פירוט השלבים

1. **גילוי (detection)** – שלב זה מורכב מניתוח התמונה ומציאה של החלקים המכילים את הכתב אותו יש לפענח.

נציין שבשלב זה נעשה שימוש בקודים שנכתבו בפרויקט של שנה שעברה עם כוונונים והתאמות לחיבורים שלנו.

תהליך הגילוי מורכב ממספר תתי שלבים עיקריים:

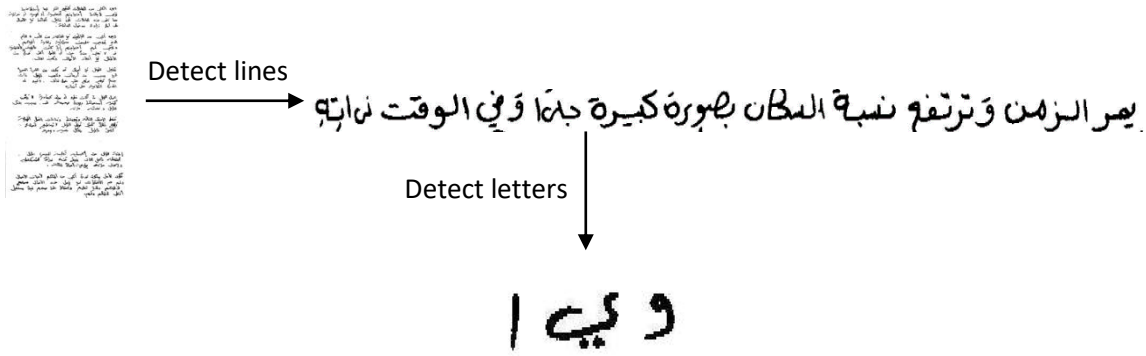
א. עיבוד מקדים - את החיבורים אנו מקבלים בפורמט tiff מהמרכז הארצי לבחינות והערכה. כל חיבור מכיל שני עמודים. בחרנו לאחד את שני העמודים לתמונה אחת על מנת שיהיה לנו עמוד אחד שיכלול את כל החיבור. לאחר מכן, נבצע חיתוך של קצוות התמונה במטרה ללכוד רק את הטקסט הנכתב על ידי הנבחן.

ב. 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).

ג. 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 – 4 איור



2. **שלב הזיהוי (recognition)** - נרצה לזהות אילו אותיות גילינו בשלב הגילוי.

לאחר שמצאנו את מיקומי השורות וגילינו את מיקום האותיות בטקסט, נפעיל מודל [2] (רשת נוירונים).

בשלב הראשוני בחרנו לקחת מהאינטרנט מודל של רשת נוירונים מוכן שכתוב בספריית TensorFlow שמזהה את 28 האותיות בערבית ואיתו מגיע dataset [1] של אותיות המחולק ל-train [3] ו-test [5].

אחוז הדיוק שרשת הנוירונים נתנה לנו בעת הבחינה [7] על ה-test הוא 93.21%.

על מנת לבצע בחינה של רשת הנוירונים על אותיות מהחיבורים שקיבלנו, ביצענו את שלב הגילוי בו קיבלנו את האותיות מכל חיבור. לאחר מכן התבצע סינון מקדים על ידנו שבו מחקנו פסיקים, צמדי אותיות מחוברות ולכלוכים שהיו בתמונה והופרדו כאותיות. לאחר סינון מקדים זה, העברנו את האותיות שנשארו לסינון נוסף אצל אדם דובר השפה הערבית, ובסינון זה הוא הוציא את חמשת האותיות על מנת שנוכל לבחון את מודל רשת הנוירונים עליהן.

לאחר בדיקה עם אדם דובר ערבית נוסף קיבלנו את הנתון שישנן רק 8 אותיות מתוך ה-28 שנכתבות במחובר רק לאות הנמצאת מימין להן ואין אותיות שנכתבות לא במחובר משני הצדדים שלהן.

כתוצאה מנתון זה הסקנו שלא נוכל לאמן ולבחון על כל 28 האותיות בשפה הערבית אלא לבדוק מי מבין 8 האותיות הנ"ל מופיעות בתדירות מספיק גבוהה על מנת שנוכל לבצע השוואות בין חיבורים שונים ולקבל תוצאה כמה שיותר אמינה. בנוסף, בכתב יד האותיות אף פעם לא נכתבות באותה צורה בדיוק ולכן נרצה שיהיו כמה שיותר אופציות להשוואה.

ביצענו בדיקה ומצאנו שישנן 5 אותיות מתוך ה-8 שמופיעות בתדירות מספיק גבוהה כדי שנוכל להשתמש בהן. בעקבות מסקנה זו בנינו dataset של אותיות שקיבלנו מהחיבורים המורכב רק מ-5 אותיות אלו. ביצענו אימון [6] של חמשת האותיות מה-dataset מהאינטרנט ובחנו על ה-dataset שהרכבנו מחמשת האותיות מהחיבורים שלנו ועדיין אחוז הדיוק היה נמוך (42.35).

בעקבות התייעצות עם המנחה ובדיקה שאחוז הדיוק אכן משתפר אם מוסיפים אימון על אותיות מהחיבורים שקיבלנו, הגענו למסקנה שהתוצאה הנמוכה של אחוז הדיוק נובעת מכך שה-dataset מהאינטרנט מהונדס ולא מספיק מתאים לאימון עבור תמונות האותיות מהחיבורים שקיבלנו.

לאחר בניית ה-dataset מהחיבורים שקיבלנו ובחינה על חמשת האותיות שבחרנו הסקנו מתוצאות ה-confusion matrix [8] שהמודל מתבלבל בין 3 מהאותיות, 2 מהן נכתבות באופן זהה ומה שמייחד כל אחת מהן הוא רק הניקוד של האות (נקודה מעל או מתחת לאות) והאות השלישית נכתבת בצורה מאוד דומה. [הצגת האותיות המדוברות באיור 5]

איור 5 – similar letters

לכן, החלטנו לוותר על שתי אותיות מתוך השלוש ולהמשיך להשתמש באות Nun שנתנה לנו את אחוז הדיוק הטוב ביותר מבין שלושת האותיות על מנת שאחוז הדיוק ישתפר והמודל יטעה פחות.

החלטנו להוסיף שני צמדי אותיות המופיעים בתדירות מספיק גבוהה כך שנוכל להשתמש בהם מכיוון שלאחר הסינון נשארנו עם שלוש אותיות בלבד. [הצגת צמדי המילים המדוברות באיור 6]

איור 6 – Letter pairs

Lam-Alif **Mim-Alif**

לאחר שהחלטנו מהן האותיות וצמדי האותיות שאיתם נעבוד בנינו dataset המורכב מ:

1. שלושת האותיות הבודדות שבחרנו מה-dataset מהאינטרנט.
2. שלושת האותיות הבודדות ובנוסף שני צמדי האותיות מה-dataset מהחיבורים שקיבלנו.

ביצענו אימון על ה-dataset שבנינו ובחנו על קבוצת test המורכבת מ-85 אותיות וקיבלנו אחוז דיוק של 68%. בשלב הסופי הוספנו ערך סף של 0.6 בבחינה אשר נותן את ההסתברות ממנה ניקח אות ולא נוותר עליה ואחוז הדיוק הסופי על קבוצת ה-test שהגענו אליו הוא 81.25%.

Confusion matrix with threshold 0.6

		predict				
		mim	**nun**	**Lam-alif**	**wow**	**Mim-alif**
real	**Mim**	12				
	Nun		11			
	Lam-alif			2	2	
	Wow				18	
	Mim-alif					

Confusion matrix without threshold

		predict				
		mim	**nun**	**Lam-alif**	**wow**	**Mim-alif**
real	**Mim**	9		4	7	
	Nun			19	1	
	Lam-alif			20	2	
	Wow			12	7	
	Mim-alif			5		

לאחר הבדיקות נדרשנו לשמור את מודל רשת הנוירונים שהגיע לתוצאות הכי טובות אך נתקלנו בקשיים בטעינתו. בעקבות התייעצות עם דר' אסף שפנייר ודר' יהודה חסין הגענו למסקנה כי עדיף להמיר את הרשת לספריית PyTorch מהסיבה שהיא מובנת וקלה למימוש.

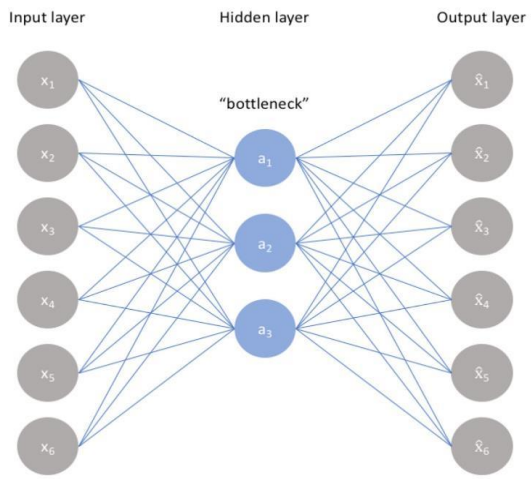
3. **שלב האימות (verification)** – שלב זה הוא השלב המרכזי בו מתקבלת ההחלטה האם שני החיבורים נכתבו על ידי אותו אדם או לא. החלטנו לפעול בשתי דרכים עיקריות: אלגוריתם Auto-Encoder ואלגוריתם 'קוף' (שלנו).

א. אלגוריתם Auto-Encoder– מודל למידה לא מפוקח (unsupervised learning) המבוסס על רשת נוירונים, כדי לבצע representation learning.

המטרה העיקרית של שימוש בסוג כזה של רשת נוירונים היא כיווץ קלט מסוים שממנו מתקבל מספר קטן יותר של פיצ'רים שבעזרתם ניתן יהיה לשחזר את הקלט המקורי.

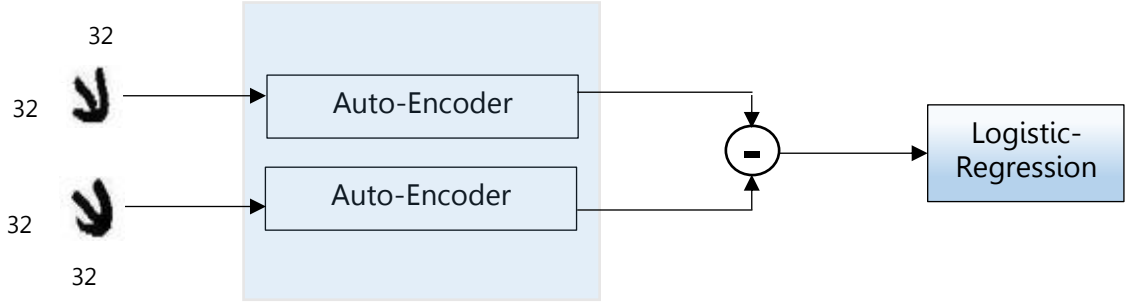
הרשת יוצרת "צוואר בקבוק" [ראה איור 7] שבאמצעותו מבוצע כיווץ של הקלט המקורי לכדי מספר קטן של פיצ'רים. תהליך זה יכול להתבצע עבור קלטים בהם יש קשר בין הפיצ'רים השונים (כלומר יש תלויות, מבנה כלשהו) אחרת זו הייתה משימה כמעט בלתי אפשרית.

הפיצ'רים שנוצרים בשכבה של צוואר הבקבוק הם הפיצ'רים החשובים של הקלט המאפיינים אותו. במקרה שלנו הקלטים הינם תמונות של אותיות בכתב יד בשפה הערבית.



איור 7 – Auto-Encoder

משלב הגילוי והזיהוי, נשארנו עם גורמי ההשוואה משני החיבורים. כעת נרצה לבצע השוואה בין אותן אותיות משני החיבורים (מים מול מים 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 (כלומר "חתימה" של האות) הכי חשובים שמאפיינים אותה שיוצגו בווקטור בגודל 32. הרעיון הוא שעבור 2 אותיות שנכתבו על ידי אותם אדם ווקטור הפיצ'רים יהיה דומה. מתוך הפיצ'רים שחילצנו האלגוריתם יבחר את הפיצ'רים החשובים ביותר ולאחר שנחלץ משתי הדגימות את הפיצ'רים, נבצע ביניהם פונקציית חיסור בערך מוחלט ואת התוצאה נעביר לאלגוריתם סיווג נוסף (Logistic Regression/CNN) שאומן מראש לזהות האם תוצאת החיסור מאפיינת דגימות של אותו מחבר או מחברים שונים. מודל הסיווג Logistic-Regression מחזיר לנו את האחוז שהחיבורים נכתבו על ידי אותו מחבר.



איור 8 – Auto-Encoder high level

הרצנו את האלגוריתם על 10 חיבורים שנכתבו על ידי אותו אדם ועל 10 חיבורים שנכתבו על ידי אנשים שונים. קיבלנו שהאלגוריתם הצליח לחזות ב-60% מהמקרים האם אותו אדבר כתב את שני החיבורים או לא.

ב. <u>אלגוריתם קוף</u> –

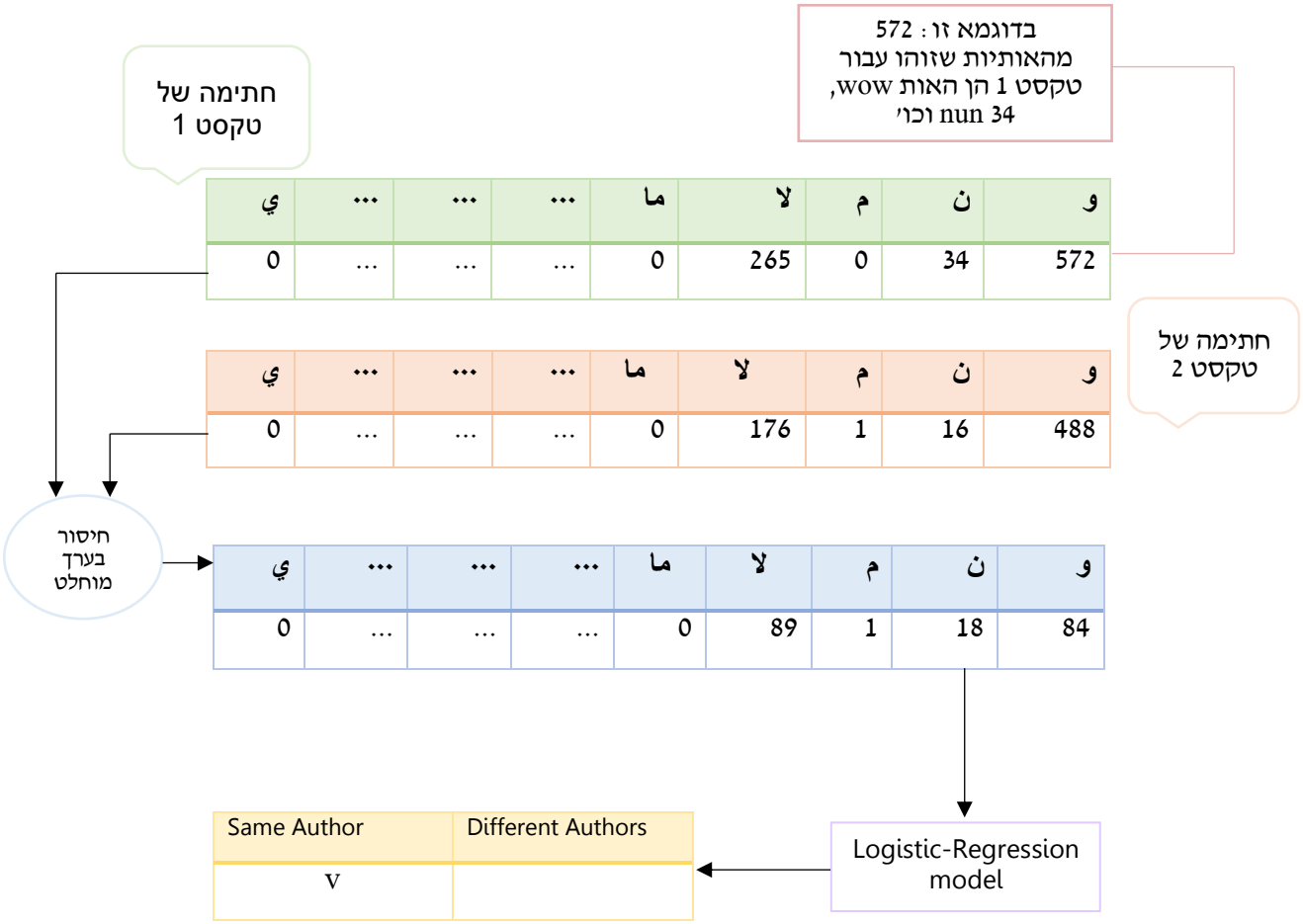
בחלק מסוים בפרויקט נדרשנו להשתמש ברעיון של אלגוריתם קוף שמומש ע"י הסטודנטים דניאל גבאי ושחר ישראלי בפרויקט גמר משנה שעברה שבוצע בהנחייתו של דר' יהודה חסין, אשר בנו מערכת להשוואת שני חיבורים בכתב יד בשפה העברית.

הרעיון של אלגוריתם קוף בנוי על כמות האותיות שרשת הנוירונים מצליחה לזהות בכל חיבור.
לדוגמא: בחיבור אחד מצאנו 10 מופעים של האות "מים" ובחיבור שני מצאנו רק 2 מופעים של אותה האות. נתון זה יכול להוות אינדיקציה על כך שהחיבורים לא נכתבו על ידי אותו אדם הרי שאם החיבורים כן היו נכתבים על ידי אותו אדם היינו מוצאות

מספר מופעים דומה לאותה אות בשני החיבורים. אנו מגדירות את כמות המופעים מכל אות כ"חתימה" של הנבחן.

[דוגמא בתרשים 9]

Example of Monkey algorithm – 9 תרשים



האלגוריתם מתבצע באופן הבא:

בשלב הראשוני חילקנו את 38 החיבורים שקיבלנו לקבוצות Train, [4] validation ו-test בצורה מסודרת כלומר, לא באופן רנדומלי.

כל האותיות שהוצאו מכל חיבור עברו ברשת הנוירונים בהתחלה ללא נרמול [9] ולאחר התייעצות עם המנחה מה ניתן לעשות על מנת להעלות את אחוז הדיוק החלטנו להעביר את תמונות האותיות נרמול וזה אכן שיפר את אחוז הדיוק.

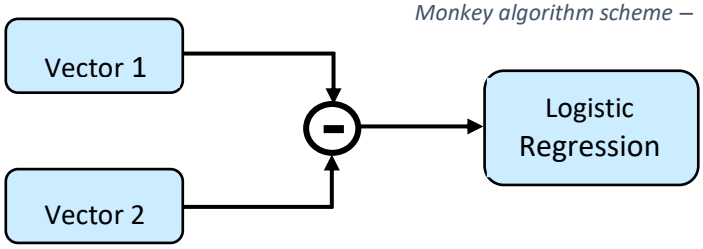
עבור כל חיבור, הגדרנו וקטור בגודל 30 (כמספר האותיות בשפה הערבית + שני צמדי האותיות שהוספנו). בחרנו להשאיר את הווקטור בגודל 30 ולא בגודל 5 (כמספר האותיות

שאנו משתמשות בהן) כדי להשאיר את הקוד גנרי במידה ובעתיד ירצו להוסיף אותיות נוספות לאימון ובחינת המודל.

בכל תא בווקטור שמרנו את כמות המופעים מכל אות שהמודל הצליח לזהות (תא 0 האות אליף וכו'). לאחר מכן, ביצענו חיסור בערך מוחלט בין שני הווקטורים משני החיבורים. את וקטור החיסור העברנו למודל סיווג נוסף (Logistic Regression). המודל אומן לזהות ווקטורי חיסור מקבוצת ה- train של חיבורים שנכתבו על ידי אותו אדם, וווקטורי חיסור של חיבורים שנכתבו על ידי אנשים שונים.

תיאור סכמתי של פעולת האלגוריתם:

תרשים 10 – Monkey algorithm scheme



לאחר האימון, ביצענו בחינה על קבוצת ה-validation שהחליטה האם אותו אדם כתב את שני החיבורים באחוז דיוק של 80% אך כשבחנו את המודל על קבוצת ה-test אחוז הדיוק היה נמוך מאוד (מתחת ל-50%).

לאחר בדיקות נוספות הגענו למסקנה שכאשר אחוז הדיוק גבוה על קבוצת ה-validation הוא בהכרח יהיה נמוך על קבוצת ה-test (מתחת ל-50%) ולהיפך. בעקבות חקירת הנושא הסקנו שכדי שנוכל להחליט האם אותו אדם כתב את שני החיבורים או לא באחוז דיוק גבוה יותר צריך לחלק את החיבורים באופן רנדומלי לקבוצות אלה. לאחר שביצענו את החלוקה בשיטה זו קיבלנו שהמערכת מצליחה לחזות האם אותו אדם כתב את שני החיבורים באחוז דיוק של 78% על ה-Test.

ביצענו אימון על 34 זוגות חיבורים שלא שייכים לאותו אדם ו-18 זוגות השייכים לאותו אדם.

ביצענו בחינה על 18 זוגות חיבורים שלא שייכים לאותו אדם ו-10 זוגות השייכים לאותו אדם ואת תוצאותיה ניתן לראות בטבלה הבאה:

Confusion matrix test

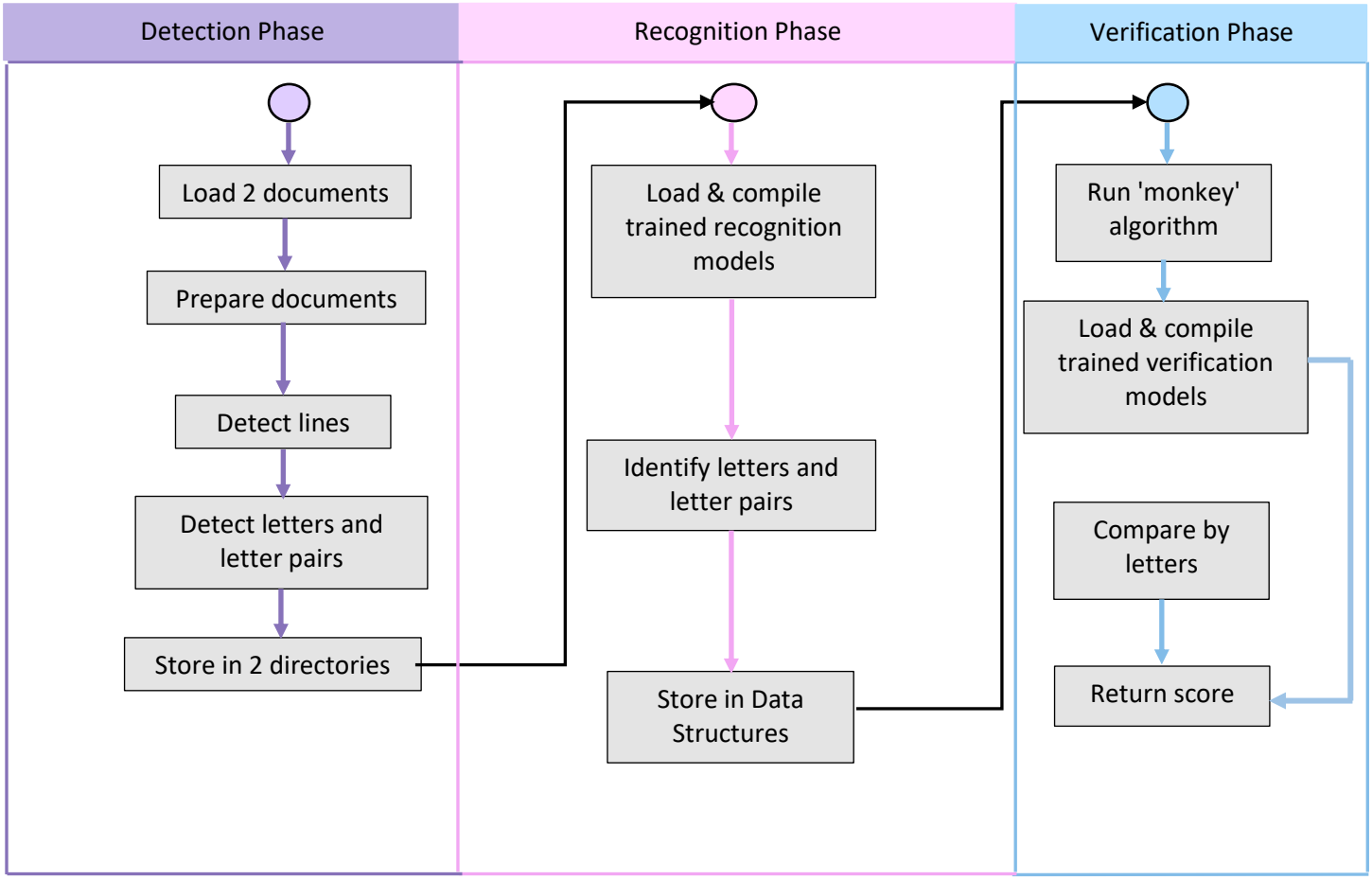
	Predicted Different	Predicted Same
Actual Different	16	2
Actual Same	4	6

בנוסף, ביצענו בחינה על החיבורים שהמודל אומן עליהם ואת תוצאותיה ניתן לראות בטבלה הבאה:

Confusion matrix train		
	Predicted Different	**Predicted Same**
Actual Different	28	6
Actual Same	2	16

ארכיטקטורת המערכת

התרשים הבא מתאר את ארכיטקטורת המערכת על כל שלביה:



טכנולוגיות

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

תוצאות

- הצלחנו לזהות בעזרת רשת הנוירונים באחוז דיוק של 81.25% את חמשת האותיות וצמדי האותיות שאיתם עבדנו (mim, nun, wow, lam-alif, mim-alif).
- בעזרת אלגוריתם קוף הצלחנו לחזות ב-78% מהמקרים האם שני חיבורים נכתבו על ידי אותו אדם או לא.

מסקנות

בפגישתנו הראשונה עם דר׳ יהודה חסין, הוצגו לנו מספר פרויקטים אפשריים. בסופו של דבר, החלטנו לבחור בפרויקט השוואת כתבי היד כיוון שנושא זה גרם לנו לעניין רב מכיוון שלא התעסקנו בתחום זה ובנוסף זהו אתגר גדול עבורנו. ממחקר ראשוני שביצענו בספרות, גילינו כי הבעיה איתה אנו מתמודדות מורכבת מאוד, קיימות מספר דרכים שונות לתקוף אותה ואין תשובה חד משמעית כיצד לפעול. בנוסף בנימה אישית, היינו מעט פסימיות לגבי היכולת שלנו לפתור בעיה מורכבת זו עם אחוז דיוק גבוה בגלל השפה המורכבת והזמן המוגבל.
בחלקו הראשון של הפרויקט ביצענו מחקר מעמיק, קראנו מאמרים, למדנו את הטכנולוגיות איתן נעבוד (python, image processing, machine learning) ובמקביל ביצענו התאמות לחלק הקוד אשר מכין את החיבורים ומבצע את חיתוך השורות והאותיות מקוד הפרויקט שבוצע בשנה שעברה ע״י דניאל גבאי ושחר ישראלי לפרויקט שלנו.

לאורך הפרויקט ביצענו חלוקה של הקוד למודולים שונים בעלי תחומי אחריות מוגדרת, השתדלנו לבצע refactor לאורך הזמן ולשמור על הקוד נקי ויציב.
אתגר מרכזי שליווה אותנו לאורך הפרויקט היה מחסור בדאטה לצורך אימון המודלים. לכמות ה-data (כלומר זוגות של חיבורים השייכים לאותו אדם) בה נשתמש לאימון המודלים יש מרכיב מרכזי בהצלחת הפרויקט מכיוון, שככל שהמודל לומד יותר כך הוא יצליח לחזות בצורה אמינה יותר על data שהוא לא אומן עליו. בתחילת הפרויקט קיבלנו ממאל״ו 38 זוגות חיבורים השייכים לאנשים

שונים (כך שכל זוג של חיבורים שייך לאותו אדם) ובכל שלבי הפרויקט חיבורים אלו היו היחידים שברשותנו. לכן, התוצאות אליהן הגענו היו נמוכות ממה שציפינו.

לאחר שסיימנו לבנות את המערכת, ביצענו ניסוי על קבוצת ה-test כפי שתואר בשלב 3 בחלק תיאור השלבים. בסופו של דבר לאחר שיפורים וכוונונים של אלגוריתם קוף הצלחנו להגיע לכ-78% הצלחה בדיוק סיווג החיבורים, תוצאה אשר נמוכה ממה שציפינו בתחילת הפרויקט.
בנוסף, מימשנו את האלגוריתם Auto-Encoder ולאחר בדיקה של 10 חיבורים של אותו אדם ו-10 של אנשים שונים האלגוריתם הצליח לחזות בכ-60% האם אותו אדם כתב את שני החיבורים או לא.
מפאת חוסר הזמן, לא הספקנו לשלב את תוצאות אלגוריתם קוף עם תוצאות האלגוריתם Auto-Encoder אך לפי אחוזי הדיוק של שני האלגוריתמים אנו משערות שאחוז המשותף יהיה גבוה.
בנוסף, אנו סבורות כי ניתן לשפר עוד את התוצאות ע"י אימון המודלים באמצעות כמות גדולה יותר של data וכן ע"י אימון מודלים נוספים וכמות גדולה יותר של אותיות שונות שכאמור במסגרת זמן הפרויקט לא הצלחנו לממש.

בנוסף, נציין כי בעקבות כך שהשפה הערבית איננה שפת האם שלנו זמן רב מהפרויקט הוקדש להבנת חוקי השפה הערבית וכללי האיות שלה. כמו כן, הזדקקנו לאדם דובר השפה שיוכל לעזור לנו בזיהוי אותיות והפרדתן לקבוצות רלוונטיות מתוך כלל הפלט של קוד הפרדת השורות לאותיות.
מציאת אדם זה וזמן עבודתו עיכבו את התקדמותנו בפרויקט.

תודות

תודה למנחה הפרויקט דר' יהודה חסין, שהיה לצידנו לאורך כל הפרויקט, הקדיש הרבה מזמנו לפגישות, עזרה ותמיכה והיה חלק ניכר מהצלחתנו. יהודה הדריך והכווין אותנו רבות במחקר, באלגוריתמים השונים ובתכנון הפרויקט. כמו כן, ניתח איתנו את התוצאות וייעץ לנו בנוגע לדרכים לשיפורן.
תודה לדר' אסף שפיינר על הקדשת הזמן בהכוונה בהמרת רשת הנוירונים לספריית PyTorch.
הערכה ותודה, לדר' פרחאת עוסמאן על הקדשת הזמן הרב בזיהוי אותיות והפרדתן לקבוצות רלוונטיות, על העצות המועילות במהלך הפרויקט, על התמיכה והעידוד.
תודה לדר' איהב אנצארי על הענקת העזרה בהבנת איפיון האותיות בתחילת מילה, במרכזה ובסופה.

מילון מונחים, סימנים וקיצורים

[1] Dataset - סט של מידע שבפרויקט זה מכיל תמונות של אותיות בערבית המחולק לקבוצות train, validation ו-test.

[2] מודל – מערכת תיאורטית פשוטה הדומה בתכונותיה למערכת מורכבת ומשמשת לבחינת רעיונות טכנולוגיים.

[3] Train set – תת קבוצה של ה-dataset שנועדה לאימון מודל.

[4] Validation set – תת קבוצה של ה-dataset שנועדה לבחינת ביניים של המודל.

[5] Test set – תת קבוצה של ה-dataset שנועדה לבחינה הסופית של המודל.

[6] אימון – פעולה שמתבצעת על המודל על מנת ללמד אותו dataset מסוים.

[7] בחינה – פעולה שמתבצעת על המודל על מנת לבחון אותו על מה שהוא אומן.

[8] Confusion matrix - פריסת טבלה המאפשרת הדמיה של ביצועי אלגוריתם.

כל שורה של המטריצה מייצגת את התוצאה בפועל ואילו כל עמודה מייצגת את התוצאה החזויה על ידי האלגוריתם.

הסבר למבנה של confusion matrix:

	Predicted Same	**Predicted Different**
Actual Same	**TP = True Positive**	**FN = False Negative**
Actual Different	**FP = False positive**	**TN = True Negative**

למעשה האלכסון שבצבע כחול מתאר את כמות הבדיקות בהן המערכת צדקה ובאדום את כמות הבדיקות בהן המערכת שגתה. חישוב Model Accuracy נעשה באופן הבא:

$$\frac{TP+TN}{TP+TN+FN+FP}$$

או במילים פשוטות, כמות הבדיקות שבהן המערכת צדקה לחלק לסה"כ הבדיקות.

[9] נרמול – פעולה שמתבצעת על ה-data שקובעת סטיית תקן וממוצע קבועים עבור כל הדאטה באופן אחיד.

רשימת ספרות

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

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

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.

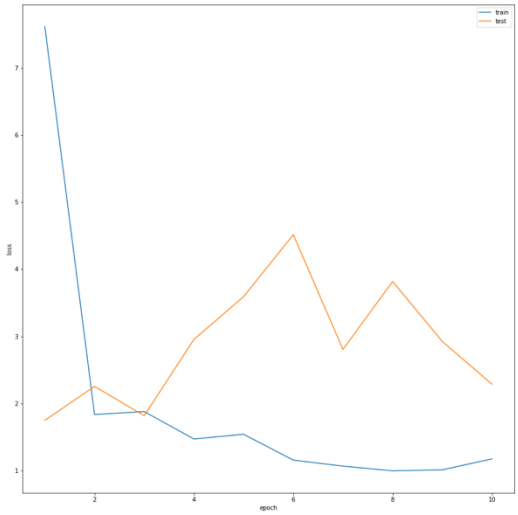
המאמר מדבר על הסוגייה והמורכבות הטכנית של זיהוי כתב יד בשפה הערבית וסוקר את השיטות שהוצעו ומומשו עד היום בנושא זה.
בנוסף, מדובר באופן ממוקד על זיהוי אותיות בתוך מסמך סרוק הרשום בכתב יד בשפה הערבית.
המאמר מתאר את תהליך הזיהוי על ידי 5 שלבים אשר בכל שלב מוצעים פתרונות שונים שמממשים את מטרת השלב.

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

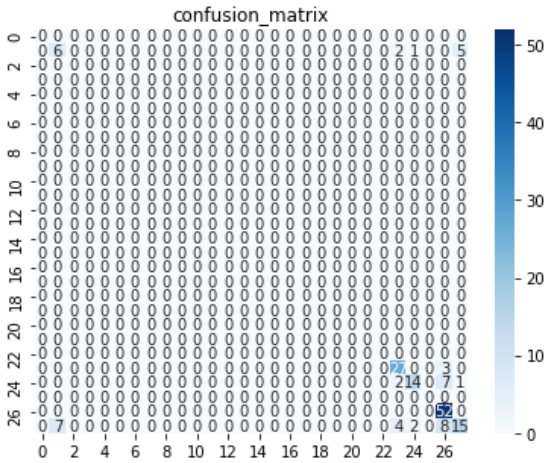
https://www.sciencedirect.com/science/article/pii/S2314717217300156#fig0020

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



גרף המתאר את ערכי ה-loss של רשת הנוירונים עבור ה-train וה-test.
ניתן לראות שערך ה-loss במגמת ירידה כפי שציפינו כלומר, השגיאה יורדת ורשת הנוירונים מצליחה ללמוד.



Confusion matrix עבור חמשת האותיות הראשוניות שבחרנו.
ניתן לראות שהרשת מתבלבלת בין המחלקות 1, 24, 27.
השורות מייצגות את המחלקה האמיתית שאליה שייכת האות והעמודות מייצגות את המחלקה שרשת הנוירונים חזתה.

Azrieli College of Engineering Jerusalem
עזריאלי מכללה אקדמית להנדסה ירושלים

טבלת סיכונים

				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.

בס"ד

Azrieli College of Engineering Jerusalem

עזריאלי מכללה אקדמית להנדסה ירושלים

Software Engineering Department

Finding handwritten sentences, words and letters of the Arabic language

by

Avigail Hila Sharbaf – 318631488

Liel Levy - 207045741

Academic Supervisor:

Dr Yehuda Hassin

בס"ד

Azrieli College of Engineering Jerusalem

עזריאלי מכללה אקדמית להנדסה ירושלים

Software Engineering Department

Finding handwritten sentences, words and letters of the Arabic language

by

Avigail Hila Sharbaf – 318631488

Liel Levy - 207045741

July 2021 (civil date) **Tamuz 5781 (Hebrew date)**