האוניברסיטה הפתוחה

המחלקה למתמטיקה ולמדעי המחשב

החטיבה למדעי המחשב

פרויקט מתקדם במדעי המחשב

פיתוח תוכנה להגהת ספר תורה

מנחה: ד״ר מיריי אביגל

מגיש:

מאיר גוטמן

ת.ז. 029519246

פברואר 2018

[הקלד טקסט]

תוכן עניינים

1 **תקציר** **1**

2 **מבוא** **1**

2.1 רקע מציאותי 1
2.2 רקע טכנולוגי 3
2.3 מבנה העבודה 5

3 **שמירת נתוני ספר התורה** **5**

3.1 נתוני ספר תורה 5
3.2 שמירת הנתונים בקבצים 5
3.3 שמירת הנתונים בזיכרון התוכנית 6
3.4 עיבוד תמונה לצורך זיהוי תווים 7
3.4.1 איכות התמונה חשובה 8
3.5 פעולות עיבוד התמונה 8
3.5.1 מציאת הסף (Threshold) של הצבע 8
3.5.2 יישור התמונה 10

4 **חילוץ התווים** **12**

4.1 אתגרים בחילוץ תווים 12
4.2 שלבי חילוץ התווים 12
4.2.1 מציאת השורות 12
4.2.2 מציאת הרווחים 13
4.2.3 מציאת התווים 15

5 **למידת התווים** **19**

5.1 תהליך זיהוי התווים 19
5.2 מעבר לתמונה בת השוואה 20
5.2.1 יצירת תמונת תו מרופדת 21
5.2.2 יצירת תמונה סופית בת השוואה 21
5.3 פעולת רשת הנוירונים 23

6 **הגהה** **24**

6.1 זיהוי תווים נוספים או חסרים 24
6.2 זיהוי רווחים 25
6.3 זיהוי התווים 25
6.4 איסוף הסטטיסטיקה 25

7 **תוצאות** **26**

7.1 תוצאות הגהת עמודה 26
7.1.1 תוצאות הגהת עמודה בדו"ח מילולי 26
7.1.2 תוצאות הגהת עמודה בתמונה 27
7.2 ניתוח ראשוני של התוצאות 29
7.3 ניתוח נוסף של הטעויות 29
7.4 הגהת עמודה עם 2 טעויות 30
7.4.1 הגהת תמונה עם טעויות – דו"ח מילולי 30
7.4.2 מדדי הצלחה 32
7.4.3 הגהת עמודה עם טעויות – תצוגה גרפית 32

[הקלד טקסט]

7.5 תוצאות של עמודות נוספות 34
8 **סיכום ומבט להמשך** **39**
ביבליוגרפיה **41**
נספח א' **42**
נספח ב' **47**

רשימת איורים

איור 1 : צילום עמודה ראשונה משני ספרי תורה 2
איור 2 : תרשים זרימה של לימוד והגהה 4
איור 3 : קטע מקובץ הטקסט עם תווים p, s, k 6
איור 4 : מבנה הנתונים של ספר התורה 7
איור 5 : אלגוריתם מציאת סף האפור 9
איור 6 : אפור מול שחור-לבן 10
איור 7 : פונקציה למציאת מפל בין השורות 10
איור 8 : אלגוריתם ליישור תמונה 11
איור 9 : תמונה לפני ואחרי יישור 11
איור 10 : אלגוריתם מציאת השורות 13
איור 11 : אלגוריתם למציאת הרווחים, ומציאת הרווח המינימאלי בין המילים 14
איור 12 : סימון שורות ורווחים 15
איור 13 : המילה הוא בGrid 16
איור 14 : פונקציית סמן את הנקודה 16
איור 15 : פונקציית סמן את כל הנקודות הקשירות 17
איור 16 : פונקציית האם החלק הקשיר החדש הוא תו חדש 17
איור 17 : אלגוריתם הוצאת תמונת התו מתמונת המילה 18
איור 18 : סימון התווים והמילים 19
איור 19 : אותיות ר ות מתוחות אופקית 20
איור 20 : תרשים זרימה המרה לתמונה בת השוואה 22
איור 21 : מבנה רשת נוירונים 23
איור 22 : תוצאה גרפית של הגהת עמודה 2 28
איור 23 : תוצאת הגהת עמודה 2 עם 2 טעויות שתולות 33
איור 24 : תוצאת הגהת עמודה 13 36
איור 25 : תוצאת הגהה לעמודה 22 37
איור 26 : תמונות תווים מקוריים וברי השוואה 42
איור 27 : דוגמת פרופיל לספר תורה 47

1 תקציר

במסגרת פרויקט זה פותחה תוכנה שמטרתה לאפשר הגהה ממוחשבת לספר תורה. התוכנה יודעת "ללמוד" את הכתב מתוך צילומים של ספר תורה שמסופקים ע"י המשתמש, ולהגיה צילומים האחרים של אותו כתב יד.

בשלב ראשון התוכנה מנתחת את צילום העמודה מפרקת את הצילום הגדול לתמונות קטנות של תווים המשויכות למילים, ולאחר מכן בונה רשת נוירונים מלאכותית (machine learning) מתוך הצילומים עבור זיהוי התווים. תפקיד רשת הנוירונים הוא לקבל כקלט תמונת תו ולהוציא כפלט את התו (טקסט). רשת הנוירונים נשמרת כקובץ.

בשלב שני מקבלת התוכנה תמונת עמודה אחרת כלשהי מספר התורה ושוב מפרקת את התמונה של העמודה לתמונות קטנות של תווים המשויכות למילים. התוכנה בודקת את העמודה ומבצעת הגהה. התוכנה בודקת את נכונות המילים התווים והרווחים בצילום העמודה, ומוציאה דו"ח על הטעויות שנמצאו בעמודה. שלב זה מורץ על כל עמודה חדשה של אותו סופר.

התוכנה הצליחה להגיע להישג נאה של זיהוי תווים בשיעור של למעלה מ99% בעמודות מסוימות, אך עוד יש כברת דרך לעבור אם ברצוננו להגיע למוצר איכותי דיו.

בפרק המבוא אתאר את הרקע המציאותי של הבעיה אותה אנחנו מנסים לפתור (מהו ספר תורה וכיצד הוא נכתב), ולאחר מכן אסביר את הרקע הטכנולוגי לבחירה שעשיתי ברשת נוירונים מלאכותית כדי לממש את זיהוי התווים. במבוא נתאר גם אילו רכיבי תוכנה הייתי חייב לפתח כדי להגיע למערכת עובדת.

2 מבוא

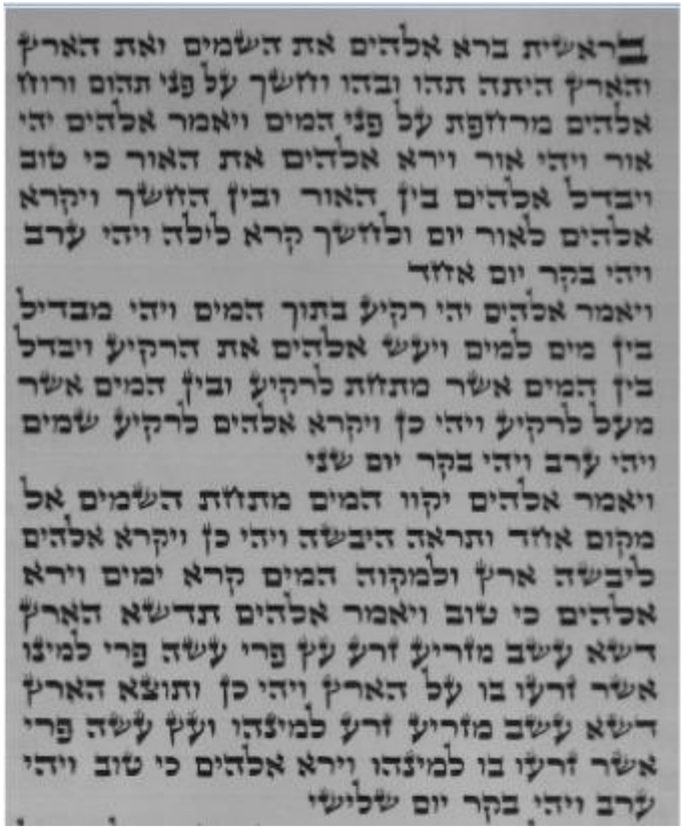
2.1 רקע מציאותי

ספר תורה הוא החפץ המקודש ביותר ביהדות. ישנם כללים (הלכות) מאוד ברורים ונוקשים לגבי שמירתו של הספר, החומרים שמהם עשוי הספר (קלף מעור מסוים, דיו מסוג מסוים) וגם לגבי אופן כתיבתו. הכתב חייב להיות כתב יד, והסופר צריך להיות מודע למטרת הכתיבה. כל התווים ורק התווים חייבים להיות כתובים בספר כלומר אם יש אות מיותרת או חסרה הספר נפסל. אם אות

כלשהי מאבדת את צורתה הספר נפסל. לתווים ישנה צורה מוגדרת היטב, אם כי יש הבדלים בין מסורות העדות השונות לגבי צורת הכתב.

התורה מחולקת לקטעים המכונים פרשיות, ובין הפרשיות יש שני סוגי רווחים: רווח עד סוף השורה (לפני פרשיה פתוחה) ורווח בן 9 תווים (לפני פרשיה סתומה). גם התווים וגם הרווחים ממוספרים ושמורים ע״פ המסורת ואין לחרוג מהן.

על מנת להמחיש על מה מדובר צירפתי תמונות של העמודה הראשונה משני ספרי תורה שונים למדיי (איור 1).


בראשית ברא אלהים את השמים ואת הארץ
והארץ היתה תהו ובהו וחשך על פני תהום ורוח
אלהים מרחפת על פני המים ויאמר אלהים יהי
אור ויהי אור וירא אלהים את האור כי טוב
ויבדל אלהים בין האור ובין החשך ויקרא
אלהים לאור יום ולחשך קרא לילה ויהי ערב
ויהי בקר יום אחד
ויאמר אלהים יהי רקיע בתוך המים ויהי מבדיל
בין מים למים ויעש אלהים את הרקיע ויבדל
בין המים אשר מתחת לרקיע ובין המים אשר
מעל לרקיע ויהי כן ויקרא אלהים לרקיע שמים
ויהי ערב ויהי בקר יום שני
ויאמר אלהים יקוו המים מתחת השמים אל
מקום אחד ותראה היבשה ויהי כן ויקרא אלהים
ליבשה ארץ ולמקוה המים קרא ימים וירא
אלהים כי טוב ויאמר אלהים תדשא הארץ
דשא עשב מזריע זרע עץ פרי עשה פרי למינו
אשר זרעו בו על הארץ ויהי כן ותוצא הארץ
דשא עשב מזריע זרע למינהו ועץ עשה פרי
אשר זרעו בו למינהו וירא אלהים כי טוב ויהי
ערב ויהי בקר יום שלישי

בראשית ברא אלהים את השמים ואת הארץ
והארץ היתה תהו ובהו וחשך על פני תהום ורוח
אלהים מרחפת על פני המים ויאמר אלהים יהי
אור ויהי אור וירא אלהים את האור כי טוב
ויבדל אלהים בין האור ובין החשך ויקרא
אלהים לאור יום ולחשך קרא לילה ויהי ערב
ויהי בקר יום אחד
ויאמר אלהים יהי רקיע בתוך המים ויהי מבדיל
בין מים למים ויעש אלהים את הרקיע ויבדל
בין המים אשר מתחת לרקיע ובין המים אשר
מעל לרקיע ויהי כן ויקרא אלהים לרקיע שמים
ויהי ערב ויהי בקר יום שני
ויאמר אלהים יקוו המים מתחת השמים אל
מקום אחד ותראה היבשה ויהי כן ויקרא אלהים
ליבשה ארץ ולמקוה המים קרא ימים וירא
אלהים כי טוב ויאמר אלהים תדשא הארץ
דשא עשב מזריע זרע עץ פרי עשה פרי למינו
אשר זרעו בו על הארץ ויהי כן ותוצא הארץ
דשא עשב מזריע זרע למינהו ועץ עשה פרי
אשר זרעו בו למינהו וירא אלהים כי טוב ויהי
ערב ויהי בקר יום שלישי

איור 1: צילום עמודה ראשונה משני ספרי תורה

מטרת התוכנה היא איפה לקבל צילום של עמודה כלשהי בספר תורה ולהוציא דו״ח שמפרט טעויות שנמצאו בעמודה זו. התוכנה אמורה לשמש כלי עזר לסופר סת״ם (כך מכונה הלבלר הכותב ספרי תורה), וגבאי בתי הכנסת הדואגים לאחזקת ושימור הספרים, שיוכלו להגיה את ספרי התורה בצורה ממוחשבת.

קיימות מספר תוכנות מסחריות שמספקות פתרון של הגהה ממוחשבת, אך כולן נסחרות והקוד איננו פתוח.

2.2 רקע טכנולוגי

על מנת להגיע לזיהוי הטעויות בספר התורה, צריך תוכנה שיודעת מה צריך להיות כתוב, שיודעת לקרא ספר תורה, ויודעת למצוא את הבעיות.

על מנת לדעת מה צריך להיות כתוב צריך לאחסן את נתוני ספר התורה: (Tora Data). זה הרכיב הראשון.

האתגר המרכזי הוא לדעת לקרא – לזהות תווים. זיהוי תווים הוא תחום רחב ומורכב במדעי המחשב הנקרא OCR (Optical Character Recognition) (Wikipedia, 2017), ותת התחום של זיהוי תווי כתב יד (handwriting recognition)(Wikipedia, 2017) .

את תחום זיהוי התווים חקרתי בסמינריון שהגשתי (גוטמן, 2016), והפרויקט מיישם חלק מהחקירה התיאורטית. כפי שמתואר בסמינריון על מנת להגיע לזיהוי התווים יש צורך (לפחות) בשלושה שלבים:

א. עיבוד מקדים לתמונה (image pre-processing)
ב. חילוץ התווים מתוך התמונה (character extraction)
ג. סיווג התווים (feature extraction, classification).

ישנם אלגוריתמים שונים שנכתבו וממשיכים להיכתב לזיהוי תווים לכל שלביו. גם בפרויקט הזה ישנו מימוש לשלבים השונים. שלב הסיווג (הכולל את חילוץ המאפיינים) הוא השלב המעניין ביותר, ויש לו קשת רחבה של שיטות ופתרונות. ארחיב מעט בעניין הסיווג הממומש בפרויקט הזה.

בשלב הסיווג השתמשתי בשיטת הזיהוי של רשת נוירונים מלאכותית (Artificial Neural Network (ANN) שהיא חלק מתחום הלמידה החישובית שהולך וצובר תאוצה בשנים האחרונות. בשיטה זו מלמדים את המערכת (את רשת הנוירונים) בעזרת דוגמאות רבות, ולאחר שלב הלימוד המערכת יודעת לזהות. הרעיון הוא לתת לתוכנה ללמוד מספר קטן יחסית של עמודות מוגהות ידנית – לייצר רשת נוירונים (השייכת לסופר או אפילו לספר התורה הזה), ואז לבצע זיהויים דרך רשת הנוירונים. שיטה זו (ANN) משיגה את התוצאות הטובות ביותר בזיהוי כתב יד בשנים האחרונות. ניתן לראות לדוגמה את התוצאות הטובות העדכניות ביותר (state of the art) באתר (Benenson, 2016) שמציג את הזיהויים על בסיס הנתונים MNIST.

לאחר שבחרתי להשתמש בשיטת הזיהוי של רשת נוירונים החלטתי גם להפעיל את הזיהוי לפי פרופיל של סופר. עקרונית אפשר היה גם לזהות תווים כלליים כלומר לייצר פרופיל כללי לכל אות בעזרת דוגמאות רבות מספרי תורה שונים או לייצר פרופיל לפי "סוגי" הכתב לפי מנהגי העדות, פרופיל לכל סוג, אך לצורך זיהוי כללי כזה היה צורך בדוגמאות רבות מספרי תורה שונים וזה לא מעשי לפרויקט ובנוסף לכך ברוב רובם של המקרים ספר תורה נכתב בידי סופר בודד ולפיכך יעיל יותר יהיה ללמוד את הכתב שלו.

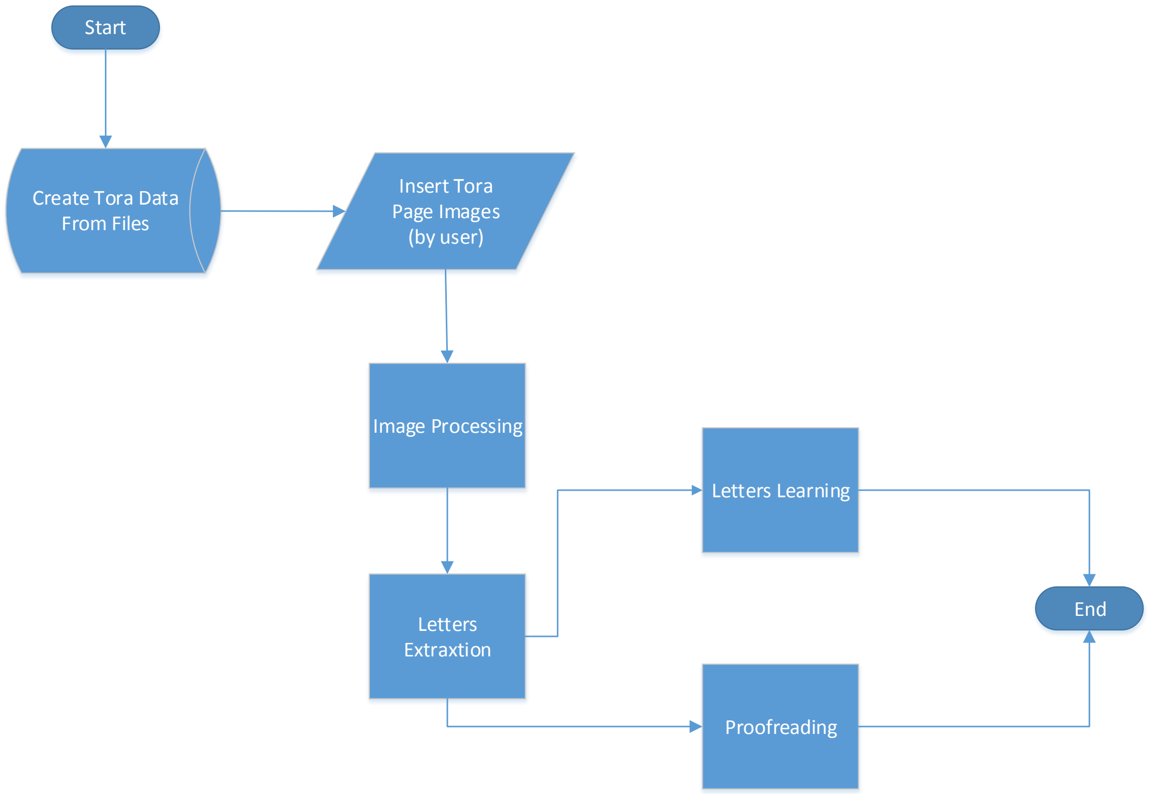
השימוש בשיטת רשת נוירונים לזיהוי התווים מחייב לחלק את המשימה של הגהת ספר תורה לשני שלבים:

1. לימוד התווים - התוכנה מקבלת מהמשתמש מספר עמודים שהיא קוראת ולומדת את הצורה של כל תו
2. בדיקת התווים – התוכנה מקבלת מהמשתמש עמוד והיא בודקת שהעמוד תקין

שני השלבים מתוארים בתרשים הזרימה שבאיור 2.

חמשת רכיבי התוכנה: יצירה ואחסון של נתוני ספר התורה (Tora Data), עיבוד התמונה של עמודת ספר תורה (Image Processing), חילוץ התווים מהתמונה (Letters Extraction), לימוד התווים (Letters Learning) והגהה (Proofreading).

כפי שניתן לראות בתרשים הזרימה (איור 2) הפעלת 3 רכיבים משותפת לשני שלבים: יצירה ואחסון של נתוני ספר התורה, עיבוד התמונה וחילוץ התווים מהתמונה. הפעלת הרכיב הרביעי משתנה בהתאם לשלב: בשלב הלימוד הרכיב הרביעי שמופעל הוא לימוד התווים, לעומת זאת בשלב ההגהה במקום הלימוד מופעל רכיב ההגהה שכולל זיהוי תווים והתאמת הזיהוי לבסיס הנתונים של התורה.



איור 2: תרשים זרימה של לימוד והגהה

2.3 מבנה העבודה

נקדיש פרק לכל אחד מחמשת רכיבי התוכנה: יצירה ואחסון של נתוני ספר התורה (Tora Data), עיבוד התמונה של עמודת ספר תורה (Image Processing), חילוץ התווים מהתמונה (Letters Extraction), לימוד התווים (Letters Learning) והגהה (Proofreading). בכל פרק נכתוב הסבר ותיאור מפורט של המודול במערכת.

לאחר תיאור רכיבי התוכנה נמצא הפרק העוסק בתוצאות שהגענו אליהן. בסיום העבודה מופיע חלק אחרון של סיכום ומבט להמשך שבו נכתוב בראשי פרקים את מה שדרוש עוד להשלמת המלאכה.

לעבודה שלשה הנספחים המכילים תמונות. בנספח הראשון תמונות של תווים בודדים שחולצו מתוך תמונת העמוד, בשני תמונת העמוד שהרצנו עליו את ההגהה לדוגמה, ובשלישי דוגמת פרופיל.

3 שמירת נתוני ספר התורה

3.1 נתוני ספר תורה

נתוני ספר התורה עצמו כוללים את התווים והרווחים. מעבר לכך יש צורך לשמור כחלק מהנתונים את החומש, הפרק, הפסוק מספר המילה ומספר התו במילה, על מנת לדעת למקם את האזור עליו נעשית הבדיקה או הלימוד, וכן במקרה של דיווח על טעות לדעת לציין את המיקום המדויק של הטעות.

כדי שיהיה בסיס להשוואה (כלומר שהתוכנה "תדע" איזה תו אמור להיות המקום בזה), צריך לשמור בסיס נתונים של כתב הספר כפי שאמור להיות. ישנה בעיה להשיג את התווים והרווחים כפי שהם מצויים בספר התורה מכיוון שהנוסח שמצוי ברוב החומשים ובספרים הרגילים אינו מתאים. יש מקומות לא מעטים של הבדל בין הקרי (איך שהמילה בתורה נקראת) לכתיב, וברוב החומשים מצוי נוסח הקרי, בעוד שאותנו (את תוכנת ההגהה) מעניין נוסח הכתיב. הנוסח כולל את האותיות והמילים ע"פ נוסח המסורה המדויק של הכתב, כולל מספר תווים מיוחדים שעל פי המסורת צריך לעשותם גדולים וקטנים, כולל רווחים בין המילים בסופי פרשיות (פרשיה פתוחה וסתומה) ובסופי החומשים.

3.2 שמירת הנתונים בקבצים

באתר: http://mechon-mamre.org של מכון ממרא מוצג ספר התורה לפי כתיב המסורה כולל הפרשיות, הפרשות והתווים המיוחדים. מהאתר הורדתי 5 קבצים, קובץ לכל חומש. את הקבצים עיבדתי "ידנית" כך שהקובץ יכיל את כל הנתונים ורק את הנתונים שרציתי. קיבלתי 5 קבצי טקסט עם סימנים מיוחדים לרווחים שונים ותווים גדולים וזעירים. בקובץ יש לכל פסוק בתחילתו את שם הפרק והפסוק לאחר מכן יבואו המילים שבפסוק. מלבד תווי המילים ומקש הרווח ישנם תווים מיוחדים של s ,p לציון רווח של פרשיה פתוחה או סתומה, k ,g לתווים גדולים או זעירים (איור 3).

<B>ב,א</B>ויכלו השמים והארץ וכל צבאם

<B>ב,ב</B>ויכל אלהים ביום השביעי מלאכתו אשר עשה וישבת ביום השביעי מכל מלאכתו אשר עשה

<B>ב,ג</B>ויברך אלהים את יום השביעי ויקדש אתו כי בו שבת מכל מלאכתו אשר ברא אלהים לעשות p

<B>ב,ד</B>אלה תולדות השמים והארץ בkהבראם ביום עשות יהוה אלהים ארץ ושמים

...

<B>ג,טו</B>ואיבה אשית בינך ובין האשה ובין זרעך ובין זרעה הוא ישופך ראש ואתה תשופנו עקב s

<B>ג,טז</B>אל האשה אמר הרבה ארבה עצבונך והרנך בעצב תלדי בנים ואל אישך תשוקתך והוא ימשל בך s

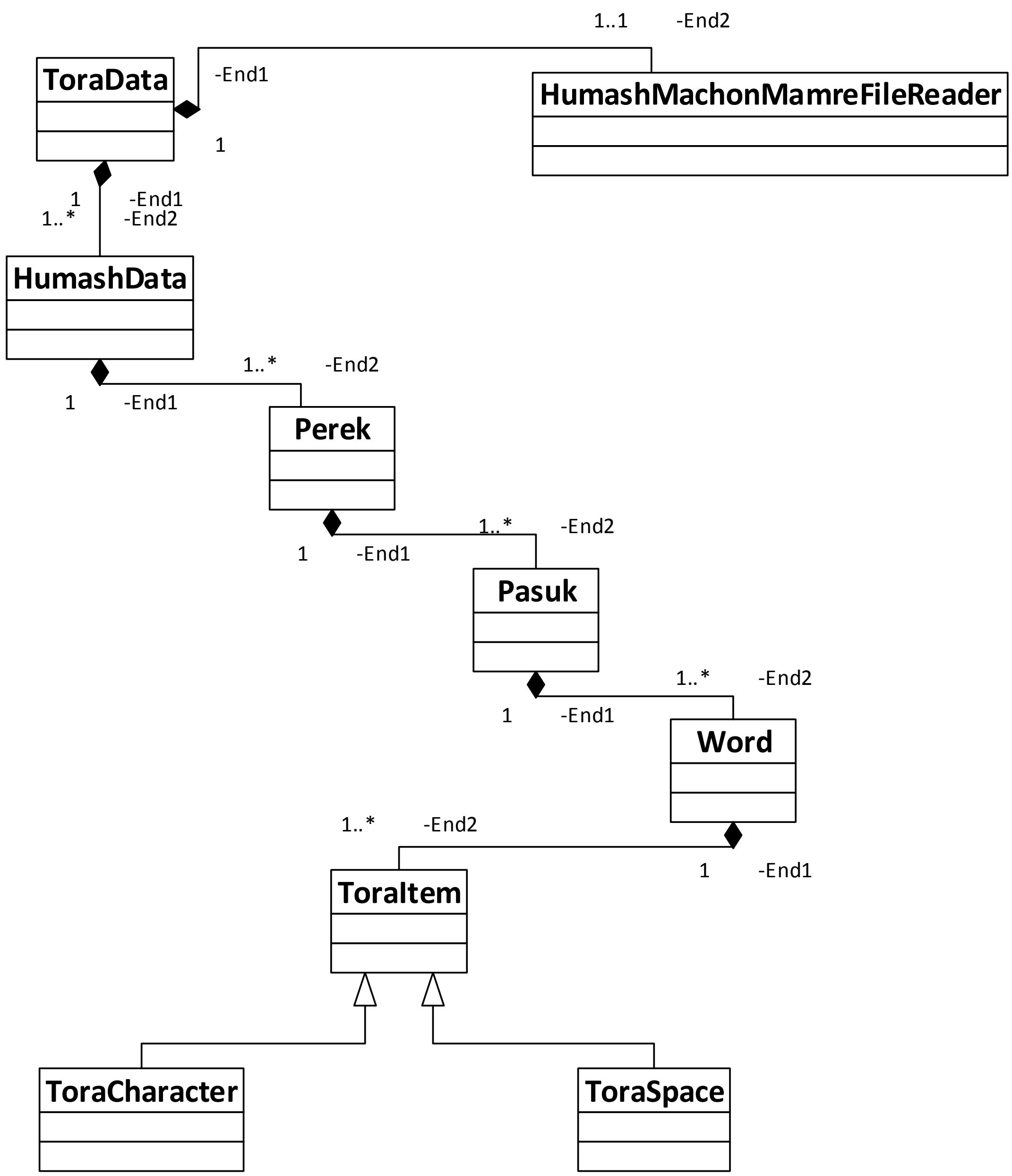
איור 3: קטע מקובץ הטקסט עם תווים P, S, K

3.3 שמירת הנתונים בזיכרון התוכנית

ע״פ המקובל היום מבנה הטקסט המקראי הוא כדלהלן: התורה מחולקת ל5 חומשים, החומש מתחלק לפרקים (לא מספר קבוע בכל חומש), הפרקים לפסוקים, הפסוקים למילים והמילים לתווים. וכך בניתי את המחלקה ToraData את האובייקטים הבאים שייצגו את הכתב כמו שהוא אמור להיות:

ToraData, HumashData, Perek, Pasuk, Word, ToraCharacter, ToraSpace ואת האובייקט שקורא את הקבצים: HumashMamreFileReader המבנה מתואר בתרשים שבאיור 4

השלב הראשון בהפעלת התוכנה הוא טעינת הנתונים מהקבצים המוכנים מראש לתוך מבני הנתונים, כך ש- ToraData מכיל מערך של חמישה חומשים (HumashData), שכל אחד מהחומשים מכיל מערך של פרקים, שכל פרק מכיל מערך של פסוקים (Pasuk), כשכל אחד מהפסוקים מכיל מערך של מילים (Word), ולכל מילה מערך של העצמים הבסיסיים ביותר: תווים ורווחים. התווים והרווחים ממששים interface של ToraItem. קריאת כל הנתונים לזיכרון מהקבצים מתבצעת תוך שניה בודדת (ו15 מאיות) בלבד.



איור 4 : מבנה הנתונים של ספר התורה

עיבוד התמונה

3.4 עיבוד תמונה לצורך זיהוי תווים

כדי לזהות את התווים צריך להוציא את התווים מתמונת עמודת ספר התורה. בדרך כלל לפני חילוץ התווים נעשית עבודה מסיבית של עיבוד תמונה הבאה להוריד כמה שיותר רעשים וכתמים, ולחדד את אזורי התווים. במקרה שלנו ניקוי הרעשים נעשה בצורה מינימאלית מאוד, מכיוון שמטרת התוכנה

היא לזהות שגיאות ובעיות וגם לכלוכים (כתמי דיו או אותיות שבורות) הן בעיות שיש צורך להתריע עליהן. התוכנה עושה פעולת ניקיון של פילטור נקודות בודדות (נקודות שחורות בודדות שמוקפות בלבן) מכיוון זהו רעש שלא נחשב אפילו כלכלוך. שתי הפעולות המשמעותיות של עיבוד התמונה הן :

א. (במקרה של תמונה באפור) מציאת ה- Threshold והפיכת התמונה לבינארית
ב. סיבוב התמונה – (אם התמונה לא ישרה זה יגרום לשיבוש בלימוד \ זיהוי התווים)

שלב עיבוד התמונה כולל את קריאת הקובץ המקורי והפיכתו לתמונה בינארית המוכנה לקראת חילוץ התווים.

3.4.1 איכות התמונה חשובה

כאן המקום לציין שאיכות הצילום משפיעה מאוד על התוצאות. במיוחד במקרה שלנו – לדוגמה אם יש חיבור כלשהו בין שני תווים סמוכים זו בעיה חמורה הפוסלת את הספר, ואם הצילום לא חד מספיק נוצרים חיבורים בין תווים סמוכים, או מקרים הפוכים בו מזהים הפרדה בין שני חלקי תו שמחוברים בחבור דק שלפעמים מיטשטש בתמונה שאינה חדה. לצלם ספר תורה בצורה טובה ומקצועית זה לא קל בשל גודל העמודה והצורך לשמור אותה ללא קיפולים וחלקה יש לצורך זה סורקים מיוחדים. כמובן שאפשר להשתמש גם במצלמות רגילות רק שהאיכות לא תהיה גבוהה.

3.5 פעולות עיבוד התמונה

3.5.1 מציאת הסף (Threshold) של הצבע

כל פיקסל בתמונה מבחינתנו הוא או שחור או לבן, ובמקרים רבים התמונה שמוכנסת כקלט היא צבעונית או אפורה. המטרה הראשונית היא להפוך את התמונה לבינארית. להפוך תמונה צבעונית לגווני אפור אנחנו משתמשים בפונקציה סטנדרטית. הבעיה היא מציאת הסף של גוון האפור שמעליו נחשב שחור ומתחתיו אפור. הבעיה פה אינה פשוטה מכיוון שהסף תלוי באיכות הצילום ובתאורה בעת הצילום.

לאחר ניסיונות רבים ולא מוצלחים למצוא את הסף הנכון לפי היסטוגרמת הגוונים (סופרים כמה נקודות יש לכל גוון ולפי ההתפלגות מנסים לאתר את הסף), הגעתי לאלגוריתם כבד במיוחד (איור 5). האלגוריתם מקבל תמונת עמודה ומוציא את הסף - הגוון האפור (מספר בין 0 ל255) הטוב ביותר.

ניחוש ראשוני:

I. צור היסטוגרמת גוונים
II. מצא את המפל הגדול ביותר בין כל 3 אינדקסים סמוכים בהיסטוגרמה (מחברים 3 אינדקסים ומשווים ל-3 שלידם וכך הלאה).
III. מצא את הנקודה שבה מתחילים מפסיקים לרדת ערכי ההיסטוגרמה מהשיא.
IV. חשב את ממוצע שתי הנקודות זהו הסף הנוכחי.
V. קח את התמונה המבוקשת הפוך אותה לבינארית ע"פ הסף הנוכחי – הפעל את חילוץ התווים.
VI. תחילת לולאה:
    i. מצא את סכום הטעויות של מספר תווים נמוך מדיי במילה ובנפרד את סכום הטעויות של מספר תווים גבוה מידיי.
    ii. אם לא נמצאו טעויות הכנס את הסף הנוכחי למשתנה הסף הטוב ביותר וצא מהלולאה.
    iii. אם מספר הטעויות הנוכחי הוא הנמוך ביותר עד עכשיו הכנס את הסף הנוכחי למשתנה הסף הטוב ביותר.
    iv. אם יש יותר טעויות של מספר תווים נמוך העלה את הסף הנוכחי, אחרת הורד את הסף הנוכחי. במידה ואין סף עליון /תחתון שמור אותם, במידה ויש העלה /הורד את הסף לערך שבין הסף הנוכחי לגבול. עדכן גבול מתאים. אם הסף הנוכחי והגבול צמודים צא מהלולאה.

חזור לתחילת הלולאה.

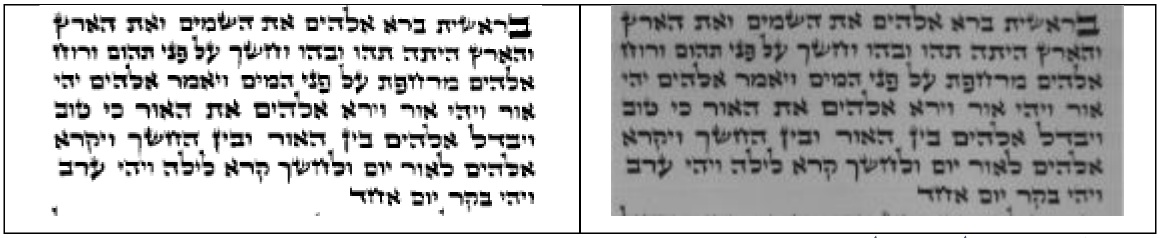
VII. החזר את הסף הטוב ביותר

איור 5: אלגוריתם מציאת סף האפור

מדובר אם כן בפעולה כבדה מאוד. על כל ניסיון אנחנו מעבדים מחדש את העמוד וסופרים טעויות מסוג תווים מחוברים או תווים מנותקים. אם יש תווים מחוברים רבים אנחנו מעלים את הסף אחרת אם יש תווים מנותקים אנחנו מורידים את הסף. ושוב חוזרים על הפעולה עד שנמצא הסף עם מספר הטעויות הנמוך ביותר.

אם ידוע ששאר הצילומים דומים באיכותם ניתן ורצוי לשמור את הסף בפרופיל המשתמש (ראה נספח ג). יש גם שמירה על קובץ נתונים נפרד של הסף לפי שם עמודה כדי שבריצה הבאה על העמודה הזו לא נחשב מחדש. להלן (איור 6) דוגמה של התוצאה.

אפור	בינארי



איור 6: אפור מול שחור-לבן

3.5.2 יישור התמונה

חשוב מאוד שהתמונה תהיה מיושרת, כדי שהזיהוי והלימוד יהיה על תווים באותה זווית. לשם כך בניתי מנגנון ליישור תמונה. הפעלת המנגנון (שהוא כבד ולוקח זמן) תלויה בפרמטר בפרופיל (נספח ג) ואפשר לנטרל את הפעלתו:

המנגנון משתמש בתהליך (פונקציה) למציאת "מפל" בין שורות (איור 7). מפל הכוונה להפרש (בערך מוחלט) בין מספר הנקודות השחורות בין שתי שורות סמוכות. ההנחה היא שבמצב אופקי ההפרשים יהיו הגדולים ביותר היות והמעבר בין שורת הפיקסלים בתחילת שורה של אותיות לשורת הפיקסלים שלפניה הוא החד ביותר. (כך גם לגבי סוף השורה רק ששם יש הרבה פחות פיקסלים בגלל מבנה האותיות.) האלגוריתם מתואר גם בעבודת הסמינריון שלי (גוטמן, 2016).

i. מצא את סכום הפיקסלים השחורים בכל שורת פיקסלים בתמונה. (התעלם משורות שחורות לגמרי.) והכנס למערך.
ii. חשב את ההפרש בין כל שתי שורות (בערך מוחלט), וסכום את התוצאות הגבוהות ביותר. מספר התוצאות הגבוהות לפי מספר השורות בעמודה. (בעמודה סטנדרטית 42 שורות אז סכום 42 תוצאות.)
iii. החזר את סכום ההפרשים.

איור 7: פונקציה למציאת מפל בין השורות

באיור 8 ניתן לראות אלגוריתם ליישור תמונה:

I. קרא את תמונת העמודה

II. מצא מפל נוכחי, ושמור ערך וזווית. הכנס ערך למפל הגבוה.

III. סובב את התמונה בזווית של 0.25 מעלות (השתמש בספריית של AWT של Java לייצור תמונה מסובבת) .

IV. מצא מפל נוכחי

V. סמן שינוי לנגד כוון שעון

VI. השווה אם מפל נוכחי גבוה יותר שמור מפל וזווית וסמן שינוי ל בכוון השעון, הכנס ערך למפל הגבוה

VII. לולאה כל עוד לא נמצאה זווית הנכונה ביותר

 i. אם אנחנו בכוון השעון העלה את הזווית בעוד 0.25, אחרת הורד 0.25

 ii. סובב תמונה בזווית הנוכחית (צור תמונה מסובבת)

 iii. מצא מפל נוכחי

 iv. בדוק אם המפל הנוכחי קטן מהמפל הגדול ביותר צא מהלולאה

VIII. החזר תמונה מיושרת

איור 8: אלגוריתם ליישור תמונה

להלן (איור 9) דוגמה של קטע עמודה לפני ואחרי יישור.

תמונה מקורית (לפני יישור)	תמונה מיושרת
 ואת כל רמש האדמה למינהו וירא אלהים כי טוב ויאמר אלהים נעשה אדם בצלמנו כדמותנו וירדו בדגת הים ובעוף השמים ובבהמה ובכל הארץ ובכל הרמש הרמש על הארץ ויברא אלהים את האדם בצלמו בצלם אלהים ברא אתו זכר ונקבה ברא אתם ויברך אתם אלהים ויאמר להם אלהים פרו ורבו ומלאו את הארץ וכבשה ורדו בדגת הים ובעוף השמים ובכל חיה הרמשת על הארץ ויאמר אלהים הנה נתתי לכם את כל עשב זרע זרע אשר על פני כל הארץ ואת כל העץ אשר בו פרי עץ זרע זרע לכם יהיה לאכלה ולכל חית הארץ ולכל עוף השמים ולכל רומש על הארץ אשר בו נפש חיה את כל ירק עשב לאכלה ויהי כן וירא אלהים את כל אשר עשה והנה טוב מאד ויהי ערב ויהי בקר יום הששי ויכלו השמים והארץ וכל צבאם ויכל אלהים ביום השביעי מלאכתו אשר עשה וישבת ביום השביעי מכל מלאכתו אשר עשה ויברך אלהים את יום השביעי ויקדש אתו כי בו שבת מכל מלאכתו אשר ברא אלהים לעשות	 ואת כל רמש האדמה למינהו וירא אלהים כי טוב ויאמר אלהים נעשה אדם בצלמנו כדמותנו וירדו בדגת הים ובעוף השמים ובבהמה ובכל הארץ ובכל הרמש הרמש על הארץ ויברא אלהים את האדם בצלמו בצלם אלהים ברא אתו זכר ונקבה ברא אתם ויברך אתם אלהים ויאמר להם אלהים פרו ורבו ומלאו את הארץ וכבשה ורדו בדגת הים ובעוף השמים ובכל חיה הרמשת על הארץ ויאמר אלהים הנה נתתי לכם את כל עשב זרע זרע אשר על פני כל הארץ ואת כל העץ אשר בו פרי עץ זרע זרע לכם יהיה לאכלה ולכל חית הארץ ולכל עוף השמים ולכל רומש על הארץ אשר בו נפש חיה את כל ירק עשב לאכלה ויהי כן וירא אלהים את כל אשר עשה והנה טוב מאד ויהי ערב ויהי בקר יום הששי ויכלו השמים והארץ וכל צבאם ויכל אלהים ביום השביעי מלאכתו אשר עשה וישבת ביום השביעי מכל מלאכתו אשר עשה ויברך אלהים את יום השביעי ויקדש אתו כי בו שבת מכל מלאכתו אשר ברא אלהים לעשות

איור 9: תמונה לפני ואחרי יישור

4 חילוץ התווים

4.1 אתגרים בחילוץ תווים

ישנו תחום באבטחת מידע הנקרא CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), כלומר מבחן האם מחשב או אדם מנסים להיכנס לתוכנה. כפי שציינתי בסמינריון (גוטמן, 2016), אחד המימושים הנפוצים לCAPTCHA הוא הצגה של מספר תווים בכתב יד ודרישת זיהוי לכתב היד. הנחת המבחן היא שיש למחשב בעיה (שטרם נפתרה) בזיהוי כתב יד. הבעיה העיקרית שיש למחשב בזיהוי כתב יד כיום היא בחילוץ התווים (בזיהוי התווים עצמם האלגוריתמים המשוכללים מגיעים לרמת דיוק גבוהה מאוד).

בעיית חילוץ התווים קשה במיוחד כשמדובר בכתב עם חפיפה או כתב מחובר, כתב עם רקע משתנה או כשיש עיוותים קלים בתווים. כלומר לעיתים מדובר במשימה קשה ביותר למחשב.

למרבה המזל במקרה שלנו אין כתב מחובר (אם יש חיבור זה פוסל את הכתב) אבל ישנה חפיפה בין התווים (כלומר יש אותיות שלמלבן החוסם שלהן נכנסות אותיות אחרות). הבעיה במקרה כזה היא האם לשייך אזור שחור לתו הקודם או לתו שבא אחריו. צריך לזכור שבאותיות כמו 'ה' ו'ק' ישנן אזורי נקודות נפרדים באותה אות, ומאידך ישנם מקרים של "חפיפה" בין אותיות כמו הרצף של 'ר' ולאחריה 'ב', 'ל' והתו שאחריה.

4.2 שלבי חילוץ התווים

כדי לחלץ את התווים מתמונת העמודה יש לזהות את מבנה העמודה הכולל שורות מילים ותווים.

השלבים בניתוח התמונה כדי להגיע לחילוץ התווים:

- לזהות את מבנה השורות של העמודה
- למצוא את הרווחים בין המילים
- למצוא את הרווחים בין התווים
- להפריד בין תווים חופפים ובין תווים מחולקים

עתה נפרט איך מבצעים כל שלב.

4.2.1 מציאת השורות

העמודה של ספר התורה בנויה משורות של מילים (לא כל השורות מלאות). כדי למצוא את התווים ע"פ הסדר השלב הראשון הוא לזהות את השורות. בנוסף מזהה חשוב לתו הוא כמה הוא בולט מעל השורה או מתחת לשורה או בתוכה. גודל התו כדאי שיישמר ביחס לשורה.

אפשר לנצל את התכונה של ספר התורה שמספר השורות בכל עמודה ידוע מראש. מספר השורות המקובל היום הוא 42, אם כי בעבר היו גם סטנדרטים אחרים (48, 72) ואין מגבלה הלכתית לשנות, אבל מכל מקום זה נתון חשוב שניתן לדרוש מראש ומקל על ניתוח העמוד. בדוגמת הפרופיל (נספח ג ניתן לראות גם את פרמטר מספר השורות. מכיוון שיש שורות שהן כמעט ריקות קשה מאוד לקבוע את המיקום רק לפי צפיפות הפיקסלים בכל שורה. כאשר ידוע מראש מספר השורות אפשר לחלק את העמוד לפי מספר השורות ובכל אזור שבו אמורה להיות שורה לבדוק את צפיפות הפיקסלים בשורות הפיקסלים ולזהות עליה וירידה בצפיות ולקבוע במדויק את התחלת וסיום השורה.

הדבר הכי בטוח ויציב בניתוח העמוד הוא הרווחים שבין השורות – המקום בו יש הכי פחות פיקסלים: אלו מרכזי הרווחים שבין השורות. אחרי מציאת הרווחים בין השורות נדע גם את מקום השורות. אלגוריתם למציאת השורות (דרך מציאת הרווחים בין השורות) יש באיור 10. האלגוריתם מקבל תמונת עמודה ומחזיר את תחומי הרווחים שבין השורות.

I. קרא את תמונת העמודה
II. חשב את מספר פיקסלים השחורים בכל שורת פיקסלים
III. מצא את התחלת אזור הכתב בסריקה מלמעלה (עד לקפיצה גדולה ראשונה במספר הפיקסלים השחורים)
IV. מצא את סוף אזור הכתב (מסוף העמוד עלה ומצא חריגה ראשונה בעליית מספר פיקסלים)
V. חשב את תחומי השורות המשוערים לפי מספר השורות הידוע
v. לולאה (לכל תחום שורה)
    i. מצא את מרכז הרווח בין השורות הנוכחי
    ii. מצא את המקסימום המקומי הקרוב בעליה למעלה (תחילת הרווח שהיא סוף השורה שלמעלה)
    iii. מצא את מקסימום הקרוב בירידה (סוף הרווח שהוא תחילת השורה הבאה)
    iv. שמור את טווח הרווח
VI. החזר את רשימת טווחי התחומים בין השורות

איור 10: אלגוריתם מציאת השורות

4.2.2 מציאת הרווחים

אחרי שמצאנו את טווחי השורות (מה שנמצא בין הרווחים), מטרתנו הבאה היא המילים:

גם כאן אנחנו מחפשים את הרווחים, ומתוך הרווחים מוצאים את המילים שמסביב לרווחים. בשלב ראשון מוצאים את כלל הרווחים בשורה ואח״כ מוצאים מהם הרווחים בין המילים ומהם הרווחים בין התווים בתוך המילים.

האתגר הוא מציאת הרווח המינימאלי בין המילים. מבחינת כללי הכתיבה הרווח אמור להיות לפחות ברוחב התו הקטן ביותר (בדרך כלל האות י). צריך להבחין בין ריווחי התווים בתוך המילים, לבין הרווחים בין המילים. לשם כך השתמשתי באלגוריתם באיור 11. האלגוריתם מקבל תחום של שורה בתוך העמודה ומחזיר את המרחק של הרווח שממנו ומעלה מדובר ברווח בין המילים. הנחת האלגוריתם שהרווחים בין התווים מסתובבים סביב המקסימום המקומי הראשון של מספר הרווחים, מהמקסימום השני ואילך מדובר ברווחים בין המילים.

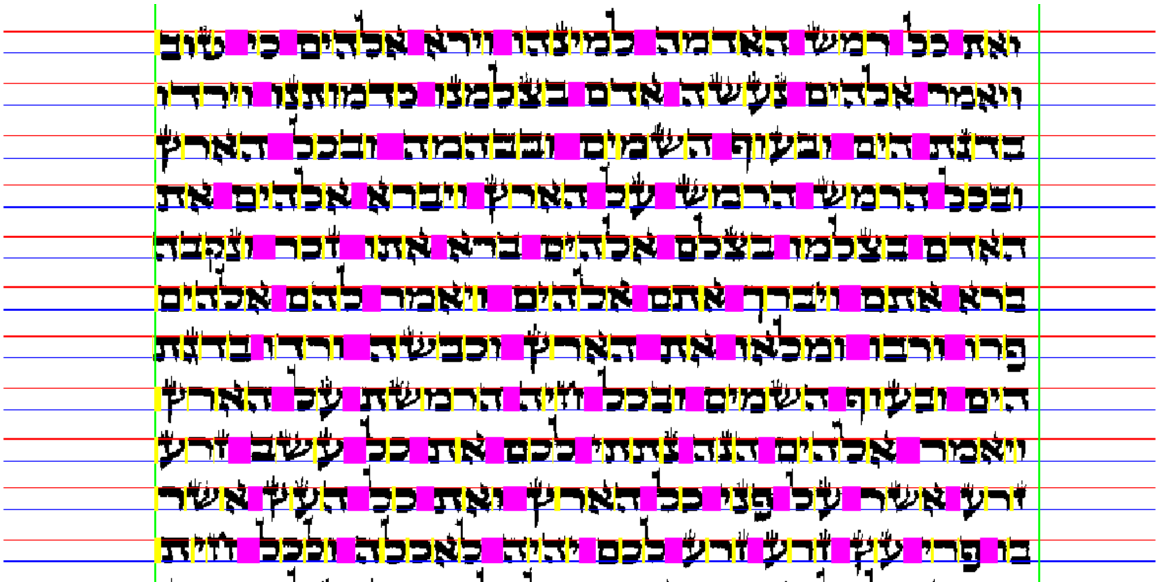
I. מסוף השורה עד תחילת השורה בדוק את עמודות הפיקסלים:
    i. בדוק בתוך השורה אם יש פיקסלים שחורים בעמודת הפיקסלים הנוכחית:
    ii. אם אין אם אנחנו בתוך רווח הוסף לרווח נוכחי,
        אחרת התחל רווח
    iii. אחרת (אם יש פיקסלים) אם היינו ברווח הכנס רווח לרשימה,
    iv. אחרת אל תעשה כלום.
II. הכנס את רוחבי הרווחים להיסטוגרמה שבה כל עמודה היא רוחב רווח.
III. מצא את המקסימום המקומי השני מתחילת ההיסטוגרמה – זהו המינימום לרווח בין המילים.

איור 11: אלגוריתם למציאת הרווחים, ומציאת הרווח המינימאלי בין המילים

ערך הרווח המינימאלי בין המילים חשוב מאוד, והוא נשמר בפרופיל הספר (נספח ג).. הגודל הזה חשוב ביותר ולא הצלחתי למצוא שיטה מדעית טובה יותר. ניתן לערוך גודל זה גם ידנית.

[אפשר היה להוסיף בדיקות מסביב לערך זה ולראות איזה ערך נותן את הקירוב הטוב ביותר למספר המילים שאמור להיות בטקסט בדומה למה שעשינו באלגוריתם למציאת הסף (איור 5).]

דוגמה צבעונית בה אפשר לראות התחלת שורה וסוף שורה (קו רוחבי אדום וכחול) רווחים בין מילים בסגול ורווחים בין תווים בצהוב ניתן לראות באיור 12.



איור 12: סימון שורות ורווחים

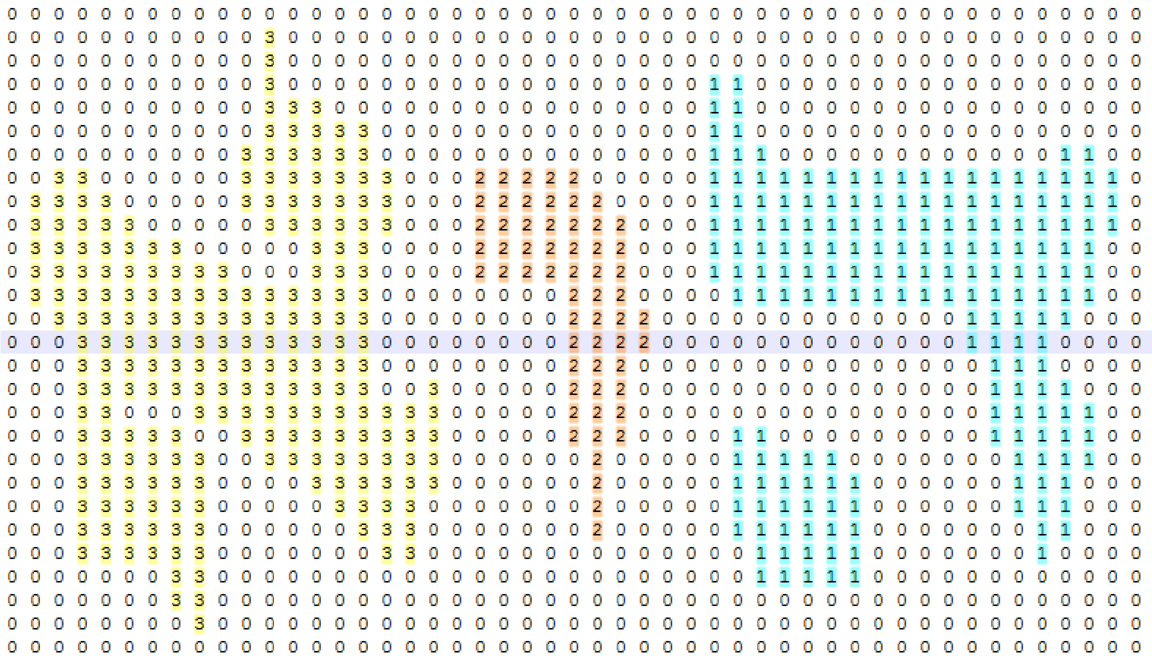
4.2.3 מציאת התווים

לאחר שקבענו את הרווח בין המילים ניתן לסמן את אזורי המילים בכל שורה. המשימה הבאה היא לפרק כל אזור מילה לתווים בודדים עם תמונה לכל תו.

באלגוריתם למציאת התו הבודד מסתמכים הרבה על הקישוריות שיש בין הפיקסלים שבכל תו. הבעייתיות היא בתווים 'ה', 'ק' שם יש חלקי תו שאינם מקושרים. אי אפשר גם לקבוע תיבה מלבנית מסביב לתו ה"שייכת" לתו כי יש מקרים (כשרים הלכתית לגמרי) בהם יש חדירה של תווים לתוך השטח "של" התווים הסמוכים כמו 'ב' שנכנסת מתחת ל'ר' שלפניה או 'ע' שגולשת מתחת ל"י" שלאחריה וכיוצא באלה.

הרעיון הוא לעבור על האזורים של תווים (המקומות שבין הרווחים שבמילים), בכל אזור כזה יש תו אחד לפחות. לעבור על כל גופי הנקודות המקושרות, כאשר בודקים כל גוף נקודות מקושר אם הוא תו עצמאי או שייך לתו הקודם.

מגדירים מערך דו ממדי (Grid) של מספרים שלמים בגודל של מלבן הפיקסלים המקיף מילה. המטרה היא לסמן במספר סידורי כל פיקסל שחור לפי המספר הסידורי של האות במילה. בסוף התהליך מוציאים את התווים מהמערך ע"פ המספר הסידורי. לדוגמה באיור 13 אפשר לראות את המילה "הוא" שבה מסומן כל פיקסל שחור ע"פ המספר הסידורי של התו. יש שלשה תווים וכל פיקסל בכל תו מסומן ע"פ המספר הסידורי של התו. האות ה מיוחדת מכיוון שיש לה שני הגופים מקושרים נפרדים ושניהם משויכים לתו הראשון.



איור 13: המילה הוא בGRID

עתה נתאר את תהליך מציאת התווים (תחילה נתאר מספר פונקציות (פרוצדורות) ולבסוף נתאר את הפעלת הפונקציות בתוך האלגוריתם):

הכן מראש מערך דו-ממדי של מספרים בגודל אזור המילה בתמונה – נקרא לו Grid .

❖ פונקציה: סמן את הנקודה (איור 14) (קלט לפרוצדורה: נקודת p(x, y), מספר תו במילה; פלט בוליאני חיובי \ שלילי):

i. בדוק אם קואורדינטות הנקודה נמצאות בשטח התמונה, אם לא החזר שלילי, אם כן:
ii. בדוק אם הנקודה צבועה, אם לא החזר שלילי, אם כן:
iii. בדוק אם הנקודה מסומנת כבר בGrid אם כן החזר שלילי אם לא:
iv. סמן הנקודה בGrid עם מספר התו והחזר חיובי

איור 14: פונקציית סמן את הנקודה

❖ פונקציה: סמן את כל הנקודות הקשירות (איור 15) (קלט לפרוצדורה: נקודת p(x, y), מספר תו במילה; פלט בוליאני חיובי/שלילי)

I. בצע פרוצדורה סמן את הנקודה

II. אם לא הצלחת לסמן את הנקודה – החזר שלילי, אחרת:

III. הוסף את הנקודה לסט נקודות חדשות:

IV. לולאה – כל עוד יש נקודות בסט הנקודות החדשות:

   i. רוקן את הסט של הנקודות החדשות, ועבור כל נקודה:

      1) הפעל פרוצדורת סמן את הנקודה על כל השכנים (קשירות 8)

      2) הוסף כל שכן שמצליח "להיסתמן" לסט הנקודות החדשות

V. החזר חיובי

איור 15: פונקציית סמן את כל הנקודות הקשירות

❖ פונקציה: בדוק האם החלק המסומן החדש הוא תו חדש (איור 16) (קלט: מספר הסימון החדש והGrid פלט בוליאני):

I. אם מספר התו הוא 1 החזר חיובי, אחרת:

   i. לולאה עבור על עמודות הפיקסלים בתו הנוכחי:

      1) ספור כמה נקודות יש בעמודה לבד של התו הקודם, כמה לבד של התו הנוכחי, וכמה משותפות.

II. במידה ויש חפיפה של יותר משני שליש מהעמודות שבהן נמצא החלק החדש או יותר משני שליש חפיפה מהעמודות של התו הישן החזר חיובי

איור 16: פונקציית האם החלק הקשיר החדש הוא תו חדש

ולבסוף (איור 17) נראה את האלגוריתם שבו משובצות הפונקציות שתוארו לעיל. בסוף האלגוריתם יהיה בידינו מערך דו ממדי שבו כל הפיקסלים של כל תו מסומנים במספר סידורי לפי המספר הסידורי של התו במילה.

I. צור מערך דו ממדי של מספרים שלמים Grid (מאותחל לאפסים)
II. חלק את שטח המילה למלבנים החשודים כתווים לפי הרווחים בין התווים:
III. עבור כל מלבן החשוד כתו (לולאה):
    I. העלה את מונה התווים באחד
    II. עבור כל נקודה משטח המלבן הנוכחי שבתוך השורה (לולאה):
    III. הפעל פרוצדורת סמן את כל הנקודות הקשירות (עם מספר התו הנוכחי)
        I. אם שלילי המשך בלולאה אחרת:
        II. הפעל את פרוצדורת בדוק אם החלק המסומן הוא תו חדש,
            I. אם חיובי העלה את המונה והמשך, אחרת:
            II. החלק החדש מוספח לתו הקודם (הנקודות החדשות בGrid יורדות בערכן לתו הקודם), והסימון הבא מתחיל במספר שהתחיל הסימון הקודם (אל תעלה את המונה).

איור 17: אלגוריתם הוצאת תמונת התו מתמונת המילה

לאחר התהליך מפעילים תהליך שמוציא מה-Grid את כל התווים הממוספרים כל תו (גם אם יש חפיפה) לתמונה נפרדת.

באיור 18 ניתן לראות את המילים והתווים מוקפים בתיבה מלבנית (התווים בכחול והמילים בירוק). ניתן לראות במילה "ויראי" שבשורה הראשונה שהתווים "ו" ו-"י" נפרדים, למרות שלא היה ביניהם רווח בתמונת הרווחים.

איור 18: סימון התווים והמילים

5 למידת התווים

5.1 תהליך זיהוי התווים

בזיהוי תמונה (image recognition) באופן כללי צריך להחליט מהם המאפיינים המזהים שאותם צריך לשמור, ובאיזו צורה לשמור. כשמדובר על זיהוי (כתב) בעזרת רשת נוירונים מדובר עקרונית בתהליך פשוט שכולל בתוכו את מציאת המאפיינים. מעבירים את תמונת התו המקורית לגודל קבוע לא גדול מדיי, ואז ממפים כל פיקסל בתמונה כנקודת כניסה לרשת הנוירונים. התכונה של התו היא הפיקסלים שלו ורשת הנוירונים מזהה אילו קבוצות פיקסלים חשובות לזיהוי. אם לדוגמה תהיה תמונת תו ברזולוציה של 30X20, יהיו לנו 600 כניסות. הבעייתיות כאן היא איך לא לאבד מידע חשוב בהעברת התמונה לגודל קבוע. לתמונה בגודל הקבוע נקרא תמונה בת השוואה. בהמשך נפרט על תהליך זה.

כל פיקסל בתמונה הוא נקודת כניסה לרשת הנוירונים (שכבת הקלט), במקרה שלנו כל פיקסל יכול להיות 0 או 1. בשלב הלמידה המערכת משנה את הערכים בשכבות הפנימיות של רשת הנוירונים כך שהרשת תדע לזהות. (בשלב הזיהוי מכניסים כל תמונה לרשת ומקבלים זיהוי מיידי.) בכל אופן בשיטה זו צריך להגיע למצב שבו גודל התמונה שנכנסת לרשת הנוירונים ללמידה או לזיהוי יהיה אחיד.

בשלב הלמידה אנחנו נותנים לתוכנה מספר עמודות של ספר תורה, כל תמונות התווים מחולצות ואנו יודעים מראש את זיהוי תמונת כל תו (בזכות נתוני הכתב השמורים). רשת הנוירונים מכיילת את

הקשרים ולומדת לזהות את התווים (שלב האימון). כחלק משלב הלמידה ישנו גם שלב הבדיקה (מקצים חלק מהתווים מראש כקבוצת בדיקה ללימוד) – לאחר הלמידה הרשת בודקת את עצמה ומראה את אחוז הזיהוי על קבוצת הבדיקה. שלב הלמידה לוקח זמן – כ3 דקות במקרה של 5 עמודות בתמונות ברזולוציה של1650 X 1275. הזמן משתנה בהתאם לכמות הנתונים ומבנה רשת הנוירונים.

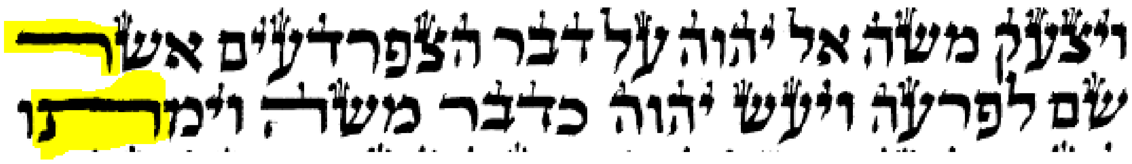
ספציפית בפרויקט הזה יש לזכור כל הזמן שהמטרה היא לזהות טעויות לכן חשוב גם לא לאבד נקודות שהן לא עוזרות לזיהוי ... לכן בניגוד לפרויקטים של זיהוי כתב, אי אפשר להגזים בהדגשת השוני על חשבון העלמת בעיות.

5.2 מעבר לתמונה בת השוואה

מכיוון שכל תמונת תו שחולצה מתמונה של עמודת ספר התורה יש לה גודל משלה, אנחנו צריכים לבצע תהליך שיעביר את תמונת התו המקורית לתמונה שהיא בת השוואה. התמונה החדשה צריכה לשמור ככל האפשר על המאפיינים של התמונה המקורית אבל כדאי שלא תהיה גדולה מידיי כי אז רשת הנוירונים תהיה מורכבת ותהליך הלמידה ארוך ומורכב יותר. אנחנו לא רוצים לבצע מתיחה או כיווץ פשוטים מכיוון שאז נאבד נתונים חשובים: למשל ההבדל בין התו ׳׳י לתו ׳ו׳ ולתו ׳ן׳ בא לידי ביטוי בעיקר באורך הקו המאונך ובמיקומו ביחס שורה. אם נבצע מתיחה לא נוכל לזהות את ההבדל ביניהם.

אם לא מתיחה – אפשר לבצע ריפוד (padding) כלומר למלא את החלל החסר בתווים הקטנים בפיקסלים לא צבועים. אבל גם כאן כדאי להיזהר לא לאבד את הנתון של גובה התו ביחס לשורה, וגם צריך להיזהר לא לאבד נתונים ע״י הוספת הרבה פיקסלים שאין להם משמעות אמתית.

בנוסף יש בעיה ייחודית לכתב דווקא בספר תורה: יש מספר תווים שיכולים להתמתח אופקית, בצורה כזו הסופר יכול למלא את השורה כך שתיגמר במקום הנכון. מדובר ב-5 תווים שבהם מותר לבצע את המתיחה ד׳, ה׳, ל׳, ר׳, ת׳. ובמילה אחת (שהפכה למונח מקצועי) להדר״ת. המיוחד בתווים הללו שיש להם קו אופקי פשוט במרכז ואפשר להאריך את התו (ואת הקו האופקי) ללא פגיעה בצורה הכללית של התו. בתהליך של המרת התמונה צריך להתחשב בתווים האלו ולכווץ אותם בצורה מושכלת. באיור 19 ניתן לראות דוגמאות של ר׳ ות׳ מתוחות (מסומנות בצהוב) לעומת ר׳ ות׳ רגילות.



איור 19: אותיות ר ות מתוחות אופקית

בנוסף ישנו ע״פ המסורת מספר מצומצם של אותיות קטנות (״זעירות״) 6 בכל התורה, או גדולות (״רבתיות״) 10 בכל התורה לאורך ספר התורה. במקרים אלו שידועים מראש צריך להעביר את התו הגדלה/מזעור לפני התחלת הזיהוי, ואז להמשיך בתהליך כרגיל.

באיור 20 מתוארים השלבים שעוברת התמונה עד שמגיעה לגודל האחיד שהוא 21X28:

תיאור שלבי המעבר:

5.2.1 יצירת תמונת תו מרופדת

מציאת הגובה המקסימאלי והמינימאלי ביחס לשורה, וכן רוחב המקסימאלי לתו רגיל: לפני ההמרה עוברים על תמונות התווים בעמודה ומוצאים בספר הזה מה הגובה הגבוה (המרחק בין תחילת גובה השורה לגובה התו הגבוה ביותר), הגובה הנמוך (המרחק בין תחילת השורה למקום הנמוך ביותר שתו מגיע) והרוחב הרחב ביותר של תו רגיל (לא כולל את תווי להדר״ת). גדלים אלו יכולים להישמר בפרופיל הספר ולא צריך למצוא אותם מחדש אם הם כבר שמורים. בכל תמונת תו שמור גם מה מיקומו ביחס לשורה, כך שאפשר לצייר את התו במקום האמתי בתוך התמונה המרופדת. תו רגיל לעולם לא יחרוג מהתמונה המרופדת ובאשר לתווי להדר״ת:

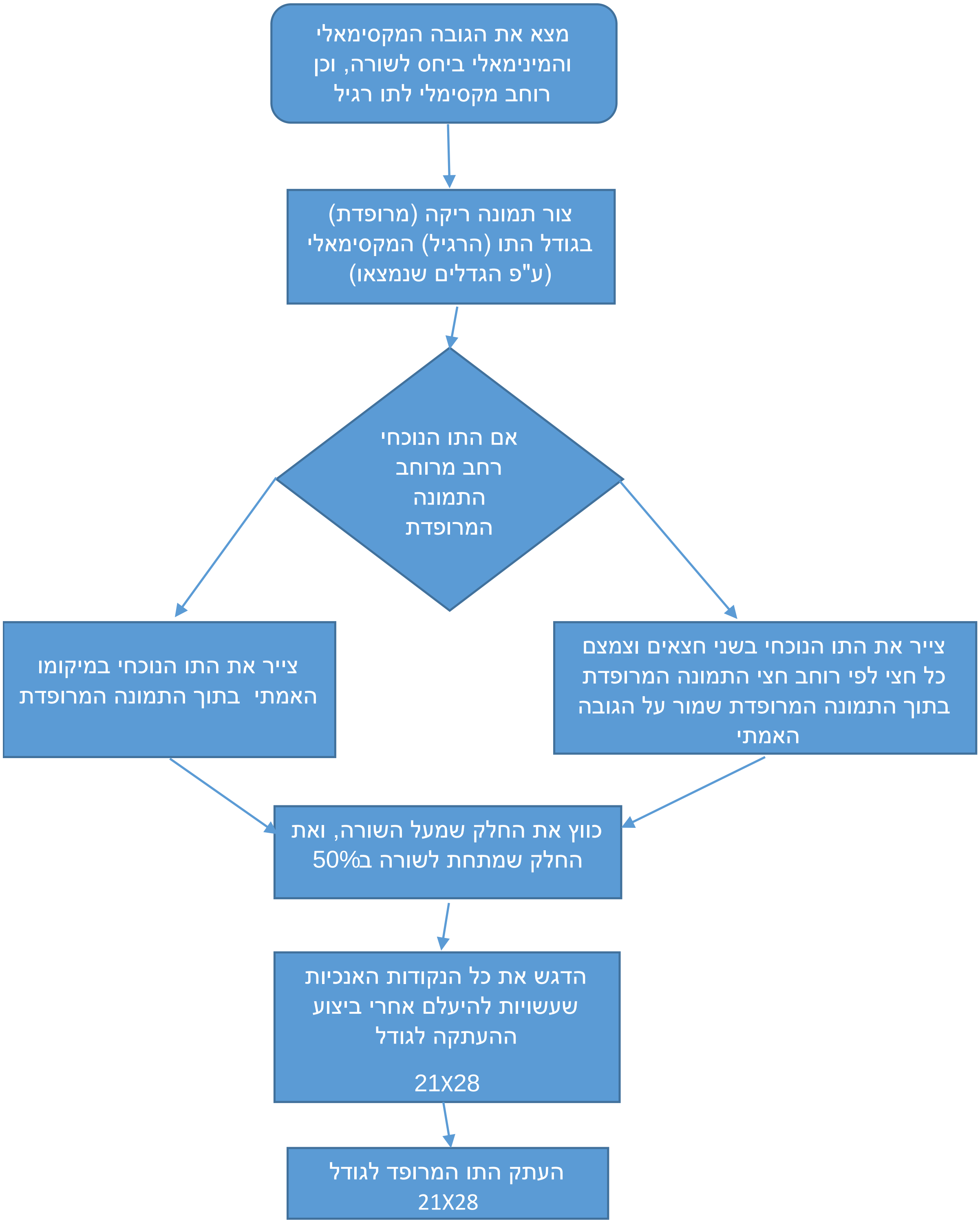
בדיקה אם התו רחב יותר מהתמונה המרופדת (תו מקבוצת להדר״ת) – כווץ את התמונה ברוחב בשתי פעימות - החלק השמאלי ואז החלק הימני. היות שהמרכז של כל התווים הוא קו אופקי הפגיעה בצורה הכללית היא מינימאלית. אם התו לא רחב מדיי העתק אותו במיקומו הנכון לתמונה המרופדת.

5.2.2 יצירת תמונה סופית בת השוואה

הבעיה בריפוד היא הוספה של כמות גדולה של פיקסלים לתמונה שלא תורמים מידע. עיקר התוספת המיותרת היא בחלקים שמעל ומתחת לשורה. על מנת להקטין משמעותית את האזורים האלו ולא לאבד את המידע עבור אותם תווים אנחנו מכווצים את האזורים בממד הגובה בחצי.

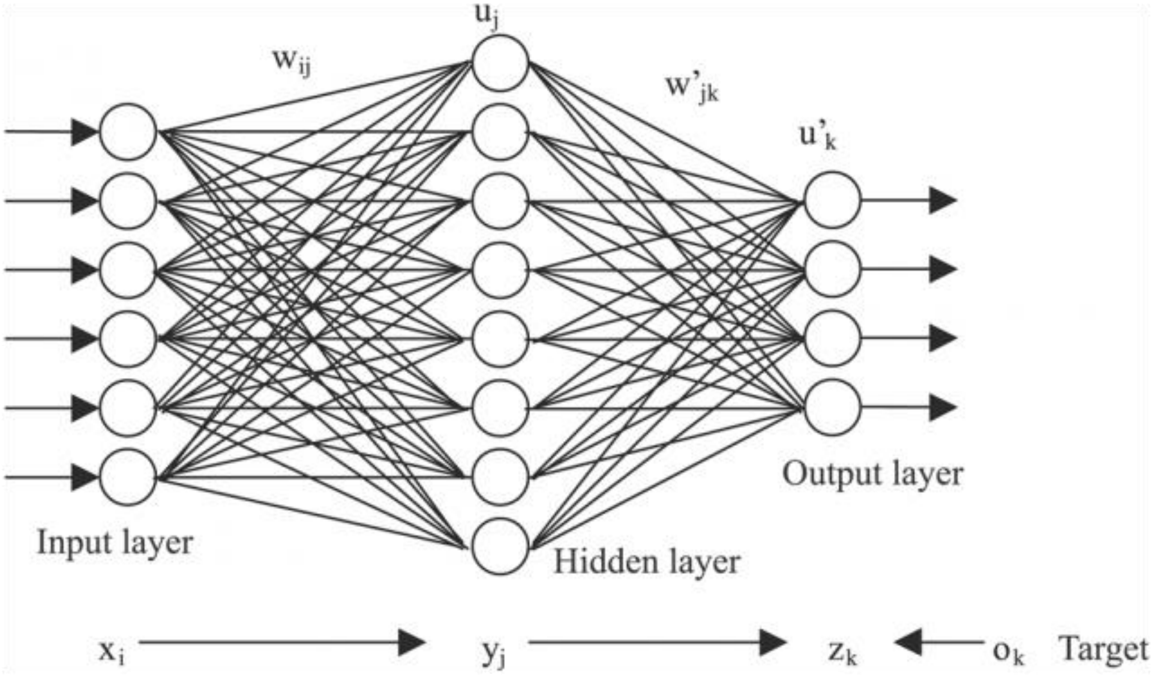
על מנת להימנע מאיבוד מידע במקרה של כווץ בקווים האנכיים (הקווים האופקיים עבים בכתיבת סת״ם) אנו עוברים על כל מקום שדק מדיי ומוסיפים פיקסלים להדגשה.

ראה בנספח א איור 26: תמונות תווים מקוריים וברי השוואה.



איור 20: תרשים זרימה המרה לתמונה בת השוואה

5.3 פעולת רשת הנוירונים



איור 21: מבנה רשת נוירונים

באיור 21 (לקוח מ(Templeton, 2015)) המשרטט מבנה כללי של כל רשת נוירונים, ניתן לראות שכבת הקלט (Input layer) שכבת הפלט (Output layer) ואת השכבה הנסתרת (Hidden layer) ואת הקשרים ביניהם.

רשת הנוירונים צריכה לקבל בשכבת הקלט את נקודות הכניסה שהם במקרה שלנו רשימה בינארית של 0/1 לפי הפיקסלים. לרשת יש נקודות יציאה בשכבת הפלט שבמקרה שלנו כל נקודת יציאה מייצגת את אחד התווים. בכל זיהוי נקודת היציאה המתאימה אמורה "להידלק".

במקרה שלנו יש 588 נקודות כניסה (הרזולוציה של תמונת כל תו היא 28 X 21) ו27 נקודות יציאה (22 אותיות בעברית + 5 סופיות).

אני משתמש בקוד פתוח של מימוש רשת נוירונים שכתב Johannes Amtén ונמצא ב: https://github.com/amten/NeuralNetwork/. זאת, לאחר שניסיתי כמה חבילות שלא היו מוצלחות.

חבילות אחרות הממשות רשת נוירונים שנוסו לצורך הפרויקט הן:

- Encog Machine Learning Framework (http://www.heatonresearch.com/encog/)

- Java Kohonen Neural Network Library (JKNNL) (http://jknnl.sourceforge.net/)

חבילות אלו, למרות שנחשבות לטובות ומומלצות, באחת (Encog) התוצאות שהגעתי אליהן היו עלובות ביותר, ואת השנייה (JKNNL) לא הצלחתי להפעיל בצורה טובה מהממשק הקיים שם.

רשת הנוירונים, לאחר שקיבלה את הקלט יוצרת 2 קבוצות, קבוצת אימון, וקבוצת בדיקה שאינה חלק מהאימון. בשלב ראשון מריצים את האימון. אורך האימון תלוי בכמות הנתונים. כאשר מדובר ב- 5 עמודות האימון רץ במשך כ-2 דקות. לאחר האימון (שבסופו מתבצעת בדיקה על חלק מקבוצת האימון – בדיקה זו אמורה תמיד להיות עם 100% הצלחה) אנחנו מריצים בדיקה על קבוצת הבדיקה ומקבלים את אחוזי ההצלחה.

ניתן "לשחק" עם כמות הנתונים ועם הקונפיגורציה של רשת הנוירונים. ברשת הנוירונים אפשר לקבוע את העומק שלה (מספר השכבות) ואת הרוחב שלה (מספר הנוירונים בכל שכבה). הקונפיגורציה שנתנה לי את התוצאות הטובות ביותר היא שכבה חבויה אחת בלבד עם 100 נוירונים. יותר שכבות של נוירונים הגדילו את זמן האימון אך לא הניבו תוצאות טובות יותר. יתכן שהסיבה לכך (לירידה בתוצאות) היא כמות הדוגמאות האימון שהשתמשתי בבדיקה. יתכן שבכמות גדולה יותר רשתות עמוקות יותר יהיו טובות יותר.

6 הגהה

בהגהה השלבים הראשונים של הפעולה דומים מאוד לשלבים הנעשים בפעולת הלימוד כמתואר בתרשים איור 2. בתחילה טוענים את נתוני ספר התורה, ולאחר מכן טוענים את התמונה של העמודה שקיבלנו כקלט ומריצים את עיבוד התמונה וחילוץ התווים רק שבשלב זה עושים שימוש בנתונים השמורים בפרופיל (ראה נספח ג) כבר לאותו ספר (או סופר). כאן ניתן לראות את חשיבות הנתונים השמורים – של הרוחב הצפוי של הרווחים בין המילים, רוחב התווים וכל גדלי האותיות המקסימאליים. ללא נתונים אלו על הספר אי אפשר היה לחלץ את התווים בהצלחה. כך מזהים את השורות, המילים התווים והרווחים. ברגע שיש את התווים והרווחים משווים לנתוני ספר התורה ובודקים את נכונות הזיהוי של התווים והרווחים. לאחר מכן מופק דו"ח המפרט את השגיאות החשודות.

שלב ההגהה מהיר מאוד באופן יחסי – כ-10 שניות בלבד לעמודה.

6.1 זיהוי תווים נוספים או חסרים

עוברים על המילים המזוהות. במקרה של חוסר התאמה במספר התווים שיש במילה מסוימת זיהינו טעות. אם מדובר בתו מיותר יש לנו "נתק" (תו אחד מזוהה כשניים), אם מדובר בתו חסר יש לנו "דבק". כל כישלון נרשם באובייקט הנתונים.

6.2 זיהוי רווחים

אחרי כל מילה שמור לנו גם סוג הרווח שבא בסיומה. אם זהו רווח מיוחד של סיום פרשיה, אנחנו בודקים אם באמת הרווח נמצא ואם זהו סוג הרווח הנכון. אם יש טעות, הטעות נרשמת באובייקט הנתונים.

6.3 זיהוי התווים

במידה ומספר התווים במילה מתאים למספר המצופה, מעבירים את תמונת התו למזהה התווים ברשת הנוירונים. רשת הנוירונים מחזירה את התו המזוהה על ידה וגם את וודאות הזיהוי. במידה וודאות הזיהוי נמוכה מ-90 אחוז, אנחנו לוקחים את 3 התוצאות הגבוהות ביותר. (רשת הנוירונים מספקת גם את הנתון הזה.)

6.4 איסוף הסטטיסטיקה

על כישלון של נתק או דבק אנחנו מעלים מונה של כישלונות במילים. אם מילה זוהתה מבחינת מספר תווים, מעלים מונה של הצלחה במילה.

על כישלון בזיהוי רווח מיוחד מעלים מונה של כישלונות ברווחים (אין הפרדה בין חוסר זיהוי של רווח קיים או זיהוי של רווח לא קיים), אם הייתה הצלחה מעלים מונה של הצלחה.

על כישלון בזיהוי תו מעלים מונה של כישלונות בזיהוי תו. בנוסף בודקים אם היה זיהוי וודאי או לא. וודאי לצורך העניין זה למעלה מ90%. סופרים את הכישלונות בזיהוי וודאי, ובזיהוי מסופק. במקביל סופרים גם את ההצלחה בתו מזוהה ודאית ובתו שאינו זוהה בוודאות. במקרה של כישלון בזיהוי לא ודאי בודקים גם את שני התווים הבאים אחרי הזיהוי, אם היה זיהוי באחד מהם, אם כן זה נספר במונה נוסף.

סופרים גם את הכישלונות וההצלחות לכל תו בנפרד (לכל תו יש מונה של הצלחות וכישלונות שמועלים אם יש הצלחה או כישלון בהתאמה.)

התוכנה תיחשב כמוצלחת לפי אחוז זיהוי הטעויות (`false negative`) אבל גם לפי אחוז זיהוי הטעיות השגוי (`false positive`). כלומר אם התוכנה מתריעה על טעויות אבל מתריעה על כתב תקין כטעות אז ישנה בעיה אע"פ שהתוכנה מוצאת טעויות. אם מדובר בהרבה מידיי התרעות שווא התוכנה לא מוצלחת.

7 תוצאות

7.1 תוצאות הגהת עמודה

אציג כאן תוצאות מפורטות של הגהת עמודה מספר תורה (ראה איור 22). מדובר על העמודה שניה של הספר. בהמשך אביא תוצאות של עמודות נוספות. הצילום של הספר נמצא באינטרנט באתר http://www1.saad.org.il/elihu, לא מצאתי עוד תמונות מסודרות שלמות של עמודות ספר תורה ברשת. בכל מקרה מדובר בתמונות אותנטיות שלא "בושלו" על ידי. הספר כתוב יפה וברור באופן יחסי, אך הצילום לא נעשה בצורה מקצועית. רבים מהעמודים בספר מצולמים כאשר הקלף לא משוטח מה שיוצר גלים ואת הבעיה זו לא פותר אלגוריתם היישור, שמתייחס למצב שהקלף משוטח אך נוטה בזווית. בנוסף התאורה לא זהה בצילומים שונים וכך גווני האפור של כל צילום משתנים, ובחלק מהמקרים בתמונה עצמה גווני האפור משתנים.

ללמד את התוכנה לקחתי 5 עמודים (אחרים) מהספר 1,3,4,5,6.

7.1.1 תוצאות הגהת עמודה בדו"ח מילולי

בטבלה 1 ניתן לראות תוצאות ההגהה:

מונה	תוצאה
תווים שזוהו בהצלחה (מתוך 1257)	1249 (99.36%)
תווים שלא זוהו (סה"כ זיהויים שגויים של תווים)	8
מילים שלא זוהו (דבק או נתק) – מספר תווים שגוי	0
מילים שזוהו	338
תווים שזוהו בוודאות (מעל 90% של זיהוי לעומת אפשרויות אחרת) בלי קשר לנכונות	1230
תווים שזוהו בזיהוי מסופק (זוהו בפחות מ90%) בלי קשר לנכונות הזיהוי	27
תווים שזוהו בוודאות ונכשלו בזיהוי (התוכנה זיהתה בוודאות אך טעתה)	0

תווים שזוהו בהצלחה אך בזיהוי מסופק (הזיהוי היה בפחות מ90% אך נכון)	19
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי (זיהויים בפחות מ90% שנכשלו)	8
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי, אך הזיהוי של התו הבא היה נכון (מקרים שהאפשרות השנייה בטיבה בזיהוי מסופק שנכשל הייתה נכונה)	6
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי, אך הזיהוי של אחד משני התווים הבאים היה נכון (מקרים שאחד משני התווים הבאים היה נכון בזיהוי מסופק שנכשל)	8
זיהוי רווחים מוצלח (זיהוי רווחים של פרשיות פתוחות וסגורות)	2
טעויות רווחים: (זיהוי רווחים לא קיימים או אי זיהוי רווחים קיימים)	0

טבלה1 : מונים מדוח הגהת עמודה 2 מספר בראשית

התפלגות הזיהויים הנכונים לפי תווים:

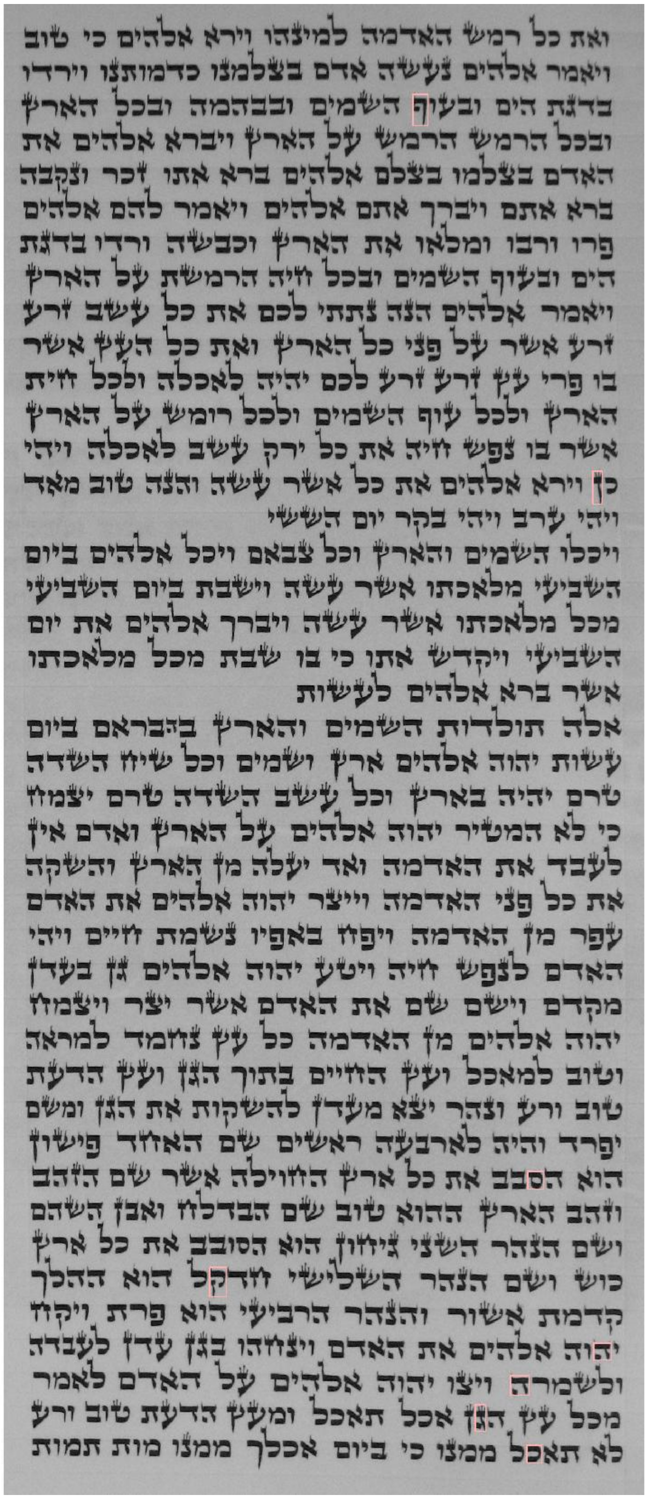
א - 118; ב - 64; ג - 7; ד - 38; ה - 152; ו - 132; ז - 7; ח - 18; ט - 10; י - 117; ך - 5; כ - 44; ל - 91; ם - 63; מ - 55; ן - 15; נ - 22; ס - 1; ע - 47; ף - 2; פ - 12; ץ - 23; צ - 10; ק - 9; ר - 76; ש - 63; ת - 48;

התפלגות הזיהויים הלא נכונים:

ס - 1; ג - 1; ף - 1; ה - 2; ק - 1; כ - 1; ן - 1;

7.1.2 תוצאות הגהת עמודה בתמונה

התוכנה מוציאה תמונה של התוצאה. בתמונה מופיעים התווים והמילים שלא זוהו. המילים במסגרת סגולה, התווים שלא זוהו אך הזיהוי היה ״מסופק״ במסגרת ורודה, ותווים שלא זוהו בזיהוי ״ודאי״ במסגרת אדומה. באיור 22 ניתן לראות את כל התווים שלא זוהו. כל הטעויות היו בזיהוי ״מסופק״.



איור 22: תוצאה גרפית של הגהת עמודה 2

7.2 ניתוח ראשוני של התוצאות

בעמוד זה אין טעויות אמתיות, כלומר שברמה אידיאלית היה צריך להיות 0 טעויות וזיהוי מלא של כל התווים. בכל זאת היו מספר תווים (8) שהתוכנה נכשלה בזיהויים – התריעה על טעות אף שלא הייתה כזו (false positive). אחוז הזיהוי של התווים הוא:

$$\frac{1249}{1249 + 8} = 99.36\%$$

זהו אחוז זיהוי מצוין לכתב יד באופן כללי. נכון שזיהוי כתב בספר תורה שונה מכיוון שכללי הכתיבה נוקשים הרבה יותר, אבל לספר תורה, כפי שציינו, קשיים משלו (תווים מאורכים לצורך התאמה לשורה ותווים דומים). בהמשך ננסה לעמוד על הסיבות לטעות באותם התווים.

כפי שהסברנו את הזיהויים אנחנו מחלקים לזיהוי ודאי (מעל 90%) וזיהוי מסופק. נתון מעניין נוסף הוא שלא הייתה שום טעות בזיהויים הוודאיים! כל 8 הטעויות היו בזיהויים מסופקים. זיהויים ודאיים היו 97.85%.

לגבי מדדי ההצלחה המקובלים ראה מדדי הצלחה.

נתון מעניין נוסף הוא שב-6 מתוך 8 הטעויות התו הנכון הוא התו הבא הקרוב לזיהוי. בשני הזיהויים השגויים הבאים התו השני הוא התו הנכון. נתון זה מאפשר לבנות שיפור משמעותי בזיהוי אם נבנה מנגנון המסוגל לבחור מבין התווים המועמדים הראשונים כאשר הזיהוי לא ודאי.

לגבי בעיות של הידבקויות ונתקים לא זוהו שגיאות - כמצופה.

מעבר לכך אנו רואים שהרווחים זוהו בהצלחה.

7.3 ניתוח נוסף של הטעויות

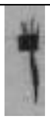
בטבלה 2 הבאתי פירוט של 8 הטעויות.

התו שלא זוהה	הזיהוי השגוי	המועמדים הבאים
ף	פ	ב, ף
ן	ג	ז, ן
ס	ם	ס, כ
ק	ח	ק, כ
ה	ת	ה, ק
ה	ת	ה, ד
ג	נ	ג, ב
כ	ב	כ, ם

טבלה 2: פירוט טעויות הזיהוי של עמוד 2 בספר בראשית

ניתן לראות שהטעויות "הגיוניות" כלומר להחליף ס ו-ם, נ ו-ג ב ו-כ אלו שגיאות צפויות. גם להחליף ה ו-ת אפשר להבין. מזה שטעות זו קרתה פעמיים ברור שיש כאן נקודה שצריך ואפשר לשפר. אפשר לשפר בתהליך נוסף שאפשר להעביר לגבי זיהוי של תווים אלו (במיוחד זיהוי מסופק) שרק יצטרך לבדוק ולהכריע בין התווים הקרובים. אפשר גם לבדוק שיפור שאפשר לבצע ביצירת התו בר ההשוואה שיתכן שמעלים אזור משמעותי לזיהוי.

מעבר לכך ניתן לראות שבמקומות של חשד לטעויות יש באמת שינוי מדרך הכתיבה הרגילה של התו בידי הסופר, וראוי להסב את תשומת ליבו לכך גם אם לא מדובר בטעות פורמאלית. לדוגמה נראה את התו 'ן' שלא זוהה בידי התוכנה. נראה בטבלה3 את התו ולאחריו 4 תווים של 'ן' שכן זוהו.

טבלה3 : ה-'ן' שלא זוהתה ולאחריה 4 דוגמאות שזוהו

אפשר להבין מדוע רשת הנוירונים טעתה "חשבה" שמדובר פה ב-'ג' או ב-'ז'.

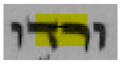
התוצאה בתמונה

המבחן הבא של התוכנה יהיה לזהות טעויות שנשתול בעמודה זו.

7.4 הגהת עמודה עם 2 טעויות

באותה עמודה שתלתי שתי טעויות:

1. במילה "ורדו" חיברתי חיבור בעובי של פיקסל בין הר לד:



2. במילה והארץ הוספתי חיבור קטן באות ה כך:

7.4.1 הגהת תמונה עם טעויות – דו"ח מילולי

נריץ את ההגהה על התמונה ונראה את התוצאות:

מונה	תוצאה
תווים שזוהו בהצלחה (מתוך 1257)	1244
תווים שלא זוהו (סה"כ זיהויים שגויים של תווים)	9

מילים שלא זוהו (דבק או נתק) – מספר תווים שגוי	1
מילים שזוהו	337
תווים שזוהו בוודאות (מעל 90% של זיהוי לעומת אפשרויות אחרת) בלי קשר לנכונות	1226
תווים שזוהו בזיהוי מסופק (זוהה בפחות מ90%) בלי קשר לנכונות הזיהוי	27
תווים שזוהו בוודאות ונכשלו בזיהוי	1
תווים שזוהו בהצלחה אך בזיהוי מסופק (הזיהוי היה בפחות מ90% אך נכון)	19
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי (זיהויים בפחות מ90% שנכשלו)	8
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי, אך הזיהוי של התו הבא היה נכון (מקרים שהאפשרות השנייה בטיבה בזיהוי מסופק שנכשל הייתה נכונה)	6
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי, אך הזיהוי של אחד משני התווים הבאים היה נכון (מקרים שאחד משני התווים הבאים היה נכון בזיהוי מסופק שנכשל)	8
זיהוי רווחים מוצלח (זיהוי רווחים של פרשיות פתוחות וסגורות)	2
טעויות רווחים: (זיהוי רווחים לא קיימים או אי זיהוי רווחים קיימים)	0

התפלגות הזיהויים הנכונים לפי תווים:

א - 118; ב - 64; ג - 7; ד - 37; ה - 151; ו - 130; ז - 7; ח - 18; ט - 10; י - 117; ך - 5; כ - 44; ל - 91; ם - 63; מ - 55; ן - 15; נ - 22; ס - 1; ע - 47; ף - 2; פ - 12; ץ - 23; צ - 10; ק - 9; ר - 75; ש - 63; ת - 48;

התפלגות הזיהויים הלא נכונים:

ס - 1; ג - 1; ף - 1; ה - 3; ק - 1; כ - 1; ן - 1;

אפשר לראות שהתוכנה זיהתה את שתי הטעויות:

במונה המילים הבלתי מזוהות יש עכשיו 1 – ובהדפסת הlogs:

text word **ורדו**, (perkek: א, pasuk: כח), index: 12 has different length than letters images (3)

במונה התווים שזוהו בוודאות ונכשלו בזיהוי יש עכשיו 1 – בזכות הטעות ששתלנו בה של המילה 'והארץ'.

אפשר לראות את הטעות באות ה בהתפלגות הזיהויים הלא נכונים והנכונים.

7.4.2 מדדי הצלחה

בדרך כלל בתוכנות למידה חישובית מקובל להגדיר את המדדים של precision, recall, accuracy שהם שקלול של ההצלחה והכישלון של הזיהוי וחוסר הזיהוי.

בתוכנה למציאת טעויות נגדיר את הטעויות שנמצאו כtrue positive (2)

האותיות שזוהו בהצלחה כtrue negative (1244)

הטעויות שזוהו בטעות false positive (8)

טעויות שלא זוהו false negative (0)

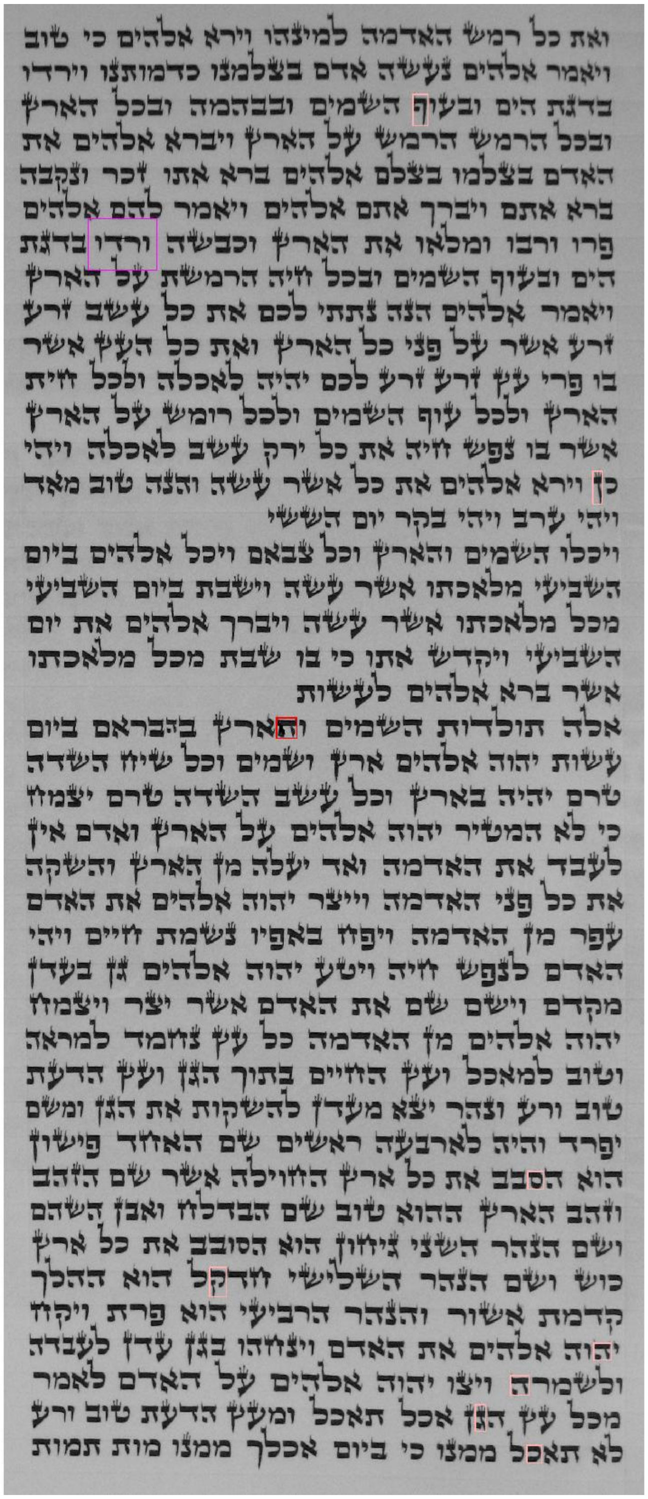
ומכאן נגזור את המדדים:

Accuracy: 0.99, Recall: 1, Precision: 0.2

במקרה זה שבו מראש מדובר על מעט טעויות שאמורות להימצא המדדים קצת מטעים (לשני הכיוונים) אך המדד הנמוך של הprecision אכן מבליט את הבעיה הכללית בתוכנה שיש הרבה התראות שווא (גם אם ההתראות הן מסופקות וגם אם מדובר ביחסית מעט טעויות לעומת כמות האותיות שלא גרמו להתראת שווא).

7.4.3 הגהת עמודה עם טעויות – תצוגה גרפית

באיור 23 ניתן לראות את המילה "ורדו" במסגרת סגולה ואת האות "ה" במילה והארץ במסגרת אדומה. (שאר הטעויות נשארו במקומם.)



ואת כל רמש האדמה למינהו וירא אלהים כי טוב
ויאמר אלהים נעשה אדם בצלמנו כדמותנו וירדו
בדגת הים ובעוף השמים ובבהמה ובכל הארץ
ובכל הרמש הרמש על הארץ ויברא אלהים את
האדם בצלמו בצלם אלהים ברא אתו זכר ונקבה
ברא אתם ויברך אתם אלהים ויאמר להם אלהים
פרו ורבו ומלאו את הארץ וכבשה ורדו בדגת
הים ובעוף השמים ובכל חיה הרמשת על הארץ
ויאמר אלהים הנה נתתי לכם את כל עשב זרע
זרע אשר על פני כל הארץ ואת כל העץ אשר
בו פרי עץ זרע זרע לכם יהיה לאכלה ולכל חית
הארץ ולכל עוף השמים ולכל רומש על הארץ
אשר בו נפש חיה את כל ירק עשב לאכלה ויהי
כן וירא אלהים את כל אשר עשה והנה טוב מאד
ויהי ערב ויהי בקר יום הששי
ויכלו השמים והארץ וכל צבאם ויכל אלהים ביום
השביעי מלאכתו אשר עשה וישבת ביום השביעי
מכל מלאכתו אשר עשה ויברך אלהים את יום
השביעי ויקדש אתו כי בו שבת מכל מלאכתו
אשר ברא אלהים לעשות
אלה תולדות השמים והארץ בהבראם ביום
עשות יהוה אלהים ארץ ושמים וכל שיח השדה
טרם יהיה בארץ וכל עשב השדה טרם יצמח
כי לא המטיר יהוה אלהים על הארץ ואדם אין
לעבד את האדמה ואד יעלה מן הארץ והשקה
את כל פני האדמה וייצר יהוה אלהים את האדם
עפר מן האדמה ויפח באפיו נשמת חיים ויהי
האדם לנפש חיה ויטע יהוה אלהים גן בעדן
מקדם וישם שם את האדם אשר יצר ויצמח
יהוה אלהים מן האדמה כל עץ נחמד למראה
וטוב למאכל ועץ החיים בתוך הגן ועץ הדעת
טוב ורע ונהר יצא מעדן להשקות את הגן ומשם
יפרד והיה לארבעה ראשים שם האחד פישון
הוא הסבב את כל ארץ החוילה אשר שם הזהב
וזהב הארץ ההוא טוב שם הבדלח ואבן השהם
ושם הנהר השני גיחון הוא הסובב את כל ארץ
כוש ושם הנהר השלישי חדקל הוא ההלך
קדמת אשור והנהר הרביעי הוא פרת ויקח
יהוה אלהים את האדם וינחהו בגן עדן לעבדה
ולשמרה ויצו יהוה אלהים על האדם לאמר
מכל עץ הגן אכל תאכל ומעץ הדעת טוב ורע
לא תאכל ממנו כי ביום אכלך ממנו מות תמות

איור 23 - תוצאת הגהת עמודה 2 עם 2 טעויות שתולות

7.5 תוצאות של עמודות נוספות

מונה	עמודה 13	עמודה 22	עמודה 30
תווים שזוהו בהצלחה	1256 / 1261 (99.60%)	1319 / 1323 99.70%	1337 / 1343 (99.55%)
תווים שלא זוהו (סה"כ זיהויים שגויים של תווים)	5	4	6
מילים שלא זוהו (דבק או נתק) – מספר תווים שגוי	0	0	1
מילים שזוהו	337	349	355
תווים שזוהו בוודאות (מעל 90% של זיהוי לעומת אפשרויות אחרת) בלי קשר לנכונות	1240	1310	1310
תווים שזוהו בזיהוי מסופק (זוהו בפחות מ90%) בלי קשר לנכונות הזיהוי	21	13	33
תווים שזוהו בוודאות ונכשלו בזיהוי (התוכנה זיהתה בוודאות אך טעתה)	1	2	0
תווים שזוהו בהצלחה אך בזיהוי מסופק (הזיהוי היה בפחות מ90% אך נכון)	17	11	27
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי (זיהויים בפחות מ90% שנכשלו)	4	2	6
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי, אך הזיהוי של התו הבא היה נכון (מקרים שהאפשרות השנייה בטיבה בזיהוי מסופק שנכשל הייתה נכונה)	2	2	5
תווים שזוהו בזיהוי מסופק ונכשלו בזיהוי, אך הזיהוי של אחד משני התווים הבאים היה נכון (מקרים שאחד משני התווים הבאים היה נכון בזיהוי מסופק שנכשל)	3	2	5

זיהוי רווחים מוצלח (זיהוי רווחים של פרשיות פתוחות וסגורות)	1	1	0
טעויות רווחים: (זיהוי רווחים לא קיימים או אי זיהוי רווחים קיימים)	0	0	0

בחרתי ב 3 עמודות נוספות של צילום טוב, ושאנחנו מגיעים בהן לאחוז זיהוי גבוה במיוחד (עד 99.7%!).

התפלגות הזיהויים הלא נכונים (בכל 4 העמודות):

ס - 2; ו - 1; ז - 2; כ - 3; ל - 1; ה - 4; י - 1; ג - 5; ק - 2; ם - 1; ן – 1; ף – 1

ניתן לראות טעויות בין התווים הדומים: קבוצת התווים הרזים ג, ו, ז, י, ן. טעויות בין ס, ם ובין ק לה.

ניתן גם לראות שרובן הגדול של הטעויות הן בזיהויים מסופקים ולא זיהויים ודאיים.

באיור24 ניתן לראות שיש זיהוי ודאי של ו במילה ויהי שנכשל. התוכנה זיהתה י. במקרה הזה התוכנה זיהתה בעיה אמתית. אמנם זה מקרה גבולי שלא בטוח שפוסל את הספר, אבל מדובר בהחלט בתיקון שצריך להיעשות להאריך את הרגל של ה״ו״ שלא יראה כמו י.

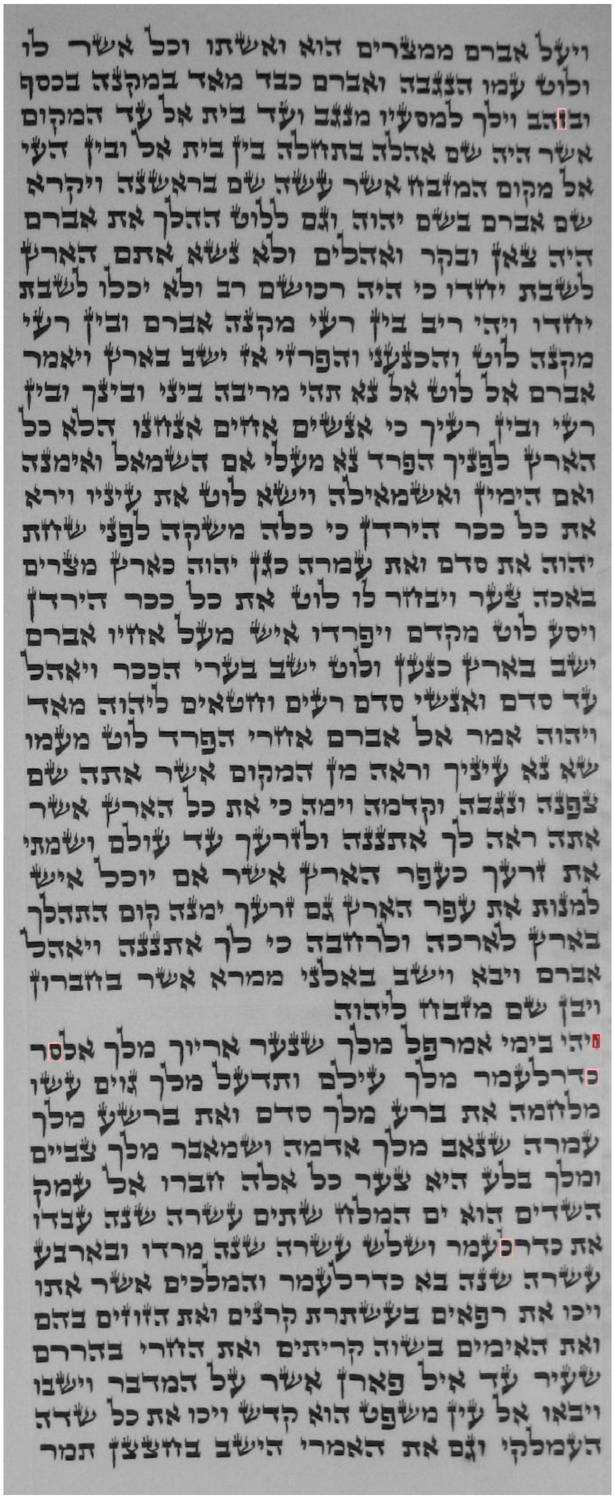
גם האות כ במילה כדרלעמר בעייתית. היא קטנה מידיי וצריכה שיפור.

האות ל לא זוהתה מכיוון שיש בה קטיעה מבחינת התוכנה. בתמונה המקורית אין קטיעה מוחלטת רק אזור דהוי של דיו שמסווג כשטח לבן. גם פה הסופר צריך להדגיש מחדש את האזור הדהוי שעשוי לדהות עוד ואז תהיה קטיעה.

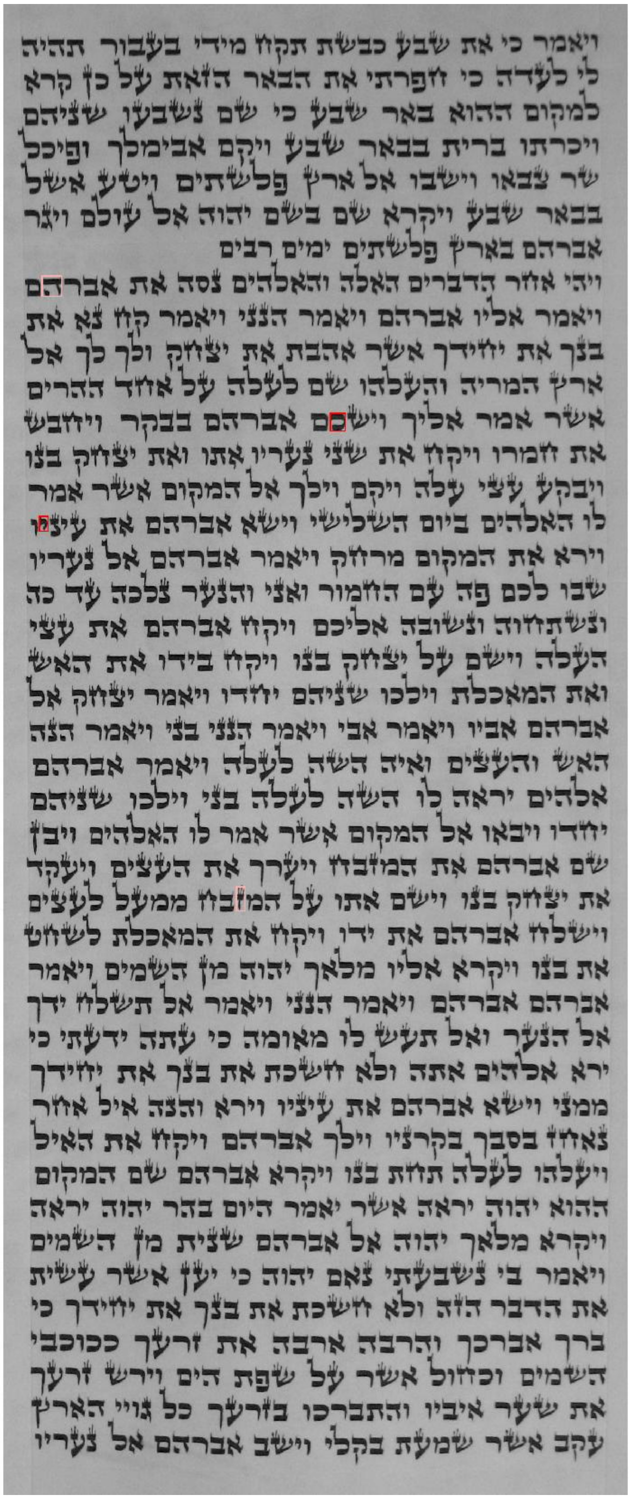
הטעויות ב״ז״ של ובזהב ו״ס״ במילה אלסר פחות ברורות אך (במיוחד ב״ז״) בהחלט ניתן לראות שינוי מהכתיבה הרגילה.

באיור 25 תוצאת הגהה לעמודה 22איור 25 הדבר בולט הוא מיעוט השגיאות שנמצאו – רק 4 מהתווים לא זוהו (אחוז זיהוי של 99.7). השגיאות שהתוכנה מצאה הן באמת זניחות כלומר שינויים קלים בלבד שלא נחשבות שגיאות אמתיות. (באות י נראה שהתוכנה רגישה לרוחב של הקו האנכי.)

ב התוכנה מתריעה על האות ק׳ (מזוהה כ ה׳), כנראה בשל המיקום של הקו האנכי התחתון שנמצא בסוף התו. מעבר לכך התוכנה לא מצליחה לזהות את האות ג׳ בארבע הופעות (17 הופעות אחרות בעמודה מזוהות נכון). בשלשה מקרים התכנה מזהה נ ובמקרה אחד (שבו הרגל של הג׳ בולטת במיוחד) התוכנה מזהה ב׳. זו דוגמה לתו שכדאי לעשות מפת נוירונים שתבחין בין ג׳ לנ׳ ואולי גם לב׳. גם לקבוצות התווים ס׳, ס׳ וו׳, ז׳, י׳, ה׳ ח׳ וק׳ ניתן להפעיל זיהוי משני שישפר עוד את התוצאות.


ויעל אברם ממצרים הוא ואשתו וכל אשר לו
ולוט עמו הנגבה ואברם כבד מאד במקנה בכסף
ובזהב וילך למסעיו מנגב ועד בית אל עד המקום
אשר היה שם אהלה בתחלה בין בית אל ובין העי
אל מקום המזבח אשר עשה שם בראשנה ויקרא
שם אברם בשם יהוה וגם ללוט ההלך את אברם
היה צאן ובקר ואהלים ולא נשא אתם הארץ
לשבת יחדו כי היה רכושם רב ולא יכלו לשבת
יחדו ויהי ריב בין רעי מקנה אברם ובין רעי
מקנה לוט והכנעני והפרזי אז ישב בארץ ויאמר
אברם אל לוט אל נא תהי מריבה ביני ובינך ובין
רעי ובין רעיך כי אנשים אחים אנחנו הלא כל
הארץ לפניך הפרד נא מעלי אם השמאל ואימנה
ואם הימין ואשמאילה וישא לוט את עיניו וירא
את כל ככר הירדן כי כלה משקה לפני שחת
יהוה את סדם ואת עמרה כגן יהוה כארץ מצרים
באכה צער ויבחר לו לוט את כל ככר הירדן
ויסע לוט מקדם ויפרדו איש מעל אחיו אברם
ישב בארץ כנען ולוט ישב בערי הככר ויאהל
עד סדם ואנשי סדם רעים וחטאים ליהוה מאד
ויהוה אמר אל אברם אחרי הפרד לוט מעמו
שא נא עיניך וראה מן המקום אשר אתה שם
צפנה ונגבה וקדמה וימה כי את כל הארץ אשר
אתה ראה לך אתננה ולזרעך עד עולם ושמתי
את זרעך כעפר הארץ אשר אם יוכל איש
למנות את עפר הארץ גם זרעך ימנה קום התהלך
בארץ לארכה ולרחבה כי לך אתננה ויאהל
אברם ויבא וישב באלני ממרא אשר בחברון
ויבן שם מזבח ליהוה
ויהי בימי אמרפל מלך שנער אריוך מלך אלסר
כדרלעמר מלך עילם ותדעל מלך גוים עשו
מלחמה את ברע מלך סדם ואת ברשע מלך
עמרה שנאב מלך אדמה ושמאבר מלך צביים
ומלך בלע היא צער כל אלה חברו אל עמק
השדים הוא ים המלח שתים עשרה שנה עבדו
את כדרלעמר ושלש עשרה שנה מרדו ובארבע
עשרה שנה בא כדרלעמר והמלכים אשר אתו
ויכו את רפאים בעשתרת קרנים ואת הזוזים בהם
ואת האימים בשוה קריתים ואת החרי בהררם
שעיר עד איל פארן אשר על המדבר וישבו
ויבאו אל עין משפט הוא קדש ויכו את כל שדה
העמלקי וגם את האמרי הישב בחצצן תמר

איור 24 תוצאת הגהת עמודה 13



ויאמר כי את שבע כבשת תקח מידי בעבור תהיה
לי לעדה כי חפרתי את הבאר הזאת על כן קרא
למקום ההוא באר שבע כי שם נשבעו שניהם
ויכרתו ברית בבאר שבע ויקם אבימלך ופיכל
שר צבאו וישבו אל ארץ פלשתים ויטע אשל
בבאר שבע ויקרא שם בשם יהוה אל עולם ויגר
אברהם בארץ פלשתים ימים רבים
ויהי אחר הדברים האלה והאלהים נסה את אברהם
ויאמר אליו אברהם ויאמר הנני ויאמר קח נא את
בנך את יחידך אשר אהבת את יצחק ולך לך אל
ארץ המריה והעלהו שם לעלה על אחד ההרים
אשר אמר אליך וישכם אברהם בבקר ויחבש
את חמרו ויקח את שני נעריו אתו ואת יצחק בנו
ויבקע עצי עלה ויקם וילך אל המקום אשר אמר
לו האלהים ביום השלישי וישא אברהם את עיניו
וירא את המקום מרחק ויאמר אברהם אל נעריו
שבו לכם פה עם החמור ואני והנער נלכה עד כה
ונשתחוה ונשובה אליכם ויקח אברהם את עצי
העלה וישם על יצחק בנו ויקח בידו את האש
ואת המאכלת וילכו שניהם יחדו ויאמר יצחק אל
אברהם אביו ויאמר אבי ויאמר הנני בני ויאמר הנה
האש והעצים ואיה השה לעלה ויאמר אברהם
אלהים יראה לו השה לעלה בני וילכו שניהם
יחדו ויבאו אל המקום אשר אמר לו האלהים ויבן
שם אברהם את המזבח ויערך את העצים ויעקד
את יצחק בנו וישם אתו על המזבח ממעל לעצים
וישלח אברהם את ידו ויקח את המאכלת לשחט
את בנו ויקרא אליו מלאך יהוה מן השמים ויאמר
אברהם אברהם ויאמר הנני ויאמר אל תשלח ידך
אל הנער ואל תעש לו מאומה כי עתה ידעתי כי
ירא אלהים אתה ולא חשכת את בנך את יחידך
ממני וישא אברהם את עיניו וירא והנה איל אחר
נאחז בסבך בקרניו וילך אברהם ויקח את האיל
ויעלהו לעלה תחת בנו ויקרא אברהם שם המקום
ההוא יהוה יראה אשר יאמר היום בהר יהוה יראה
ויקרא מלאך יהוה אל אברהם שנית מן השמים
ויאמר בי נשבעתי נאם יהוה כי יען אשר עשית
את הדבר הזה ולא חשכת את בנך את יחידך כי
ברך אברכך והרבה ארבה את זרעך ככוכבי
השמים וכחול אשר על שפת הים וירש זרעך
את שער איביו והתברכו בזרעך כל גויי הארץ
עקב אשר שמעת בקלי וישב אברהם אל נעריו

איור 25 תוצאת הגהה לעמודה 22

ואמשך בני האתה זה בני עשו אם לא ויגש יעקב
אל יצחק אביו וימשהו ויאמר הקל קול יעקב
והידים ידי עשו ולא הכירו כי היו ידיו כידי עשו
אחיו שערת ויברכהו ויאמר אתה זה בני עשו
ויאמר אני ויאמר הגשה לי ואכלה מציד בני למען
תברכך נפשי ויגש לו ויאכל ויבא לו יין וישת
ויאמר אליו יצחק אביו גשה נא ושקה לי בני
ויגש וישק לו וירח את ריח בגדיו ויברכהו ויאמר
ראה ריח בני כריח שדה אשר ברכו יהוה ויתן
לך האלהים מטל השמים ומשמני הארץ ורב דגן
ותירש יעבדוך עמים וישתחוו לך לאמים הוה
גביר לאחיך וישתחוו לך בני אמך ארריך ארור
ומברכיך ברוך ויהי כאשר כלה יצחק לברך את
יעקב ויהי אך יצא יצא יעקב מאת פני יצחק אביו
ועשו אחיו בא מצידו ויעש גם הוא מטעמים ויבא
לאביו ויאמר לאביו יקם אבי ויאכל מציד בנו
בעבר תברכני נפשך ויאמר לו יצחק אביו מי אתה
ויאמר אני בנך בכרך עשו ויחרד יצחק חרדה
גדלה עד מאד ויאמר מי אפוא הוא הצד ציד ויבא
לי ואכל מכל בטרם תבוא ואברכהו גם ברוך יהיה
כשמע עשו את דברי אביו ויצעק צעקה גדלה
ומרה עד מאד ויאמר לאביו ברכני גם אני אבי
ויאמר בא אחיך במרמה ויקח ברכתך ויאמר הכי
קרא שמו יעקב ויעקבני זה פעמים את בכרתי
לקח והנה עתה לקח ברכתי ויאמר הלא אצלת לי
ברכה ויען יצחק ויאמר לעשו הן גביר שמתיו לך
ואת כל אחיו נתתי לו לעבדים ודגן ותירש סמכתיו
ולכה אפוא מה אעשה בני ויאמר עשו אל אביו
הברכה אחת הוא לך אבי ברכני גם אני אבי וישא
עשו קלו ויבך ויען יצחק אביו ויאמר אליו הנה
משמני הארץ יהיה מושבך ומטל השמים מעל
ועל חרבך תחיה ואת אחיך תעבד והיה כאשר
תריד ופרקת עלו מעל צוארך וישטם עשו את
יעקב על הברכה אשר ברכו אביו ויאמר עשו
בלבו יקרבו ימי אבל אבי ואהרגה את יעקב אחי
ויגד לרבקה את דברי עשו בנה הגדל ותשלח
ותקרא ליעקב בנה הקטן ותאמר אליו הנה עשו
אחיך מתנחם לך להרגך ועתה בני שמע בקלי
וקום ברח לך אל לבן אחי חרנה וישבת עמו ימים
אחדים עד אשר תשוב חמת אחיך עד שוב אף
אחיך ממך ושכח את אשר עשית לו ושלחתי
ולקחתיך משם למה אשכל גם שניכם יום אחד

8 סיכום ומבט להמשך

הצלחנו להגיע למערכת שמחזיקה את המבנה הנתונים של ספר תורה ומצליחה לקרא צילום של עמודה של ספר תורה. המערכת כוללת פעולות של עיבוד תמונה ומאפשרת לחלץ מתוך העמוד את התווים המילים והרווחים.

כאשר המערכת מקבלת עמוד להגהה המערכת יודעת לזהות את התווים הכתובים בכתב היד, ולהצביע על טעויות.

למרות שפיתחנו כאן מערכת שהגיעה להישגים משמעותיים בהגהה ממוחשבת, יש עוד כברת דרך רבה כדי להגיע למוצר טוב ונוח שיוכל לעשות לספר תורה הגהה ממוחשבת ברמה טובה עוד יותר:

א. לא בניתי ממשק למשתמש. ממשק למשתמש צריך לכלול ממשק נפרד לפעולת הלמידה (אימון) וממשק לפעולת ההגהה. הממשק לפעולת הלמידה צריך לכלול חלון של browser לבחירת קבצים כדי לממש אפשרות נוחה להוספת קבצים. הממשק צריך לכלול תצוגה של הפרופיל שאליו משויכת הלמידה. צריכה גם להיות אפשרות לראות את כל הפרמטרים שנמצאים בפרופיל (נספח ב') לצורך אימון וגם לאפשר לפחות לחלק מהפרמטרים שינוי ידני. כחלק מביצוע הלמידה צריך להוסיף אפשרויות של יישור תמונה ומציאת סף לכל עמודה שמכניסים. בממשק ההגהה צריכה להיות תצוגה גרפית של התוצאה וגם תוצאה מילולית. כאן ישנה קשת רחבה של אפשרויות של תצוגת התוצאה. הן התוצאה בצורה גרפית: הסוגים השונים של החשדות לטעויות, ואיך להציג מידע שקשור לטעות כמו איזה תו זוהה, והן הדו"ח המילולי – סטאטיסטי.

ב. כחלק מהממשק למשתמש צריך לבנות מערכת לניהול הפרופיל של הסופר. כרגע הפרופיל קיים כאובייקט אך אין ניהול של הפרופילים. סביר שיהיו לסופר כמה פרופילים אם הוא כותב ספרים בגדלים שונים או שנעשה שימוש בשיטת צילום שונה של עמודות הספר.

ג. לא עשיתי מספיק ניסויים כדי לאבחן את כמות העמודות האידיאלית הנדרשת ללמידה. עליתי בהדרגה עד שהגעתי ל5 עמודות ששם הזיהוי החל להיות סביר, אבל לא המשכתי הלאה. די ברור לי ששימוש בעמודות נוספות ללמידה היה משפר עוד את הזיהוי במיוחד בתווים הבעייתיים.

ד. צריך לפתח עוד שכבה של זיהוי בין תווים קרובים. לאחר הזיהוי הראשוני בתווים שיש להם תווים קרובים צריכה להגיע שלב זיהוי נוסף שאמור לבדוק בתוך קבוצת התווים הזו את הזיהוי. גם כאן יש כיוונים שונים שאפשר לפתח. אפשר ליצור זיהוי משני בעזרת עוד מפות נוירונים שמקבלות כקלט רק את התווים הדומים, או להיעזר בבדיקות שונות לגמרי לצורך בדיקת תכונות כמו מציאת \ אי מציאת פיקסלים באזורים מסוימים.

ה. אפשר עוד לשפר את התמונה בת ההשוואה ע"י בדיקה עם עוד ספרי תורה שונים. יתכן מאוד שכאשר נבדוק על ספרי תורה נוספים נגלה ששם צריך להפעיל מניפולציות אחרות או נוספות על מה שעשינו כדי לייצר תמונה בת השוואה יעילה לספר הזה או לסוג הכתב (כיוון שיש מספר מנהגים לצורות הכתב). ייתכן ששם גם נרצה רזולוציה אחרת שתשקף טוב יותר את הכתב. הנקודה הרגישה הזו של יצירת התמונה בת השוואה לא מוצתה עדיין.

ו. לא פיתחתי זיהוי לתווים המוקטנים (זעירים) והמוגדלים (אות רבתי). התשתית מוכנה לזה, כלומר יש סימון לכל תו שכזה בבסיס הנתונים אבל צריך בדיקה שזה מוקטן \ מוגדל ואז להתאים לתו בר השוואה ומכאן להמשיך.

ז. צריך לנסות את המערכת על ספרי תורה נוספים. אני הרצתי את התוכנה על שני סוגים של כתב. אבל את שלב הניסויים המשמעותי ערכתי על סוג אחד. את הסוג השני לא בדקתי בצורה יסודית. בדיקה

ראשונית שעשיתי אכן הניבה תוצאות סבירות אך פחות טובות. ברור שאפשר יהיה לשכלל עוד את המערכת כך שתתמוך בכתבי יד נוספים אך צריך זמן משמעותי נוסף לחקור את ההבדלים.

ח. אפשר יהיה לבנות רשת נוירונים מוכנה לכל נוסח מבלי להאכיל בנתונים של הסופר. לצורך זה יש להכין בסיס נתונים רחב של דוגמאות למידה מסופרים שונים. מאוד יתכן שבכמות כזו של למידה נצטרך קונפיגורציה שונה של מפת הנוירונים.

ביבליוגרפיה

Benenson, R. (2016). *Classification datasets results*. Retrieved from Rodrigo Benenson's webpage: http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html#4d4e495354

Templeton, G. (2015, 10 12). *Artificial neural networks are changing the world. What are they?* Retrieved from Extreme Tech: https://www.extremetech.com/extreme/215170-artificial-neural-networks-are-changing-the-world-what-are-they

Wikipedia. (2017, October 19). Retrieved from Optical character recognition: https://en.wikipedia.org/wiki/Optical_character_recognition

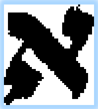
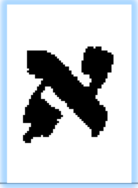
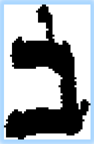
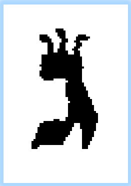
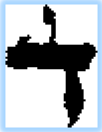
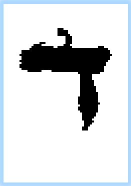
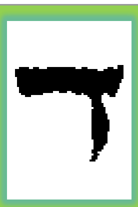
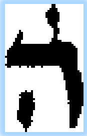
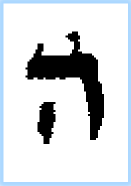
Wikipedia. (2017, September 12). Retrieved from Handwriting recognition: https://en.wikipedia.org/wiki/Handwriting_recognition

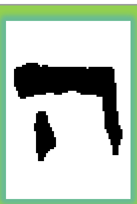
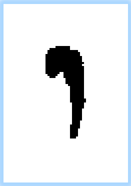
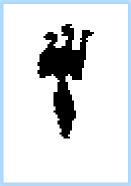
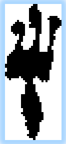
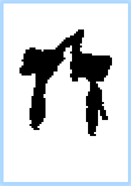
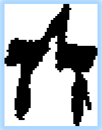
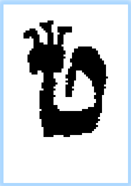
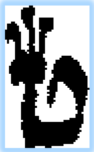
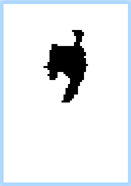
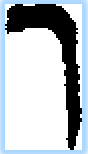
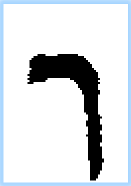
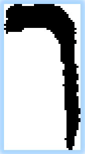
גוטמן, מ. (2016). זיהוי תווי כתב יד לא מקוון. *זיהוי תווי כתב יד לא מקוון*. האוניברסיטה הפתוחה.

נספח א'

טבלת המרת תווים מתמונת התו המקורית לתמונה ברת השוואה [דוגמאות אותיות להדר"ת ארוכות ברקע ירוק]

איור 26: תמונות תווים מקוריים וברי השוואה

תמונת תו מקורי	תמונת תו מרופד (62X89)	תמונת תו סופית (21X28)
45 X 50 		
43 X 67 		
33 X 70		
45 X 61 		
116 X 64		
39 X 64 		

		70X66
		20 X 47
		28 X 68 
		46 X 61 
		42 X 71 
		22 X 45
		38 X 73 

כ	כ	36 X 42 כ
ל	ל	41 X 79 ל
ל	ל	75 X 82 ל
ם	ם	45 X 46 ם
מ	מ	45 X 46 מ
ז	ז	26 X 93 ז
נ	נ	27 X 74 נ

ס	ס	36 X 50 ס
ע	ע	38 X 84 ע
ף	ף	41 X 69 ף
פ	פ	41 X 69 פ
ץ	ץ	44 X 91 ץ
צ	צ	45 X 75 צ
ק	ק	38 X 87 ק

ר	ר	42 X 48 ר
ר	ר	68 X 42 ר
ש	ש	58 X 77 ש
ת	ת	40 X 49 ת
ת	ת	98X51 ת

נספח ב'

נתאר כאן את הפרופיל שיצרנו בעזרת טבלת שדות הפרופיל שיצרנו בספר התורה של דוגמת ההרצה, ולכל שורה ניתן הסבר לפרמטר של הפרופיל.

איור 27 : דוגמת פרופיל לספר תורה

שם השדה	ערך השדה (דוגמה)	הסבר
name	SaadImproved	שם הפרופיל
lines number	42	מספר השורות בעמודה
isBlackAndWhiteImages	false	האם מדובר בתמונה שחור לבן (בינארית)
blackThreshold	170	סף הצבע של הפיקסל שממנה נחשב כשחור (מתוך 255 רמות אפור)
neuralNetworkFileName	saad_improved_neural_network	שם הקובץ של רשת הנוירונים
regularLetterMaxWidth	29	רוחב מקסימאלי של תו רגיל (בפיקסלים)
maxAboveLineHeight	19	גובה מעל לקו עליון של שורת התווים (בפיקסלים)
maxUnderLineHeight	38	גובה מתחת לקו העליון של השורה (בפיקסלים)
minSpaceBetweenWords	6	רוחב מינימאלי בין מילים (בפיקסלים)
minLetterWidth	7	רוחב תו מינימאלי (בפיקסלים)
shouldCheckRotation	false	האם צריך לבדוק יישור תמונה לפני הרצה