**המחלקה להנדסת תוכנה**

**פרויקט גמר – תש"פ**

ניתוח כתב יד על מנת לזהות זיופים בבחינה הפסיכומטרית

Handwriting analysis to detect forgeries in the psychometric exams

**מאת**

**דניאל גבאי**

**שחר ישראלי**

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

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

מערכות ניהול הפרויקט:

|  |  |  |
| --- | --- | --- |
| # | מערכת | מיקום |
| 1 | מאגר קוד | <https://github.com/DanielGabay/Author-verification-by-handwriting-samples> |
| 2 | יומן | <https://trello.com/b/7jKnxmLL/author-verification-by-handwriting-samples> |
| 3 | סרטון גרסת אלפא |  |

נאום המעלית

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

מבוא

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

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

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

בעקבות כך, המרכז הארצי לבחינות ולהערכה מעסיק מומחים לזיהוי כתבי יד, בכדי לנסות לבצע השוואה בין שני מועדי בחינה של נבחנים מסוימים אשר מוגדרים כחשודים, על ידי בדיקה של כתב ידו של הנבחן במטלת החיבור. ההשוואה אמורה לאמת/להפריך חשד עבור נבחן שרימה בבחינה.

השוואת כתב ידו של נבחן בין שני המועדים מתבססת על ההנחה שכתב ידו של כל אדם הוא ייחודי ושניתן על פיו לזהות את כותבו. בנוסף יש הסכמה כי אין לשני בני אדם כתב יד זהה (לחלוטין) ואף האדם אינו יכול לכתוב בשנית באופן טבעי דברים שכתב בעבר בצורה זהה לגמרי. עם זאת, סגנון הכתיבה של האדם מאופיין, מוגדר, אישי וייחודי לכל אדם. פעולת הכתיבה היא תוצאה של שיתוף פעולה בין מערכות אחדות בגוף, כגון הראייה, התפיסה ועיבוד הנתונים, וכן פעולות מוטוריות של שרירים רבים. שיתוף כל אלה לפעולות שגרתיות המתבצעות מדי יום, מפתח מיומנות שהיא הבסיס לסגנון כתב היד על תכונותיו הייחודיות.

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

מתוך ההנחה כי קשה להכריע ששני כתבי יד זהים, נרצה לבנות תוכנית שמקבלת כקלט שני חיבורים סרוקים של אותו נבחן (במועדים שונים) ומטרתה לאמוד את הסיכוי שהחיבורים נכתבו על ידי שני אנשים שונים.

תיאור הבעיה

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

במטרה לייעל את תהליך אימות הנבחן ולצמצם את מספר הבדיקות הנעשות ע"י המומחים, המרכז הארצי לבחינות והערכה זקוק למערכת ממוחשבת שתייעל את תהליך הבדיקה.

באמצעות המערכת ניתן יהיה לצמצם את כמות הבדיקות אשר מגיעות לידי המומחים כך שרק הנבחנים שזוהו ע"י המערכת כחשודים, יגיעו לידיהם לבדיקה ידנית מעמיקה.

## דרישות ואפיון הבעיה

## הבעיה מבחינת הנדסת תוכנה

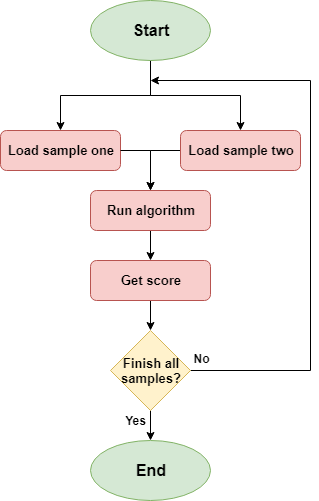
על מנת לבצע השוואה בין שני כתבי יד ואימות המחבר, נדרש תהליך המורכב משלושה חלקים מרכזיים איתן נתמודד בפרויקט: **גילוי, זיהוי ואימות**, עליהם נפרט בתיאור הפתרון.

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

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

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

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

תיאור הפתרון

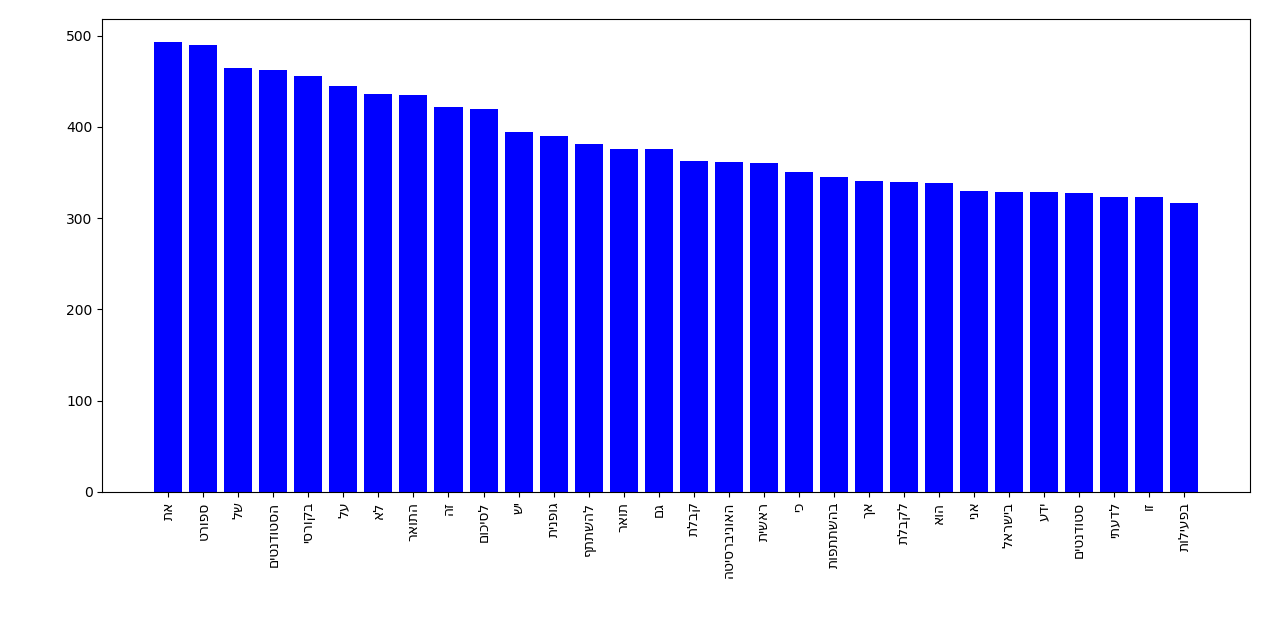
הפתרון הוא יצירת תוכנת מחשב, אשר תקבל כקלט מהמשתמש שני קבצים המכילים טקסט הכתוב בכתב יד. עם קבלת פקודה, התוכנה תבצע ניתוח והשוואה של החיבורים, ותחזיר כפלט את הסיכוי שהחיבורים נכתבו על ידי אותו אדם.

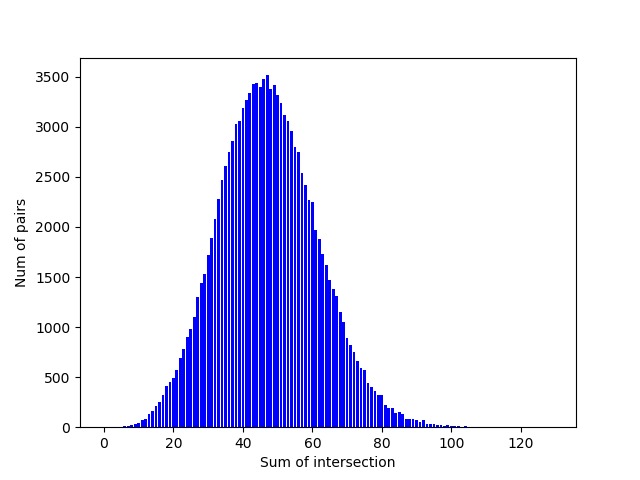
על מנת לבצע את תהליך הבדיקה, כלומר לפתור את הבעיות המרכזיות שהוזכרו תחת 'תיאור הבעיה', שלבי הגילוי, הזיהוי והאימות יחולקו לתתי בעיות. בכל שלב בתהליך, נבחן אלגוריתמים שונים מהספרות ובמידת הצורך ניצור בעצמנו אלגוריתמים במטרה לממש כל שלב בדרך לפתרון בצורה אופטימלית.

ארכיטקטורת המערכת בנויה משלושה שלבים עיקריים: גילוי, זיהוי, ואימות. עבור כל נבחן, יש לטעון חיבור משני מועדי הבחינה, וכל חיבור יעבור את כל אחד מהשלבים.

יש לציין כי בחלק מהשלבים, נעזרנו בפרויקט גמר משנה שעברה אשר בוצע ע"י הסטודנטים מורן זרגרי ואיתי חפץ אשר חקרו את נושא זיהוי כתב יד והמרתו לטקסט ממוחשב.

1. **גילוי detection))** – בשלב זה ננתח את החיבור הסרוק, במטרה לחלץ גורמי השוואה. תהליך הגילוי מורכב ממספר תתי שלבים עיקריים:
2. עיבוד מקדים של החיבור – את החיבורים אנו מקבלים בפורמט tiff מהמרכז הארצי לבחינות והערכה. כל חיבור מכיל עמוד אחד או שניים, לכן בחרנו לאחד את העמודים לתמונה אחת. לאחר מכן, נבצע חיתוך של קצוות התמונה במטרה ללכוד רק את הטקסט הנכתב על ידי הנבחן. בנוסף, יש צורך ביישור התמונה והפחתת "רעש" המקשה על זיהוי הטקסט.
3. גילוי שורות - מציאת השורות בהן מופיע הטקסט הכתוב של הנבחן.
4. גילוי המילים - הפרדת השורות למילים בודדות.
5. גילוי אותיות – הפרדת השורות לאותיות בודדות.
6. **שלב הזיהוי (recognition)** – לאחר שחילצנו את גורמי ההשוואה בשלב הגילוי, נרצה לזהות אותם, כלומר לזהות את המילים והאותיות שמצאנו:
7. זיהוי אותיות: לאחר שמצאנו את מיקומי השורות, נפעיל מודל שאומן לזהות אותיות בשפה העברית (מהפרויקט של שנה שעברה). נרצה לשמור לשלב הבא רק אותיות שהאלגוריתם זיהה בהסתברות גבוהה (כרגע ה-threshold הוא 99%). רמת הבטחון בזיהוי האות חשובה לנו כיוון שבשלב ההשוואה נרצה להשוות בין אותיות **זהות**.
8. זיהוי מילים: אנו נסתפק בזיהוי של חלק מהמילים בחיבור ולא כולן. כלומר, בשלב זה נרצה לסנן את המילים אשר באמצעותן נערוך את תהליך האימות. בחרנו לפעול בדרך זאת על מנת שנוכל "לאמן מכונה" במספר מצומצם של מילים נבחרות, אשר בדקנו כי חוזרות בכל החיבורים. בכך נייעל את תהליך איסוף ה- data שישמש בלמידת המכונה.

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

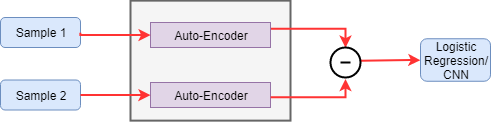
לאחר מכן, כדי לוודא שאכן יש 'חיתוך' בין טקסטים שונים לבין המילים המופיעות בטקסט, כתבנו תוכנית נוספת שמבצעת את ההשוואה הבאה: מכל טקסט חילצנו את המילים הנפוצות שמופיעות בו, וביצענו חיתוך עם כל שאר הטקסטים וספרנו את מספר המילים השוות. קיבלנו את הגרף הבא:

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

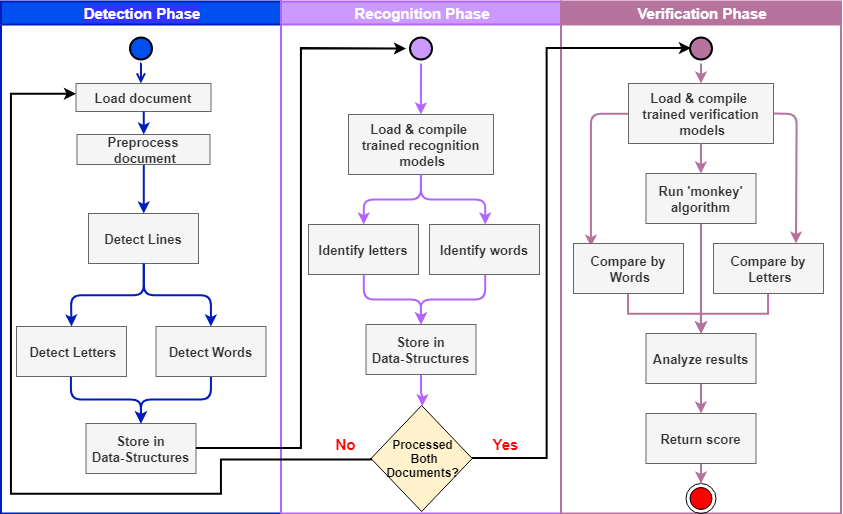
בכדי לאמן מכונה לזהות את המילים שבחרנו, אספנו כ-300 דפים של מילים בכתב יד (3600 מכל מילה). לאחר שנבצע אימון למודל, נפריד את השורות למילים. כל מילה נעביר למודל בכדי לבדוק האם היא חלק מהמילים הנפוצות. את המילים שזיהינו נשמור במבנה נתונים לשלב הבא.

1. **שלב האימות (verification)** – שלב זה הוא השלב המרכזי בו מתקבלת ההחלטה האם שני החיבורים נכתבו על ידי אותו אדם או לא. נכון להיום, החלטנו לפעול בשלושה דרכים עיקריות: השוואת מילים, השוואת אותיות ואלגוריתם 'קוף'.
2. השוואת מילים ואותיות – משלב הגילוי והזיהוי, נשארנו עם גורמי ההשוואה משני החיבורים. כעת נרצה לבצע השוואה בין אותן אותיות ומילים משני החיבורים ('א' מול 'א', 'של' מול 'של' וכו'). כלומר, מכל חיבור נשאיר רק את האותיות והמילים שהצלחנו לזהות בשני החיבורים. את ההשוואה נבצע באופן הבא:

נשתמש באלגוריתם Auto-Encoder של Keras שהוא אלגוריתם שעושה compress ו- decompress ל- data. באמצעות האלגוריתם ניתן לחלץ מכל מילה/אות ('הדגימה') את הפיצ'רים הכי חשובים שמאפיינים אותה. את הפיצ'רים שחילצנו משני הדגימות באמצעות ה- encoder, נעביר בפונקציית חיסור ואת התוצאה נעביר לאלגוריתם סיווג נוסף (Logistic Regression/CNN) שאומן מראש לזהות האם תוצאת החיסור מאפיינת דגימות של אותו מחבר או מחברים שונים.



1. אלגוריתם קוף – במהלך המחקר, לאחר שביצענו זיהוי האותיות גילינו כי המודל שבו אנו משתמשים לזיהוי אותיות מצליח לזהות בהסתברות גבוהה אותיות שונות עבור אנשים שונים. מה הכוונה? שמנו לב כי עבור נבחן א' נזהה בדיוק רב את האותיות א', ח', מ', ש' לדוגמא, ועבור נבחן ב' נזהה בדיוק רב אותיות אחרות כמו ב', ג', ה', ק'. הבנו מתוך תוצאות אלו, כי ניתן לנצל את חוסר הדיוק של המודל בכך שאם בטקסט מסוים מצאנו סט של אותיות, הרי שהאלגוריתם שלנו אמור לדעת לזהות את סט אותיות אלו גם בחיבור אחר שנכתב על ידי אותו אדם. אם בחיבור הנוסף האותיות שזוהו אינן תאומות, או תואמות חלקית, הרי שזה יכול להעיד על כך שהחיבורים לא נכתבו על ידי אותו אדם. קראנו לאלגוריתם זה 'קוף' מכיוון שאנו מנצלים את חוסר היכולת של המודל לזהות בדיוק של 100% את כל האותיות. כלומר, לא נזדקק להשוואה אמיתית בין הגורמים השונים בכדי אולי להגיע להכרעה האם מדובר באותו אדם או לא. ביצוע האלגוריתם: עבור כל חיבור, נגדיר וקטור בגודל 27 (כמספר האותיות בשפה העברית) שבכל תא נשמור את כמות המופעים (באחוזים) מכל אות שהמודל הצליח לזהות (תא 0 האות א' וכו'). לאחר מכן, נבצע חיסור בערך מוחלט בין שני הווקטורים משני החיבורים. את וקטור החיסור נעביר למודל סיווג נוסף (Logistic Regression), שאומן לזהות וקטורי חיסור של חיבורים שנכתבו על ידי אותו אדם, וחיבורים שנכתבו על ידי אנשים שונים.

****תיאור ארכיטקטורת המערכת לפי השלבים השונים:

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

סקירת עבודות דומות \ בספרות והשוואה \ סקר שוק

Writer verification based on a single handwriting word samples

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

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

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

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

Matching Handwritten Document Images

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

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

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

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

Hybrid Feature Learning for Handwriting Verification

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

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

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

נספחים

## **תכנון הפרויקט**

|  |  |
| --- | --- |
| 8.9.19 | פגישה ראשונה עם המנחה והצגת רעיון הפרויקט |
| 11.9.19 | פגישה עם המנחה והנציגים של המרכז הארצי לבחינות והערכה, חידוד הדרישות של הפרויקט. |
| 6.10.19 | פגישה עם המנחה, הצגת התקדמות והצבת מטרות נוספות לפגישה הבאה. |
| 6.11.19 | פגישה עם המנחה, הצגת התקדמות והצבת מטרות נוספות לפגישה הבאה. |
| 8.12.19 | הגשת דו"ח הצעה |
| 26.01.20 | הגשת דו"ח אלפא |
| 21.3.20 | הגשת לוז עבודה לסמסטר ב' |
| 26.4.20 | הגשת דו"ח בטא |
| 26.6.20 | הגשת דו"ח סופי |
| 5-7.7.20 | הצגת הפרויקט |
| 16.7.20 | הצגת פוסטר בכנס פוסטרים |
| 20.7.20 | העברה של הפרויקט |

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

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

**טבלת דרישות (User Requirement Document)**

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

עם זאת הדרישות היחידות כרגע מהלקוח הן:

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