

אוניברסיטת תל אביב
המחלקה למתימטיקה ומדעי המחשב

סיווג אוטומטי של שמות עצם פרטיים
באמצעות אלגוריתמי למידה

אודליה דיין

מוגש כמילוי חלקי של הדרישות למוסמך במדעי המחשב
תל-אביב תשס"ה

העבודה בוצעה בהנחייתו של פרופ' נחום דרשוביץ'.

תודות

ברצוני להודות לפרופסור נחום דרשוביץ על תמיכתו, צידודו
והנחייתו המסורית.

תודה לד"ר יעל וילנה על עזרתה הרבה.

תודה לד"ר צידו דאן על הליווי בתחילת הדרכ.

ולבסוף, תודה מיוחדת לבן זוגי אבי וילדי, נעמה, רעות,
מצול ורצן, על התמיכה האדירה, הצידוד וההשקעה לכל
אורך הדרכ...

תוכן עניינים

6.....	1. תקציר
7.....	2. הגדרת הבעיה
7.....	2.1 תאור הבעיה
8.....	2.2 מטרת העבודה
9.....	3. סקר ספרות (3)
9.....	3.1 ה-MUC (Message Understanding Conference)
9.....	3.2 עבודות מבוססות חוקים
9.....	3.3 עבודות מבוססות סטטיסטיקה
10.....	4. דרך הפתרון
10.....	4.1 תאור הפתרון
11.....	4.2 פירמוט הטקסט
11.....	4.2.1 קורפוס באנגלית – מבנה ואופן ניתוח
13.....	4.2.2 קורפוס בעברית – מבנה ואופן ניתוח
15.....	4.3 הוצאת מאפיינים
15.....	4.3.1 סוגי מאפיינים
15.....	4.3.2 צירופי מאפיינים
16.....	4.3.3 דוגמאות להמחשה
18.....	4.4 בניית מסמכים וירטואליים
19.....	4.5 אימון / סיווג
19.....	4.5.1 תאור רידוד מאפיינים
20.....	4.5.2 תאור האלגוריתם (Bayes)
21.....	5. תוצאות
21.....	5.1 כללי
21.....	5.1.1 מדדי הערכה בסיסיים – הגדרות
22.....	5.1.2 כמויות
22.....	5.1.3 תוצאות אנגלית
23.....	5.1.4 תוצאות עברית
24.....	5.2 מציאת סף
24.....	5.2.1 אנגלית
24.....	5.2.2 עברית
25.....	5.3 מאפיינים חזקים
25.....	5.3.1 אנגלית
34.....	5.3.2 עברית
40.....	5.4 משפחות מאפיינים

41.....	5.5. בחינת חסינות (Robustness)	41.....
41.....	5.5.1. אנגלית	41.....
42.....	5.5.2. עברית	42.....
43.....	6. סיכום ומסקנות	43.....
44.....	7. כיוונים עתידיים	44.....
44.....	7.1. צירופי מאפיינים	44.....
44.....	7.2. אלגוריתמים נוספים	44.....
44.....	7.3. סיווג מסמכים לנושאים	44.....
45.....	7.4. איחוד מופעי שם ברמת מסמך	45.....
45.....	7.5. סיווג קשרים באופן אוטומטי	45.....
46.....	8. ביבליוגרפיה (1)	46.....

1. תקציר

טיפול יעיל בכמויות המידע האדירות המצויות לפתחנו בעידן המידע מחייב הסתייעות בכלים אוטומטיים לעיבוד, ניתוח, והבנה של טקסט בשפה טבעית. כלים אלו נדרשים בין היתר למלא משימות דומות לאלו הנופלות בנחלתם של מומחים. תיוג שמות עצם פרטיים בצורה אוטומטית מהווה דוגמה לכך. סיווג שמות העצם הפרטיים, מאפשר הוצאת מידע על ישויות מהטקסט, ומהווה נכס חיוני המסייע ביישומים שונים של אחזור מידע ועיבוד שפה טבעית, כגון מנועי חיפוש. עיקר העבודה בתחום מתחלקת בין עבודה המבוססת על חוקים הנבנים ידנית ע"י מומחים אנושיים, לבין עבודה המבוססת על סטטיסטיקות. כלים המתבססים על חוקים ידניים אמנם מדויקים יותר, אך חסרונם הגדול הוא הצורך במומחים אנושיים, התלות בשפה והצורך בתחזוקה מתמדת של אוסף החוקים. הגישה האוטומטית מתבססת ע"פ רוב, על סטטיסטיקות שונות הנאספות ביחס לקורפוס הנתון. האימון המתקבל בהתאם ע"י כך הינו יחסי למהות הנושאים המיוצגים בקורפוס. יצוין כי האפשרות למצוא את סוג שם העצם הפרטי, מבוססת על ההנחה לפיה סוג השם קשורה בדפוס הופעותיו המשותפות עם המלים האחרות ולכן לשמות מאותה קטגוריה יהיו מן הסתם דפוסים ומאפיינים דומים.

עבודה זו מתמקדת בבעיית סיווג שמות עצם פרטיים, בצורה אוטומטית מבוססת סטטיסטיקות. העבודה מוציאה מאפיינים של כל שם, ומסווגת על סמך איתור אוטומטי של מאפיינים חזקים לכל קטגורית שמות. בהנתן סט המאפיינים הרלוונטי מבוצע הסיווג באופן דומה לסיווג מילות מפתח למסמך. המאפיינים המתקבלים יכולים להיות מסוגים שונים ובעיקר מתבססים על המילים השכנות לשם. מאפיין יכול להיות סמנטי כגון הופעת "פרופ" לפני שם איש, או למשל הקשרי (contextual) כגון "בתוך השם".

במסגרת עבודה זו מוצג תהליך להוצאת מאפיינים ואיתור המאפיינים הרלוונטיים לכל קטגורית שמות. על מנת לאתר את מאפיינים אלו, נערכים ניסויים מקיפים הבוחנים חסם אופטימאלי לציון המאפיינים. בהנתן רשימת מאפיינים שניתן להשתמש בהם, מבוצע סיווג על פיהם. אלגוריתם הסיווג בו נעשה שימוש, הינו רדוקציה של הגישה בייסיאנית לאלגוריתם לסיווג מילות מפתח למסמך לאלגוריתם לסיווג קטגורית שם למופע שם עצם פרטי.

לצורך הוכחת יתרונות השימוש בשיטה הסטטיסטית, מבוצעת השוואת תוצאות בין השפות אנגלית ועברית. העבודה מתבססת מעט ככל האפשר על חוקי השפה, ולכן מתאפשר מעבר בין השפות כמעט ללא שינויי קוד.

2. הגדרת הבעיה

מסקנות ועדת החקירה של הקונגרס האמריקאי לבדיקת מחדלי המודיעין במניעת אירועי ה-11 לספטמבר 2001 חשפו את מה שכל בירוקרט מנסה להסתיר. המערכת קיבלה מספיק פיסות מידע מבודדות, שהיו צריכות לעורר את פעמוני האזהרה בזמן, אבל אף אחד לא חיבר את הפיסות זו לזו על מנת להתרשם מהתמונה הכוללת.

אכן, "ידע הוא כח", אך הבעיה הגדולה אינה השגת המידע הזורם באוטוסטרדת המידע, אלא הפיכתו לידע.

בעידן האינטרנט, שבו כמויות עצומות של מידע זורמות ברשת של תקשורת מחשבים מכל מקום אל כל מקום בעולם. קצב זרימת המידע, כמויות המידע וכמות המחשבים המחוברים לרשת גדלים בקצב מסחרר והופכים את העולם שלנו ל"כפר גלובלי" שבו המידע נגיש יותר מאי-פעם. דבר זה הופך את בעיית השגת הידע, באופן פראדוקסאלי, לחמורה יותר.

כריית מידע מטקסט, מוגדרת כתהליכי ניתוח טקסט ואיפיון תבניות שפה לצורך מיצוי מידע. המידע המופק בתהליך זה יכול להיות שם מחבר, כותרת המאמר, תאריך פרסום המידע, זיהוי מגמות, קשרים בין יישויות ועוד.

2.1 תאור הבעיה

הבעיה בה מתמקדת עבודה זו, שייכת לתחום כריית הטקסט. הבעיה עוסקת בסיווג שמות עצם פרטיים הנתונים בטקסט לקטגוריות (שם איש, שם ארגון, שם מקום).

עבודה זו מטפלת בתת משימה בנושא, המוגדרת ב-MUC 7 (Message Understanding Conference).

להלן הגדרת הבעיה:

בהנתן קורפוס מתוייג בשפה האנגלית, יש לסווג כל ביטוי המופיע בגוף המאמר ומתוייג כשם ישות, לאחת מ-3 הקטגוריות הבאות:

- שם איש – שם אדם או משפחה.
- שם ארגון – חברות, גופים ממשלתיים וכיו"ב.
- שם מקום - ערים, פרובינציות, מדינות, אזורים בינלאומיים, ימים ואגמים, הרים וכיו"ב.

לצורך בחינת כלליות הפתרון, נבחנה המערכת בנוסף, גם על קורפוס בשפה העברית.

2.2. מטרת העבודה

העבודה מציגה שיטה כללית ואוטומטית לסיווג מופעי שמות עצם פרטיים נתונים בטקסט, לאחת ממספר קטגוריות שם ידועות מראש. העבודה עושה שימוש במאפייני המופע של שם העצם הפרטי (למשל: מילים שהופיעו לפני ואחרי השם, סימני פיסוק ומילים שהופיעו בתוך השם). לאחר הוצאת כל מאפייני השמות האפשריים, נעשה רידוד המאפיינים באופן אוטומטי, לרשימת מאפיינים רלוונטים לכל קטגוריה.

לבסוף, מתבצע סיווג של מופע השם לקטגוריית שם ע"פ אלגוריתם בייסיאני המשמש לסיווג מילות מפתח למסמך.

העבודה מתבססת על ההנחות הבאות:

1. שימוש בשיטות סטטיסטיות בלבד, ללא שימוש בחוקים ידניים.
 2. שימוש מינימאלי ככל האפשר בהנחות המבוססות על השפה בה כתוב הטקסט.
 3. לא מבוצע שימוש בתזאורוס מכל סוג.
- העבודה סוקרת סוגי מאפיינים שונים ואת השפעתם על התוצאות. כן, מבצעת העבודה השוואה בין השפות אנגלית ועברית.

3. סקר ספרות (3)

בעיה זו מכונה בספרות המקצועית בשם Named Entity Recognition), וקיימות עבודות רבות העוסקות בה. ניתן לחלק את העבודות לשני סוגים עיקריים:

1. Rule Based – עבודות המסתמכות על חוקים הנכתבים בצורה ידנית (למשל: אם בתחילת השם או לפניו מופיעה המילה "Doctor", אזי בהסתברות 80% מדובר בשם איש).

3.1 ה-MUC (Message Understanding)

(Conference

3.2 עבודות מבוססות חוקים

3.3 עבודות מבוססות סטטיסטיקה

4. דרך הפתרון

4.1. תאור הפתרון

עבודה זו מבצעת את הסיווג באמצעות שלב אימון ושלב סיווג.

הן במסגרת האימון והן במסגרת הסיווג מבוצעים השלבים הבאים :



להלן יתוארו שלבים אלו בהרחבה.

4.2 פירמוט הטקסט

בשלב פירמוט הטקסט מבוצעת המרה מטקסט חופשי לייצוג טבלאי של מסמכים המורכבים מ-Token-ים, כולל מידע על כל Token.

מבוצעים השלבים הבאים :

1. חלוקה למסמכים.
2. חלוקת כל מסמך למשפטים.
3. חלוקת כל משפט ל-Token-ים.
4. ניתוח כל Token ושמירת מידע כדלקמן :

- צורה מקורית (OriginalForm) – ה-Token בצורה בה הופיע בטקסט.

- צורה בסיסית (BaseForm) – ה-Token לאחר Stemming.

- סוג ה-Token (POS) – צורה פונטית או חלק דיבר (במקרים בהם ניתן להסיק – ראה פירוט בהמשך).

- מיקום ה-Token במשפט.

- סיווג שם עצם פרטי (אם מדובר בשם).

שלב זה, הוא השלב העיקרי המתבסס על השפה. להלן נתאר את מבנה הקלט עבור כל שפה שנבחנה (אנגלית ועברית) ואת ההנחות המתבססות על השפה בכל אחד מהמקרים.

4.2.1 קורפוס באנגלית – מבנה ואופן ניתוח

מבנה הקורפוס באנגלית:

עבור סיווג השמות בשפה האנגלית, נעשה שימוש בקורפוס Train ו-Test, שתוייגו עבור MUC 7.

בכל אחד מהמאגרים קיימים 100 מאמרים בשפה האנגלית.

העבודה מטפלת בסיווג שמות המופיעים בגוף המאמר עצמו ולא בחלקיו האחרים (גוף המאמר מתויג בטקסט בסימון SGML : <TEXT>).

היות והעבודה מתמקדת בתת המשימה של שמות עצם פרטיים, סווגו שמות בעלי תיוג מתאים המסומן ENAMEX.

אופן התיוג הניתן בטקסט :

```
<ELEMENT-NAME ATTR-NAME="ATTR-VALUE" ...>text-string  
</ELEMENT-NAME>
```

דוגמה :

```
<ENAMEX TYPE="ORGANIZATION">Taga Co.</ENAMEX>
```

צורה בסיסית באנגלית :

- עבור מילה (רצף אותיות או אותיות ומספרים) : הפעלת אלגוריתם Stemming של Porter.
- עבור מספר (רצף ספרות או Token שתיוג בטקסט כמספר) : "Number".
- עבור תאריך (Token שתיוג בטקסט כתאריך) : "Data".

סוגי ה-Token באנגלית :

עבור השפה האנגלית לא בוצע ניתוח תחבירי. סוגי ה-Token-ים :

- | | | |
|--|---------------------------|---|
| Num | מספר | • |
| Date | תאריך | • |
| Punct | סימן ניקוד | • |
| EndOfSentence | סימן סוף משפט | • |
| "מוערך" ע"פ הלוגיקה : "." שאחריה רוח ומילה שמתחילה באות גדולה. | | |
| Sign | סימן אחר | • |
| Name | שם עצם פרטי | • |
| Preposition | מילת יחס | • |
| ע"פ רשימת מילות יחס באנגלית בה מופיעות כ-20 מילים. | | |
| UpperLetters | אותיות בלבד - כולן גדולות | • |
| LowerLetters | אותיות בלבד – כולן קטנות | • |
| Letters | אותיות בלבד – מעורב | • |

- UpperAndSigns אותיות גדולות וסימנים
- LettersAndSigns אותיות מעורב וסימנים
- UpperAndDigits אותיות גדולות וספרות
- LettersAndDigits אותיות מעורב וספרות
- DigitsAndSigns ספרות וסימנים
- UpperDigitsAndSigns אותיות גדולות, ספרות וסימנים
- LettersDigitsAndSigns אותיות, ספרות וסימנים (ב"מ)

4.2.2. קורפוס בעברית – מבנה ואופן ניתוח

מבנה הקורפוס בעברית:

עבור סיווג השמות בשפה העברית, נעשה שימוש בקורפוס שנבנה במסגרת פרויקט מילה בטכניון (<http://mila.cs.technion.ac.il>). הקורפוס מכיל 1892 משפטים בעברית מעיתון "הארץ". הקורפוס מגיע במבנה XML בו המשפטים מתוייגים ידנית סינטקטית ומורפולוגית.

בכל אחת מההרצות נלקחו 1692 משפטים עבור אימון ו-200 עבור סיווג. הקורפוס אינו מחולק למאמרים אלא מכיל רשימת משפטים מנושאים שונים. לא תמיד קים רצף המחבר אותם למאמר מסוים. היות והעבודה מתמקדת בתת המשימה של שמות עצם פרטיים, סווגו שמות בעלי תיוג מתאים המסומן ProperName. אופן התיוג הניתן בטקסט ראה:

http://mila.cs.technion.ac.il/corpora/2000sentences/hebrew_corpus.xsd

דוגמה:

```
<token id="5" surface="לישראל" transliterated="lieral">
  <analysis id="1">
    <prefix id="1" function="preposition" surface="ל" transliterated="l"/>
    <base lexiconItem="ישראל" transliteratedLexiconItem="ieral">
      <properName gender="masculine and feminine" type="location"/>
    </base>
  </analysis>
</token>
```

צורה בסיסית בעברית:

מופיעה בקורפוס עצמו (Tag=base, Attr=lexiconItem)

סוגי ה-Token בעברית:

מופיע בקורפוס ב-Tag נפרד תחת base. סוגי ה-Token-ים:

Noun	• שם עצם
Verb	• פועל
Name	• שם עצם פרטי
AuxVerb	• פועל
Adjective	• תואר השם
Adverb	• תואר הפועל
Num	• מספר
Particle	• מילית
Conjunction	• מילת חיבור
Preposition	• מילת יחס
Interrogative	• מילת שאלה
Pronoun	• כינוי גוף
Punct	• סימן ניקוד
Negation	• מילת שלילה
EndOfSentence	• סימן סוף משפט

סומן ידנית לצורך פישוט.

4.3. הוצאת מאפיינים

בשלב זה נבנה עבור כל Token המסומן כשם עצם פרטי, סט מאפיינים מסוגים שונים וברמות הכללה שונות (ע"פ הניתוח שבוצע בשלב הפירמוט הראשוני).

המאפיינים נבנים ע"פ השכנים של שם הישות (ה-Token-ים המקיפים אותו) וכן ע"פ מאפייני השם עצמו.

4.3.1. סוגי מאפיינים

ORIGINAL	צורת הופעה מקורית בטקסט.
BASE	צורה בסיסית.
POS	סוג ה-Token.

4.3.2. צירופי מאפיינים

המאפיינים הנ"ל מורכבים על Token-ים שכנים לשם שאת סוגו מנסים לגלות, בגבולות משפט (כלומר, בתנאי שלא גלשנו למשפט הקודם/הבא), ועל Token-ים המרכיבים את שם הישות עצמו.

NeighborBefore	3 שכנים לפני השם
NeighborBefore1	שכן צמוד לפני השם
NeighborBefore2	שכן שני לפני השם
NeighborBefore3	שכן שלישי לפני השם
PartOf_Start	חלק מהשם – התחלה
PartOf_Mid	חלק מהשם – אמצע
PartOf_End	חלק מהשם – סוף
PartOf	חלק מהשם – כללי
NeighborAfter	3 שכנים אחרי השם
NeighborAfter1	שכן צמוד אחרי השם

NeighborAfter2	שכן שני אחרי השם
NeighborAfter3	שכן שלישי אחרי השם

בנוסף קיימים מאפיינים אודות אורך שם העצם הפרטי ב-Token-ים :

Part Of_Len_LessThan4	אורך 1-3 Token-ים
PartOf_Len_4To5	אורך 4-5 Token-ים
PartOf_Len_MoreThan5	יותר מ-5 Token-ים

4.3.3. דוגמאות להמחשה

אנגלית:

המשפט בטקסט :

Scott, 27, is married and lives in **Boston**.

המאפיינים שנבנו עבור השם Boston :

NeighboreBefore1_Base_in
NeighboreBefore1_Original_in
NeighboreBefore1_POS_Preposition
NeighboreBefore2_Base_live
NeighboreBefore2_Original_lives
NeighboreBefore2_POS_Letters
NeighboreBefore3_Base_and
NeighboreBefore3_Original_and
NeighboreBefore3_POS_Letters
PartOf_Len_LessThan4
PartOf_BOSTON
PartOf_POS_Letters

עברית:

המשפט בטקסט:

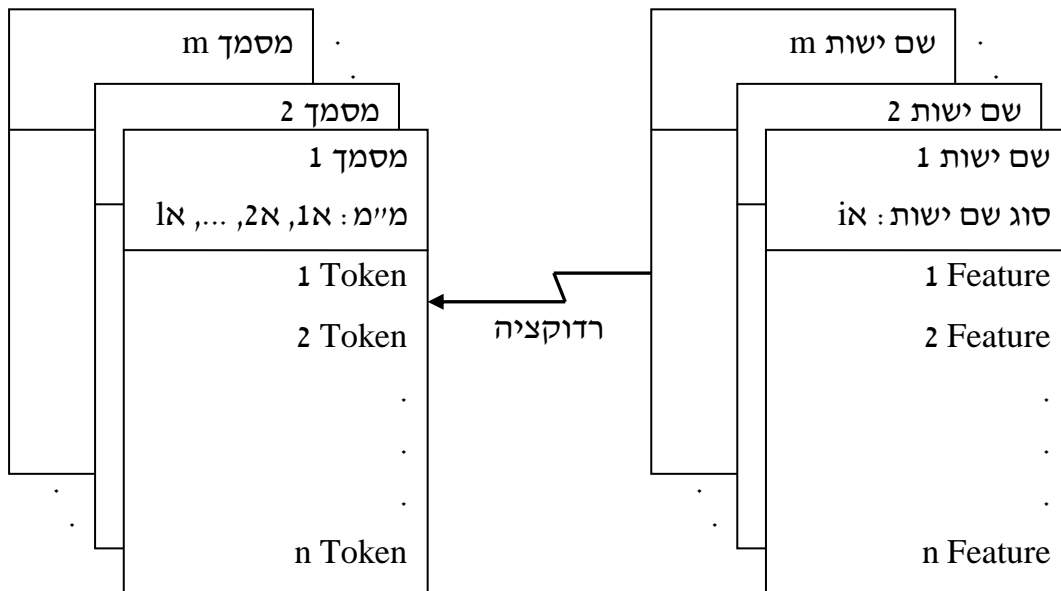
בכפר בלעא בנפת טול כרם תקפו שלושה רעולי פנים בסכינים רופא מקומי ד"ר עפ"י ברברח, בן 03, שאושפז במצב קשה בבית החולים אל איתיחאד בשכם.

המאפיינים שנבנו עבור השם: עפ"י ברברח.

NeighboreBefore1_Base_ד"ר
NeighboreBefore1_Original_ד"ר
NeighboreBefore1_POS_Noun
NeighboreBefore2_Base_מקומי
NeighboreBefore2_Original_מקומי
NeighboreBefore2_POS_Adjective
NeighboreBefore3_Base_רופא
NeighboreBefore3_Original_רופא
NeighboreBefore3_POS_Noun
PartOf_Len_LessThan4
PartOf_Start_עפ"י
PartOf_Start_POS_Null
PartOf_End_ברברח
PartOf_End_POS_Null
NeighboreAfter1_Base_,
NeighboreAfter1_Original_,
NeighboreAfter1_POS_Punct
NeighboreAfter2_Base_בן
NeighboreAfter2_Original_בן
NeighboreAfter2_POS_Noun
NeighboreAfter3_Base_Num
NeighboreAfter3_Original_03
NeighboreAfter3_POS_Num

4.4. בניית מסמכים וירטואליים

בסוף שלב הוצאת החוקים מתקבל מצב בו לכל שם עצם פרטי יש סט מאפיינים, והבעיה היא סיווג השם + סט המאפיינים לקטגוריות השונות. כדי להשתמש באלגוריתמי סיווג רגילים (סיווג מסמכים למילות מפתח) מבוצעת רדוקציה כדלקמן:



כלומר, לכל שם ישות נבנה מסמך וירטואלי המכיל את המאפיינים שנבנו כ-Token-ים. בתהליך האימון מתייחסים ל"מסמך" כאילו סווג לקטגוריה הנתונה, ובשלב הסיווג מנסים למצוא לאיזו קטגוריה מתאים ה"מסמך" שנוצר.

4.5. אימון / סיווג

4.5.1. תאור רידוד מאפיינים

בשלב הוצאת המאפיינים, נבנים מאפיינים מסוגים שונים ללא תהליך בחירה של המאפיינים הרלוונטים יותר או פחות. בתום השלב קיימים הרבה מאפיינים אשר חלקם הגדול אינו רלוונטי למשימה ואין טעם להשתמש בהם.

המטרה היא לרדד מאפיינים שאינם חשובים, כלומר כאלו שהשפעתם על כל הקטגוריות דומה.

לצורך העניין נבחן את הערך: $p(c/T=ti)$. אם קיימת קטגוריה עבורה ערך זה גבוה מהאחרות הרי שהמאפיין תורם לקטגוריה והינו משמעותי.

עם זאת, יש לקחת בחשבון גם את $p(c)$. קטגוריה שכיחה בעלת מאפיין המצביע עליה יותר מאשר על אחרות, אינה כמשקל קטגוריה "נדירה" יותר, בעלת מאפיין בולט.

$\frac{p(c/T=ti)}{p(c)}$ לכן, עבור כל קטגוריה, מחושב הערך: באשר:

$$p(c/T=ti) = \frac{\text{מספר מופעי } ti \text{ במסמכים מסווגים ל-} c}{\text{מספר מופעי } ti \text{ ב-Corpus}}$$
$$p(c) = \frac{\text{מספר מסמכים מסווגים ל-} c}{\text{מספר מסמכים}}$$

כדי להחליט איזה ערך סף ינתן לערך המחושב, מבוצעת הרצה של מסמכי האימון, עבור סט של ערכים (בשיטת אריה במדבר) עד לקבלת ערך אידיאלי (ראה פירוט בפרק התוצאות).

לסיכום, לפני שלב הסיווג, מרודדים מאפיינים כדלהלן:

- מאפיינים שהופיעו פחות מ-3 פעמים.
- מאפיינים עבורם באף אחת מהקטגוריות לא עלה הערך $p(c/T=ti)/p(c)$ על הסף שנקבע.

4.5.2. תאור האלגוריתם (Bayes)

כאמור, נעשה שימוש באלגוריתם לסיווג מסמכים לקטגוריות.

לכל קטגוריה c , מחשבים את ההסתברות לקבלת הקטגוריה בהנתן מסמך וירטואלי (מופע של שם ישות) d . הקטגוריה עברה מתקבל ערך מכסימאלי היא הנבחרת.

חישוב ההסתברות מתבצע ע"פ הנוסחה:

$$p(c | d) = p(c) \cdot \sum_{ti} \frac{p(T = ti | c) \cdot p(T = ti | d)}{p(T = ti)}$$

עבור c – קטגוריה

d – מסמך וירטואלי = מופע שם ישות.

ti – Token במסמך (במקרה שלנו הכוונה לחוק - Feature).

באשר:

$$p(c) = \frac{\text{מספר מסמכים מסווגים ל-}c}{\text{מספר מסמכים}}$$

$$p(T=ti/c) = \frac{\text{מספר מופעי } ti \text{ במסמכים המסווגים ל-}c}{\text{מספר ביטויים במסמכים המסווגים ל-}c}$$

$$p(T=ti/d) = \frac{\text{מספר מופעי } ti \text{ ב-}d}{\text{מספר ביטויים ב-}d}$$

$$p(T=ti) = \frac{\text{מספר מופעי } ti \text{ ב-Corpus}}{\text{מספר ביטויים ב-Corpus}}$$

היות וכל חוק מופיע פעם אחת בכל מסמך: $p(T = ti | d) = 1$

$$Score = \frac{p(c) \cdot p(T = ti | c)}{p(T = ti)}$$

לכן נקבל:

5. תוצאות

5.1. כללי

5.1.1. מדדי הערכה בסיסיים – הגדרות

מדידת טיב התוצאות נעשתה ע"י מדדים מקובלים להערכת איכות בתחומים של עיבוד שפה טבעית ואחזור מידע. מאחר והשימוש במדדי הערכה אלו נפוץ במגוון רחב של משימות יינתנו תחילה הגדרות כלליות עבורם.

תהי X שאילתא נתונה, תהי A קבוצת התשובות שאוחזרו עבורה ותהי B קבוצת כל התשובות הרלוונטיות לאותה שאילתא. להערכת איכות המענה שמספקת קבוצת התשובות A לשאילתא X מחושב מדד דיוק (precision) וממד כיסוי (recall).

הדיוק מוגדר כיחס בין מספר התשובות הרלוונטיות שאוחזרו למספר התשובות שאוחזרו בכלל.

$$precision = \frac{\|A \cap B\|}{\|A\|}$$

הכיסוי מוגדר כיחס בין מספר התשובות הרלוונטיות שאוחזרו למספר התשובות הרלוונטיות בכלל.

$$recall = \frac{\|A \cap B\|}{\|B\|}$$

כל אחד מן המדדים הנ"ל נותן ציון להיבט מסוים בלבד של איכות האחזור. לכן הוגדר מדד המכמת את איכות האחזור הכוללת ומוצג ע"י Rijsbergen (1979). הממד המכונה F_β , משקלל את זוג הערכים המתקבלים ממדדי הדיוק והכיסוי לכדי ערך יחיד, כאשר β משמש כפרמטר הקובע את אופן השקלול של ערכים אלו.

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{recall + \beta^2 \times precision}$$

מדד F_β מוגדר באופן הבא:

עבור β השווה ל-1 מתקבל מדד שמשקלל באופן זהה את הדיוק והכיסוי

(מתקבל למעשה הממוצע ההרמוני שביניהם):

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

5.1.2. כמויות

להלן טבלאות המתארות את כמויות הנתונים בכל שפה.

עברית		אנגלית		
Test	Train	Test	Train	
199	1693	21	100	מספר מסמכים / משפטים
159	1727	673	4300	מספר שמות עצם פרטיים (מסמכים וירטואלים)
3827		4218		מספר מאפיינים לפני רידוד
267		810		מספר מאפיינים לאחר רידוד

5.1.3. תוצאות אנגלית

דיוק כללי (אחוז השמות שסווגו נכון): 74.14%

סה"כ	ארגון	מקום	איש	תוצאת סיווג
				סיווג נכון
151	26	22	103	איש
212	45	164	3	מקום
310	232	53	25	ארגון
673	303	239	131	סה"כ

מדדי דיוק וכיסוי לקטגוריות:

F1	כיסוי	דיוק	קטגוריה
73.05%	68.21%	78.63%	איש
72.73%	77.36%	68.62%	מקום
75.70%	74.84%	76.57%	ארגון

5.1.4. תוצאות עברית

דיוק כללי (אחוז השמות שסווגו נכון): 76.72%

סה"כ	ארגון	מקום	איש	תוצאת סיווג
				סיווג נכון
22	0	9	13	איש
58	1	54	3	מקום
79	55	24	0	ארגון
159	56	87	16	סה"כ

מדדי דיוק וכיסוי לקטגוריות:

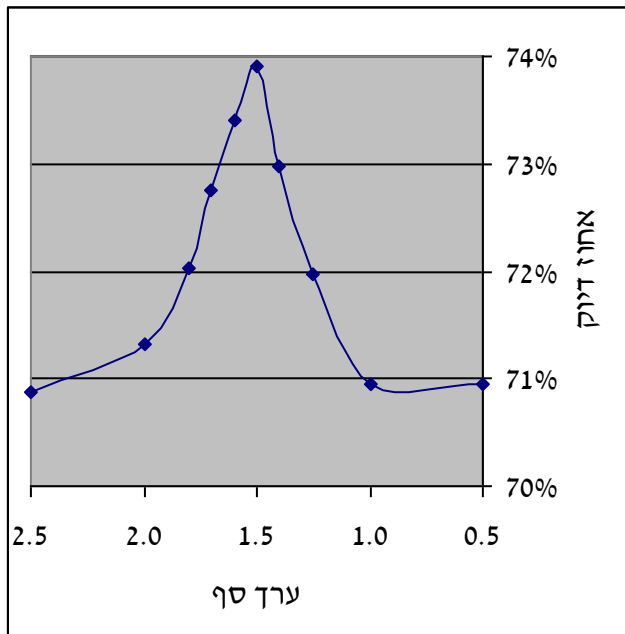
F1	כיסוי	דיוק	קטגוריה
68.42%	59.09%	81.25%	איש
74.48%	93.10%	62.07%	מקום
81.48%	69.62%	98.21%	ארגון

5.2. מצייאת סף

לצורך רידוד מאפיינים המתואר לעיל, נמצא סף מתאים עבור הערך $p(c/T=ti)/p(c)$. מצייאת הסף נעשתה על קורפוס Test (שונה כמובן מהקורפוס עליו מבוצע ה-Test בפועל), מולו הורץ סיווג עבור ספים שונים, ואותר הסף הנותן תוצאה טובה ביותר.

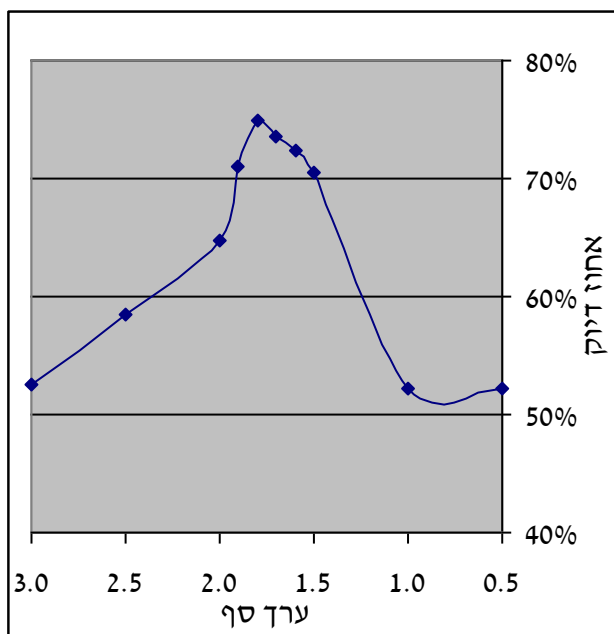
להלן מתוארות תוצאות ההרצה בכל שפה עבור אוסף ערכי סף שנבחנו.

5.2.1. אנגלית



ערך סף	אחוז דיוק
0.5	70.95%
1.0	70.95%
1.3	71.98%
1.4	72.97%
1.5	73.90%
1.6	73.40%
1.7	72.75%
1.8	72.03%
2.0	71.32%
2.5	70.88%

5.2.2. עברית



ערך סף	אחוז דיוק
0.5	52.20%
1.0	52.20%
1.5	70.44%
1.6	72.32%
1.7	73.58%
1.8	74.84%
1.9	71.06%
2.0	64.78%
2.5	58.49%
3.0	52.57%

5.3 מאפיינים חזקים

במסגרת תהליך האבליואציה, נבחנו מגוון צירופי מאפיינים עבור הנתונים. להלן דוגמאות למאפיינים חזקים בכל שפה.

5.3.1 אנגלית

דוגמאות למאפיינים חזקים עבור קטגוריה: איש

מאפיין	משקל
NeighboreAfter1_Base_added	3.84
NeighboreAfter1_Base_administration	3.84
NeighboreAfter1_Base_described	3.84
NeighboreAfter1_Base_family	3.84
NeighboreAfter1_Base_recalled	3.84
NeighboreAfter1_Base_replied	3.84
NeighboreAfter2_Base_commander	3.84
NeighboreAfter2_Base_deputy	3.84
NeighboreAfter2_Base_executive	3.84
NeighboreAfter2_Base_former	3.84
NeighboreAfter2_Base_himself	3.84
NeighboreAfter2_Base_she	3.84
NeighboreAfter2_Base_spokesman	3.84
NeighboreAfter3_Base_administrator	3.84
NeighboreAfter3_Base_analyst	3.84
NeighboreAfter3_Base_assistant	3.84
NeighboreAfter3_Base_chairman	3.84
NeighboreAfter3_Base_vice	3.84
NeighboreAfter3_Base_wife	3.84
NeighboreBefore1_Base_chairman	3.84
NeighboreBefore1_Base_director	3.84
NeighboreBefore1_Base_spokesman	3.84
NeighboreBefore1_Base_spokeswoman	3.84
NeighboreBefore1_Original_Administrator	3.84
NeighboreBefore1_Original_asked	3.84
NeighboreBefore1_Original_Engineer	3.84
NeighboreBefore1_Original_General	3.84
NeighboreBefore1_Original_Minister	3.84
NeighboreBefore1_Original_Mr	3.84
NeighboreBefore1_Original_President	3.84
NeighboreBefore1_Original_says	3.84
NeighboreBefore1_Original_Secretary	3.84

NeighboreBefore2_Base_father	3.84
NeighboreBefore2_Base_husband	3.84
NeighboreBefore2_Base_vice	3.84
NeighboreBefore2_Base_wife	3.84
NeighboreBefore2_Original_corporate	3.84
PartOf_Arthur	3.84
PartOf_Brown	3.84
PartOf_Campbell	3.84
PartOf_End_Clinton	3.84
PartOf_End_Erickson	3.84
PartOf_End_Johnson	3.84
PartOf_End_Jordan	3.84
PartOf_End_Jr.	3.84
PartOf_End_Smith	3.84
PartOf_End_Snyder	3.84
PartOf_Levy	3.84
PartOf_Mid_A.	3.84
PartOf_Mid_E.	3.84
PartOf_Mid_L.	3.84
PartOf_Mid_M.	3.84
PartOf_Mid_R.	3.84
PartOf_Mid_S.	3.84
PartOf_Start_Bill	3.84
PartOf_Start_Bob	3.84
PartOf_Start_James	3.84
PartOf_Start_Peter	3.84
NeighboreAfter2_Base_director	3.29
NeighboreAfter2_Original_director	3.29
NeighboreAfter1_Base_said	3.10
NeighboreAfter3_Base_agency	3.07
NeighboreAfter3_Base_chief	3.07
NeighboreAfter3_Base_director	3.07
NeighboreBefore2_Base_daughter	3.07
NeighboreAfter2_Base_his	3.01
NeighboreBefore3_Base_her	3.01
NeighboreBefore1_Base_said	2.99
NeighboreAfter1_Base_noted	2.88
NeighboreAfter2_Base_president	2.88
NeighboreAfter3_Base_company	2.88
NeighboreAfter3_Original_knew	2.88
NeighboreBefore1_Base_;	2.88
NeighboreBefore1_Base_but	2.88
NeighboreBefore1_Original_was	2.88
NeighboreBefore3_Original_this	2.88
NeighboreAfter1_Base_of	2.85

NeighboreAfter3_Base_former	2.74
PartOf_Mid_POS_UpperAndSigns	2.69
NeighboreAfter2_Base_who	2.67
NeighboreAfter2_Base_her	2.66
NeighboreBefore3_Original_his	2.66
NeighboreAfter1_Base_asked	2.56
NeighboreAfter1_Base_denied	2.56
NeighboreAfter1_Base_suggested	2.56
NeighboreAfter1_Base_told	2.56
NeighboreAfter1_Base_took	2.56
NeighboreAfter1_Base_wrote	2.56
NeighboreAfter2_Base_:	2.56
NeighboreAfter3_Base_professor	2.56
NeighboreAfter3_Original_team	2.56
NeighboreBefore1_Base_were	2.56
NeighboreBefore1_Original_visited	2.56
NeighboreBefore2_Original_daughter	2.56
NeighboreAfter2_Base_he	2.49
NeighboreAfter3_Base_of	2.36
NeighboreAfter2_Base_all	2.30
NeighboreAfter1_Base_also	1.92
NeighboreAfter1_Base_called	1.92
NeighboreBefore1_Base_``	1.76
NeighboreAfter3_Base_'	1.74
NeighboreBefore1_POS_Punct	1.71
NeighboreAfter2_Base_the	1.71
NeighboreAfter2_Original_the	1.71
NeighboreBefore2_POS_Punct	1.68
PartOf_Mid_and	1.60
NeighboreAfter3_Base_``	1.59
NeighboreAfter1_Base_(1.53
NeighboreBefore2_Original_people	1.53
NeighboreBefore2_POS_LettersAndSigns	1.53
NeighboreBefore3_Base_``	1.53
NeighboreBefore3_Original_killed	1.53
NeighboreBefore2_Base_,	1.52

דוגמאות למאפיינים חזקים עבור קטגוריה : מקום

מאפיין	משקל
NeighboreAfter1_Base_airline	3.00
NeighboreAfter1_Base_airport	3.00
NeighboreAfter1_Base_area	3.00
NeighboreAfter1_Base_border	3.00
NeighboreAfter1_Base_carriers	3.00
NeighboreAfter1_Base_forces	3.00
NeighboreAfter1_Base_medical	3.00
NeighboreAfter1_Base_military	3.00
NeighboreAfter1_Base_region	3.00
NeighboreAfter1_Original_coast	3.00
NeighboreAfter2_Base_d.c.	3.00
NeighboreAfter2_Base_ind.	3.00
NeighboreAfter2_Base_n.c.	3.00
NeighboreAfter2_Base_n.y.	3.00
NeighboreAfter2_Base_north carolina	3.00
NeighboreAfter2_Base_ohio	3.00
NeighboreAfter2_Base_san diego	3.00
NeighboreAfter2_Base_scotland	3.00
NeighboreAfter2_Base_south carolina	3.00
NeighboreAfter3_Base_(3.00
NeighboreAfter3_Base_border	3.00
NeighboreAfter3_Base_miles	3.00
NeighboreAfter3_Original_border	3.00
NeighboreBefore1_Base_around	3.00
NeighboreBefore1_Base_into	3.00
NeighboreBefore1_Base_near	3.00
NeighboreBefore1_Original_approached	3.00
NeighboreBefore2_Base_approach	3.00
NeighboreBefore2_Base_arriv	3.00
NeighboreBefore2_Base_bound	3.00
NeighboreBefore2_Base_headquart	3.00
NeighboreBefore2_Base_land	3.00
NeighboreBefore2_Base_mountain	3.00
NeighboreBefore2_Base_north	3.00
NeighboreBefore2_Base_south	3.00
NeighboreBefore2_Base_takeoff	3.00
NeighboreBefore2_Original_back	3.00
NeighboreBefore2_Original_bound	3.00
NeighboreBefore2_Original_runway	3.00
NeighboreBefore2_Original_south	3.00
NeighboreBefore2_Original_waters	3.00
NeighboreBefore3_Base_area	3.00

NeighboreBefore3_Base_left	3.00
NeighboreBefore3_Base_mile	3.00
NeighboreBefore3_Base_mountain	3.00
NeighboreBefore3_Base_olympic	3.00
NeighboreBefore3_Base_state	3.00
NeighboreBefore3_Base_train	3.00
NeighboreBefore3_Base_unit	3.00
PartOf_Africa	3.00
PartOf_America	3.00
PartOf_Balkans	3.00
PartOf_Bosnia	3.00
PartOf_BOSTON	3.00
PartOf_California	3.00
PartOf_D.C.	3.00
PartOf_Dallas	3.00
PartOf_End_Bay	3.00
PartOf_End_City	3.00
PartOf_End_County	3.00
PartOf_End_Desert	3.00
PartOf_End_East	3.00
PartOf_End_Gulf	3.00
PartOf_End_Harbor	3.00
PartOf_End_InternationalAirport	3.00
PartOf_End_Ocean	3.00
PartOf_End_Republic	3.00
PartOf_End_Sea	3.00
PartOf_End_States	3.00
PartOf_End_Station	3.00
PartOf_End_Strait	3.00
PartOf_End_Worth	3.00
PartOf_End_York	3.00
PartOf_Europe	3.00
PartOf_France	3.00
PartOf_Germany	3.00
PartOf_Greece	3.00
PartOf_Ind.	3.00
PartOf_Israel	3.00
PartOf_Japan	3.00
PartOf_Miami	3.00
PartOf_Paris	3.00
PartOf_Start_CAMP	3.00
PartOf_Start_East	3.00
PartOf_Start_Interstate	3.00
PartOf_Start_North	3.00
PartOf_Start_San	3.00

PartOf_Start_south	3.00
PartOf_Start_St.	3.00
PartOf_Start_West	3.00
PartOf_U.S.	3.00
PartOf_End_Airport	2.93
PartOf_POS_UpperAndSigns	2.92
NeighboreBefore2_Base_into	2.86
NeighboreBefore1_Base_in	2.83
NeighboreBefore1_Base_over	2.79
NeighboreBefore2_Base_coast	2.77
NeighboreAfter1_Base_-	2.68
PartOf_Start_New	2.64
NeighboreBefore2_POS_Num	2.60
PartOf_End_Beach	2.57
PartOf_POS_LettersAndSigns	2.53
NeighboreAfter1_Base_;	2.50
NeighboreAfter2_Base_california	2.50
NeighboreBefore3_Original_military	2.50
PartOf_Start_Atlantic	2.50
NeighboreBefore3_Base_land	2.40
NeighboreBefore3_Base_went	2.40
PartOf_Mid_International	2.31
NeighboreAfter1_POS_Date	2.25
NeighboreBefore1_Base_'s	2.25
PartOf_Start_United	2.15
NeighboreAfter2_POS_Date	2.11
PartOf_End_POS_UpperLetters	2.06
NeighboreBefore1_Base_to	2.03
NeighboreAfter1_Base_home	2.00
NeighboreAfter3_Base_near	2.00
NeighboreBefore1_Base_between	2.00
NeighboreBefore2_Base_center	2.00
NeighboreAfter1_Base_to	1.98
NeighboreBefore3_Base_at	1.95
NeighboreBefore1_POS_Preposition	1.93
NeighboreAfter1_Base_on	1.90
NeighboreBefore3_Base_for	1.90
NeighboreBefore3_POS_Num	1.83
NeighboreAfter1_Base_a	1.80
NeighboreBefore3_Base_between	1.80
PartOf_Start_POS_LowerLetters	1.69
NeighboreAfter1_POS_Preposition	1.55
NeighboreAfter3_Base_,	1.54

דוגמאות למאפיינים חזקים עבור קטגוריה : ארגון

מאפיין	משקל
NeighboreAfter1_Base_administrator	2.46
NeighboreAfter1_Base_agents	2.46
NeighboreAfter1_Base_aircraft	2.46
NeighboreAfter1_Base_airliner	2.46
NeighboreAfter1_Base_chairman	2.46
NeighboreAfter1_Base_employees	2.46
NeighboreAfter1_Base_executive	2.46
NeighboreAfter1_Base_inspector	2.46
NeighboreAfter1_Base_investigation	2.46
NeighboreAfter1_Base_officer	2.46
NeighboreAfter1_Base_passengers	2.46
NeighboreAfter1_Base_plane	2.46
NeighboreAfter1_Base_president	2.46
NeighboreAfter1_Base_report	2.46
NeighboreAfter1_Base_secretary	2.46
NeighboreAfter1_Base_vice	2.46
NeighboreAfter1_Original_Administrator	2.46
NeighboreAfter1_POS_UpperDigitsAndSigns	2.46
NeighboreAfter2_Base_-	2.46
NeighboreAfter2_Base_)	2.46
NeighboreAfter2_Base_army	2.46
NeighboreAfter2_Base_confirmed	2.46
NeighboreAfter2_Base_grounded	2.46
NeighboreAfter2_Base_investigating	2.46
NeighboreAfter2_Base_showed	2.46
NeighboreAfter3_Original_problems	2.46
NeighboreBefore1_Original_based	2.46
NeighboreBefore1_Original_called	2.46
NeighboreBefore1_Original_firm	2.46
NeighboreBefore2_Base_``	2.46
NeighboreBefore2_Base_analyst	2.46
NeighboreBefore2_Base_official	2.46
NeighboreBefore2_Base_week	2.46
NeighboreBefore2_Original_told	2.46
NeighboreBefore3_Base_director	2.46
NeighboreBefore3_Original_According	2.46
NeighboreBefore3_POS_DigitsAndSigns	2.46
PartOf_American	2.46
PartOf_Army	2.46
PartOf_Commerce	2.46
PartOf_Congress	2.46
PartOf_Continental	2.46

PartOf_defense	2.46
PartOf_End_Administration	2.46
PartOf_End_Agency	2.46
PartOf_End_Airlines	2.46
PartOf_End_Association	2.46
PartOf_End_Authority	2.46
PartOf_End_Center	2.46
PartOf_End_Club	2.46
PartOf_End_Co.	2.46
PartOf_End_Committee	2.46
PartOf_End_Corp.	2.46
PartOf_End_Corps	2.46
PartOf_End_Council	2.46
PartOf_End_Department	2.46
PartOf_End_Engineers	2.46
PartOf_End_Group	2.46
PartOf_End_House	2.46
PartOf_End_Inc.	2.46
PartOf_End_Ministry	2.46
PartOf_End_Nation	2.46
PartOf_End_Nations	2.46
PartOf_End_Navy	2.46
PartOf_End_Press	2.46
PartOf_End_Service	2.46
PartOf_End_Staff	2.46
PartOf_End_Tech	2.46
PartOf_End_Times	2.46
PartOf_End_Transportation	2.46
PartOf_End_University	2.46
PartOf_FBI	2.46
PartOf_Mid_Airlines	2.46
PartOf_Mid_Associated	2.46
PartOf_Mid_Aviation	2.46
PartOf_Mid_Bureau	2.46
PartOf_Mid_Coast	2.46
PartOf_Mid_County	2.46
PartOf_Mid_Department	2.46
PartOf_Mid_for	2.46
PartOf_Mid_Institute	2.46
PartOf_Mid_News	2.46
PartOf_Mid_Pacific	2.46
PartOf_Mid_Security	2.46
PartOf_Mid_Space	2.46
PartOf_Mid_World	2.46
PartOf_POS_LowerLetters	2.46

PartOf_Reuters	2.46
PartOf_Start_Boston	2.46
PartOf_Start_Defense	2.46
PartOf_Start_European	2.46
PartOf_Start_Host	2.46
PartOf_Start_House	2.46
PartOf_Start_National	2.46
PartOf_Start_POS_LettersAndDigits	2.46
NeighboreAfter1_POS_Num	2.43
NeighboreAfter1_Base_official	2.30
NeighboreBefore2_Original_president	2.24
PartOf_Mid_National	2.24
PartOf_POS_UpperLetters	2.24
NeighboreBefore1_Base_an	2.23
PartOf_Mid_POS_LettersAndSigns	2.15
PartOf_Mid_of	2.10
PartOf_Mid_POS_Preposition	2.08
NeighboreAfter2_POS_DigitsAndSigns	2.05
NeighboreBefore1_Base_a	2.03
PartOf_Len_4To5	1.98
NeighboreAfter3_Base_first	1.97
NeighboreBefore1_POS_Date	1.97
NeighboreBefore2_Base_an	1.97
PartOf_End_POS_LettersAndSigns	1.96
PartOf_Mid_POS_Letters	1.94
NeighboreBefore1_Base_(1.90
PartOf_End_POS_Num	1.88
PartOf_Start_POS_UpperAndSigns	1.85
NeighboreBefore2_Base_that	1.82
NeighboreAfter1_Base_who	1.76
PartOf_Start_the	1.76
NeighboreBefore1_Base_the	1.75
NeighboreAfter3_Base_are	1.64
NeighboreAfter3_Base_help	1.64
NeighboreAfter3_Base_least	1.64
NeighboreAfter3_Base_major	1.64
NeighboreAfter3_Base_most	1.64
NeighboreBefore1_Original_Some	1.64
NeighboreBefore2_Original_group	1.64
NeighboreBefore2_Original_security	1.64
NeighboreBefore3_Original_president	1.64
NeighboreBefore3_Original_vice	1.64
NeighboreBefore3_POS_Date	1.63
NeighboreAfter1_POS_Name	1.54

דוגמאות למאפיינים חזקים עבור קטגוריה : איש

מאפיין	משקל
NeighboreAfter1_Base_חזר	2.12
NeighboreAfter1_Base_מחפש	2.12
NeighboreAfter1_Base_עלול	2.12
NeighboreAfter1_Base_שלה	2.12
NeighboreAfter2_Base_איש	2.12
NeighboreAfter2_Base_אשר	2.12
NeighboreBefore1_Base_אמר	2.12
NeighboreBefore1_Base_גנרל	2.12
NeighboreBefore1_Base_ר"ד	2.12
NeighboreBefore1_Base_ך"ח	2.12
NeighboreBefore1_Base_כש	2.12
NeighboreBefore1_Base_כתב	2.12
NeighboreBefore1_Base_לשעבר	2.12
NeighboreBefore1_Base_מחנה	2.12
NeighboreBefore1_Base_נזכר	2.12
NeighboreBefore1_Base_ניצב	2.12
NeighboreBefore1_Base_עשי	2.12
NeighboreBefore1_Base_פקד	2.12
NeighboreBefore1_Base_רב	2.12
NeighboreBefore1_Base_שר	2.12
NeighboreBefore1_Original_הגנרל	2.12
NeighboreBefore1_Original_הוא	2.12
NeighboreBefore1_Original_היה	2.12
NeighboreBefore1_Original_המאמן	2.12
NeighboreBefore1_Original_הנשיא	2.12
NeighboreBefore1_Original_הרב	2.12
NeighboreBefore1_Original_השר	2.12
NeighboreBefore1_Original_כש	2.12
NeighboreBefore2_Base_הוא	2.12
NeighboreBefore2_Base_החליף	2.12
NeighboreBefore2_Base_יחד	2.12
NeighboreBefore2_Base_מסוים	2.12
NeighboreBefore2_Base_רפובליקאי	2.12
NeighboreBefore2_Base_שם	2.12
NeighboreBefore2_Original_אחרת	2.12
NeighboreBefore2_Original_אנשים	2.12
NeighboreBefore2_Original_זו	2.12
NeighboreBefore3_Original_שר	2.12
PartOf_End_וייט	2.12
PartOf_End_ויליאמס	2.12

PartOf_End_חסיין	2.12
PartOf_End_מובארק	2.12
PartOf_End_פולק	2.12
PartOf_End_רייגן	2.12
PartOf_End_שריד	2.12
PartOf_Mid_בן	2.12
PartOf_Start_גיימס	2.12
PartOf_Start_דוד	2.12
PartOf_Start_ויליאם	2.12
PartOf_Start_חיים	2.12
PartOf_Start_יורק	2.12
PartOf_Start_משה	2.12
PartOf_Start_סדאם	2.12
PartOf_Start_פול	2.12
PartOf_Start_פטריק	2.12
PartOf_Start_צבי	2.12
PartOf_בוש	2.12
PartOf_וייט	2.12
PartOf_שמיר	2.12
PartOf_שרון	2.12
NeighboreAfter1_Base_היה	1.93
NeighboreAfter1_Original_היה	1.93
NeighboreAfter1_Base_אמר	1.93
NeighboreAfter1_Original_אמר	1.93
NeighboreBefore1_POS_Adverb	1.88
NeighboreBefore1_POS_Verb	1.87
NeighboreBefore2_Base_זה	1.85
NeighboreAfter2_Base_כ	1.82
NeighboreBefore1_Base_כ	1.82
NeighboreBefore1_Base_מאמן	1.82
NeighboreBefore1_Base_נשיא	1.82

דוגמאות למאפיינים חזקים עבור קטגוריה : מקום

מאפיין	משקל
NeighboreBefore1_Base_אוניברסיטה	3.56
NeighboreBefore1_Base_ארץ	3.56
NeighboreBefore1_Base_בית	3.56
NeighboreBefore1_Base_מזרח	3.56
NeighboreBefore1_Base_מחוז	3.56
NeighboreBefore1_Base_מערב	3.56
NeighboreBefore1_Base_משטרה	3.56
NeighboreBefore1_Base_עיר	3.56
NeighboreBefore1_Base_עירייה	3.56
NeighboreBefore1_Base_רחוב	3.56
NeighboreBefore1_Base_רצועה	3.56
NeighboreBefore1_Original_במדינת	3.56
NeighboreBefore1_Original_במזרח	3.56
NeighboreBefore1_Original_ברחוב	3.56
NeighboreBefore1_Original_ברצועת	3.56
NeighboreBefore1_Original_יהודי	3.56
NeighboreBefore1_Original_מדינת	3.56
NeighboreBefore1_Original_משטרת	3.56
NeighboreBefore1_Original_עיריית	3.56
NeighboreBefore2_Base_הגיע	3.56
NeighboreBefore2_Base_יצא	3.56
NeighboreBefore2_Original_בא	3.56
NeighboreBefore2_Original_במלון	3.56
NeighboreBefore3_Original_בית	3.56
PartOf_End_אנגלס	3.56
PartOf_End_גרסי	3.56
PartOf_End_זילנד	3.56
PartOf_End_יורק	3.56
PartOf_Start_בלוס	3.56
PartOf_Start_ניו	3.56
PartOf_אוסטרליה	3.56
PartOf_ב"ארה	3.56
PartOf_בירושלים	3.56
PartOf_בישראל	3.56
PartOf_בלונדון	3.56
PartOf_בשכמ	3.56
PartOf_לפולין	3.56
PartOf_קרקוב	3.56
NeighboreBefore1_Base_מדינה	3.24
NeighboreAfter1_Base_;	2.97
NeighboreBefore2_Base_בא	2.97
NeighboreAfter1_Base_היא	2.85

NeighboreBefore2_Base_יהודי	2.85
NeighboreBefore1_Base_ב	2.78
NeighboreBefore1_Original_ב	2.78
NeighboreBefore1_Base_יהודי	2.67
NeighboreBefore1_Base_שכונה	2.67
NeighboreBefore1_Original_בשכונת	2.67
NeighboreBefore1_Base_מ	2.41
NeighboreAfter2_Base_סמוך	2.37
NeighboreBefore2_Base_אתמול	2.37
NeighboreBefore2_Base_מ	2.37
NeighboreBefore2_Original_מ	2.37
NeighboreBefore2_Original_ראש	2.37
NeighboreBefore2_Original_רק	2.37
NeighboreBefore2_POS_AuxVerb	2.37
NeighboreBefore3_Base_מדינה	2.37
NeighboreBefore3_Base_עיר	2.37
NeighboreBefore3_Original_השבוע	2.37
NeighboreBefore3_POS_AuxVerb	2.37
NeighboreAfter2_Base_של	2.23
NeighboreAfter2_Original_-	2.23
NeighboreAfter1_Base_היתה	2.14
NeighboreBefore2_Base_"	2.08
NeighboreBefore2_Original_"	2.08
NeighboreBefore2_Base_ב	1.84

דוגמאות למאפיינים חזקים עבור קטגוריה : ארגון

מאפיין	משקל
NeighboreAfter1_Base_מסר	5.45
NeighboreBefore1_Base_איש	5.45
NeighboreBefore1_Base_דובר	5.45
NeighboreBefore1_Base_הנהלה	5.45
NeighboreBefore1_Base_חברה	5.45
NeighboreBefore1_Base_קרן	5.45
NeighboreBefore1_Base_שחקן	5.45
NeighboreBefore1_Original_אנשי	5.45
NeighboreBefore1_Original_דובר	5.45
NeighboreBefore1_Original_החברה	5.45
NeighboreBefore2_Base_כות	5.45
NeighboreBefore2_Base_משחק	5.45
NeighboreBefore2_Base_ניצחון	5.45
NeighboreBefore2_Base_עיתון	5.45
NeighboreBefore2_Original_תנועת	5.45
PartOf_End_גלוב	5.45
PartOf_End_גן	5.45
PartOf_End_חיפה	5.45
PartOf_End_יוניטד	5.45
PartOf_End_ירושלים	5.45
PartOf_End_ישראל	5.45
PartOf_Len_MoreThan5	5.45
PartOf_Start_א	5.45
PartOf_Start_בוסטון	5.45
PartOf_Start_הפועל	5.45
PartOf_Start_מכבי	5.45
PartOf_Start_ראשון	5.45
PartOf_אש"ף	5.45
PartOf_הארץ	5.45
PartOf_כך	5.45
PartOf_פורד	5.45
PartOf_רוקפלר	5.45
PartOf_Mid_.	4.77
NeighboreBefore2_Base_תנועה	4.77
NeighboreAfter2_Base_ניצחון	4.36
NeighboreBefore2_Base_אירח	3.63
NeighboreBefore2_Original_היה	3.63
NeighboreBefore3_Base_סיום	3.63
NeighboreBefore1_Base_יד	3.11
NeighboreBefore3_Base_:	3.11
NeighboreBefore1_Base_"	2.96
NeighboreBefore1_Original_"	2.96

NeighboreAfter1_Base_"	2.85
NeighboreAfter1_Original_"	2.85
NeighboreBefore1_Base_מרכז	2.72
NeighboreBefore3_Base_בית	2.72
NeighboreAfter2_Base_Num	2.49
NeighboreAfter2_POS_Num	2.49
PartOf_Mid_-	2.26
NeighboreAfter2_Original_,	2.18
NeighboreAfter2_Original_היו	2.18
NeighboreBefore2_Base_(2.18
NeighboreBefore3_Base_)	2.18
NeighboreAfter1_POS_Name	2.04
NeighboreAfter2_Base_כי	2.04
NeighboreBefore1_Base_(1.98
NeighboreBefore2_Original_ל	1.98
PartOf_Len_4To5	1.98
NeighboreAfter2_Base_ה	1.82
NeighboreAfter2_Base_ולא	1.82
NeighboreAfter3_Base_(1.82
NeighboreAfter3_Original_(1.82
NeighboreBefore2_Base_כי	1.82
NeighboreBefore2_Base_ל	1.82
NeighboreBefore2_Base_לאומי	1.82
NeighboreBefore2_Original_החוץ	1.82
NeighboreBefore2_Original_ראש	1.82

5.4. משפחות מאפיינים

אחת התובנות שהתקבלו במסגרת תהליך האבליואציה, הינה השפעת סוגי המאפיינים השונים על התוצאות. כלומר, אם נביט ברשימת המאפיינים המשמעותיים לשתי השפות, אלו סוגי מאפיינים מככבים ברשימות אלו.

קטגוריית שם איש:

בשתי השפות ניתן להבחין, כי המאפיינים החזקים הם הצורה הבסיסית של השכנים לפני ואחרי השם. מדובר בעיקר בפעלים המתאימים לבני אדם (כגון: Described, חזר) או בתארי שמות (כגון: Commander, גנרל).

כמו כן, במאפיינים מסוג PartOf מופיעים (כצפוי) בעיקר שמות אנשים נפוצים – שמות פרטיים ושמות משפחה. לעיתים מופיעים גם תארים כגון Dr.

קטגוריית שם מקום:

בקטגוריה זו קים הבדל בין השפות. הן בשפה האנגלית והן בשפה העברית מאפיינים חזקים הינם השכנים לפני השם. מדובר בעיקר במילים הקשורים למקומות (בעיקר שמות עצם מעט תארים כמעט ואין פעלים) כגון: אוניברסיטה, ארץ, מחוז, עיר, מזרח. אך באנגלית קימת תרומה משמעותית גם לשכנים שמופיעים אחרי שם המקום, למשל: State, Area.

כמו כן, בשתי השפות מאפייני PartOf, הינם שמות מקומות נפוצים, אך באנגלית נוספים גם קידומות וסיומות מקובלים (מילים הקשורים למקומות) כגון: East, Country, City.

קטגוריית שם ארגון:

בקטגוריית שם ארגון, הן באנגלית והן בעברית, קיים מספר קטן יחסית של מאפיינים מסוג שכנים לפני או אחרי. המילים המקיפות את השם שנבחרו כמאפיינים, אינן בהכרח קשורות לארגונים. בולטים בעיקר מאפיינים הבוחנים כללי ניקוד וצורה. למשל: גרשיים מקיפות, מספר שכן, תאריך שכן.

לגבי חוקי ה-PartOf, להבדיל משמות אנשים ומקומות, בקטגוריה זו אין הרבה שמות נפוצים מספיק כדי לשרוד כמאפיין חזק.

מלבד שמות ארגונים נפוצים, מופיעות במאפיינים חזקים מסוג PartOf באנגלית, מילים הקשורות לארגון כגון: Agency, Corp., Group, Club, Department. כמו כן מופיעים באנגלית מאפייני צורה יותר מהקטגוריות

האחרות. כגון: מכיל ספרות, מכיל סימנים, מכיל Token שבו כל האותיות גדולות, מכיל מספר Token-ים גבוה.

5.5 בחינת חסינות (Robustness)

במסגרת תהליך האבליואציה, נבחנה מידת חסינות (Robustness) השיטה. בוצעה השוואה בין שני סטים של מסמכים לסיווג. ההשוואה כללה השוואת משפחות מאפיינים שהתקבלו, והן השוואת תוצאות.

5.5.1 אנגלית

נבחן סט של 20 מסמכים שונים ולהלן התוצאות שהתקבלו:

דיוק כללי (אחוז השמות שסווגו נכון): 73.77%

תוצאת סיווג / סיווג נכון	איש	מקום	ארגון	סה"כ
איש	109	23	29	161
מקום	10	228	48	286
ארגון	27	52	192	271
סה"כ	146	303	269	718

מדדי דיוק וכיסוי לקטגוריות:

קטגוריה	דיוק	כיסוי	F1
איש	74.66%	67.70%	71.00%
מקום	75.25%	79.72%	77.42%
ארגון	71.38%	70.85%	71.11%

5.5.2. עברית

נבחן סט של 200 משפטים שונים ולהלן התוצאות שהתקבלו :

דיוק כללי (אחוז השמות שסווגו נכון) : 72.10%

סה"כ	ארגון	מקום	איש	תוצאת סיווג סיווג נכון
78	2	11	65	איש
45	1	27	17	מקום
24	14	9	1	ארגון
147	17	47	83	סה"כ

מדדי דיוק וכיסוי לקטגוריות :

F1	כיסוי	דיוק	קטגוריה
80.75%	83.33%	78.31%	איש
58.70%	60.00%	57.47%	מקום
68.29%	58.33%	82.24%	ארגון

6. סיכום ומסקנות

עבודה זו מראה שיטה לסיווג אוטומטי של שמות עצם פרטיים כמעט ללא תלות בשפה. גישה זו איפשרה בחינה של מספר משפחות מאפיינים ואת מידת השפעתם. בשיטה זו התקבלה עבור כל קטגורית שם עצם פרטי, רשימת מאפיינים שניתן להשתמש בהם (ע"פ ציון סף התואם סט מסמכי אימון). ע"פ רשימה זו נבנו מסמכים וירטואלים עבור כל מופע שם בטקסט, המכילים מאפיינים מתאימים לכל שם. מסמכים וירטואלים אלו סווגו ע"פ אלגוריתם סיווג לקטגורית שם העצם המתאימה.

במסגרת עבודה זו נבחנו מסמכים בשתי שפות שונות: אנגלית ועברית, בהם התקבלו תוצאות דומות, ללא צורך כמעט בשינויי קוד.

כמו כן נבחנה מידת חסינות השיטה בכל אחת מהשפות ונמצאו תוצאות דומות גם באספקט זה.

רשימת המאפיינים שהתקבלה לכל קטגורית שם יכולה להעיד לגבי משפחות וסוגי מאפיינים שכדאי יותר להשתמש בהם ולגבי מאפיינים שמיותר להתייחס אליהם. גם מבחינה זו דומות התוצאות בין השפות שנבחנו (אם כי אינן זהות). תובנה כזו יכולה אף לאפשר נקודת מוצא טובה לעבודה משולבת חוקים ידניים.

7. כיוונים עתידיים

7.1. צירופי מאפיינים

במסגרת עבודה זו נבחנו סוגי מאפיינים שונים. עם זאת ניתן לבחון שימוש בצירופי מאפיינים כגון: ה-Token לפני ואחרי השם, שני ה-Token-ים הקודמים לשם, שני ה-Token-ים העוקבים אחר השם ועוד.

בשיטה כזו ניתן להשתמש בביטויים המכילים מספר מילים כמאפיינים. כמו כן, שיטות פיסוק כגון " לפני ואחרי השם יופיעו כמאפיין. בדוגמה זו למשל, יקבל המאפיין ציון גבוה לשם ארגון, פחות לשם מקום ועוד פחות לשם איש. כך יובלט השימוש בגרשיים לסימון ארגון, לעומת שימוש במאפייני גרשיים לפני ואחרי השם בנפרד, המוסיפים רעש של שמות שהופיעו כחלק מציטוט וכיו"ב.

7.2. אלגוריתמים נוספים

כובד המשקל בעבודה זו הושם על שיטת הוצאת המאפיינים ודירוגם. האלגוריתם בו נעשה שימוש הינו אלגוריתם בייסיאני בסיסי המשרת מטרות אלו בצורה טובה.

עם זאת, ניתן לבצע בחינה של התוצאות שיתקבלו מול אלגוריתם סיווג ממשפחה שונה. לצורך הנושא ניתן לבחון את אלגוריתם ה-SVM.

7.3. סיווג מסמכים לנושאים

ברור כי למסמכים שונים קיימות תכונות שונות. בבואנו משימת עיבוד טקסט חשוב לבחון את השפעת תכונות אלו על התוצאות. סגנון הכתיבה של כותב המאמר, למשל, מכתיב את המאפיינים שיופיעו סביב השם. אך נתון מכריע וניתן לחישוב עוד יותר הינו נושא המסמך.

למשל, במאמר העוסק בספורט ניתן להניח שנמצא הרבה שמות קבוצות, המסווגים כארגונים, ושמות שחקנים המסווגים כאנשים. גם במסמך העוסק בפוליטיקה יופיעו ארגונים ואנשים, אלא שהפעם הארגונים יהיו מפלגות והאנשים יהיו חברי כנסת ושרים.

סגנון הכתיבה והמאפיינים המתאימים לקטגוריות בשני המקרים שונים בתכלית. לכן מומלץ לבצע תחילה חלוקה של המסמכים הנתונים לנושאים ואימון כל נושא בנפרד. כך, בתהליך הסיווג, יתבצע תחילה סיווג של המסמך הנתון לנושא, ורק לאחר מכן סיווג שמות העצם הפרטיים המופיעים בו לפי סט המאפיינים שהתקבלו באימון לפי נושא המסמך הרלוונטי.

7.4. איחוד מופעי שם ברמת מסמך

מסמך אחד מכיל מספר מופעי שמות עצם פרטיים. הנחה סבירה היא, כי אם מופיעות אותן מילים המרכיבות את השם, מספר פעמים באותו מסמך, המדובר באותו שם. כלומר, אם המסמך מכיל מספר פעמים את המילים "תל אביב" כשם עצם פרטי, יש להניח שמדובר באותו שם. לכן, אם ברוב הפעמים סווג השם לקטגורית שם מקום, נשנה את שאר הסיווגים ע"פ "החלטת הרוב".

הבעיה בגישה זו היא לגבי שמות "בעייתיים" למשל: "ישראל ודני גרים בירושלים". בשנת 2004 הם עזבו את ישראל ונסעו לארה"ב". כאן השיטה תקלקל את אחד הסיווגים. כמו כן, במקרים שם החברה הוא כשם המוצר העיקרי שלה, שוב תתעורר בעיה דומה. בכל זאת יתכן וניתן להשתמש בגישה זו עבור שמות המכילים מספר Token-ים.

7.5. סיווג קשרים באופן אוטומטי

ניתן לבחון את האפשרות לתקוף בעיה מורכבת יותר בעזרת שיטה זו – בעית סיווג קשרים בטקסט. כלומר בהינתן צמד ישויות המופיעות בטקסט, איתור סוג הקשר ביניהן אם קיים.

ניתן לאתר ולסווג את הישויות במסמך בעזרת השיטה שהוצגה, ולהשתמש בסט מאפיינים חדש, המתאר את הקשר, לצורך סיווג.

למשל, עבור המשפט: "ישראל ודני גרים בירושלים." תתקבל בסוף השלב הראשון התבנית הבאה: <שם איש> ו<שם איש> גרים ב<שם מקום>.

ומכאן, לקשר מסוג כתובת מגורים (בין איש למקום) יופיעו מאפיינים כגון: <שם איש>, <שם מקום>, "גרים ב" מופיעים יחד באותו משפט/פסקה.

גם בבעיה זו, כמו בבעית סיווג שמות, נוציא את מירב המאפיינים האפשריים ונרדדס ע"פ מידת תרומתם לסיווג.

8. ביבליוגרפיה (1)