**זיהוי Named Entities ע"י חיפושים מקוונים**

**פרויקט בעיבוד שפה טבעית**

**נגה צוקר 300859402**

**אביגיל פגי 304756067**

תוכן עניינים

[מבוא- מטרת הפרויקט 3](#_Toc428221424)

[תיאור פעולת הכלי 4](#_Toc428221425)

[תיאור המסווגים 6](#_Toc428221426)

[1. מסווג DuckDuckGo כללי 6](#_Toc428221427)

[2. מסווג DuckDuckGo מכוון חברות 6](#_Toc428221428)

[3. מסווג Facebook כללי 7](#_Toc428221429)

[4. מסווג על פי אות גדולה 8](#_Toc428221430)

[5. מסווג על פי חיפוש במילון 8](#_Toc428221431)

[Training – כיול מסווגים בנפרד 9](#_Toc428221433)

[1. מסווג DuckDuckGo כללי 10](#_Toc428221434)

[2. מסווג DuckDuckGo לחברות 11](#_Toc428221435)

[3. מסווג פייסבוק + מילון 12](#_Toc428221436)

[4. מסווג אותיות גדולות 13](#_Toc428221437)

[5. מסווג מילון 14](#_Toc428221438)

[Training - שילוב המסווגים 16](#_Toc428221439)

[Testing 18](#_Toc428221440)

[פוטנציאל בשפות נוספות 20](#_Toc428221442)

[סיכום ומסקנות 21](#_Toc428221443)

[הצעות לשיפור המערכת 21](#_Toc428221444)

[נספחים 22](#_Toc428221445)

[נספח א'- 22](#_Toc428221446)

[מילות מפתח duckduckgo 22](#_Toc428221447)

[קטגוריות חיפוש Facebook 22](#_Toc428221448)

[נספח ב'- ביבליוגרפיה 23](#_Toc428221449)

[קבצים נבדקים 23](#_Toc428221450)

[כלים נוספים 23](#_Toc428221451)

# מבוא- מטרת הפרויקט

מטרת הפרויקט לבחון האם ניתן לזהות Named Entity בטקסט לפי תוצאות החיפוש שחוזרות מול מנועי חיפוש אינטרנטיים ורשתות חברתיות. הסיווג מתבצע לפי אחת מ-5 קטגוריות-

1. שם של אדם
2. שם של חברה / ארגון
3. שם של מיקום
4. שם כלשהו (NE כללי)
5. מילה רגילה

להערכתנו אם ניתן לזהות בהצלחה את ה-NE-ים בשיטה זו, יש בכך מספר יתרונות מרכזיים:

1. השיטות שהוצגו בכיתה מבוססות בעיקר על שימוש באות גדולה בתחילת שמות, לעומת זאת בשיטה זו ניתן לבדוק את כל המילים שמופיעים במאגר ולכן היא תהיה רלוונטית גם למאגרים פחות פורמליים כמו ברשתות חברתיות שלא נוהגים להקפיד על האות הגדולה.
2. מרגע שזוהתה מילה רלוונטית, כגון שם של בן אדם, ניתן לנסות להגדיל את החלון ולזהות רצפים יותר מורכבים של NE-ים.
3. שיטה זו עשוייה להתאים גם לשפות נוספות ולא רק אנגלית על ידי התאמות יחסית פשוטות כמו התייחסות לרשתות החברתיות שפופולריות במקומות שונים (לדוגמא ברוסיה נפוץ יותר השימוש ברשת חברתית מקומית ולא פייסבוק).
4. יתכן שבשיטה זו ניתן לשמור מידע נוסף על אותו ה-named entity כגון – tag של זכר\נקבה, שיכול לעזור בהמשך ניתוח המידע באפליקציות שונות לדוגמא:
   1. בתרגום מאנגלית לעיתים הדרך היחידה להסיק האם מדובר בזכר או נקבה זה לפי שם האדם, וכשבאים לתרגם לשפה אחרת מידע זה נחוץ לצורך תרגום נכון והטיית כלל המילים במשפט.
   2. כאשר מנסים לעשות סיווג לטקסט - זיהוי שרוב הדמויות בכתבה הם מתחום מסויים (לדוגמא- פוליטיקאים) מגדיל את הסבירות שמדובר בכתבה שעוסקת באותו תחום.

# תיאור פעולת הכלי

בהנתן קלט של טקסט כלשהו מתבצעים השלבים הבאים:

1. **טוקניזציה של הקלט - חלוקה לשורות והפרדת סימני פיסוק.**

מילים שבהן יש יותר מנקודה אחת (כגון U.S.A) ומילים ששייכות לתארים (כגון Dr., Mrs וכו') לא עוברות שינוי. כל שאר סימני הפיסוק (כולל מרכאות ו-apostrophe שמסמן שייכות) מרופדים עם רווחים לפני ואחרי כך שלאחר פיצול השורה לטוקנים לפי רווחים נשארים עם 2 סוגי טוקנים – סימני פיסוק ומילים.

1. **בחירת הביטויים שנרצה לחפש עליהם**

אורך ה-ngram של ביטויים ההתחלתיים שמעניינים לחפש הוא קונפיגורבילי ומאפשר לבדוק בסדר יורד את כל הביטויים מאורך n ועד אורך 1 , כאשר לא נבדוק מחדש ביטוי קצר שמוכל בתוך ביטוי ארוך שכבר סווג כ- NE כלשהו.

* + במקרה ש-ngram=1:
    - מסננים החוצה סימני פיסוק ו-stop words (רשימה מלאה של ה-stop words מופיעה בהמשך) מכיוון שאין הצדקה לבצע עליהם חיפוש.
  + במקרה ש-ngram>1:
    - לא מחפשים ביטוי שמכיל סימן פיסוק בתוכו. (לדוגמא במקרה ש-ngram=3 ומקבלים apples,oranges,grapes נחפש כל מילה בנפרד בלי הפסיק באמצע)
    - אם הביטוי מסתיים ב-stop word מרחיבים אותו (לדוגמא עבור ngram=2 הביטוי "United States of America" יופרד ל-
      * United States
      * States of America (במקום ""States of ו-""of America)
    - לא מחפשים ביטוי שמתחיל ב-stop word ( כי לפי הכלל הקודם כבר הרחבנו את החלון כשהמילה האחרונה היתה stop word).
    - היוצאים מן הכלל לשני הכללים הקודמים הם ה-apostrophe של שייכות ומקף (hyphen) שמתייחסים אליהם בתור stop words – כלומר לא נחפש בנפרד את הטוקן של ה-apostrophe או הטוקן של "s" אבל כן נחפש ביטויים כמו Diner Tony's כביטוי אחד.

1. **בחינת כל ביטוי בנפרד אל מול כלל המסווגים הפעילים** (ניתן לבחור להפעיל רק חלק מהמסווגים במהלך הרצה או לתת למסווגים מסויימים יותר משקל על פני אחרים):
   * מסווג duckduckgo
   * מסווג duckduckgo מכוון חברות
   * מסווג facebook
   * מסווג על פי חיפוש במילון
   * מסווג אות גדולה

כאשר כל מסווג מנקד בנפרד את הסבירות שמדובר ב-Name,Company,Location או Regular (מילה רגילה) ומחזיר את התוצאה שלו.

נדגיש כי כלל המסווגים הם דטרמיניסטיים ויחזירו את אותם התוצאות לא משנה באיזה הקשר מחפשים אותם. לכן, כדי למנוע עומס בקשות שעשוי לגרום לחסימות של Facebook\Duckduckgo אנחנו משתמשים ב-cache ששומר את הציונים של ביטויים שכבר חיפשנו.

1. **שקלול כלל הציונים של המסווגים הפעילים וקביעת סוג הביטוי.**

באופן כללי, סוג התיוג שקיבל את הציון הגבוה ביותר לאחר שקלול הציונים שניתנו על ידי המסווגים (לפי המשקל היחסי שלהם) הוא התיוג המנצח. עם זאת, ברור כי אם נשתמש במסווגים שונים עם משקלות שונים, ייתכן שנגיע לסיווגים קרובים מאוד זה לזה, ואז סטיה קטנה לכיוון מסויים לא אמורה לשנות את ההחלטה מקצה לקצה. לכן, אם יש כמה ציונים גבוהים קרובים זה לזה, אזי המונח מקבל את התיוג- "NE", כלומר- מדובר בשם, לא ידוע מאיזה סוג.

לכן, אנו לוקחים את הציון הגבוה ביותר, מחשבים אחוז מסויים ממנו (זהו פרמטר קונפיגורבילי- כרגע 90%), ואם יש יותר ממסווג אחד שקיבל ציון גבוה מסף זה, המונח יתוייג כ-NE.

1. **הרחבת הביטוי**:

ללא תלות באורך הביטוי המקורי שחיפשנו, במקרה שנקבע שהביטוי המקורי הוא NE כלשהו ננסה להרחיב אותו, ראשית עם המילים שהופיעו לפניו במשפט (בתנאי שהם לא סימני פיסוק או stop words) ולאחר מכן עם המילים שהופיעו אחריו (כאשר כמו מקודם נדאג לא לחפש ביטוי שמכיל סימני פיסוק או נגמר ב-stop word).

במקרה שהביטוי המקורי סווג בקטגוריה אחת והביטוי המורחב סווג בקטגוריה אחרת – אנחנו מתייחסים לקטגוריה של הביטוי הכי ארוך שהצלחנו למצוא.

בשיטה זו מספיק לזהות מילה אחת כפוטנציאל ל-NE בשביל ללמוד את הביטוי המלא ולזהות בהצלחה שם + שם משפחה (גם כאשר לא מדובר באנשים מפורסמים) , או ביטויים יותר מורכבים כגון George Washington Bridge (שצריך לזהות כמקום ולא כאדם) , United States of America או Harvard Law School.

1. **כתיבת הפלט**

מייצאים לקובץ את אותו הקטע כאשר לכל מילה מוצמד התיוג שלה:

* /O – מילה רגילה
* /NAME – שם של אדם
* /COMPANY – שם של חברה
* /LOCATION – שם של מקום
* /NE – ביטוי שזוהה בתור שם אבל לא ניתן לקבוע לאיזה מהקטגוריות הוא שייך

# תיאור המסווגים

## מסווג DuckDuckGo כללי

DuckDuckGo הינו מנוע חיפוש המאפשר שימוש ב-API חינמי וללא הגבלות נראות לעין. בניגוד לחיפוש של גוגל, לא חוזרת ממנו רשימה של לינקים, כי אם ישויות מסוגים אחרים (ייתכן שלא יחזרו כולם, ואפילו לרוב זה המצב)-

* + abstract- תיאור כללי של המונח. לעיתים, הפסקה הראשונה מתוך ערך הוויקיפדיה המתאים
  + definition- הגדרה מילונית של המונח המבוקש
  + answer- עבור חיפוש תשובה לשאלה
  + results- אתרים קשורים שונים (בדומה לגוגל)- לרוב רק אחד
  + related- מונחים קשורים. בפרט עבור מילים עם כמה משמעויות שונות, יופיעו כמה ערכים פוטנציאליים מתאימים (למשל- apple החברה ו-apple הפרי)
  + redirect- העברה לנושא רלוונטי אחר

הממשק תומך בריבוי שפות לחיפוש, אך לא כל החיפושים קיימים בכל השפות, ואם חיפוש לא קיים, חוזרת התוצאה באנגלית.

הקוד שלנו מבצע בנוסף לחיפוש רגיל, פונ' חיפוש נוספת בסגנון "I Feel Lucky" שמחזירה תוצאה אחת אחידה שמוצעת כטקסט הרלוונטי ביותר ומאחד את התוצאה לכדי פסקה מאוחדת אחת המתארת את המונח המבוקש.

קביעת הציונים לסיווג הביטוי

* הביטויים מסווגים על ידי חיפוש תדירות מילות מפתח אינדיקטיביות בהתאם לקטגוריה. למשל, אם נחפש שם של אדם, כנראה יופיעו בפסקה חלק מהמילים- הוא/היא, גיל, נולד, בן/בת. באופן דומה, אם נחפש שם של חברה נקבל מילים שקשורות לעסקים, למטה החברה ולייסוד שלה, ואם נחפש שמות מקומות נמצא מילים אינדיקטיביות לערים, איזורים, ארצות וכו'.
* נציין שאם המשתמש חיפש כמונח את אחת ממילות המפתח שמשמשות לזיהוי, אזי נתייחס אליה בתור מילה רגילה. למשל, המילה age היא אינדיקטיבית לאדם, אבל אם המשתמש חיפש את המילה age בעצמה, הוא יקבל הרבה תוצאות לא רלוונטיות. המילים האינדיקטיביות אינן שמות בעצמן (אבל כן יכולות להיות חלק משם, כמו New York City), ולכן ניתן הציון בהתאם.
* פירוט כלל מילות המפתח שמשתמשים בהם ניתן למצוא בנספח א'.
* לבסוף בודקים את התדירויות שנמצאו בכל קטגוריה מול רף מינימלי (קונפיגורבילי - יכוייל בפרק ייעודי בהמשך):
  + אם בקטגוריה מסוימת יש פחות מכמות זו של מילים, הציון בקטגוריה הזו נחתך לאפס.
  + אם לבסוף בכל הקטגוריות המונח קיבל ציון אפס, אזי כנראה מדובר במילים רגילות ולא שם, וניתן ציון גבוה בקטגורית "מילה רגילה".

## מסווג DuckDuckGo מכוון חברות

* מסווג זה מסתמך על מנוע החיפוש DuckDuckGo, אך משתמש בו באופן אחר מהחיפוש הקודם.
* מצאנו כי שמות של חברות עשויות להחזיר disambiguation של אפשרויות שונות (למשל- Apple הינו חברה וגם תפוח). עם זאת, אם נחפש Apple Inc. או Apple company, נגיע לפסקת מידע ייעודית לגבי החברה. מעניין עוד יותר, שאם נחפש מילה אחרת בתוספת Inc או company, שאינה מתארת שם של חברה, נקבל תוצאה ריקה בחיפוש.
* לכן אנו מבצעות חיפוש דומה לקבלת אינדיקציה האם מונח מייצג חברה או לא- מצרפים לו אחת מרשימת מילות מפתח רלוונטיות, ובודקים כמה מהן גרמו להחזרת תשובה כלשהי ממנוע החיפוש.
* מונח שכאשר מצרפים לו מילת מפתח כמו inc לא מחזיר תשובה בחיפוש, הוא כנראה לא שם של חברה.

## מסווג Facebook כללי

Facebook מייצגים את כלל היישויות ברשת שלהם והקשרים ביניהם באמצעות גרף ומאפשרים קריאה, וכתיבה (במגוון נרחב של שפות ) דרך ממשק ה-Graph API.

אנחנו מבצעות חיפושים על שלושה סוגים של יישויות:

* page – מייצג דפים רשמיים עבור אנשים (לדוגמא: Beatles fan page) ,חברות מסעדות קבוצות וכו'. כל רשומה מקושרת לקטגוריה (כגון government official או local business), ולפעמים לרשימה של תתי קטגוריות.
* user – מייצג זהויות של פרופילים שאמורים להיות שייכים לאנשים פרטיים בלבד (עם זאת, ועל אף שפייסבוק משתדלים לאכוף את הנהלים שלהם, יש הרבה חברות שפותחות דפי פרופיל במקום page).
* place – מייצג כל דבר שניתן לתת לו נ"צ. ניתן למצוא שם ערכים לערים ושכונות אבל גם לאתרים כמו מגדל אייפל או ה-Grand Canyon.כל רשומה מקושרת לקטגוריה (כגון landmark או historical monument) ולפעמים לרשימה של תתי קטגוריות.

יש לציין שהתוצאות שחוזרות אינן בהכרח זהות לביטויים שחיפשנו ויתכן שאפילו לא דומות – לדוגמא חיפוש של bob יחזיר גם Robert ו-Bobby. כדי לגשר על הפער הזה אנחנו שולפים מספר גבוה של ערכים (זהו פרמטר קונפיגורבלי) והם נספרים רק אם הם מכילים במדוייק לפחות מילה אחת מבין המילים בביטוי שחיפשנו. ערכים שזהים לגמרי לביטוי שחיפשנו נספרים עם משקל יותר גבוה לציון.

חשוב לציין שעל אף שמבחינת ההגדרות הרשמיות של המגבלות של Facebook איננו מבצעים פעילות אסורה, העובדה שאנחנו פועלים מיוזר חדש, אנונימי לגמרי שמייצר הרבה פעילות ברשת, יחד עם זה שאנחנו מבצעים שליפות גדולות יחסית גורמים לכך שהם נוטים לחשוד שזהו חשבון ספאם ולחסום את המשתמש שלנו לאחר זמן מה. זהו אילוץ טכני שכמובן ניתן להתגבר עליו במקרה שמעוניינים בתוכנה בשימוש נרחב (סביר להניח שאפילו אם נגדיל את נפח הפעילות באופן הדרגתי לא תהיה להם בעיה עם זה) אבל זה בהחלט כן הקשה עלינו בשלב זה של ההוכחת ההיתכנות והשפיע על גודל הקורפוס שיכולנו לעבוד עליו. למעשה זו הסיבה המרכזית שבגללה אנחנו צריכים לשמור cache של התוצאות שכבר שלפנו (מכיוון שתמיד יחזרו אותם תוצאות עבור אותם ביטויים אין טעם לחפש אותם פעמיים) וזה הגביל מאוד את כמות המידע שיכולנו לעבוד עליו.

קביעת הציונים לסיווג הביטוי

הערכים שחוזרים מתשאול page ו-place נספרים בהתאם לקטגוריה ששוייכו, ולכן יותר אמינות מהערכים שחוזרים מחיפוש של דפי פרופילים (user) על אף שבמקרים רבים בדפי הפרופיל מופיעים הרבה יותר תוצאות. מסיבה זו הן מקבלות משקל יותר גבוה, במיוחד אם הביטוי שחיפשנו זהה לביטוי שמופיע שם, ובנוסף הגדרנו רף (קונפיגורבילי) שבמקרה שנספרו בקטגוריות ארגון ומקום יותר תוצאות מאותו רף חותכים את התוצאות שחזרו מיוזר.

לבסוף קבענו פרמטר נוסף (גם כן קונפיגורבילי) שמגדיר סף שבמקרה שלא הגענו למספר מינימלי של ערכים שנספרו – נתייחס לביטוי כאלה מילה רגילה.

את כלל הקטגוריות שבודקים ניתן למצוא בנספח א'.

## מסווג על פי אות גדולה

* מסווג זה בא לבדוק את המאפיין הבסיסי והשכיח ביותר של שמות באנגלית- התחלת כל מילה ב-Capital Letter. בחיפוש על ביטוי שמכיל יותר ממילה אחת אנו כן מאפשרות למילה שהיא stop word להתחיל באות קטנה (למשל- United States of America).
* מסווג זה לא מבחין בין סוגי ה-NE השונים ונועד רק להגדיל את ההסתברות שמדובר ב-NE כלשהו ולכן הוא מחזיר ציון אחיד לאדם, חברה ומיקום (הציון הינו קונפיגורבלי – ונקבע לאחר ניסויים). במידה ויש מילה שאינה חוקית, המסווג נותן ציון קבוע (גם כן קונפיגורבלי) לקטגורית "מילה רגילה".

## מסווג על פי חיפוש במילון

* בעוד שהמסווגים האחרים מצליחים לזהות נכון סוגים שונים של NE, כאשר הביטוי שמחפשים מורכב ממילה בודדת ישנו קושי לזהות דווקא מילים רגילות (כלומר מתקבלים הרבה תוצאות שהם false positive), לדוגמא, על פי מסווג Facebook: garden ,anywhere ,capital ,district ,north כולם מזוהים כ-location.
* עבודה מול Google היתה מאפשרת הכנסה של מסווג מוכוון מילון שעדיין פועל מול מנוע חיפוש (בדומה למסווג מוכוון חברות של Duckduckgo) אך מכיוון שאין ב-Duckduckgo או Facebook פיצ'רים רלוונטיים הוספנו מסווג שעובד מול קובץ מילון.
* מילון זה מכיל באופן חלקי בלבד הטיות שונות של מילים, ואינו מכיל את צורות הרבים ולכן מילים שמסתיימות ב-s אנחנו נחפש גם בלי האות האחרונה. כמו כן המילון מכיל גם מילים שהם NE (כפי שמילון בעברית יכיל גם מילים שהן שמות כמו גל וטל) ואלה מופיעים עם אות גדולה. מכאן שמסווג זה אינו יודע להבדיל בין הקטגוריות השונות של NE אך הוא נותן אינדיקציה חזקה האם מדובר במילה רגילה או מילה שיש לה פוטנציאל להיות NE (כי היא לא נמצאת במילון או מופיעה עם אות גדולה).
* המילון מורכב ממילים בודדות ולא מחפשים בו ביטויים ארוכים יותר (אבל אותם דווקא מצליחים לסווג נכון במסווגים האחרים גם כשהם ביטויים רגילים), כמו כן אם הביטוי שמחפשים הוא מספר נחזיר שזו מילה רגילה (כן מחפשים מספרים בשאר המסווגים כי לדוגמא נרצה לזהות ש- 9th avenue זה מקום).

# Training – כיול מסווגים בנפרד

בחלק זה, ביצענו ניסויים על מנת לכייל פרמטרים שונים בכל מסווג בנפרד.

השתמשנו בקורפוס של כתבות ב- 4 קטגוריות: חדשות, בידור, נסיעות וויקיפדיה. השתדלנו לבחור כתבות עתירות שמות (מסוגים שונים), על מנת לבצע ניתוח משמעותי.

כל אחת מהכתבות הללו גם תויגה ידנית על ידינו, וכתבנו סקריפט אוטומטי להשוואה בין הרצוי למצוי, וחישוב מדדים שונים-

|  |  |  |  |
| --- | --- | --- | --- |
| **מדד** | **תיוג ידני** | **תיוג מערכת** | **הערות** |
| **Success** | תיוג זהה בשניהם | |  |
| **Partial success** | אדם/חברה/מיקום | NE | המערכת זיהתה בהצלחה שלא מדובר במילה רגילה אך התוצאות קרובות מדי בקטגוריות השונות ובחרנו לא להבדיל ביניהם |
| **Confused** | אדם/חברה/מיקום | אדם/חברה/מיקום (אחר) | המערכת זיהתה בהצלחה שלא מדובר במילה רגילה אך טעתה בקטגוריה |
| **Confused NE** | NE | אדם/חברה/מיקום | בתיוג הידני סימנו גם NE במקרים שבהם זה Named Entitity שלא מתאים במדוייק לקטגוריות האחרות אבל המערכת כן יודעת לזהות אותם – כמו שמות של ספרים\סרטים (לכן מדובר בטעות פחות חמורה) |
| **False Negative** | אדם/חברה/מיקום | מילה רגילה |  |
| **False Negative NE** | NE | מילה רגילה |  |
| **False Positive** | מילה רגילה | NE/אדם/חברה/מיקום |  |

1. ביצענו טיובים שונים למסווגים, ובחנו את השפעתם על המדדים הללו.
   1. ניסינו להגדיל את success ולהקטין את false negative, false positive ו-confused.
   2. כמו כן ניסינו במידה פחותה להגדיל את partial success ולהקטין את confused NE ואת False Negative NE (שגיאה חמורה פחות).
2. לבסוף חישבנו ציון סופי בשלוש קטגוריות
   1. **Accuracy**- הצלחה כללית- כמה מתוך כלל התיוגים (שמות או מילה רגילה) הם נכונים
   2. **Name Precision**- כמה מתוך המילים שתויגו כשמות הם אכן שמות (כאשר מאפשרים בלבול בין NE כללי ובין שם מסוג אחר).
   3. **Name Recall**- כמה מתוך השמות הקיימים במאגרים נמצאו כשמות מתאימים ע"י המערכת (כאשר מאפשרים פספוס של NE כלליים).
3. בהמשך, השתמשנו בקורפוס זה והמדדים הללו על מנת לבצע משקול חכם בין המסווגים השונים.
4. במהלך העבודה נתקלנו בבעיה להריץ הרבה בקשות מול פייסבוק (שנעלו את המשתמשים שפתחנו לאחר זמן קצר) ולכן נאלצנו להשתמש בקורפוס יותר קטן בשביל כל ניסוי שבו השתמשנו בפייסבוק, ונמנענו מלהריץ את זה על ngram גדול מ-1 (כן משתמשים בפונקציית הגדלת החלון).

סה"כ בכלל המסמכים יחד קיימים-

|  |  |  |
| --- | --- | --- |
|  | Full Corpus | Facebook Corpus |
| LOCATION | 173 | 93 |
| PERSON | 243 | 174 |
| ORGANIZATION | 99 | 66 |
| NE | 77 | 51 |
| O (מילה רגילה) | 3119 | 1967 |

## מסווג DuckDuckGo כללי

* הערך לכיול הינו סף הנרמול של כל אחת מהקטגוריות- לאחר חיפוש מילים אינדיקטיביות, מתבצע "חיתוך" בהתאם לסף קונפיגורבילי- למשל, אם לקטגוריה "אדם" היו פחות מ-X מילים אינדיקטיביות, אזי ניתן לו ציון 0 (במקום ציון לפי מספר המילים האינדיקטיביות שנמצאו). לאחר הקיצוץ, במידה וכל הקטגוריות נקצצו, המילה תוכרז כמילה רגילה.
* נציג תוצאות שונות עבור ערכי סף שונים (הורץ עבור ngram=1 עם הגדלות חלון פוטנציאליות):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **5-5-5** | **10-10-10** | **10-15-15** | **10-15-10** | **10-15-5** | **8-8-10** | **20-20-20** | **Person-company-place** |
| 2946 | 3168 | 3188 | 3185 | 3120 | 3141 | 3191 | **Success** |
| 5 | 4 | 1 | 2 | 2 | 4 | 0 | **Partial success** |
| 67 | 39 | 24 | 35 | 56 | 42 | 17 | **Confused** |
| 160 | 270 | 330 | 290 | 230 | 245 | 402 | **False negative** |
| 456 | 153 | 91 | 122 | 226 | 202 | 24 | **False positive** |
| 37 | 66 | 69 | 67 | 57 | 49 | 76 | **False negative ne** |
| 40 | 11 | 8 | 10 | 20 | 28 | 1 | **Confused ne** |
| 683 | 462 | 445 | 447 | 512 | 489 | 443 | **Total error**  **(FP +FN + Confuse)** |
| 0.80 | 0.85 | 0.86 | 0.86 | 0.84 | 0.85 | 0.86 | **Accuracy** |
| 0.39 | 0.53 | 0.60 | 0.56 | 0.47 | 0.51 | 0.70 | **Name Precision** |
| 0.73 | 0.54 | 0.44 | 0.51 | 0.61 | 0.59 | 0.32 | **Name Recall** |

**ניתוח תוצאות**

ניתן לראות שאפשר למצוא נקודות עבודה טובות עבור recall גבוה או עבור precision גבוה, אך לא עבור שניהם- מדובר ב-trade-off עמוק בין המדדים. ה-precision הכללי בקושי מושפע משינויים אלו, משום שסה"כ כמות המילים הרגילות גדולה בהרבה מכמות השמות בקורפוס.

ננסה להסביר התנהגויות שונות שהביאו לפספוסים ולטעויות.

1. duckduckgo מחזיר פעמים רבות disambiguation בין אפשרויות משמעות שונות. בהחלט ייתכן שאחת האפשרויות הינה יותר סבירה מהאחרות. עם זאת, הבדיקה שלנו לא נותנת משקל לרלוונטיות, ולכן מילת תיאור של משמעות שולית תקבל את אותו המשקל של מילה תיאור של המשמעות הסבירה יותר. למשל, בחיפוש apple, תקבל אותו משקל מילה שמתארת את החברה, ומילה שמתארת נובלה בשם apple.
2. אנו כן נותנים משקל ל**כמות** מופעים, ולכן אם למילה יופיעו כמה פירושים דומים, ובהם יופיעו מילות אינדיקטיביות- כן נצליח לתפוס את הזיהוי.
3. מקרה מבלבל נוסף הוא שם של אדם או משפחה שיש מקומות רבים שקרוים על שמו (כמו משפחת קנדי). כפי שהוסבר קודם, מספיק שיש יותר מקומות על שם קנדי מאשר אנשים ידועים בשם קנדי, וכבר נקבל הטיה בתוצאות.
4. כחלק מטיוב קודם, ביצענו איפוס ציון עבור מונחים שהם עצמם מילים אינדיקטיביות- מדובר במילים שאינן שמות בפני עצמן, אך עלולות להחזיר מילים אינדיקטיביות רבות הנגזרות מהן. למשל, המילה company שהיא אינדיקטיבית לחברה, תחזיר בעצמה מופעים רבים של המילה company, ולא נרצה שהיא תתויג בטעות. עם זאת, מילה אינדיקטיבית יכולה כן להיות חלק משם בן כמה מילים, אך אם המילים האחרות אינן לבדן מעידות על שם, ומסתכלים על חלון של ngram=1, הדבר יביא לפספוסים.
5. הכלי כן עלול למצוא גם מילים רגילות שקשורות לנושא הרלוונטי (ואינן מילות מפתח), אבל אינן שמות בעצמן, למשל- businessman, subsidiary.
6. תופעה נוספת בכלי היא שאין אחידות בתוצאות, ולכן קשה לנרמל- ייתכן מונח עם עשרות מילים אינדיקטיביות בתוצאה, ומונח עם מילים בודדות, ואולי שניהם נכונים. השוני נוביע מכפל המשמעות הפוטנציאלי שלהם, ופירושים נוספים שאינם קשורים. לכן קשה לקבוע סף, וקשה לנרמל לפי הסף הזה.

**הגדלה ל-ngram=2**

נראה כי הגדלת ה-ngram הראשוני ל-ngram=2 (בנוסף להגדלת החלון הרגילה בזמן ריצה במידה וזוהה שם) מבצעת את האפקט הבא- מגדילה את כמות השמות החוזרים, ובכך מקטינה את ה-false negative ומגדילה את ה-false positive. נראה כי השפעתה החיובית גדולה יותר- בסופו של דבר, מקבלים name precision דומה אך recall גבוה משמעותית!

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **5-5-5** | | **10-10-10** | | **10-15-15** | | **20-20-20** | | **Person-company-place** |
| **Ngram=1** | **Ngram=2** | **Ngram=1** | **Ngram=2** | **Ngram=1** | **Ngram=2** | **Ngram=1** | **Ngram=2** |  |
| 2946 | 2931 | 3168 | 3160 | 3188 | 3196 | 3191 | 3198 | **Success** |
| 160 | 145 | 270 | 243 | 330 | 292 | 402 | 386 | **False negative** |
| 456 | 483 | 153 | 181 | 91 | 117 | 24 | 33 | **False positive** |
| 0.39 | 0.38 | 0.53 | 0.52 | 0.60 | 0.59 | 0.70 | 0.69 | **Name Precision** |
| 0.73 | 0.76 | 0.54 | 0.59 | 0.44 | 0.51 | 0.32 | 0.35 | **Name Recall** |

## מסווג DuckDuckGo לחברות

* הערך לכיול- סף הנרמול- לאחר חיפוש מילים אינדיקטיביות, מתבצע "חיתוך" בהתאם לסף קונפיגורבילי- אם כמות המילים האינדיקטיביות נמוכה מסף זה, הציון יהיה 0.
* המדדים דומים לאלו של הסעיף הקודם, למעט ההבדל הבא- מאחר ומסווג זה מחזיר רק ציון חיובי עבור company (או אפס בכל המדדים), נתייחס רק למדדים הרלוונטיים לקטגוריה זו (ולא למשל- סיווג מילה רגילה במקום person)
* נציג תוצאות שונות עבור ערכי סף שונים:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Min=1**  **Ngram=1** | **Min=2**  **Ngram=1** | **Min=1**  **Ngram=2** | **Min=2**  **Ngram=2** |  |
| 2950 | 3089 | 2946 | 3085 | **Success** |
| 75 | 14 | 79 | 14 | **Confuse (company instead of other)** |
| 51 | 79 | 51 | 79 | **False negative (company)** |
| 217 | 50 | 221 | 54 | **False positive (company)** |
| 15 | 1 | 15 | 1 | **Confuse ne (company instead of NE)** |
| 343 | 85 | 351 | 147 | **Total error (FP +FN + Confuse)** |
| 0.79 | 0.83 | 0.79 | 0.83 | **Accuracy** |
| 0.18 | 0.25 | 0.17 | 0.24 | **Name Precision** |
| 0.48 | 0.2 | 0.48 | 0.20 | **Name Recall** |

**ניתוח תוצאות**

נראה כי המסווג אינו יודע להכריע באופן טוב בין חברה ובין מילה רגילה. יתכן שבשילוב עם מסווגים אחרים, ועם נקודת עבודה מוכוונת recall אולי נוכל למצוא את אותן חברות שהמסווגים האחרים פספסו.

נראה גם שינוי ל-ngram=2 אינו משפיע על הריצה כמעט, מאחר ורוב החברות בקורפוס הן בנות מילה אחת בלבד.

לגבי התאמתויות שוא, מספיק שישנה חברה שולית עם מונח מסוים בשמה, וכבר יחזרו תוצאות מתאימות. עם זאת, נראה שגם בכיוון ההפוך, לא כל שם של חברה חוזר ממנוע החיפוש יחד עם המילה האינדיקטיבית המתאימה.

## מסווג פייסבוק + מילון

בעוד שמסווג הפייסבוק מגלה היטב את ה-NE השונים יש לו נטייה לשייך גם מילים רגילות ל-NE (זה קורה בעיקר בגלל ששמות של דפים עלולים להיות ארוכים ועיתים מכילים בתוכם גם תיאורים וכו'). כתוצאה מכך אנחנו מייצרים הרבה שאילתות מול פייסבוק (מכיוון שמנסים להרחיב את החלון) והם נועלים את החשבון שאנחנו עובדים ממנו. לכן נעשו הרצות קצרות עם מסווג הפייסבוק לבדו, ולאחר מכן הרצנו אותו רק עם המילון. נדגיש שהתוצאות שחוזרות מפייסבוק לא מושפעות מכך שמריצים את זה עם המילון, רק הסיווג הסופי שניתן לביטוי מאוזן יותר ולכן מתבצעים פחות הרחבות חלון מיותרות וכמות השאילתות פוחתת.

פרמטרים לכיול:

1. סף סכום התוצאות הכולל שפחות ממנו יגרור התייחסות למילה כ-regular.
2. סף סכום התוצאות ששוייכו ל-company ו-location שמעליו צריך להעניק להם יותר משקל על פני person (כי יש הטייה לא נכונה כלפי person) – זהו פרמטר שמשמעותי עבור טיוב התוצאות וזיהוי נכון של סוג ה-NE אך לא בא לידי ביטוי במדדים הכלליים שאנחנו בודקים ולכן כויל בנפרד.

תוצאות:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **REGULAR <100** | **REGULAR <60** | **REGULAR <30** | **REGULAR<10** |  |
| 1876 | 1895 | 1908 | 1906 | **Success** |
| 86 | 43 | 17 | 13 | **Partial success** |
| 12 | 22 | 24 | 26 | **Confused** |
| 65 | 63 | 59 | 59 | **False negative** |
| 204 | 215 | 229 | 232 | **False positive** |
| 23 | 21 | 20 | 20 | **False negative ne** |
| 19 | 26 | 28 | 29 | **Confused ne** |
| 281 | 300 | 312 | 317 | **Total error (FP +FN + Confuse)** |
| 0.86 | 0.85 | 0.84 | 0.84 | **Accuracy** |
| 0.57 | 0.54 | 0.53 | 0.52 | **Name Precision** |
| 0.83 | 0.84 | 0.85 | 0.85 | **Name Recall** |

ניתוח התוצאות:

* ניתן לראות שנקודת החולשה בסיווג בעזרת פייסבוק+מילון הינו ריבוי false positives, אך כן מקבלים ציון recall מוצלח ואין הבדל משמעותי בין תוצאות הקנפוגים השונים.
* נדגיש כי השמות האנושיים שאנו מזהים אינם בהכרח של אנשים מוכרים או מפורסמים- מספיק שהתאמתנו על שם פרטי או משפחה בלבד, וחלון החיפוש כבר יוגדל על מנת להכיל את השם המלא. לכן פייסבוק היא פלטפורמה טובה במיוחד לחיפושים בקטגוריה הזו, ועשוי לכפר על פערים במסווגים אחרים.
* ישנם פערים גדולים בין כמויות התוצאות עבור ביטויים שונים (כתוצאה מכך שאם ביטוי מקבל תוצאות זהות באובייקטים אמינים הם מקבלים משקל גבוה כפי שהסברנו בתיאור המסווג) ועל כן נבחר סף באמצע (45).

## מסווג אותיות גדולות

במקרה זה אין ערך לכיול, נציג את תוצאות המסווג כפי שהן על פני הקורפוס.

מסווג זה יודע לסמן רק NE או מילה רגילה, נתייחס לתוצאות בהתאם- זיהוי שם או מילה רגילה, ללא התייחסות לסוג.

|  |  |
| --- | --- |
| **Results** |  |
| 2908 | **Success** |
| 510 | **Partial success** |
| 5 | **False negative** |
| 285 | **False positive** |
| 3 | **False negative ne** |
| 290 | **Total error (FP +FN)** |
| 0.92 | **Accuracy** |
| 0.67 | **Name Precision** |
| 0.99 | **Name Recall** |

**ניתוח תוצאות**

מאחר והשתמשנו בקורפוסים עם הקפדה על אותיות גדולות, אכן כמעט ולא היו שמות שפוספסו, וה-Recall כמעט מושלם. עם זאת, אם היינו מבצעים את אותה הבדיקה על קורפוס שכולו ב-lower case, היינו מקבלים Recall של 0.

מבחינת precision ניתן לראות פספוסים רבים- מילים בתחילת משפט, שמות ימים וחודשים, מספרים וכו'.

בנוסף, מדובר כאן בזיהוי ראשוני בלבד- שם או לא שם- ללא סיווג מתקדם של סוג השם. ייתכן ששילוב עם המסווגים האחרים יוכל לאפשר סיווג ספציפי יותר, עם סכנה פחותה של פספוסי שמות.

## מסווג מילון

בדומה למסווג אותיות גדולות גם כאן אין ערכים לכיול ומתבצע זיהוי רק בין NE כללי למילה רגילה, בלי הבחנה בין סוגי ה-NE.

## 

|  |  |
| --- | --- |
| **Results** |  |
| 2760 | **Success** |
| 329 | **Partial success** |
| 169 | **False negative** |
| 242 | **False positive** |
| 54 | **False negative ne** |
| 411 | **Total error (FP +FN)** |
| 0.87 | **Accuracy** |
| 0.59 | **Name Precision** |
| 0.71 | **Name Recall** |

**ניתוח תוצאות**

יש ריבוי של false positive – מילים רגילות שמזוהות בטעות בתור NE - שנגרמים כתוצאה מכך שאנחנו מחפשים את המילה במדוייק ואין את כלל ההטיות במילון. סביר שבדיקה לפי צורת ה-lemma של המילים היה עוזר לצמצם את זה משמעותית.

עם זאת זה כן נותן אינדיקציה טובה למילים שהם מילים רגילות שיותר קשה לזהות בעזרת המסווגים האחרים.

# Training - שילוב המסווגים

בחלק זה ניסינו לבצע שילוב אופטימלי של המסווגים השונים

1. בחירת מאפייני הכיול של duckduckgo הרגיל בהתאם לשילוב עם היתר (כאשר לכולם משקל זהה)
2. קביעת משקל לכל אחד מהמסווגים
   1. משקל זהה
   2. משקל גדול יותר ל-duckduck go ו-facebook לעומת uppercase ומילון (2 לעומת 1)
   3. השוואה ללא uppercase וכל היתר שווים
   4. משקל גדול יותר ל-facebook (3) וכל היתר זהים (1)

נבצע הרצה על הקורפוס המצומצם (מותאם פייסבוק- ראו הסבר בפרק הקודם), ונשווה את המדדים

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Training 6** | **Training 5** | **Training 4** | **Training 3** | **Training 2** | **Training 1** |  |
| Facebook has more emphasis than DuckDuckGo | Without uppercase classifier | More emphasis on DuckDuckGo and Facebook | DuckDuckGo – 10,15,10 | DuckDuckGo – 5,5,5 | DuckDuckGo – 20,20,20 | **Description** |
| 1847 | 1981 | 1930 | 1981 | 1977 | 1998 | **Success** |
| 10 | 16 | 11 | 34 | 28 | 13 | **Partial success** |
| 30 | 17 | 20 | 18 | 20 | 21 | **Confused** |
| 53 | 18 | 68 | 59 | 58 | 85 | **False negative** |
| 294 | 142 | 206 | 147 | 154 | 121 | **False positive** |
| 21 | 26 | 24 | 20 | 20 | 25 | **False negative ne** |
| 30 | 23 | 26 | 26 | 28 | 22 | **Confused ne** |
| 377 | 239 | 294 | 224 | 232 | 227 | **Total error (FP +FN + Confuse)** |
| 0.81 | 0.87 | 0.85 | 0.88 | 0.88 | 0.88 | **Accuracy** |
| 0.47 | 0.62 | 0.55 | 0.64 | 0.62 | 0.64 | **Name Precision** |
| 0.86 | 0.79 | 0.82 | 0.85 | 0.85 | 0.78 | **Name Recall** |

**ניתוח תוצאות**

* מ-3 הניסויים הראשונים ניתן לראות כי המסווגים מכפרים זה על חולשותיו של זה, וסה"כ השלם שווה יותר מסכום חלקיו. בחרנו את המשקול של duckduckgo שנותן את התוצאה הטובה ביותר, והוא זה שנותן את ערכי הביניים.
* נראה כי העלאת המשקל של facebook לעומת היתר מוציאה את המערכת מאיזון- משפרת ממש במעט את ה-recall אבל ה-name precision יורד אל מתחת ל-50%.
* באופן דומה ניתן לראות כי הרצה לא מאוזנת של המסווגים מרעה אחד או יותר מן המדדים, ולכן נבחר במשקל זהה לכל מסווג.
  + במידה וקיימות הנחות ידועות מראש על הקורפוס, ייתכן ויהיה צורך בשינוי (למשל- קורפוס שכולו אותיות קטנות). עם זאת, נראה בהמשך, כי גם כאן מקבלים תוצאות יפות גם בלי שינוי.

# Testing

## תיאור הבדיקה

* שלב המבחן מתבצע על 4 כתבות חדשות, שלא נבדקו קודם לכן (אחת מכל קטגוריה), כל אחת ב-3 גרסאות- רגילה, lower-case, upper-case (תחילת כל מילה).
* קונפיגורצית המערכת היא בהתאם לכלל הטיובים שבוצעו במהלך הניסויים הקודמים.
* התוצאה מושווית לתיוג ידני שלנו, לקבלת מדדי הצלחה. בנוסף , ביצענו מדידה של כלי NER של סטנפורד בהשוואה לתיוג הידני על מנת שנוכל להשוות בין ביצועי הכלים.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Our Regular** | **Stanford Regular** | **Our Lower** | **Stanford Lower** | **Our Upper** | **Stanford Upper** |
| **Success** | 544 | 611 | 536 | 519 | 525 | 586 |
| **Partial success** | 6 | 0 | 5 | 0 | 6 | 0 |
| **Confuse** | 12 | 8 | 12 | 0 | 12 | 7 |
| **False negative** | 33 | 16 | 41 | 120 | 33 | 34 |
| **False positive** | 44 | 4 | 46 | 0 | 63 | 12 |
| **False negative ne** | 5 | 6 | 6 | 13 | 5 | 11 |
| **Confuse ne** | 8 | 7 | 6 | 0 | 8 | 2 |
| **Total error (FP +FN + Confuse)** | 89 | 28 | 99 | 120 | 108 | 53 |
| **Accuracy** | 0.84 | 0.94 | 0.83 | 0.80 | 0.81 | 0.90 |
| **Name Precision** | 0.63 | 0.91 | 0.60 | 1.00 | 0.56 | 0.83 |
| **Name Recall** | 0.78 | 0.89 | 0.72 | 0.18 | 0.78 | 0.77 |

**ניתוח תוצאות**

* ניתן לראות כי במערכת שלנו מקבלים תוצאות דומות עבור שלוש גרסאות הקבצים- רגיל, upper ו-lower – למרות שאנחנו כן מתייחסים למסווג ה-uppercase באופן זהה בשלושת ההרצות. זה נובע מכך שהעבודה מול מנועי החיפוש איננה תלוייה ב-case ולכן רוב המסווגים אינם מושפעים מזה.
* ניתן לראות הבדל קל מאוד ב-upper- עליה ב-recall על חשבון ה-precision לאור מציאת מילים חדשות שאין להם מאפיין אחר מבלבד אות גדולה בהתחלה.
* עבור טקסט סטנדרטי (עם הקפדה על אותיות גדולות במקומות הנכונים), הכלי של סטנפורד מחזיר תוצאות מדוייקות יותר בכל המדדים, ובפרט במדד ה-precision. אנו מאמינות שניתן לגשר על הפער הזה ולקרב את התוצאות שלנו למדדים שלהם על ידי תוספות למערכת שיפורטו בסיכום.
* בטקסט ב-lower case סטנפורד החזירו רק 10% מהשמות (+8% שתויגו ידנית כ-NE, ואפשרנו פספוס שלהם). כולם אכן שמות נכונים, ולכן ה-precision מצוין, אבל בפועל ה-recall לא מאפשר שימוש בכלי זה עבור טקסטים כאלה (שנפוצים בשימוש לא פורמלי של השפה, כמו ברשתות החברתיות) ולכן במקרה זה יש יתרון משמעותי למערכת שלנו.
* בטקסט ב-upper case יש לסטנפורד הרעה משמעותית בתוצאות, אך התוצאה שלנו כמעט ללא שינוי. עם זאת, התוצאה של סטנפורד עדיין טובה יותר.

# פוטנציאל בשפות נוספות

נרצה לבחון את יכולות המערכת על פני קורפוסים בשפה נוספת, בפרט כזו שבה אין uppercase שיכול לסייע במציאת השמות. נשתמש בשפה העברית.

מבין המסווגים במערכת, נשתמש ב-Duckduckgo הכללי וב-Facebook. נציין של-Duckduckgo אין תמיכה טובה בעברית, אך יש לשער שבעבודה מול גוגל יהיה ניתן לקבל תמיכה בכלל השפות, ובפרט בעברית.

בשל הגבלה זו נציג תוצאות רק על מילים מייצגות לשם הבנת הפוטנציאל והוכחת היתכנות.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Word** | **Person** | **Company** | **Regular** | **Place** |
| ירושלים | 11 | 17 | 0 | 73 |
| חיפה | 13 | 31 | 0 | 62 |
| תל אביב | 10.9 | 17 | 0 | 18 |
| ישראל | 1 | 29 | 0 | 97 |
| הרצליה | 14 | 22 | 0 | 10 |
| נתניהו | 16 | 1 | 0 | 2 |
| זאב | 13 | 11 | 0 | 10 |
| מיכאל | 4 | 11 | 0 | 2 |
| אל-על | 0 | 15 | 0 | 1 |
| ארומה | 3 | 48 | 0 | 11 |
| מכבי חיפה | 18.3 | 51 | 0 | 0 |
| הליכוד | 0 | 8 | 0 | 0 |
| יאיר לפיד | 50 | 0 | 0 | 0 |
| טוש | 4 | 2 | 0 | 0 |
| כסא | 1 | 1 | 0 | 0 |
| קודם | 0 | 3 | 0 | 0 |

ניתן לראות ש-named entities זוהו בהצלחה, ורובם אף שוייכו לקטגוריה הנכונה.

מילים כלליות חוזרות עם מעט תוצאות (צבוע באפור), שניתן לקבוע סף שיאפשר להבדיל ביניהן ובין שמות.

# סיכום ומסקנות

* הצלחנו לקבל מסווג ברמת דיוק דיי טובה, שמתאים הן לקורפוסים פורמליים והן לקורפוסים לא פורמליים באנגלית (נבדל לפי הקפדה על case).
* בנוסף ראינו פוטנציאל שימוש בשפה נוספת שאין לה מענה ממשי היום שאנו מכירות.
* ההגבלה העיקרית במערכת נוצרה בשל השימוש ב-api של פייסבוק, ולכן התוצאות על קורפוס קטן יחסית, אך אנו מאמינות שעם הרשאות מתאימות ניתן יהיה לעבוד ללא מגבלות.
* שימוש במנוע חיפוש מקיף יותר (בתוכן ובשפה) כמו גוגל, צפוי לטייב משמעותית את התוצאות, ולאפשר מסווגים נוספים שלא יכולנו לממש.

## הצעות לשיפור המערכת

* קיימות תשתיות חיפוש אחרות מקיפות יותר מהתשתיות שעבדנו איתן שעשויות לעמוד בפני עצמן, או לכפר על הבעיות הטכניות שנתקלנו בהן:
  + שימוש בגוגל (+maps)
  + שימוש ב- linkedin
  + שימוש בויקיפדיה
* ניתן לבחון השפעה של מאפיינים תלויי הקשר-
  + ניתן לאסוף סטטיסטיקות על המילים שמופיעות לפני ה- NE שמזהים נכון (כגון with,tell, to) ולהשתמש בהן כאינדיקטור נוסף לכך שמדובר ב-NE.
* מאפיינים בלשניים נוספים שיכולים להקטין את מספר ה-false positives-
  + ניתוח תחבירי (שם עצם יותר סביר כשם)
  + שימוש במילון שמכיל את צורת ה-lemma של המילים.
* שיפורים נוספים
  + הרצה על ngram גדול יותר (נמנע בשל מגבלות טכניות אך ראינו שזה משפר את התוצאות של Duckduckgo).
  + למידה של מילים אינדיקטיביות נוספות עבור Duckduckgo
  + יישום בשפות אחרות לפי אתרים מקומיים (ומילונים רלוונטים)

# נספחים

## נספח א'-

### מילות מפתח duckduckgo

**חיפוש כללי**

|  |  |
| --- | --- |
|  |  |
| **'he'**, **'she'**, **'born'**, **'age'**, **'person'**, **'people'**, **'his'**, **'her'**, **'raised'**, **'man'**, **'men'**, **'woman'**, **'women'**, **'male'**, **'female'**, **'family'**, **'jr'**, **'politician'**, **'artist'**, **'writer'**, **'athlete'**, **'musician'**, **'entertainer'**, **'actor'**, **'director'**, **'author'**, **'comedian'**, **'husband'**, **'wife'**, **'son'**, **'sons'**, **'daughter'**, **'daughters'**, **'brother'**, **'brothers'**, **'sister'**, **'sisters'** | person |
| **'corp'**, **'corporation'**, **'company'**, **'companies'**, **'inc'**, **'conglomerate'**, **'founded'**, **'headquarters'**, **'headquartered'**, **'business'**, **'product'**, **'products'**, **'chairman'**, **'founder'**, **'founders'**, **'businesses'**, **'org'**, **'technology'**, **'service'**, **'services'**, **'owned'**, **'owner'** | Company |
| **'located'**, **'place'**, **'places'**, **'city'**, **'cities'**, **'country'**, **'countries'**, **'continent'**, **'province'**, **'area'**, **'areas'**, **'region'**, **'regions'**, **'landmark'**, **'landmarks'**, **'travel'**, **'park'**, **'parks'**, **'population'**, **'populated'**, **'district'**, **'districts'**, **'center'**, **'centers'**, **'island'** | place |

**חיפוש חברות**

מילים אינדיקטיביות-

**'corp'**, **'corporation'**, **'company'**, **'inc'**, **'headquarters'**

### קטגוריות חיפוש Facebook

|  |  |
| --- | --- |
| organization | company, "non-profit organization" ,"organization", "professional services", "product/service", "government organization", "church/religious organization" |
| place | country, "city", "landmark", "public places", "historical place", "tours/sightseeing", "travel/leisure", "national park", "neighborhood" |
| person | politician, "public figure", "government official", "writer", "athlete", "artist", "musician/band", "news personality", "entertainer", "actor/director", "author", "comedian" |

## נספח ב'- ביבליוגרפיה

### קבצים נבדקים

|  |  |  |
| --- | --- | --- |
| **Filename** | **Source** | **Site** |
| Entertainment1 | Entertainment Weekly | <http://www.ew.com/article/2015/08/01/game-thrones-ian-mcshane> |
| Entertainment2 | People Magazine | http://www.people.com/article/jennifer-aniston-justin-theroux-married-wedding-bel-air-home |
| Entertainment3 | NY post | - http://nypost.com/2015/08/08/jon-stewarts-daily-show-exit-caps-off-the-week-that-was/ |
| Entertainment4 | Jerusalem post | <http://www.jpost.com/Israel-News/Culture/Mariah-Carey-Nobody-tried-to-pressure-me-not-to-come-here-412427> |
| News1 | NY post | http://nypost.com/2015/08/10/google-to-become-alphabet-inc-in-structure-shakeup/ |
| News2 | NY times | http://www.nytimes.com/2015/08/14/world/asia/tianjin-china-explosions.html?hp&action=click&pgtype=Homepage&module=second-column-region&region=top-news&WT.nav=top-news&\_r=0 |
| News3 | NY times | http://www.nytimes.com/politics/first-draft/2015/08/13/republicans-meet-with-netanyahu-to-discuss-iran-nuclear-deal/ |
| News4 | Washington Post | https://www.washingtonpost.com/news/early-lead/wp/2015/08/13/lebron-james-wont-commit-to-play-for-team-usa-at-rio-olympics/ |
| Travel1 | Lonelyplanet | http://www.lonelyplanet.com/world/activities/food-drink/eiffel-tower-paris-moulin-rouge-show-seine-river-cruise/item-v-2050TEM-id |
| Travel2 | Telegraph | http://www.telegraph.co.uk/travel/hotels/11310206/Top-10-European-city-breaks-for-2015.html |
| Travel3 | nycgo | <http://www.nycgo.com/articles/new-york-city-in-three-days> |
| Travel4 | dc.about.com | <http://dc.about.com/od/touristattractions/ss/Top-10-Things-to-Do-in-Washington-DC-Must-Sees.htm#showall> |
| Wiki1 | Wikipedia | https://en.wikipedia.org/wiki/Kennedy\_family |
| Wiki2 | Wikipedia | https://en.wikipedia.org/wiki/Albert\_Einstein |
| Wiki3 | Wikipedia | https://en.wikipedia.org/wiki/Apple\_Inc. |
| Wiki4 | Wikipedia | https://en.wikipedia.org/wiki/Barack\_Obama |

### כלים נוספים

**Stanford NER**- http://nlp.stanford.edu/software/CRF-NER.shtml

**DuckDuckGo for Python**- https://pypi.python.org/pypi/duckduckgo2

**FacePy (Facebook Python API)**- https://facepy.readthedocs.org/en/latest/