**מטרת הפרויקט**

לבחון האם ניתן לזהות Named Entity בטקסט לפי תוצאות החיפוש שחוזרים בחיפוש 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 טוויטר שלא נוהגים להקפיד על האות הגדולה.
2. מרגע שזוהתה מילה רלוונטית, כגון שם של בן אדם, ניתן לנסות להגדיל את החלון ולזהות רצפים של שם ושם משפחה ביחד.
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 לא תמיד ברור אם מדובר בזכר או נקבה אבל מידע זה נחוץ לשפות אחרות.
   2. כאשר מנסים לעשות סיווג לטקסט - זיהוי שרוב הדמויות בכתבה הם מתחום מסויים (נניח 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 שבהם יש יותר מנקודה אחת (כגון U.S.A) ומילים ששייכים לתארים (כגון Dr., Mrs וכו') נשארים כמו שהם. 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.

* + במקרה ש-ngram=1:
    - מסננים החוצה סימני פיסוק ו-stop words (רשימה מלאה של ה-stop words מופיעה בהמשך) מכיוון שאין טעם 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):
   * 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.
2. **הרחבת הביטוי**:

ללא תלות באורך הביטוי המקורי שחיפשנו, במקרה שנקבע שהביטוי המקורי הוא NE כלשהו ננסה להרחיב. 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).

נציין שיתכן שהביטוי המקורי יסווג בקטגוריה אחת והביטוי המורחב יסווג בקטגוריה אחרת – אנחנו מתייחסים לקטגוריה של הביטוי הכי רחב שהצלחנו למצוא.

דוגמאות

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**

**1. 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 מילה אינדיקטיבית בעצמה, 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 - שקבענו לאחר הרצות על ביטויים ומונחים שונים):
  + 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".

1. **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.

1. **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: lady gaga 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 (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) וסופרים אותם רק אם מכילים במדוייק לפחות מילה אחת מבין המילים בביטוי שחיפשנו.

חשוב לציין שעל אף שמבחינת ההגדרות הרשמיות של המגבלות של facebook איננו מבצעים פעילות אסורה, העובדה שאנחנו פועלים מיוזר חדש, אנונימי לגמרי שישר מייצר 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 ה-proof of concept 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).

קביעת הציונים לסיווג הביטוי

**4. 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".

**5. מסווג על פי חיפוש במילון**

* בעוד שהמסווגים האחרים מצליחים לזהות נכן 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 שהורדנו מ-
* מילון זה מכיל הטיות שונות של מילים אך לא 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 את המילים 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)

**שכלול כל תוצאות המסווגים וקביעת התיוג המנצח**

* באופן כללי, סוג התיוג שקיבל את הציון הגבוה ביותר 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.

הבדיקה נעשית לפי המדד של כמה מסווגים נתנו ציון שהוא מעל לאחוז מסוים (קונפיגובילי) מהציון המקסימלי.

**ניסויים וטיובים**

**מסווג מילים שכיחות ב-duckduck go**

כבדיקת שפיות התחלנו עם 3-6 מילות מפתח לכל קטגוריה, ובחנו על מספר מילים מייצגות מכל קטגוריה. קיבלנו את התוצאות הבאות:

Barack Obama:

{'person': 6, 'company': 0, 'place': 0}

Lady Gaga:

{'person': 3, 'company': 0, 'place': 1}

Microsoft:

{'person': 0, 'company': 2, 'place': 0}

Apple:

{'person': 0, 'company': 8, 'place': 1}

San Francisco:

{'person': 2, 'company': 1, 'place': 17}

Grand Canyon:

{'person': 0, 'company': 0, 'place': 11}

לאט לאט הרחבנו את מאגר המילים האינדיקטיביות ואת מספר המילים הנבדקות, והוספנו גם מילים "רגילות" כדי לבדוק שאין תוצאות שווא. קבוצת המדגם-

**['Barack Obama'**, **'Lady Gaga'**, **'Microsoft'**, **'Apple'**, **'San Francisco'**, **'Grand Canyon'**, **'Haifa'**, **'Book'**, **'Cup'**, **'Ground'**, **'Piece'**, **'Ball'**, **'Tony']**

ראינו תוצאות טובות למילים אלו עם סף של 9 מילים אינדיקטיביות עבור המילים האינדיקטיביות הבאות-

|  |  |
| --- | --- |
| קטגוריה | מילים אינדיקטיביות |
| Person | 'he', 'she', 'born', 'age', 'person', 'people', 'his', 'her', 'raised', 'man', 'men', 'woman', 'women', 'male', 'female', 'politician', 'artist', 'writer', 'athlete', 'musician', 'entertainer', 'actor', 'director', 'author', 'comedian', 'husband', 'wife', 'son', 'daughter', 'brother', 'sister' |
| Company | 'corp', 'corporation', 'company', 'companies', 'inc', 'conglomerate', 'founded', 'headquarters', 'headquartered', 'business', 'product', 'products' |
| Place | 'located', 'place', 'places', 'city', 'cities', 'country', 'countries', 'area', 'areas', 'region', 'regions', 'landmark', 'landmarks', 'travel', 'park', 'parks', 'population', 'populated', 'district', 'districts' |

לאחר מכן עברנו לבדוק על קטעי טקסט מלאים, ואז ביצענו אופטימיזציות נוספות-

* **איפוס הציון 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
* **הוספת מילים אינדיקטביות נוספות וסינון מילים מזבלות**
  + למשל, מילים כמו West, East אמנם אינדיקטיביות בעת זיהוי מיקום, אבל מוצאות גם המון דברים אחרים (מיקום של חברה, סוג של שפה או לאום, איזור רלוונטי לאדם מתואר וכו')
* **סף שונה עבור כל קטגוריה**- מאפשר כיול נפרד

ניתוח סיבות אפשריות לשגיאות במערכת

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 שנכתבה ע"י Richard Milward.

אנו כן נותנים משקל ל**כמות** מופעים, ולכן אם למילה יופיעו כמה פירושים דומים, ובהם יופיעו מילות אינדיקטיביות- כן נצליח לתפוס את הזיהוי.

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.

הכלי מוצא גם מילים רגילות שקשורות לנושא הרלוונטי (ואינן מילות מפתח), אבל אינן שמות בעצמן, למשל- businessman, subsidery.

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