

Github:

<https://github.com/EmanuelMaximov/MasksPleaseDHP>

מסמך מתאר ומסכם

נושא הפרויקט

במסגרת הפרויקט נעסוק בניתוח תחבירי אודות שלטים המופיעים ברחבי הארץ המבקשים מהקהל הרחב לעטות מסיכות בעת הכניסה למקומות ציבוריים. המידע ינותח ממאגר שנוצר על ידי ד"ר יעל נצר המפורט להלן:

אודות אוסף המידע

תהליך איסוף המידע

מאז פרוץ מגפת הקורונה ד"ר יעל נצר צילמה מעל 2,800 שלטים של "מסכה בבקשה" שנתלו ברחבי הארץ ואף בצרפת במקומות כמו: מרפאות, חנויות, מרכזי קניות, מסעדות וכו'. השלטים שצולמו נכתבו בשלל שפות, ובפרט בעברית.

תהליך עיבוד המידע

ד"ר יעל נצר תמללה על-ידי "קריאה צמודה" את המלל ואת הפרטים הטכניים אודות השלטים בתמונות ובכך הפכה את המידע ל-Structured Data כטבלת csv, כאשר כל שורה בטבלה מייצגת מידע אודות תמונה אחת של שלט. וכמו כן, על-ידי סקריפט בפייתון ד"ר יעל נצר חילצה את המטא-דאטה של השלטים הכולל את שם קובץ התמונה, את המיקום הגאוגרפי שלהם הכולל כתובת ונקודות ציון, ואת תאריך צילום התמונה, והוסיפה את המידע הנ"ל למאגר.

*הקישור לאוסף מצורף בתחתית העמוד

תיאור הפרויקט – מסלול "Smart Data"

החתכים לניתוח התחבירי שנבצע בפרויקט הם:

1. מין הפנייה בשלטים: זכר/נקבה/שני המינים/ללא
2. אוריינטציה בפנייה בשלטים: על דרך החיוב/השלילה
3. ריבוי הפנייה בשלטים: רבים/יחיד/ללא
4. סוג הפנייה בשלטים: בקשה/ציווי

ניתוח המידע על פי החתכים

ננתח את המידע שברשותנו על ידי "קריאה מרחוק", כלומר, התמלול של השלט שמופיע תחת עמודת "description" בעברית במאגר המידע, יינתן כקלט לאלגוריתם שיבצע ניתוח מורפולוגי של הטקסט באופן הבא:

- זיהוי ריבוי הפנייה ומין הפנייה בשלטים יתבצע באלגוריתם על ידי חילוץ המגדר על פי חלקי הדיבר - שמות התואר, כינויי גוף והפעלים המופיעים בטקסט.
- זיהוי סוג הפנייה ואוריינטציית הפנייה בשלטים יתבצע באלגוריתם על ידי זיהוי חלקי הדיבר: Interjection, Modal, Existential, Negation, Preposition. אם לא יזוהו חלקי הדיבר הנ"ל, האלגוריתם יבצע השוואת המילים בטקסט למילים שמורות המצינות את סוג הפנייה ואוריינטציית הפנייה.
* האלגוריתם שנשתמש בו הוא של הכלי "[The Dicta Nakdan](#)"

הרצת הפרויקט

על מנת להריץ את הפרויקט יש להתקין את הספריות הבאות:
requests, json, pandas, dash, plotly.express, webbrowser,
googleapiclient.discovery, google.oauth2
קלט התוכנית: עבור פונקציית ה-run ניתנו כפרמטרים שם הלשונית וה-id של הלשונית הרצויה ב-spreadsheet שממנה ייקראו הנתונים, וגם מספר העמודה ממנה רוצים לקרוא והעמודה אליה רוצים לכתוב.

תיאור העבודה על הפרויקט

בפרויקט זה השתמשנו במאגר שלטי "מסכה בבקשה" שנוצר ע"י ד"ר יעל נצר.
כל קטעי הקוד נרשמו בשפת Python.
בתחילת התהליך השתמשנו ב-[Google API](#) שהתחבר ל-spreadsheet של המאגר על מנת לקרוא את כל התאים של עמודת ה-description (התמלול של השלטים) ואותם שלחנו אחד בכל פעם ל-[Dicta Nakdan API](#) שהחזיר לנו קובץ JSON עם נתונים אודות תחביר הטקסט ומהנתונים הרבים שהיו בקובץ אנחנו שמרנו במערך של מילונים (כל מילון עבור מילה מהטקסט) את שדה ה-Bitmask שלמפתח שלו קראנו 'morph' ושדה עבור המילה שהמפתח שלו נקרא 'word'.
לדוגמה:

```
[{'word': 'כניסה', 'morph': '0x0000000041460000'},  
{ 'word': 'עם', 'morph': '0x0000000000080000'},  
{ 'word': 'מסכה', 'morph': '0x0000000041460000'},  
{ 'word': 'חובה', 'morph': '0x0000000041460000'}]
```

ביצענו עיבוד-מקדים על הטקסט ובו הסרנו מהטקסט שני אלמנטים:

1. סגירה / closing – כל החלק של הסגירה בטקסט שהתחיל במילות סגירה כמו "בתודה..." או "בברכה..." הוסר, וזאת על מנת להפחית בשגיאות בסיווג הטקסט. דוגמה למקרה בעייתי: "בתודה, מספרת מאיר", כאשר המילה 'מספרת' מסווגת כפועל בנקבה.

2. Definite article & Verb – הסרנו את כל צמדי המילים שהתחילו בשם עצם עם ה' הידיעה ואחריו פועל. דוגמה למקרה בעייתי: "החנות מכילה", כאשר המילה 'מכילה' מסווגת כפועל בנקבה, למרות שלא הייתה כאן נימת פנייה אלא תיאור שם העצם.

לאחר העיבוד המקדים סיווגנו את הטקסט לפי ארבעת החתכים כאשר בכל פעם רצנו על המילים בטקסט וניתחנו את שדה ה-'morph' שלהם לפי bitfields מסוימים באופן הבא:

1. מין הפנייה בשלטים: זכר/נקבה/שני המינים/ללא –

על מנת לזהות את מין הפנייה, ראשית, חיפשנו פעלים בלשון ציווי (לדוגמה: "כנסו לחנות") או בלשון עתידי בגוף שני (לדוגמה: "תיכנס לחנות") או בלשון הווה בגוף שני (לדוגמה: "אתם מתבקשים"). שנית, אם לא נמצא פועל בטקסט חיפשנו פנייה לפי רצף של שם עצם ולאחריו שם תואר מתוך מאגר מילים שמור שיצרנו, לדוגמה: "לקוחות (שם עצם) יקרים (שם תואר)" ולפי המין בשם התואר סיווגנו את הפנייה בטקסט כולו. במידה ולא זוהה אף אחד מהמקרים, מין הפנייה סווג כ-"ללא" (none).

2. אוריינטציה בפנייה בשלטים: על דרך החיוב/השלילה –

על מנת לזהות את אוריינטציית הפנייה, חיפשנו מילות יחס (Preposition) כמו 'בלי' או מילות קיום (Existential) כמו 'אין' והרצנו גם השוואות על מאגר מילים משלנו על מנת לסווג על דרך השלילה, ואם מילים כאלה לא נמצאו – סיווגנו על דרך החיוב.

3. ריבוי הפנייה בשלטים: רבים/יחיד/ללא –

בדומה לסיווג על פי מין הפנייה בשלטים

על מנת לזהות את ריבוי הפנייה, ראשית, חיפשנו פעלים בלשון ציווי (לדוגמה: "כנסו לחנות") או בלשון עתידי בגוף שני (לדוגמה: "תיכנס לחנות") או בלשון הווה בגוף שני (לדוגמה: "אתם מתבקשים"). שנית, אם לא נמצא פועל בטקסט חיפשנו פנייה לפי רצף של שם עצם ולאחריו שם תואר מתוך מאגר מילים שמור שיצרנו, לדוגמה: "לקוחות (שם עצם) יקרים (שם תואר)" ולפי הריבוי בשם התואר סיווגנו את הפנייה בטקסט כולו. במידה ולא זוהה אף אחד מהמקרים, ריבוי הפנייה סווג כ-"ללא" (none).

4. סוג הפנייה בשלטים: בקשה/ציווי –

על מנת לזהות אם הפנייה היא בציווי, חיפשנו פעלים בלשון **ציווי** (לדוגמה: "כנסו לחנות") או בלשון **עמיד בגוף שני** (לדוגמה: "תיכנסו לחנות"), אם פעלים כאלה לא נמצאו, הטקסט סווג בפנייה ב-"בקשה" (not imperative).

לאחר הסיווג על פי החתכים השתמשנו ב-API של Google על מנת לכתוב את הפלט של האלגוריתם שלנו - רביעייה המכילה את תוצאות הסיווג עבור כל חתך - לתוך ה-spreadsheet של המאגר בתאים המתאימים.

הצגת המידע

השתמשנו בספריית plotly של python על מנת להציג את המידע מטבלאות csv ב-spreadsheet בצורת דיאגרמת עוגה. המידע המוצג על ידי דיאגרמות העוגה מחובר ישירות למאגר המידע, כך שכל שינוי במאגר המידע בעמודות הרלוונטיות ישתקף במידע המוצג בדיאגרמות. בכל הרצה של התוכנית, האלגוריתם רץ רק על המידע הנוסף שהתווסף למאגר מאז העדכון האחרון, אם היה כזה, על מנת לייעל את זמן הריצה של התוכנית. את הממצאים הצגנו דרך ממשק שיצרנו על ידי שימוש בספריית dash. בלחיצה על דיאגרמות העוגה נפתח הקישור למאגר המידע ובו מוצגות התוצאות המסוננות על פי הסיווג הרצוי.

מה למדנו

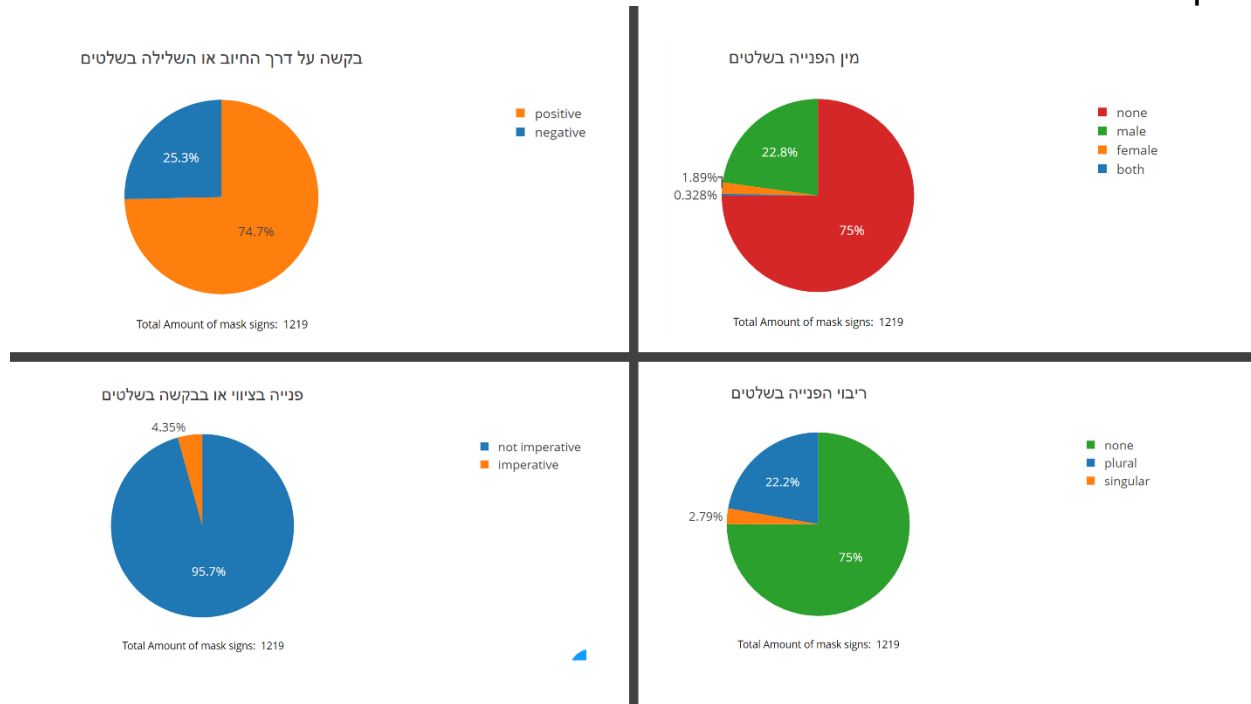
- למדנו להשתמש ב-API של Google על מנת לתקשר עם ה-spreadsheet כדי לקרוא ממנו ולכתוב לתוכו, וליצור מסננים למידע המופיע בו על פי תנאים רצויים
- ספריות שלמדנו להשתמש בהן ב-Python:
 - ✓ **Pandas** – קריאת המידע מה-spreadsheet לאחר המרתו לקובץ csv
 - ✓ **Plotly** - יצירת דיאגרמות עוגה
 - ✓ **Dash** - יצירת ממשק של dashboard להצגת דיאגרמות העוגה בצורה נוחה עם תפריט נגלל, ועל מנת ליצור callbacks שבעת ההקלקה על דיאגרמת העוגה יפתח מאגר המידע המסונן
- למדנו בלשנות ואת חלקי הדיבר וניתוח תחבירי של משפטים
- למדנו את חשיבות הנגשת המידע למשתמש לקריאה ולהורדה בפרויקט המדעי הרוח הדיגיטליים על מנת להוציא את המיטב מהמידע המעובד

השערת התוצאות

- עבור מין וריבוי הפנייה - אנחנו שיערנו שכל הנראה ברוב השלטים לא תהיה פנייה מפורשת לאחד המינים, ובשלטים בהם כן תהיה פנייה לאחד המינים, הפנייה תהיה למין זכר וברבים
- עבור אוריינטציית הפניה – אנחנו שיערנו שברוב השלטים הפנייה תהיה על דרך השלילה, מפני ששלטים אלה נועדו לאסור כניסה ללא מסכה
- עבור סוג הפנייה – בבקשה או ציווי – אנחנו שיערנו שברוב השלטים הפנייה תהיה בבקשה, מפני שזאת דרך נעימה יותר לפנות לקהל הלקוחות

תוצאות

להלן התוצאות לאחר הרצת התוכנית על המאגר שהכיל 1219 רשומות:



מסקנות

- על פי התוצאות ניתן לראות שצדקנו במין הפנייה ואכן 75% מהשלטים הם ללא פנייה מפורשת למין מסוים, ו-22.8% מהשלטים הם בפנייה לזכר כפי שבדרך כלל מקובל לפנות מבחינה חברתית, ורק ב-1.89% מהשלטים הפנייה היא בנקבה ולאחר הסתכלות במאגר המידע על שלטים אלה ניתן לראות שהם נתלו בחנויות בגדי נשים ובמספרות, שכאמור במקומות אלה רוב הלקוחות הן נשים
- ניתן לראות שלמרות השערותנו שרוב הפניות בשלטים יהיו על דרך השלילה, בפועל 74.7% מהשלטים הם על דרך החיוב וגם 95.7% מהשלטים הם בפנייה בבקשה ולא בציווי, אנו מסיקים מכך שבעלי החנויות מעדיפים לפנות

בצורה נעימה יותר לקהל הלקוחות בכך שהם פונים על דרך החיוב ובלשון בקשה ללקוחותיהם

- על פי התוצאות ניתן לראות ש-2.79% מהפניות הן בפנייה ביחיד, ולאחר הסתכלות במאגר המידע על שלטים אלה ניתן לראות שרובם נתלו בחנויות קטנות ובבוטיקים, כלומר, בחנויות בהן יש חשיבות ליחס האישי עם הלקוחות.

מדוע הפרויקט הוא פרויקט במדעי הרוח הדיגיטליים?

המידע אותו אנו מעבדים בפרויקט שלנו הוא אוסף מקוון של אלפי רשומות כאשר בכל רשומה יש מידע אודות שלט של "מסכה בבקשה" וחומר הגלם אותו אנו מנתחים הוא הטקסט המופיע בכל שלט בעברית. הניתוח משלב בתוכו ניתוח תחבירי מתחום הבלשנות ע"י "קריאה מרחוק" ואנו משתמשים בכלי החזיית מידע (דיאגרמות עוגה) על מנת להציג את תוצאות הניתוח שלנו. בנוסף, הפרויקט שלנו דוגל בפתיחות ושיתוף פעולה, הן בהנגשת המידע לקהל הרחב והזמנתו להוסיף למידע הקיים ע"י הוספת רשומות למאגר המידע ב-spreadsheet, והן בשיתוף הפרויקט בפלטפורמה שיתופית ב-Github בה כל אחד יכול לעשות שימוש חוזר בכלים של מדעי הרוח הדיגיטליים ובכך מתאפשרת הדמוקרטיזציה של המידע.

האתגרים בעבודה על הפרויקט

האתגר שלנו התחלק לשני חלקים:

1. האתגר הטכני – האתגר בשימוש בכלים לקריאת, כתיבת וסינון המידע במאגר המידע, ובעבודת התכנות והשימוש בספריות בפייתון על מנת ליצור תצוגה אינטראקטיבית עבור המשתמש.
2. האתגר הבלשני – האתגר בהבנת כל חלקי הדיבר המשפיעים על סיווג הטקסט על פי החתכים הרצויים, והתאמת העבודה מול הפלט של DICTA NAKDAN והעבודה עם ה-BITMASK.


דוגמאות לאי-הצלחה בסיווג של האלגוריתם שלנו

שגיאות כתיב בשלטים:

 טקסט: בס"ד בעסק זה מקפידים על בדיקת חום יאין כניסה ללא מסכה נא לשמור 2 מטר אחד מהשני

הערה: המילה עם שגיאת הכתיב היא "יאין" והיא אובחנה כפועל בנקבה, ולכן הטקסט זוהה כפנייה

ב"נקבה" כשפועל הפנייה היא "none".

 טקסט: לציבור המבקרים בחנות הכניסה לחנות מותרת בחבישת מסכה בלבד נא לשמור על מרחק של

שני מטר בין הלקוחות גם מחוץ לחנות לשמור מרחק אחד מהשני נא לשמור על כללי הייגיינה העסק עובד


לפי התו הסגול

הערה: המילה עם שגיאת הכתיב היא "הייגיינה" והיא אובחנה כפועל בנקבה, ולכן הטקסט זוהה כפנייה

ב"נקבה" כשפועל הפנייה היא "both".



זיהוי לא נכון (אי-זיהוי) של המילה ב-Dicta Nakdan:

 טקסט: לק"י כניסה לחנות עם מסכה!

הערה: לק"י = לקוחות יקרים, המילה לא זוהתה בגלל ראשי התיבות