מסמך מתאר ומסכם

בפרויקט שלנו, ניתחנו את עמודת ה- description במאגר:
https://docs.google.com/spreadsheets/d/1Q6HUEabx4SCCu6r4l-KBhbcI3XDcke4Wy4-xC4jFSM8/edit#gid=0
המציין את הטקסט הכתוב בשלטים שנתלו בכניסה לחנויות המבקשים מקהל הלקוחות לעטות מסיכה בזמן שהותם בחנות. אנו בחנו שני דברים:

1. כמות הפניות בלשון זכר לעומת כמות הפניות הנטולות מגדר
2. כמות השלטים המבקשים בגישה חיובית לעומת כמות השלטים המבקשים בגישה שלילית

בהקשר של מדעי הרוח הדיגיטליים, המסלול שלנו הוא Big data מכיוון שעבדנו על אוסף של טקסט. יצרנו כלי שמבצע שני ניתוחים תחביריים לכל משפט, על ידי שני כלים שונים ומשווה בין הניתוחים שהתקבלו ובכך יוצר מידע הנגיש לקריאה ע"י בני אדם.

תרומתנו למדעי הרוח הדיגיטליים הוא כלי נגיש ופשוט לשימוש שיכול לעזור בלשפר כלים קיימים כמו Yapp שבהמשך נראה כי התקבלו הרבה ניתוחים שאינם מדויקים.

הרחבת ההצעה

עברנו באופן איטרטיבי על כל המשפטים כך שעבור כל משפט ביצענו:

1. ניתוח מורפולוגי של yapp ועליו ניתוח תחבירי של yapp בעזרת סקריפט שכתבנו
2. ניתוח מורפולוגי של dicta ועליו ניתוח תחבירי של yapp בעזרת סקריפט שכתבנו
3. השווינו בין שני הניתוחים –

- אם התקבל ניתוח זהה, נריץ בדיקה:

  - האם הפניה נעשית בלשון זכר בלבד? גם בלשון זכר וגם בלשון נקבה? נטולת מגדר?
    נעשה זאת באופן הבא: אם אחד מחלקי הדיבר זוהה כ-VERB אז נבדוק מה המגדר שלו. אם אין לו מגדר או שאין VERB נשתמש בבנק מילים שהכנו מראש המבוסס על עמודת ה- salutation ("לקוחות יקרים", "שומרים על ההנחיות", "לקוח יקר", "לקוחו יקרים/יקרות", "לקוחות יקרות", "כמה טוב לראות אתכם שוב", "שימו לב", "כיף לראות אתכם", "אורחים יקרים", "אורחים/דיירים יקרים", "עובד/מבקר יקר", "היי טוב שבאת", "לקוחות נכבדים", "מטופלים יקרים", "התגעגעתם?", "ברוכים הבאים") ולפיו נבחין בין פניה לזכר ופניה לנקבה.

    אם גם אין salutation אז הפניה היא נטולת מגדר

  - האם הגישה חיובית או שלילית?
    נעשה זאת באופן הבא: ראשית נחפש חלק דיבר שזוהה כ- Negation ואם לא קיים נשתמש בבנק מילים שהגדרנו מראש המכיל מילות שלילה ("בלי", "אי", "ללא", "אין", "אל", "בל", "בלתי", "לא", "לבלתי", "מבלי", "מבלתי")
    אם אף חלק דיבר לא זוהה כ- negation וגם אף מילה מבנק המילים אינה מופיעה, אז הגישה חיובית.

- אם שני הניתוחים שהתקבלו אינם זהים, בדקנו ידנית ובעזרת שני מקורות עזר המפורטים מטה, מי הניתוח הנכון ועליו הסקנו את שתי המסקנות.

הכלים שהיו בשימוש

yapp - לניתוח תחבירי ומורפולוגי של אוסף המשפטים

Dicta - השתמשנו באתר של dicta על מנת לבצע ניתוח מורפולוגי

ספריות python: pyautogui, selenium

מקורות עזר להבנה של מושגים:

O:/research/tagging-guideline/tagging-guideline.dvi (bgu.ac.il)

UD_Hebrew-HTB (universaldependencies.org)

מה למדנו

למדנו כי ניתוח תחבירי של משפט בעברית זו פעולה שיכולה להיות קשה לביצוע על ידי המחשב. הכלי של Yapp ניתח בצורה שגויה קרוב לחצי מהמשפטים לעומת הכלי של Dicta שהצליח את הרוב.

דבר נוסף שראינו על דרך הפעולה של שני הכלים, הוא שכאשר ניתחנו טקסט המורכב מכמה משפטים המופרדים ע"י סימן פיסוק כלשהו, Dicta התייחס לטקסט כאל כמה משפטים נפרדים לעומת Yapp אשר הצמיד את סימן הפיסוק למילה שבאה לפניו והתייחס לכל הטקסט כאל משפט אחד. כאשר יש סימן פיסוק בסוף הטקסט, Dicta יצר משפט ריק נוסף, מיותר.

בנוסף למדנו לכתוב תכנית שמבצעת ניתוח מורפולוגי ותחבירי באופן אוטומטי ולאחר מכן מנתחת את התוצאה שהתקבלה בהתאם למסקנות שרצינו להסיק.

אתגרים מולם התמודדנו:

הכלי של Yapp לניתוחים מורפולוגים ותחביריים ניתן להפעלה גם מה- command line לעומת הכלי של Dicta . לכן היינו צריכות למצוא דרך להשתמש באתר של Dicta באופן אוטומטי. הפתרון שמצאנו היה להשתמש בספריה של Python המדמה שימוש ידני במקלדת.

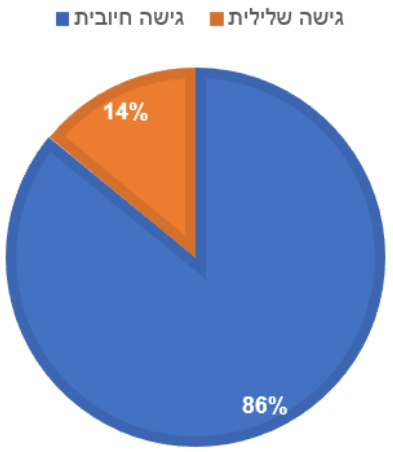
אתגר נוסף שעמדנו מולו הוא שאינטרפרטר של הגרסה 3.8 של Python לא עבד עם משפטים בעברית לכן על מנת לשלוח את המשפטים בעברית לניתוח נאלצנו לעשות זאת בגרסה 3.7 של Python שכן תומכת בעברית.

אתגר נוסף היה ש-Dicta מחזיר את המילים מנוקדות ולכן לעתים הוריד אותיות ניקוד כמו י' ו- ו' ו- Yapp לא. לכן כשהשווינו בין שני הניתוחים, נאלצנו להשוות בין מילים זהות שכתובות עם שינוי בניהן מה שהקשה על תהליך ההשוואה. על מנת לפתור את הבעיה כתבנו פונקציה שמורידה ניקוד מכל מילה, ואם השוני בין שני מילים הוא רק באחת מן האותיות י' ו-ו' החשבנו אותן כשוות.

מסקנות

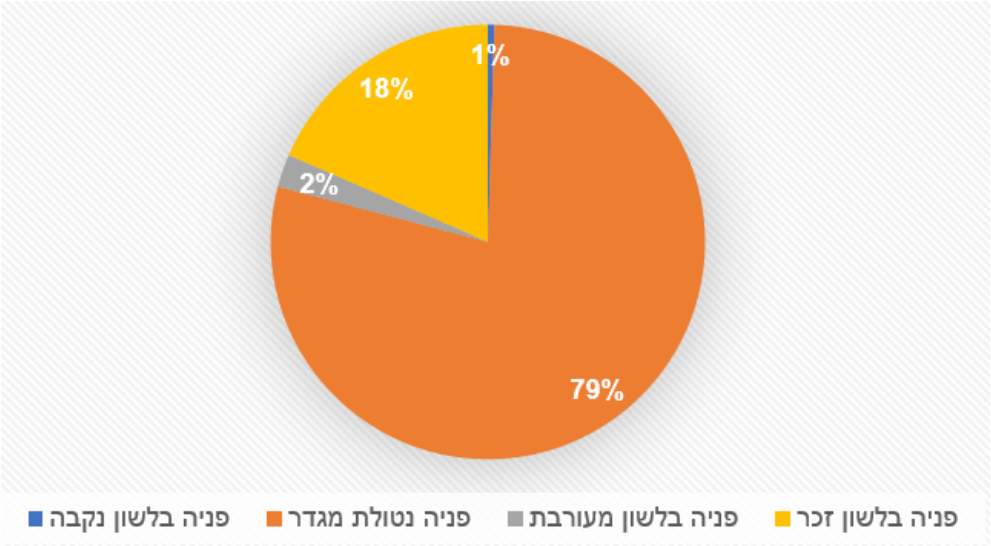
מתוך 206 המשפטים שנבדקו:

מצאנו כי 177 מתוכם מבקשים לעטות מסיכה לפני הכניסה לחנות בגישה חיובית, ו- 29 בגישה שלילית



מהתוצאה שהתקבלה אנו מסיקות כי הרוב המוחלט של החנויות והעסקים העדיפו לבקש עטיית מסיכה בגישה חיובית על פני גישה שלילית. לדעתנו, הסיבה לכך יכולה להיות הרצון של בעלי העסקים והחנויות לעודד לקוחות להיכנס לתוך העסק/חנות ע"י בקשה נעימה יותר ולא כמו הצבת תנאי של עטיית מסיכה בשביל להיכנס.

בנוסף מצאנו כי 162 מהפניות נעשו בלשון נטולת מגדר, 38 מהפניות נעשו בלשון זכר, 5 מהפניות נעשו בלשון פניה מעורבת ופניה אחת בלבד נעשתה בלשון נקבה.



מהתוצאה שהתקבלה אנו מסיקות כי הרוב המוחלט של החנויות והעסקים העדיפו לבקש עטיית מסיכה בלשון נטולת מגדר, לדעתנו על מנת להימנע מאי-נעימות כיוון שכיום עלתה המודעות לכך שיש צורות פניה נוספות מעבר לזכר ונקבה בלבד, ולכן החליטו על פניה בצורה כללית.

בנוסף גם ראינו שעסקים וחנויות שלא בחרו להשתמש בלשון נטולת מגדר, העדיפו לבקש לרוב בלשון זכר, ומעטים ובודדים השתמשו בלשון מעורבת. אנו מניחות כי זה נובע מהרגל של פניה בלשון זכר כאשר רוצים לפנות אל קבוצה של אנשים (לקוחות), גם אם יכול להיות רוב נשי.

מסקנות שהסקנו על הכלים בהם השתמשנו:

הכלי של Yapp אמנם נגיש לשימוש מה- command line אך טעה בפירוש של חלק נכבד מהמילים, למשל:

בפירוש המילה "מַסֵּכָה" אשר תורגמה כמילה "מִסְכָּה" (גוף שלישי נקבה, מיסוך)

gen=F|num=S|per=3|tense=PAST	VB	VB	מיסך	מסכה

בפירוש המילה "לָשִׁים" אשר תורגמה כמילה "לָשִׁים" (לשים בצק)

gen=M|num=P|per=A	BN	BN	לש	לשים

והכלי של Dicta גרם לבעיה בהשוואה בין שני הניתוחים על ידי כך שהוסיף ניקוד לכל מילה וגם הוריד י' ו-ו' , דבר שדרש התייחסות ספציפית על מנת שהתוכנית בכל זאת תזהה את המילים כזהות.