קבוצת המט"ריסטים: קסם לופו 318604691, טל דקלו 209326370, שיר שגב 318146032, אברהם מרון 208882001, נווה מבורך 318284569

פרויקט סיום במסגרת הקורס "ביג דאטה ובינה עסקית למדעי המדינה"- 56146

המוטיבציה לפרויקט

מדד השקט הביטחוני שבנינו נועד לשקף את העיסוק העולמי (התקשורתי) בנושא: ביטחון, במטרה להבין עד כמה העולם היה בטוח בתאריך מסוים. המדד מאפשר לנו לזהות את רגעי "השקט" או את רגעי "הרעש" העולמיים על פני ציר זמן ולזהות אירועים ביטחוניים חריגים בעזרת ניתוח כתבות.

עבודת המחקר

המשפך ההפוך (ראה נספח א'):

geg_gcnlapi- התחלנו לעבוד עם המאגר geg_gcnlapi- דאטה סט של gdelt המכיל כ-257 מיליון שורות, בגודל 1.07 TB. טבלה זו היא הטבלה הראשונה שמבוססת על עיבוד שפה טבעית (NLP). באמצעות טכנולוגיית ה-NLP ניתן לזהות את ההקשר של המילים בטקסט, וכך ניתן להסיק את התיוגים של הישויות והאירועים בטקסט ואת רמת המובהקות שלהם שנעה בין 0-1, ככל שהציון גבוה יותר, החשיבות של המילה בטקסט היא גבוהה יותר.

מאגר זה בו השתמשנו מכיל כתבות ב-11 שפות. כל ישות בטקסט מקבלת סיווג, לדוגמה- ראש מדינה, ארגון, מיקום, אירוע, וכו'. ניתן לזהות אירועים בזמן אמת כמו למשל: התפרצות של מגפות, מלחמות וכו'. בנוסף, כפי שציינו, לכל ישות שזוהתה ניתן ציון הבולטות.

source_urls- בשלב הזה בחרנו לבצע 2 סינונים: כתבות באנגלית בלבד, וסינון תאריכים מינואר 21 עד יוני 22. בחרנו בשפה האנגלית משום שזוהי השפה הבינלאומית הנפוצה ביותר בעולם. כתבה חשובה בנושא בטחון שנרצה לזהות, סיכוי גבוה שתופיע באנגלית. נציג סיבה נוספת לבחירה בשפה האנגלית בשלב הבא. בשלב זה של המשפך צמצמנו את מספר השורות ל-52 מיליון שורות, מדובר בצמצום של כ80% מהטבלה הקודמת.

source_url_no_covid- בשלב הזה בחרנו להסיר את כל הכתבות במאגר שלנו, שנמצאות גם במאגר הקורונה של gdelt- covid19. כך אנחנו נמנעים מלספור כתבות שעלולות להשתמע ככתבות ביטחוניות, אך בפועל מדברות על הקורונה. לדוגמה- כתבה על קורונה יכולה להכיל משפט כמו: "covid 19 attack", והקוד שלנו יזהה את הכתבה ככתבה ביטחונית בגלל המילה "attack", מה שיצור הרבה "זבל" בדאטה שהיינו רוצים להימנע ממנו. מספר השורות כעת עומד על 44 מיליון שורות, צמצום של 16% מהטבלה הקודמת.

base_table- בשלב זה השתמשנו בשתי רשימות מילים שיצרנו: black list ו-white list. ה white list זוהי רשימה של 239 מילים בנושא בטחון (מלחמה, הפצצות, טרור...) שהיינו רוצים שיופיעו בכתבות, כי אם יופיעו, הכתבה ככל הנראה מדברת על נושא ביטחוני. ה-black list זוהי רשימה של 39 מילים שלא היינו רוצים שיופיעו בכתבה, כי אם יופיעו, כנראה שהכתבה לא עוסקת בנושא ביטחוני (לדוגמה היסטוריה). בטבלה הזו בחרנו להציג כתבות שיש בהן לפחות מילה אחת מרשימת המילים הלבנות ברמת בולטות מעל 0.001, ואף מילה מרשימת המילים השחורות. מספר השורות בטבלה זו הוא 53 מיליון. מספר זה גבוה ממספר השורות בשלב הקודם, משום שכתבה אחת

שנמצאה ומופיעה כשורה בטבלה, מתפצלת לעוד שורות ומציגה את כל הישויות שזוהו מרשימת המילים הלבנות, וגם את כל הישויות (מסוג אירוע) שזוהו בטקסט שהן ברמת בולטות מעל 0.001.

base_table_select_events- בשלב זה השארנו רק את המילים הלבנות שנמצאו. כל שורה בטבלה מכילה את ה- URL של הכתבה, את המילה הלבנה שנמצאה בה ואת הרמת הבולטות של המילה (הסרנו את האירועים האחרים שזוהו בטבלה הקודמת ואינם קשורים לנושא ביטחון). כך הגענו ל9 מיליון שורות.

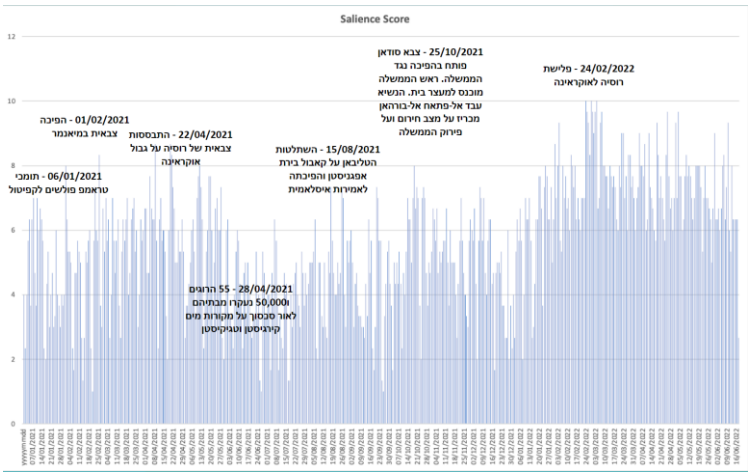
base_table_select_events_summary- זוהי הטבלה האחרונה והשלב האחרון במשפך שלנו. טבלה זו מכילה 534 שורות. הצמצום נובע מהעובדה שקיבצנו את כל הכתבות לפי תאריכים, כך כל שורה בטבלה מייצגת תאריך. ביצענו נורמליזציה לנתונים הבאים: כמות הכתבות ביום, כמות האזכורים של המילים הלבנות ו"מד השקט". הנורמליזציה עוזרת לנו להבין את הפרופורציות של הנתונים בפרויקט, מה נחשב הרבה ומה נחשב מעט, ולהצליח להציג את הנתונים על גרף אחד כך שנוכל להשוות ביניהם.

עמודות הטבלה הסופית שלנו הן: תאריך, מספר הכתבות באותו התאריך שעוסקות בביטחון, מספר האזכורים של המילים הלבנות בכתבו ורמת בולטות ממוצעת (מד השקט). ככל שהמדד הבטחוני גבוה יותר, יש פחות בטחון. מקרה חריג של חוסר בטחון יתבטא בשילוב של ערכים גבוהים לכל עמודה: מספר כתבות גבוה, מספר אזכורים גבוה, ומדד גבוה.

כדי להסיק מסקנות, ביצענו ויזואליזציות על הטבלה הסופית.

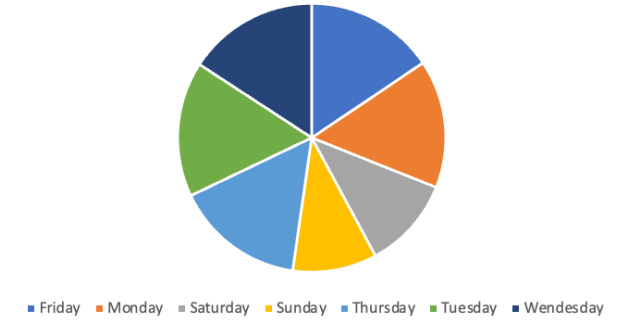
<u>ממצאים</u>

התוצר הגרפי המרכזי שבחרנו להציג הוא גרף טורים על פני התאריכים שהגדרנו, כאשר כל טור בגרף הוא שקלול של מספר הכתבות, מספר האזכורים והמדד הבטחוני באותו תאריך. נתנו משקל שונה לכל פרמטר שנורמל, לפי רמת החשיבות שלו בעינינו. ניסינו כמה וריאציות של שקלולים, ובחרנו לבסוף בשקלול הבא: מדד השקט שהוא מדד החשיבות הגבוה ביותר קיבל 50 אחוז, כמות הכתבות 25 אחוז וכמות האזכורים 25 אחוז. שקלול זה קיבל את הדיוק הקרוב ביותר למציאות. זיהינו את התאריכים בהם היה פחות ביטחון לפי גובה הטור (גבוה יותר- פחות בטחון). הצלבנו את נתוני הגרף שקיבלנו עם אירועים ביטחוניים שקרו בעולם, ואכן הגענו ברוב המקרים להתאמה:



*מקרה ביטחוני חריג שמד השקט שלנו לא זיהה והיה נמוך, היה בתאריך 28.4.21.

התוצר הגרפי השני שבחרנו להציג הוא גרף עוגה המציג התפלגות של הכתבות הביטחוניות על פני ימי השבוע. המסקנה העיקרית מהגרף היא שבכל ימות השבוע יש התפלגות יחסית אחידה של הכתבות הביטחוניות, מלבד הימים: שבת וראשון בהם יש פחות אירועים ביטחוניים. השערתנו היא כי בימים אלה יש פחות כתבות שיוצאות לתקשורת באופן כללי, ובנוסף גם פחות אירועים בטחונים מתרחשים בעולם.



<u>משימות להמשך</u>

1. דיוק של רשימת המילים הלבנות.
2. בחינת התנאים שהגדרנו לסינון הכתבות, הרצת מספר וריאציות של תנאים ובחירת המודל הטוב ביותר.
3. סיווג הדאטה לפי מדינות והצגת ממצאים על גבי מפה (נוכל להסיק מסקנות מעניינות נוספות אם פרמטר המיקום יהיה קיים).
4. חלוקת המילים הלבנות לפי נושאים, לדוגמה- פיגועים, תאונות, אלימות, מלחמות וכו'. כך נוכל להעשיר את הפרויקט ואת מסקנותינו ובנוסף לזיהוי עד כמה העולם היה בטוח, נוכל להגיד גם באיזה תחום.

עמוד נספחים

נספח א' - המשפך ההפוך