מדד השקט הביטחוני העולמי

פרויקט סיום במסגרת הקורס "ביג דאטה ובינה עסקית למדע המדינה" - 56146
קבוצת המט"ריסטים

מגישים: קסם לופו, טל דקלו, שיר שגב, אברהם מרון, נווה מבורך

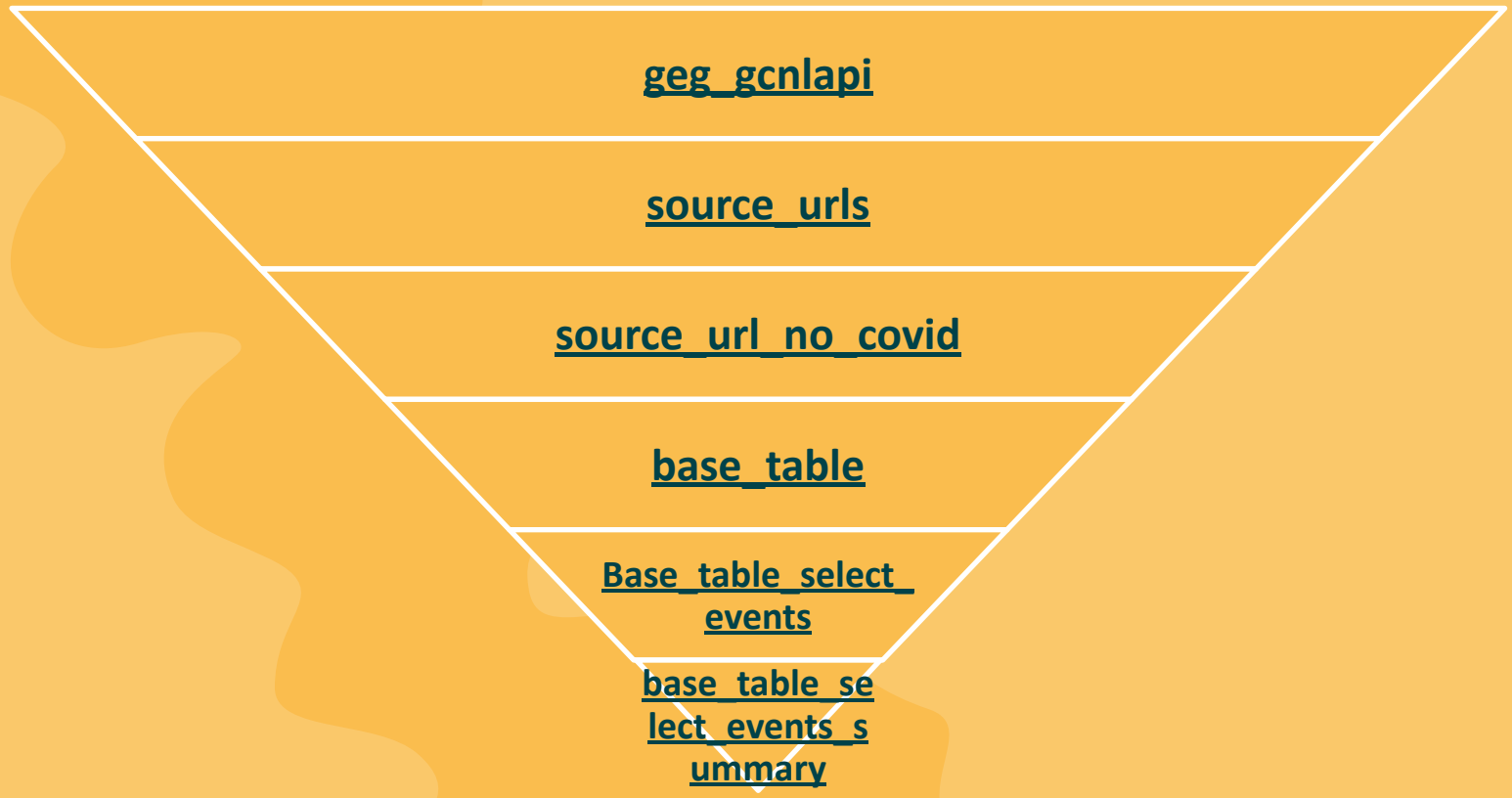
המוטיבציה והרקע לפרויקט

מדד השקט הביטחוני נבנה במטרה לשקף את העיסוק העולמי (התקשורתי) בנושא: ביטחון.

כך נוכל 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 חריגים בעזרת ניתוח כתבות.

המשפך ההפוך של הפרויקט



geg_gcnlapi

geg_gcnlapi

↓

source_urls

↓

source_url_no_covid

↓

base_table

↓

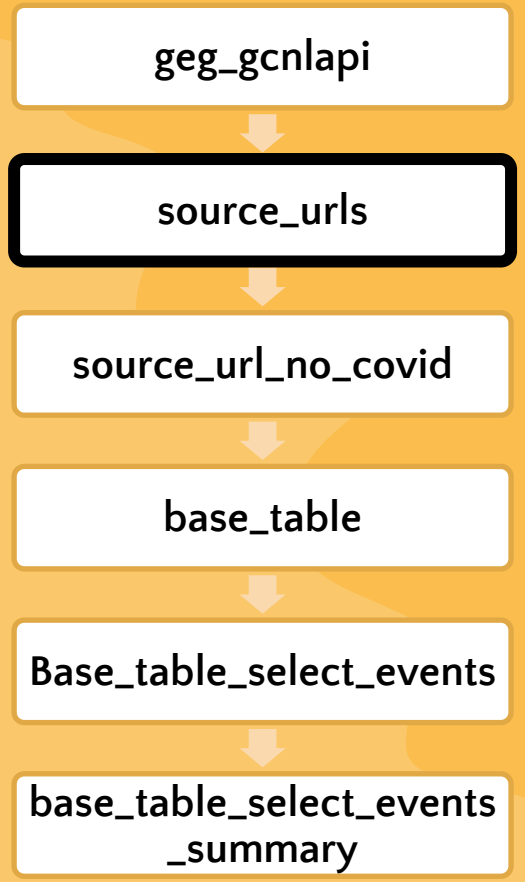
Base_table_select_events

↓

base_table_select_events
_summary

1. 256M שורות.
2. NLP - הטבלה הראשונה שמבוססת על עיבוד שפה טבעי
3. כל ישות שמוזכרת מקבלת סיווג- ראש מדינה, ארגון, מיקום, אירוע, תאריך וכו'.
4. 11 שפות
5. ניתן לזהות אירועים בזמן אמת: מגפות, מלחמות...
6. לכל ישות ניתן ציון בולטות בטווח 0-1

source_urls



1. 52M שורות.
2. סינון לפי שפה: אנגלית בלבד.
3. סינון תאריכים: ינואר 21 – יוני 22.

צמצום של 80% מהטבלה הקודמת.

source_url_no_covid

geg_gcnlapi

source_urls

source_url_no_covid

base_table

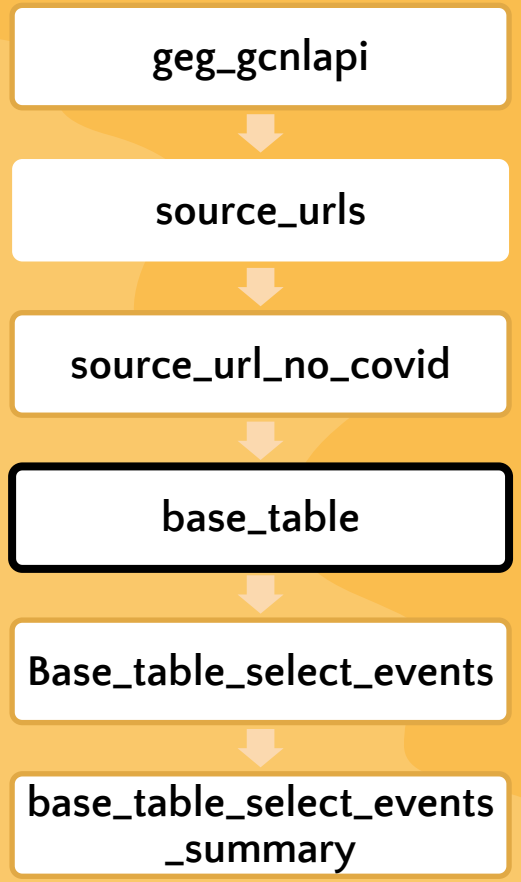
Base_table_select_events

base_table_select_events
_summary

1. 44M שורות.
2. הסרת כתבות בנושא הקורונה- ממאגר הכתבות COVID 19

צמצום של 16% מהטבלה הקודמת.

Base_table

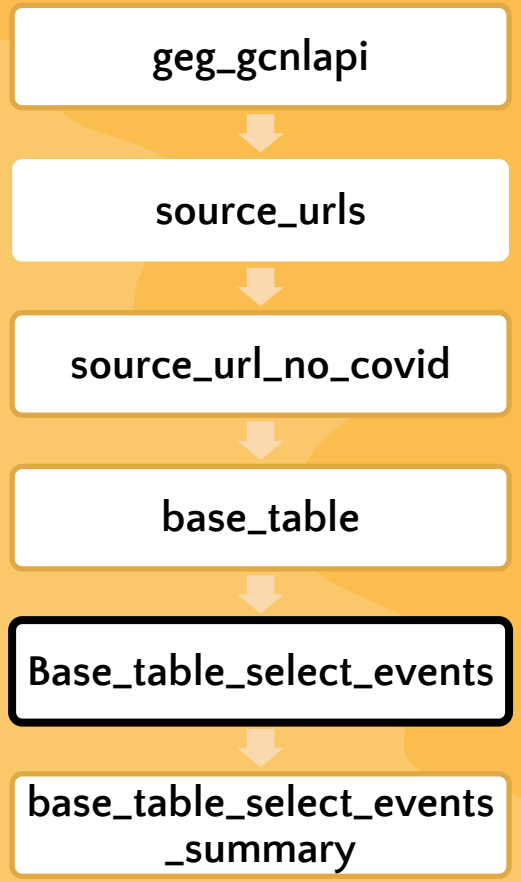


White list: רשימת מילים שנרצה שיופיעו בכתבה - בנושא ביטחון
Black list: רשימת מילים שאסור שיופיעו בכתבה

1. 53M שורות.
2. ישות מסוג אירוע.
3. אין מילה שמופיעה בblack list.
4. לפחות מילה 1 מופיעה בwhite list.
5. רמת בולטות 0.001 של האירוע.
6. כל שורה – מכילה את כל האירועים

של הכתבה ברמת בולטות מעל 0.001 (לכן גדלנו במספר)

Base_table_select_events



1. 9M שורות.
2. כל שורה מכילה URL, אירוע מwhite list, רמת בולטות.
3. לוקחים רק את האירועים הרלוונטיים.

base_table_select_events_summary

geg_gcnlapi

↓

source_urls

↓

source_url_no_covid

↓

base_table

↓

Base_table_select_events

↓

base_table_select_events_summary

1. 534 שורות.
2. מקבצים את הכתבות לפי תאריכים.
3. כל שורה = יום.

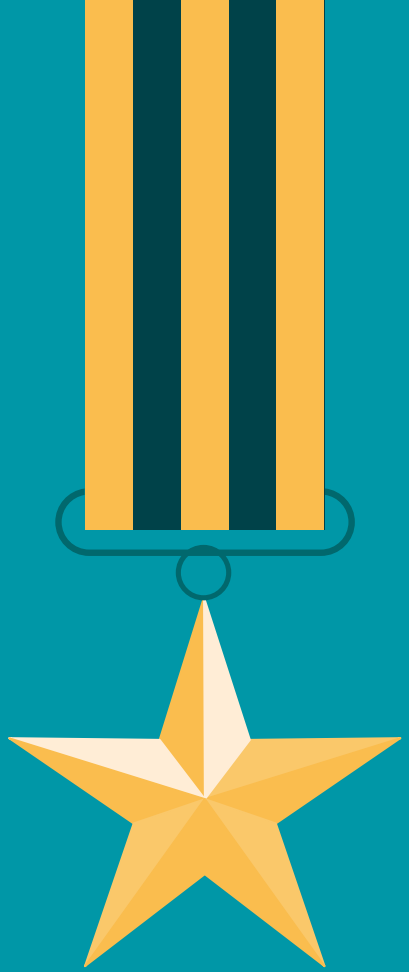
base table select events summary

Row	yyyyymmdd	count_distinct_urls	sum_numMentions	avg_avgSalience
1	2021-01-01	8384	14714	0.011296849944236036
2	2021-01-02	7546	13667	0.009286360058309037
3	2021-01-03	8361	14850	0.010949649659182038
4	2021-01-04	11182	18935	0.009956830177245874
5	2021-01-05	11768	20055	0.01020762288013797
6	2021-01-06	11480	19339	0.008743130456963778

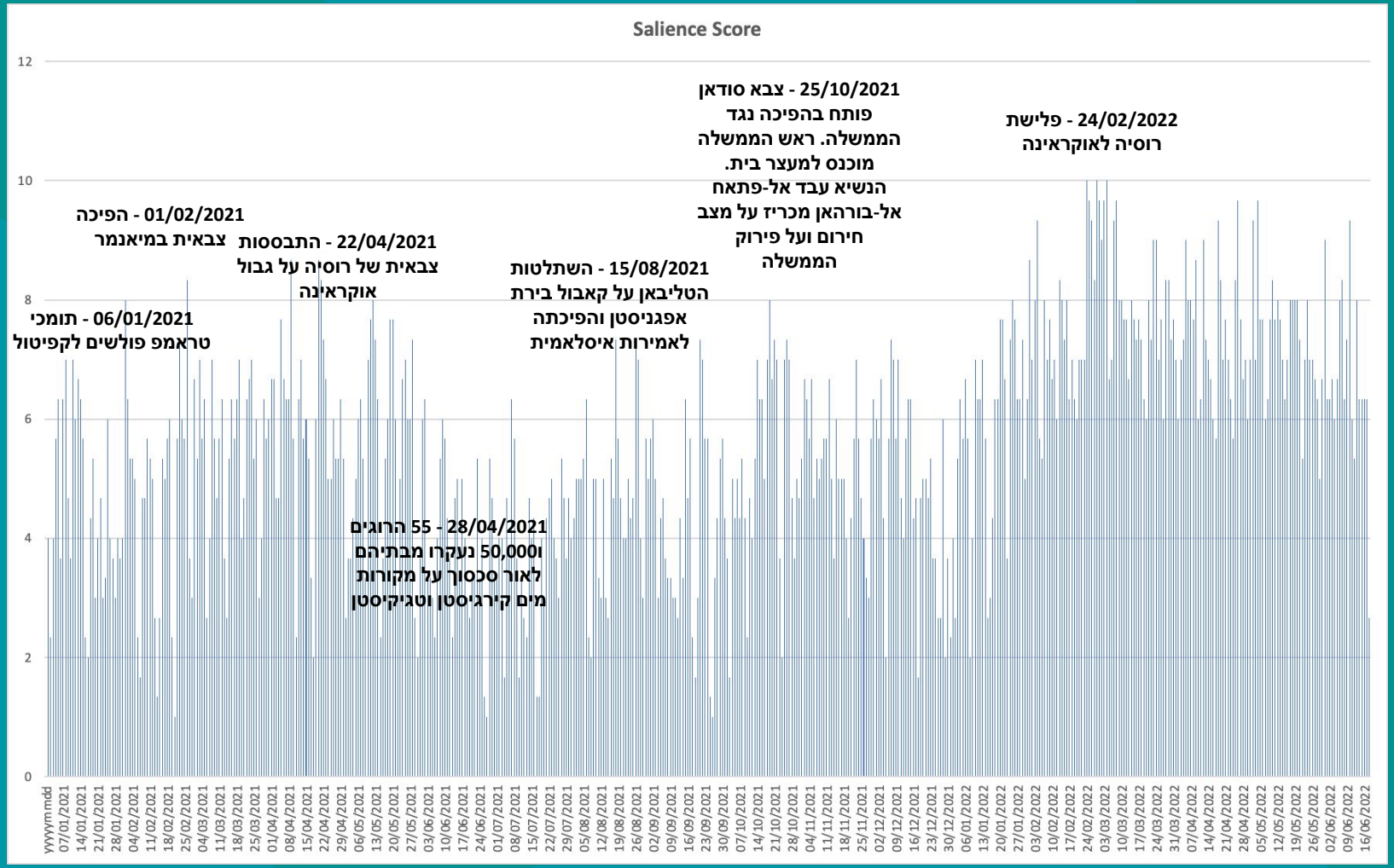
normalized

count_distinct_urls	count_distinct_urls_normalized	sum_numMentions	sum_numMentions_normalized
8384	1	14714	1
7546	1	13667	1
8361	1	14850	1
11182	4	18935	4
11768	6	20055	4
11480	5	19339	4
14477	8	25617	8

1. תאריך
2. מספר הכתבות באותו היום בנושא ביטחון
3. מספר האזכורים של white words בכתבות
4. רמת הבולטות הממוצעת- ממוצע השקט
5. **ככל שמדד הרעש גבוה יותר- יש פחות בטחון.**
6. **מקרה חריג: קפיצה בכמות הכתבות, קפיצה במספר האזכורים, וקפיצה במדד.**



ממצאים


Salience Score
12
10
8
6
4
2
0
06/01/2021 - תומכי טראמפ פולשים לקפיטול
01/02/2021 - הפיכה צבאית במיאנמר
22/04/2021 - התבססות צבאית של רוסיה על גבול אוקראינה
28/04/2021 - 55 הרוגים ו50,000 נעקרו מבתיהם לאור סכסוך על מקורות מים קירגיסטן וטגיקיסטן
15/08/2021 - השתלטות הטליבאן על קאבול בירת אפגניסטן והפיכתה לאמירות איסלאמית
25/10/2021 - צבא סודאן פותח בהפיכה נגד הממשלה. ראש הממשלה מוכנס למעצר בית. הנשיא עבד אל-פתאח אל-בורהאן מכריז על מצב חירום ועל פירוק הממשלה
24/02/2022 - פלישת רוסיה לאוקראינה
yyyyymmdd
07/01/2021
14/01/2021
21/01/2021
28/01/2021
04/02/2021
11/02/2021
18/02/2021
25/02/2021
04/03/2021
11/03/2021
18/03/2021
25/03/2021
01/04/2021
08/04/2021
15/04/2021
22/04/2021
29/04/2021
06/05/2021
13/05/2021
20/05/2021
27/05/2021
03/06/2021
10/06/2021
17/06/2021
24/06/2021
01/07/2021
08/07/2021
15/07/2021
22/07/2021
29/07/2021
05/08/2021
12/08/2021
19/08/2021
26/08/2021
02/09/2021
09/09/2021
16/09/2021
23/09/2021
30/09/2021
07/10/2021
14/10/2021
21/10/2021
28/10/2021
04/11/2021
11/11/2021
18/11/2021
25/11/2021
02/12/2021
09/12/2021
16/12/2021
23/12/2021
30/12/2021
06/01/2022
13/01/2022
20/01/2022
27/01/2022
03/02/2022
10/02/2022
17/02/2022
24/02/2022
03/03/2022
10/03/2022
17/03/2022
24/03/2022
31/03/2022
07/04/2022
14/04/2022
21/04/2022
28/04/2022
05/05/2022
12/05/2022
19/05/2022
26/05/2022
02/06/2022
09/06/2022
16/06/2022

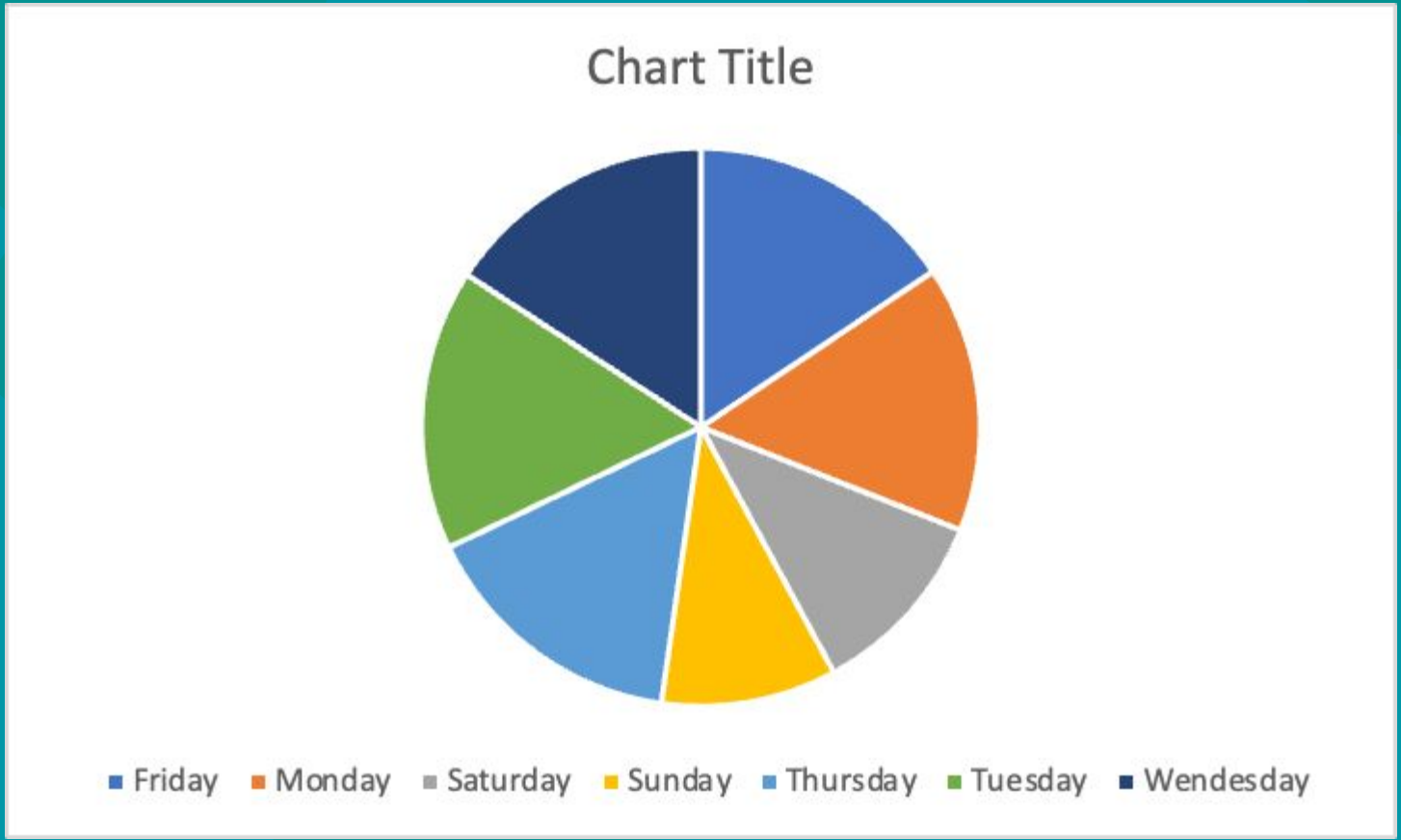


Chart Title
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wendesday

צעדים להמשך

01

בחינת רשימת המילים מחדש white words. דרישה של יותר ממילה אחת בכתבה

02

סיווג הדאטה לפי מדינות והצגת ממצאים על גבי מפה

03

סיווג המילים עצמן לנושאים – פיגועים, תאונות, אלימות ומלחמות..