מבחן מסכם - מגמת DATA - קודקוד 2

תאריך התחלה: 18.12.24
משך: 5 ימים מלאים

תיאור הפרויקט: פלטפורמת ניתוח נתוני טרור

הפרויקט נועד לתכנן וליישם פלטפורמה אנליטית לניתוח והצגת נתונים על טרור עולמי בעבר ממספר מקורות, בין השנים 1970 ל-2022 וגם להזין נתונים בזמן אמת שמגיעים ממקור של API של אתר חדשות.

הפלטפורמה תשלב ארכיטקטורת מיקרו-שירותים, ועליכם לבחור את המסדי נתונים המתאימים למשימות המתאימות, מערכות מסרים (messaging) מתאימים במידת הצורך ורכיבי עיבוד מידע מתאימים (Pandas, Numpy, Spark וכו').

בנוסף בפרויקט הזה יהיה לכם אתגר לייצר ממשק המציג מפה עם יכולות סינון מידע לפי הדרישות במסמך.

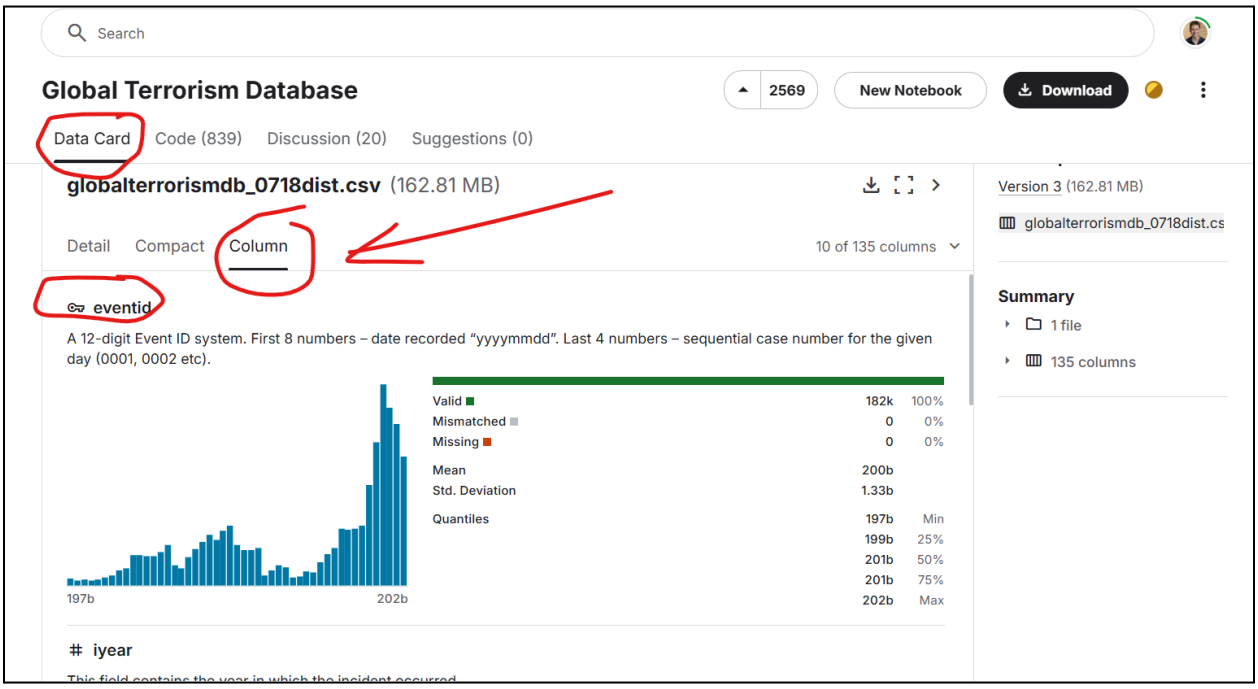
הפרויקט מחולק ל-5 שלבים:

1. **ניקוי מקור המידע הראשוני - קובץ CSV, בניית סכימת נתונים והזנתם לבסיס נתונים לבחירתכם**
2. **ביצוע חישובים סטטיסטיים ע"פ דרישות והנגשתם באמצעות Flask**
3. **הצגת חישובים במפות וגרפים**
4. **מיזוג מקור מידע נוסף של נתוני טרור - CSV נוסף**
5. **משיכת נתונים בזמן אמת ממקור API של חדשות (משיכת מידע כל 2 דקות) ועיבוי ע"י מציאת מיקום הידיעה וסיווגה ע"פ אירוע טרור או ידיעה כללית**
6. **הוספת יכולת לחיפוש טקסטואלי יעיל מכל מקורות המידע (מידע הסטורי ומידע שמגיע בזמן אמת)**

מבחן מסכם - מגמת DATA - קודקוד 2 **1**
תיאור הפרויקט: פלטפורמת ניתוח נתוני טרור 1
שלב 1: ניקוי מקור המידע הראשוני - קובץ CSV, בניית סכימות והזנה ל-DB 2
שלב 2: ביצוע חישובים סטטיסטיים ע"פ דרישות והנגשתם באמצעות Flask 3
קבוצת שאלות א 3
קבוצת שאלות ב 4
שלב 3: הצגת חישובים במפות וגרפים 5
שלב 4: מיזוג מקור מידע נוסף של נתוני טרור - CSV נוסף 8
שלב 5: משיכת נתונים בזמן אמת ממקור API של חדשות (משיכת מידע כל 2 דקות) ועיבוי ע"י מציאת מיקום הידיעה וסיווגה ע"פ אירוע טרור או ידיעה כללית 9
משיכת מידע בזמן אמת מ-newsapi.ai 9
סיווג הודעת חדשות וחילוץ שם של איזור לאירוע - GroqAPI 11
איתור נקודות ציון - OpenCage API 11
שלב 6: הוספת יכולת לחיפוש טקסטואלי יעיל מכל מקורות המידע (מידע הסטורי ומידע שמגיע בזמן אמת) 14

שלב 1: ניקוי מקור המידע הראשוני - קובץ CSV, בניית סכימות והזנה ל-DB

1. **הבנת מבנה הנתונים של מקורות הטרור**
   - דוגמא של בסיס הנתונים (1,000 רשומות)
   https://drive.google.com/file/d/17ZHQQXUGroh_aIUx0F7tHk23f-q650zO/view?usp=drive_link
   - דאטה-סט מלא (180,000 רשומות)
   https://drive.google.com/file/d/1zPn9zzjXPqHwemKb05bT2NJQrxelcdwr/view?usp=drive_link
   - הסבר מפורט על כל שדה בדאטה-סט
   https://www.kaggle.com/datasets/START-UMD/gtd/data



2. הבנת היישויות העיקריות בדאטה-סט (רמז: תאריך, מיקום, קבוצת טרור, סוג ההתקפה, כמות נפגעים)
3. ניקוי וסידור הנתונים
4. בחירת בסיס נתונים לשמירת הנתונים לביצוע שאילות בשלבים הבאים
5. עליכם להחליט החלטות לגבי מתי ואיפה מתבצעים חישובים ולפי זה לבחור בסיס נתונים מתאים ואת הסכימה המתאימה עבורם

שלב 2: ביצוע חישובים סטטיסטיים ע"פ דרישות והנגשתם באמצעות Flask

עליכם להוסיף נקודות קצה עבור 10 מתוך ה-19 שאילתות הבאות (אתם תחליטו כמה נקודות קצה לממש). אתם גם תחליטו איך ומתי לממש את הנקודות קצה האלה.

קבוצת שאלות א

<u>עליכם לבחור לפחות 5 שאילתות לממש מהרשימה למטה (לפחות 3 שאילתות עם תצוגה של מפות)</u>

1. **סוגי ההתקפה הקטלניים ביותר.** "קטלני ביותר" = הסוגים עם כמות הנפגעים (הרוגים ופצועים) הכי גדולה, כאשר נפגע = נקודה 1 והרוג שווה 2 נקודות מבחינת החישוב.
   a. אפשרות סינון לפי Top-5 או הכל
2. **ממוצע נפגעים לפי איזור.** ממוצע אחוז נפגעים (כמות הנפגעים לפי החישוב מהשאלה הקודמת) לאירוע לפי איזור.
   a. הצגה על מפה (יש לוודא שאפשר להבחין בין הגדלים באמצעות צבעים)
   b. אפשרות סינון לפי Top-5 או הכל
3. **חמש הקבוצות עם הכי הרבה נפגעים לאורך השנים**
4. **בחינת קורלאציה בין סוגי ההתקפה לסוג המטרות**
5. **מגמות שנתיות וחודשיות בתדירות ההתקפות.** "תדירות התקפות" = יחס כמות אירועים ייחודית לתקופת זמן.
   a. הצגה ב-2 גרפים, לתקופה של שנים וחודשים לפי שנה מסוימת
6. **אחוז שינוי במספר הפיגועים בין שנים לפי אזור.**
   a. אפשרות סינון: Top-5 או הכל
   b. הצגה על המפה
7. **מוקדי טרור גיאוגרפיים**. הצגת מוקדי אירועי טרור על המפה באמצעות שיטת **Heatmap** (הצגת כמות גדולה של אירועים במיקומים שונים, וה-Plugin של Folium מציג קבוצה קרבה של אירועים כגוש אחד)
   a. הצגה על המפה
   b. https://python-visualization.github.io/folium/latest/reference.html#folium.plugins.HeatMap
   c. הצגת כל האירועים לחודש, שנה, 3 שנים ו-5 שנים
   d. **בונוס**: אפשרות להציג את המידע בצורה של Time Animation, כלומר, לבנות רשימה של רשימות של מוקדי טרור (Heatmaps). במבנה של רשימה של רשימות, הרשימה החיצונים היא רשימה של פרקי זמן לאנימציה של הנתונים על המפה, והרשימות הפנימיות, הם ה-Heatmaps של אירועים קרובים אחד לשני: https://python-visualization.github.io/folium/latest/reference.html#folium.plugins.HeatMapWithTime
8. **הקבוצות הפעילות ביותר באזור מסוים.** *קיבוץ כמות אירועים לפי קבוצות טרור.*
   a. סינון לפי איזור ספציפי או כל האיזורים
   b. הצגה של הקבוצה הכי פעילה באיזור במפה. ע"י לחיצה על ה-marker במפה, להציג את ה-Top-5 קבוצות פעילות
9. **קורלציה בין מספר הפוגעים למספר הנפגעים.** *רמז: חישוב קורלציה בין nperps לסה"כ ההרוגים.*
10. **קורלציה של כמות אירועים מול מספר הנפגעים לפי אזור.**
    a. סינון לפי איזור (איזור ספציפי או כל האיזורים).
    b. הצגת קורלציה על המפה באיזור. יש להשתמש בצבעים להמחיש את הקורלציות הגבוהות לעומת הנמוכות

קבוצת שאלות ב

<u>עליכם לבחור לפחות 5 שאילתות לממש מהרשימה למטה (לפחות 3 שאילתות עם תצוגה של מפות)</u>

11. **זיהוי קבוצות עם מטרות משותפות באותו אזור. שאלה מנחה**: אילו קבוצות טרור תקפו את אותם מטרות באזור מסוים?
    a. תצוגה במפה.
    b. אפשרות סינון איזור: region או country
    c. הצגת marker עם כמות הקבוצות הגדולה ביותר עם מטרה משותפת באותו איזור
    d. בלחיצה על ה-marker אפשר לראות את רשימת כל הקבוצות עם מטרות משותפות באותו האיזור
12. **מעקב אחר פעילות קבוצות במספר אזורים לאורך זמן. שאלה מנחה**: אילו קבוצות הרחיבו את הפעילות שלהן לאזורים חדשים לאורך השנים? **לוגיקה**: מציאת קבוצות שנכנסו לאזורים חדשים לפי שנים.
13. **איתור קבוצות שהשתתפו באותן תקיפות. שאלה**: אילו קבוצות טרור היו מעורבות באותה תקיפה?
14. **זיהוי אזורים עם אסטרטגיות תקיפה משותפות בין קבוצות. שאלה**: באילו אזורים קבוצות שונות משתמשות באותן סוגי התקפות?
    a. תצוגה במפה
    b. אפשרות סינון איזור: region או country
    c. הצגת marker עם סוג ההתקפות עם כמות הקבוצות הייחודיות הכי גדולה באותו איזור
    d. בלחיצה על ה-marker, להראות את רשימת הקבוצות
15. **איתור קבוצות עם העדפות דומות למטרות. שאלה**: אילו קבוצות תוקפות באופן תדיר את אותם סוגי מטרות (למשל אזרחים, ממשלה)?
16. **זיהוי אזורים עם פעילות בין-קבוצתית גבוהה. שאלה**: באילו אזורים קיימת פעילות של המספר הגדול ביותר של קבוצות שונות - מגוון קבוצות ולא אותן קבוצות? **לוגיקה**: ספירת קבוצות ייחודיות הפעילות בכל אזור.
    a. תצוגה במפה
    b. אפשרות סינון איזור: region או country
    c. הצגת marker עם כמות קבוצות ייחודיות פעילות באיזור
    d. לחיצה על ה-marker תציג את רשימת הקבוצות
17. **ניתוח דפוסי נדידת קבוצות לאורך זמן. שאלה**: כיצד קבוצות משנות את אזורי הפעילות שלהן לאורך שנים?
18. **זיהוי קבוצות בעלות השפעה רחבה. שאלה**: אילו קבוצות הם גורמי השפעה עיקריים על אזורים וסוגי מטרות? **לוגיקה:** קבוצות עם הקישוריות הגדולה ביותר לאיזורים וסוגי מטרות
    a. תצוגה במפה
    b. אפשרות סינון איזור: region או country
    c. הצגת marker עם כמות הקישוריות הגדולה ביותר לאורך כל התקופה לאיזור מסוים
    d. לחיצה על ה-marker תציג את רשימת הקבוצות
19. **זיהוי קשרים בין קבוצות עם מטרות משותפות באותו זמן. שאלה**: אילו קבוצות תוקפות מטרות זהות באותה שנה?

שלב 3: הצגת חישובים במפות וגרפים

1. **איך מציגים מפות בפרויקט**
   a. אנחנו נשתמש בספרייה Folium (חינם, open source מבוסס על leaflit של JS)
   b. מידע ראשוני, התקנה ושימוש בסיסי:
   https://python-visualization.github.io/folium/latest/getting_started.html
   c. הוספת מרקרים (markers) נקודות ציון על המפה:
   https://python-visualization.github.io/folium/latest/getting_started.html#Adding-markers
2. **הצגת מפות**
   a. בניית ממשק UI ב-HTML להנגשת השאילתות שבניתם
   b. הממשק יהיה בנוי מדף HTML רגיל עם טופס אשר מאפשר (1) בחירת שאילתה ו-(2) הזנת נתונים שהשאילתות דורשות. עליכם לבנות את הפקדים הנדרשים עבור השאילתות שבחרתם. פקדים משמעותם:
      i. <input type="text">
      ii. <select> <option></option></select>
      iii. וכו'
   c. בתוך הדף HTML, יהיה IFrame שמציג את המפה. זאת כדי לאפשר לדף הראשי לרנדר את המפה
   d. דוגמא לדף **index.html**

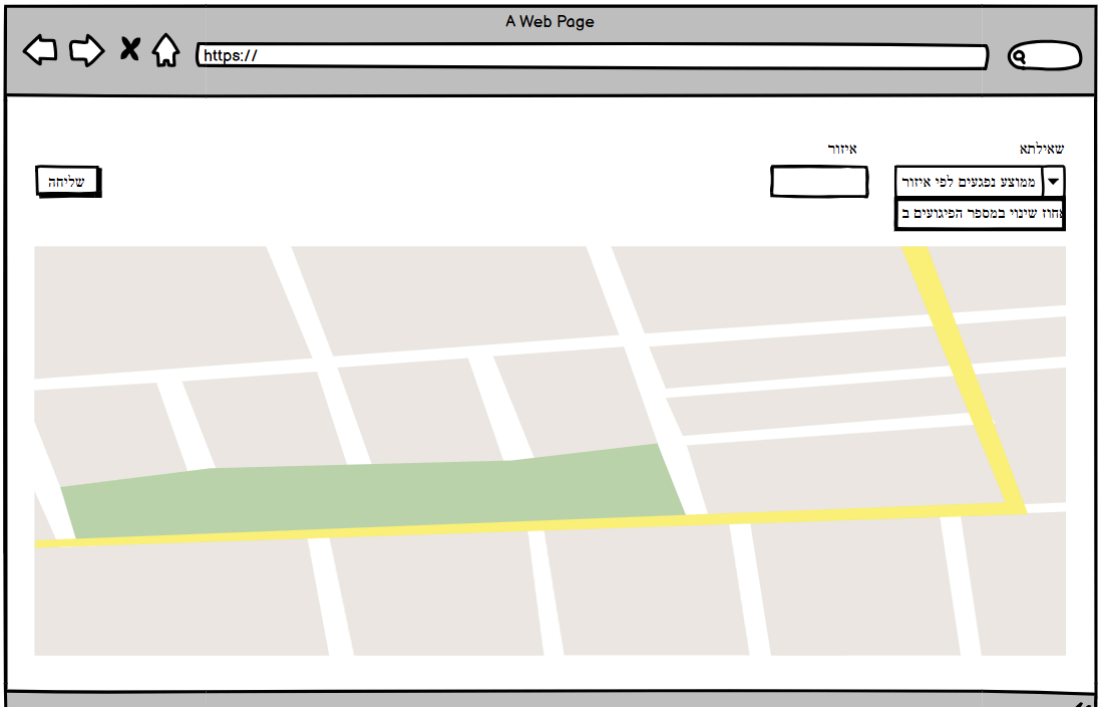
```python
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Interactive Map</title>
</head>
<body>
    <h1>Interactive Map with Flask and Folium</h1>

    <!-- User input form -->
    <form method="POST" action="/">

        <!-- Select field to choose query -->
        <!-- ... -->

        <!-- Region input field -->
        <label for="region">Region:</label>
        <input type="text" id="region" name="region" placeholder="Enter Region
Name" required><br>

        <!-- Other fields -->
```

```
        <!-- ... -->

        <!-- User input form -->
        <button type="submit">Update Map</button>
    </form>

    <!-- Embed the updated map -->
    <iframe src="{{ url_for('render_map') }}" width="100%"
height="600px"></iframe>
</body>
</html>
```

3. תבנו Flask endpoint אשר תקלוט את ה-submit של הטופס.
   a. דוגמא לנקודת קצה של Flask שמייצרת מפה ומרנדרת את הדף HTML הראשי (הכוללת את ה-IFRAME של המפה עצמה)

```python
from flask import Flask, render_template, request
import folium
import os

app = Flask(__name__)

@app.route("/", methods=["GET", "POST"])
def home():
    # Capture input from the user's HTML Form
    const region = request.form.get("region", region)

    # Capture additional fields
    # ...

    # Folium map definition and logic
    # ...

    # Save map to the 'map.html' file and render the main 'index.html' file
    map_path = os.path.join("templates", "map.html")
    map.save(map_path)

    return render_template("index.html")
```

4. **סקיצה.** המחשה של איך הממשק יראה (עם שאילתה אחת)



5. **הצגת גרפים**
   a. שאילתות שלא מציגות במפות, ייתכן וניתן להציג את המידע בגרפים.
   b. במידה ובחרתם שאילתא שניתן להציג בגרף, עליכם לממש נקודות קצה נוספת ב-Flask ולייצר HTML גרף אינטרקטיבי
   c. אתם יכולים לבחור את ספריית הגרפים שאתם רוצים - (מומלץ להשתמש ב-Bar Chart, Histogram או Pie Chart - תצוגות נפוצות)

שלב 4: מיזוג מקור מידע נוסף של נתוני טרור - CSV נוסף

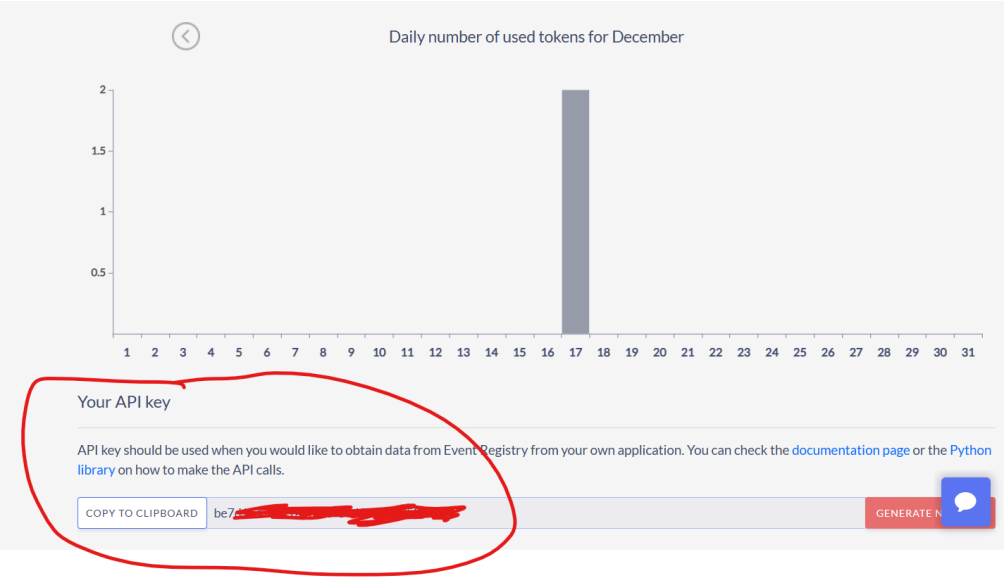
1. עליכם למזג דאטהסט נוסף על אירועי טרור למבנה נתונים שיצרתם.
2. צריך לוודא שכל השאילתות שבחרתם כוללות את המידע שצירפתם (במידה והמידע הרלוונטי קיים).
3. הדאטה-סט הנוסף הוא לא מלא כמו הדאטה-סט הראשי ויש בו חוסרים
4. מידע על הדאטה:
   a. **דוגמא (5,000 רשומות)**
   https://drive.google.com/file/d/1Qc3SjvrKt1RX5KKiEwQr6-lGZQpVpTdG/view?usp=drive_link
   b. **דאטהסט מלא**
   https://drive.google.com/file/d/1ydTQsk7II2h6sNTiZALvpuF4SoKKXyVH/view?usp=drive_link

---

שלב 5: משיכת נתונים בזמן אמת ממקור API של חדשות (משיכת מידע כל 2 דקות) ועיבוי ע"י מציאת מיקום הידיעה וסיווגה ע"פ אירוע טרור או ידיעה כללית

משיכת מידע בזמן אמת מ-newsapi.ai

1. עליכם להשתמש ב-API של חדשות חיצוני כדי לשאוב מידע נוסף בזמן אמת (קרוב לזמן אמת), להוסיף מידע חסר (תוך שימוש בשירותי API נוספים) ולשמור בבסיס נתונים לבחירתכם
2. ה-API שתשתמשו בו הוא **newsapi.ai**
   a. רישום חינם ומאפשר הפעלת 2,000 חיפושים בחינם
   b. חילוץ ה-API KEY לשימוש ב-API שלהם מהדף הראשי שלהם (אחרי רישום). לגלול למטה עד הסוף



3. מידע כללי על ה-API.
   a. אנחנו נשתמש בנקודת קצה אחת למשיכת הודעות חדשות https://newsapi.ai/documentation?tab=searchArticles
4. אתם תקראו ל-API הזה כל 2 דקות (כמובן להגביל שימוש בזמן הפיתוח כי יש הגבלה של 2,000 חיפושים לחשבון חינאמי) ותבקשו 100 תוצאות
   a. שימוש לב בדוגמא למטה מה שמסומן בצהוב. כל 2 דקות צריך לשנות את המשתנה articlesPage למספר הבא לקבל את ה-batch הבא של תוצאות
5. דוגמא לשימוש ב-API
   a. Request: **POST** **https://eventregistry.org/api/v1/article/getArticles**
   b. Body: (למטה)

```python
{
  "action": "getArticles",
  "keyword": "terror attack",
```

```
  "ignoreSourceGroupUri": "paywall/paywalled_sources",
  "articlesPage": 1,
  "articlesCount": 100,
  "articlesSortBy": "socialScore",
  "articlesSortByAsc": false,
  "dataType": [
    "news",
    "pr"
  ],
  "forceMaxDataTimeWindow": 31,
  "resultType": "articles",
  "apiKey": "be7d1e47-d51a-46d5-8440-4b6b75304261"
}
```

c. דוגמא לתוצאה שחוזרת (סיווג לאירוע טרור הסטורי)

Python

```
            {
                "uri": "8429097876",
                "lang": "eng",
                "isDuplicate": false,
                "date": "2024-11-26",
                "time": "05:50:27",
                "dateTime": "2024-11-26T05:50:27Z",
                "dateTimePub": "2024-11-26T05:49:25Z",
                "dataType": "news",
                "sim": 0,
                "url":
"https://aninews.in/videos/national/2611-terror-attack-maha-cm-shinde-deputy-cm
s-fadnavis-ajit-pawar-pay-tribute-on-anniversary/",
                "title": "Asia's Leading News Site - India News, Business &
Political, National & International, Bollywood, Sports | ANI News",
                "body": "Mumbai, Nov 26 (ANI): November 26, 2024 marked the
16th anniversary of the 26/11 Mumbai terror attack that shook the nation. On
```

```
the occasion, Maharashtra Guv Radhakrishnan paid floral tributes to Bravehearts
at Martyrs' Memorial on premises of Police Commissioner's Office. Maharashtra
CM Eknath Shinde also paid tribute to the Bravehearts at the Memorial on the
16th anniversary. Further, Maharashtra Deputy CMs Devendra Fadnavis, Ajit Pawar
paid homage to the Bravehearts.",
                "source": {
                    "uri": "aninews.in",
                    "dataType": "news",
                    "title": "Asian News International (ANI)"
                },
```


6. עבור כל תוצאה שחוזרת, אנחנו רוצים לחלץ נקודות ציון של האירוע המדובר כדי שנוכל להציג על המפה marker עם המידע המתאים. לכן נבצע את הפעולות הבאות:
   a. לסווג את ההודעה ל-3 קטגוריות באמצעות שימוש ב-**GroqAPI**: חדשות כללי, אירוע טרור היסטורי, אירוע טרור עכשווי
   b. <u>עבור אירוע טרור הסטורי או עכשווי</u>
      i. יש לבצע זיהוי מיקום - עיר, מדינה, איזור, וכו' (תוך שימוש ב-GroqAPI גם)
      ii. לבצע איתור נקודות ציון על המפה באמצעות שירות נוסף OpenCage GeoCoding API

סיווג הודעת חדשות וחילוץ שם של איזור לאירוע - GroqAPI

1. **מידע על** groq.com **- GroqAPI**
   a. מערכת קוד פתוח מקבילה ל-ChatGPT בחינם והשימוש בספרייה של groq היא דומה לשימוש בספרייה של openai
   b. צריך להירשם וליצור API Key לשימוש בנקודות קצה
2. הסבר ראשוני ודוגמאת קוד פשוטה בפייתון לשימוש ב-:GroqAPI
   https://console.groq.com/docs/quickstart

איתור נקודות ציון - OpenCage API

1. מידע על opencagedata.com - **OpenCage**
   a. מערכת חינאמית לאתר נקודות ציון במפה על בסיס שם של מקום (איזור, עיר וכדו')
2. הסבר ראשוני: https://opencagedata.com/api#quickstart
3. דוגמא לקריאת API

```python
https://api.opencagedata.com/geocode/v1/json?q=Gaza&key=d5b30e940b0d4b9fbe645fdeec29c10e&page=1
```

תשובה רצויה

```python
            "bounds": {
                "northeast": {
                    "lat": 31.5487199,
                    "lng": 34.5181367
                },
                "southwest": {
                    "lat": 31.4644954,
                    "lng": 34.3961515
                }
            },
            "components": {
                "ISO_3166-1_alpha-2": "PS",
                "ISO_3166-1_alpha-3": "PSE",
                "ISO_3166-2": [
                    "PS-GZA"
                ],
                "_category": "place",
                "_normalized_city": "Gaza",
                "_type": "city",
                "city": "Gaza",
                "continent": "Asia",
                "country": "Palestinian Territory",


              ...


            "confidence": 6,
            "formatted": "Gaza, Gaza Governorate, Palestinian Territory",
            "geometry": {
```

```
                "lat": 31.5128679,
                "lng": 34.4581358
            }
```

שלב 6: הוספת יכולת לחיפוש טקסטואלי יעיל מכל מקורות המידע (מידע הסטורי ומידע שמגיע בזמן אמת)

1. **הערות כלליות**
   - לכל endpoint תהיה האפשרות להגביל את כמות התוצאות באמצעות שדה limit
   - החיפוש על דאטה הסטורי מתכוון לשני ה-CSV שטענתם במהלך הפרויקט
2. **תאור נקודות הקצה הרצויות**
   - `/search/keywords`: חיפוש בכל מקורות הנתונים.
   - `/search/news`: חיפוש בנתוני חדשות בזמן אמת.
   - `/search/historic`: חיפוש בנתונים ההיסטוריים.
   - `/search/combined`: חיפוש כולל עם אפשרות לסינון תאריכים.
     i. כולל פרמטרים של תאריך התחלה ותאריך סיום
3. **תוצאות**
   - התוצאות יכללו מידע על אירועים ולכן התוצאות יוצגו על המפה

בהצלחה!!!!