תרגיל 2 – data scraping/crawling & data exploration

בתרגיל זה נתרגל את שלב איסוף והבנת הנתונים.

לעומת החיים האמיתיים, כאן נאפשר לכם להפוך את הסדר הטבעי של התהליך ולנתח את הנתונים לפני שבוצעה הבנה עסקית והוגדרה שאלת מחקר.

על כל זוג מגישים לבחור/לאסוף שני סטים של נתונים. האחד יהיה קובץ נתונים הנמצא על ה Web בפורמט מובנה, כגון: csv, טבלה, וכד’. על סט הנתונים השני להיאסף ע"י שימוש ב API של אתר מוכר.

אתם רשאים להחליט מהם תנאי הסינון להבאת הנתונים, אך עליכם לוודא שבידכם נתונים מגוונים מספיק בכדי לחלץ מתוכם תובנות. בנוסף עליכם לוודא שלפחות אחד מקבצי הנתונים כולל גם נתוני מיקום שניתן למקם על מפה.

הצעות לAPIs אפשריים: twitter, ebay, facebook, amazon, Instagram, Pinterest

מינימום מספר רשומות בכל אחד מהקבצים: 500.

בשלב הבא אתם נדרשים לנתח **בנפרד** את כל אחד משני קבצי הנתונים בכדי להבינם ולזהות מגמות מעניינות בנתונים. היעזרו בכלים שלמדנו כדי לתעד את תהליך הניתוח (שימוש בגרפים, ניתוח חשיבות מאפיינים שונים, מציאת קורלציה וכו'. אין צורך לבצע קיבוץ נתונים, מאחר וטרם נלמד) וליצור דו"ח הכתוב בשפת Markdown המסכם את תהליך הבנת הנתונים עבור כל אחד מקבצי הנתונים.

סכמו את הדו"ח בהמלצה על כיווני פעולה להמשך.

* פורמט הגשה:
  + יוקמו שני מאגרי GitHub לטובת התרגיל (אחד לכל סט נתונים)
  + כל מאגר יכיל את הנתונים לצד קוד רלוונטי ודו"ח מסכם
  + הדו"ח יכתב בשפת markdown, ישמש כקובץ README למאגר
    - כשילוב של מלל לצד קוד ותוצאה (ניתן להסתייע בRMARKDOWN)
  + הדו"ח כולל:
    - קריאת רשומות ראשונות בקבצי הנתונים.
    - תיאור הנתונים.
    - ניתוח הנתונים לפי קריטריונים שונים.
    - יש לצרף לפחות מפה אחת שנוצרה מתוך נתוני המיקום.
    - סיכום, מסקנות והמלצות להמשך מחקר.

רכיבי הציון:

* ייחודיות הנתונים (אין פרויקטים נוספים בקורס שהשתמשו ב API של אותו אתר להורדת נתונים): 5% בונוס (הציון המקסימאלי עדין 100, אבל 5 הנקודות יכפרו על חוסרים אחרים)
* שלמות ותקינות (הדו"ח מכיל את כל השלבים הנדרשים, כל שלב בוצע באופן תקין): 40%
* הבנת הנתונים (הודגמה הבנת הנתונים באופן משכנע): 25%
* יכולת אנליטית (ניתוח הנתונים בצורה אפקטיבית תוך הפגנת שליטה בכלי עזר רלוונטיים): 25%
* איתור מגמות מעניינות: 10%

יש להגיש את התרגיל בזוגות עד יום חמישי 14/4/2016.

תרגילים שיוגשו באיחור "יקנסו" באופן הבא: 5 נקודות יורדו מציון התרגיל לכל יום איחור.

ההגשה תתבצע בתיקיה המתאימה באתר הקורס באופן הבא:

1. על כל אחד מבני הזוג להגיש את העבודה.

2. יש להגיש קובץ ובו קישור לשני מאגרי ה Githubשל התרגיל וכן שמות השותפים לתרגיל. שם הקובץ יהיה מהסוג ID1\_ID2.zip (תעודות הזהות של המגישים מופרדות באמצעות קו תחתון).

בהצלחה ובהנאה!