המכללה האקדמית להנדסה ע"ש סמי שמעון

המחלקה להנדסת תוכנה

קורס: מבוא למדעי הנתונים – תשפ"ב

מטלה 1 – Data Collection

תאריך הגשה: 25/11/2021

מטרת המטלה

התנסות הסטודנטים באיסוף מידע מאתר (Crawling) לצד הבאת מידע מ-API, סידורו והכנתו כדאטאסט לניתוח.

תוכן המטלה

הנכם מתבקשים לבנות דאטאסט של שמות, כתיבתם בצורות שונות ופירושיהם. לשם כך עליכם לפרסר את כל השמות המצויים באתר "20,000 שמות" שכתובתו https://www.20000-names.com/index.htm ולסדר את השמות בקבצי CSV קלים ונוחים לשימוש.

<u>שלב א': בניית קרולר (50 נקודות)</u>

עליכם להשתמש באחת מחבילות הפייתון שנלמדו בכתה כמו BeautifulSoup, Scrapy וכ"ו על מנת לבנות את שני קבצי CSV (מקבילים לקבצי excel) הבאים:

1. name_descriptions.csv
2. names_other_languages.csv

הקובץ הראשון צריך להכיל את השמות המקוריים הנכתבים באנגלית, לצד מידע כללי על השם ופירושו, יחד עם פרטים נוספים כמו מין והמקור ממנו מגיע השם.

הקובץ הראשון הנקרא name_descriptions.csv נדרש להכיל את העמודות הבאות:

1. שם (Original_Name) - עמודה השומרת את השם מהאתר.
2. תיאור (Description) -עמודה זו שומרת את המלל המצורף לצד השם באתר.
3. מין (Gender) – עמודה שנותנת אינדיקציה האם מדובר בשם לבן או לבת.
4. מיקום (Location) – עמודה בה נשמור את המיקום בעולם (שם אלבני, יהודי, אפריקאי וכ"ו)

הקובץ השני הנקרא names_other_languages.csv נדרש לכלול את העמודות הבאות:

1. שם (Original Name) - עמודה השומרת את השם מהאתר.
2. שם בשפת המקור (Name in Other Language) – עמודה השומרת את השם בשפת המקור (יכולה להיות ריקה במקרה ואין נתון שכזה).
3. השפה בה כתוב השם (Language) – עמודה השומרת על השפה בה כתוב השם.

4. מיקום (Location) – עמודה בה נשמור את המיקום הגיאוגרפי או הלאום ממנו הגיע השם (Polish, Irish, Hebrew).

נדגים זאת באמצעות מספר דוגמאות:

1. AI (Chinese: 1: 蔼, 2: 爱, Japanese: 1: 藍, 2: 愛): Japanese name meaning 1) "indigo" or 2) "love." Compare with another form of Ai.

   דוגמה זו הגיעה מהדף - http://www.20000-names.com/female_japanese_names.htm

2. ABIYSHAG (אֲבִישַׁג): Hebrew name meaning "my father is a wanderer" or "father of error." In the bible, this is the name of a young girl who cared for David in his old age. Also spelled Avishag.

   דוגמה זו הגיעה מהדף - http://www.20000-names.com/female_hebrew_names.htm

במקרה כאן עליכם לבנות את הקובץ הראשון הנקרא name_descriptions.csv בצורה הבאה:

Original Name	Description	Gender	Location
AI	Japanese name meaning 1) "indigo" or 2) "love." Compare with another form of Ai.	Female	Japanese
ABIYSHAG	Hebrew name meaning "my father is a wanderer" or "father of error." In the bible, this is the name of a young girl who cared for David in his old age. Also spelled Avishag.	Female	Hebrew

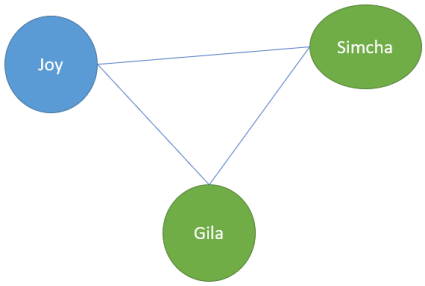
לגבי הקובץ השני הנקרא names_other_languages.csv הנכם נדרשים למלא את הקובץ עבור אותה הדוגמה בצורה הבאה:

Original Name	Name in Other Language	Language	Location
AI	蔼	Chinese	Japanese
AI	爱	Chinese	Japanese
AI	藍	Japanese	Japanese
AI	愛	Japanese	Japanese
ABIYSHAG	אֲבִישַׁג	Hebrew	Hebrew

שלב ב': בניית גרפים (30 נקודות)

בשלב זה בתור Data Scientists מתחילים, עליכם לבחון שמות בעלי משמעות זהה, אך ממקורות שונים. לשם כך הנכם נדרשים לבנות גרף מבוסס משמעות. גרף זה יכלול שמות ממקורות שונים שיש להם את אותה משמעות.

לדוגמה: נניח שיש את השמות Gila, Joy ו-Simcha שלהם משמעות זהה על פי האתר. במקרה זה עליכם להיעזר בקוד שהוצג בכיתה ולבנות גרף בצורת קליקה (כולם מחוברים לכולם). להלן אילוסטרציה פשוטה:



שלב ג': Twitter API (20 נקודות)

בשלב זה הנכם נדרשים לאחזר טוויטים מטוויטר בהינתן מילות מפתח שונות. תחילה עליכם לקבל access token מטוויטר על מנת להתחבר ל-API. לשם כך, הנכם נדרשים לקרוא את הלינק הבא:

https://developer.twitter.com/en/docs/authentication/oauth-1-0a/obtaining-user-access-tokens

ולקבל חשבון מפתח בטוויטר עפ"י ההוראות. לאחר שהשגתם את המפתח הדרוש, הנכם נדרשים לאסוף מידע לגבי הטוויטים וכותביהם.

עליכם לאסוף טווטים המאוחזרים כתוצאה מהביטויים הבאים:

1. "Climate Change Crisis"
2. "Champions League" (ליגת האלופות בכדורגל האירופית)
3. "The Witcher" (הסדרה "המכשף")

הנכם נדרשים לאסוף 100 טוויטים אחרונים לגבי כל ביטוי. בסה"כ עליכם להכין שני קבצי csv כאשר אחד מהם נקרא tweets.csv המכיל את הטוויטים שאוחזרו. לצד המידע לגבי כל טוויט עליכם להוסיף עמודה בשם keyword המציינת את הביטוי שהוכנס ובגינו הגיע הטוויט הנתון. הקובץ השני הינו authors.csv והוא מכיל את כותבי הטוויטים שפירסמו את הטוויטים השמורים בקובץ tweets.csv.

הוראות הגשה

1. עליכם להגיש קובץ ZIP הכולל את מספרי תעודות הזהות של שני המגישים עם קו תחתון מפריד ביניהם בצורה הבאה: ID1_ID2.ZIP.
2. קובץ ה-ZIP נדרש להכיל את תוצרי הפרויקט: קבצי ה-CSV המתוארים, מחברת Jupyter Notebook או Google Collab.
3. את קובץ ה-ZIP עליכם להעלות לתיבת ההגשה.
4. הגשה **בזוגות בלבד**.
5. תאריך ההגשה המעודכן ומיקום ההגשה יופיעו באתר הקורס.