המכללה האקדמית להנדסה ע"ש סמי שמעון

המחלקה למדעי המחשב

קורס: מבוא לכריית הנתונים – תשפ"ג

מטלה 1 – **Data Collection**

תאריך הגשה: 27/03/2023

מטרת המטלה

התנסות הסטודנטים באיסוף מידע מאתר (Crawling), סידורו והכנתו כדאטאסט לניתוח.

תוכן המטלה

בקורס כריית נתונים נלמד להוציא תובנות ותבניות שכיחות ומעניינות ממידע נתון. עם זאת, במקרים רבים אין ברשותנו דאטאסט (סט מידע) מתאים לניתוח. לכן, במקרים שכאלו, אנו נדרשים לאסוף את המידע בעצמנו ולשמור אותו בצורה טובה ונוחה.

במשימה זו הנכם מתבקשים לבנות דאטאסט של שמות, ואת צורת כתיבתן השונה בשפות השונות.

לשם כך עליכם להוציא את **כל** השמות הפרטיים והמשפחה המצויים באתר ויקיפדיה תחת העמוד שכתובתו https://en.wikipedia.org/wiki/Category:Given_names

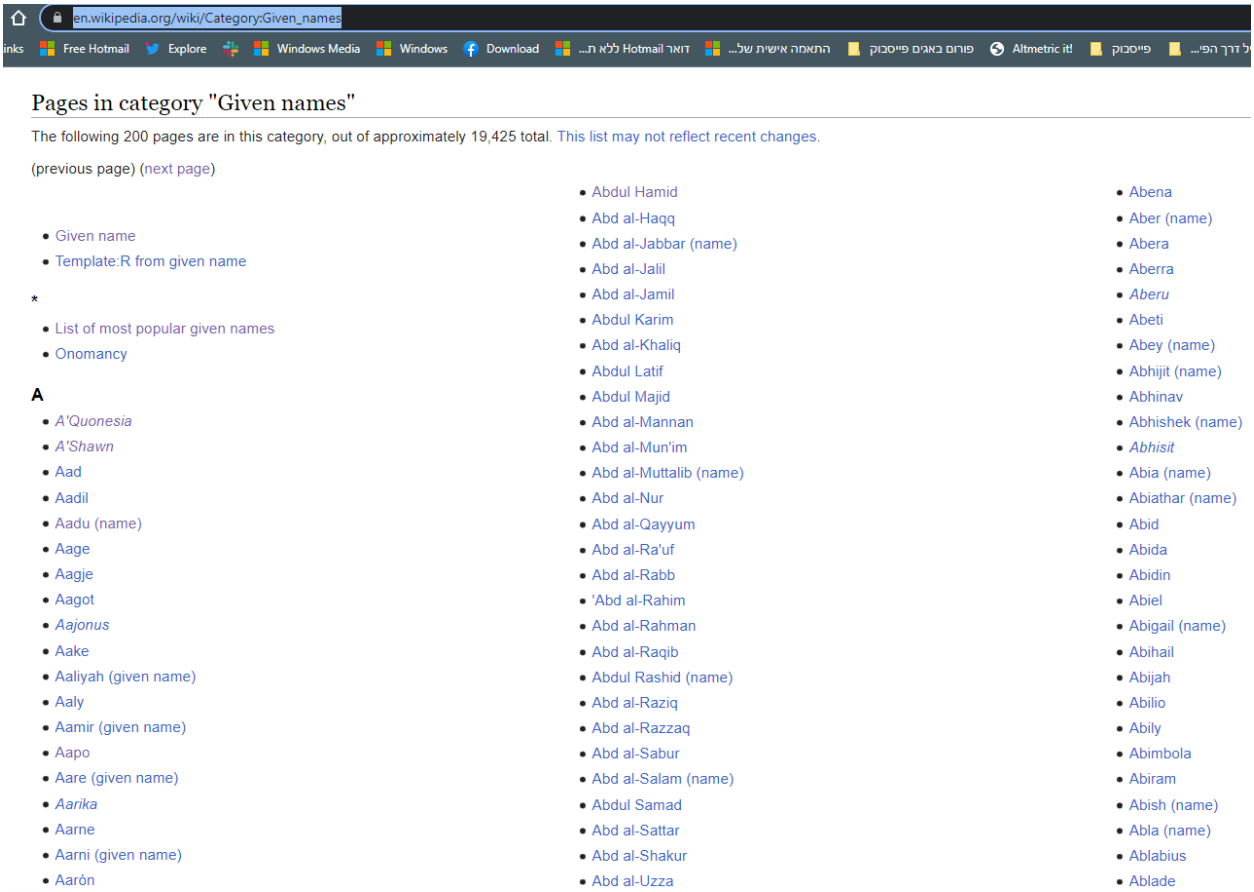
https://en.wikipedia.org/wiki/Category:Surnames

ולסדר את השמות בקבצי CSV קל ונוח לשימוש.

<u>הסבר על התהליך</u>:

במקרה כאן תצטרכו לחפש את השמות בוויקיפדיה. ולעבור בין השמות באמצעות הלינקים השונים.

1. לשם ההדגמה הלינק ממנו יש להתחיל הנו https://en.wikipedia.org/wiki/Category:Given_names אחרי שתסיימו עם השמות הפרטיים עברו לשמות המשפחה בלינק הבא: https://en.wikipedia.org/wiki/Category:Surnames
2. דפים אלו בנויים ממספר רב של לינקים של השמות כאשר כל השמות עצמם מסודרים בסדר ABC.



3. עליכם תחילה להביא דף זה באמצעות requests ולהוציא את השמות של אותו דף באמצעות הדוגמאות שהראנו בכיתה (שימוש בחבילת requests, beautifulSoup וכ"ו). יש לזכור שעליכם לעבור על כל השמות. אתם נמצאים ב-A וצריכים לעבור ל-B, C, D וכך הלאה עד הסוף.
4. עבור כל שם עליכם לחפש את המקבילה עבורו באתר WikiData.

מצורפת לכם דוגמה לקוד שמוציא את ה-WikiData ID בהינתן ערך מויקיפדיה.

יכול להיות שלדף אין ערך ב-WikiData. במקרה זה עליכם לדלג על דף זה ולהמשיך לשם הבא.

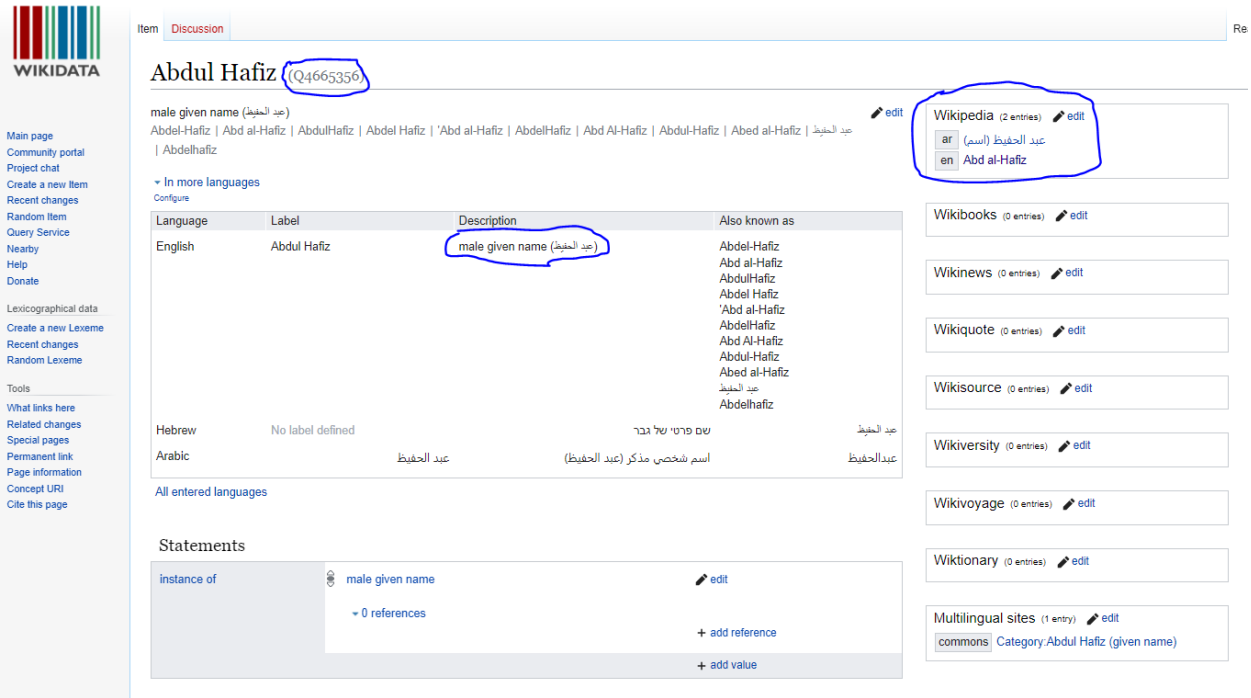
דוגמה לקוד ניתן למצוא מטה.

```python
import requests
import json

# Replace 'page_title' with the title of the Wikipedia page you want to extract the Wikidata ID for
page_title = 'Abd al-Hafiz'
#page_title = "Boarman"
#page_title = "Bobert"
#page_title = "Aabrekk"

# Construct the API URL
api_url = f'https://en.wikipedia.org/w/api.php?action=query&titles={page_title}&prop=pageprops&format=json'

# Make the API request and convert the response to JSON
response = requests.get(api_url).json()

# Extract the page ID from the JSON response
page_id = next(iter(response['query']['pages'].values()))['pageprops']['wikibase_item']

print(f"The Wikidata page ID for {page_title} is {page_id}")
```

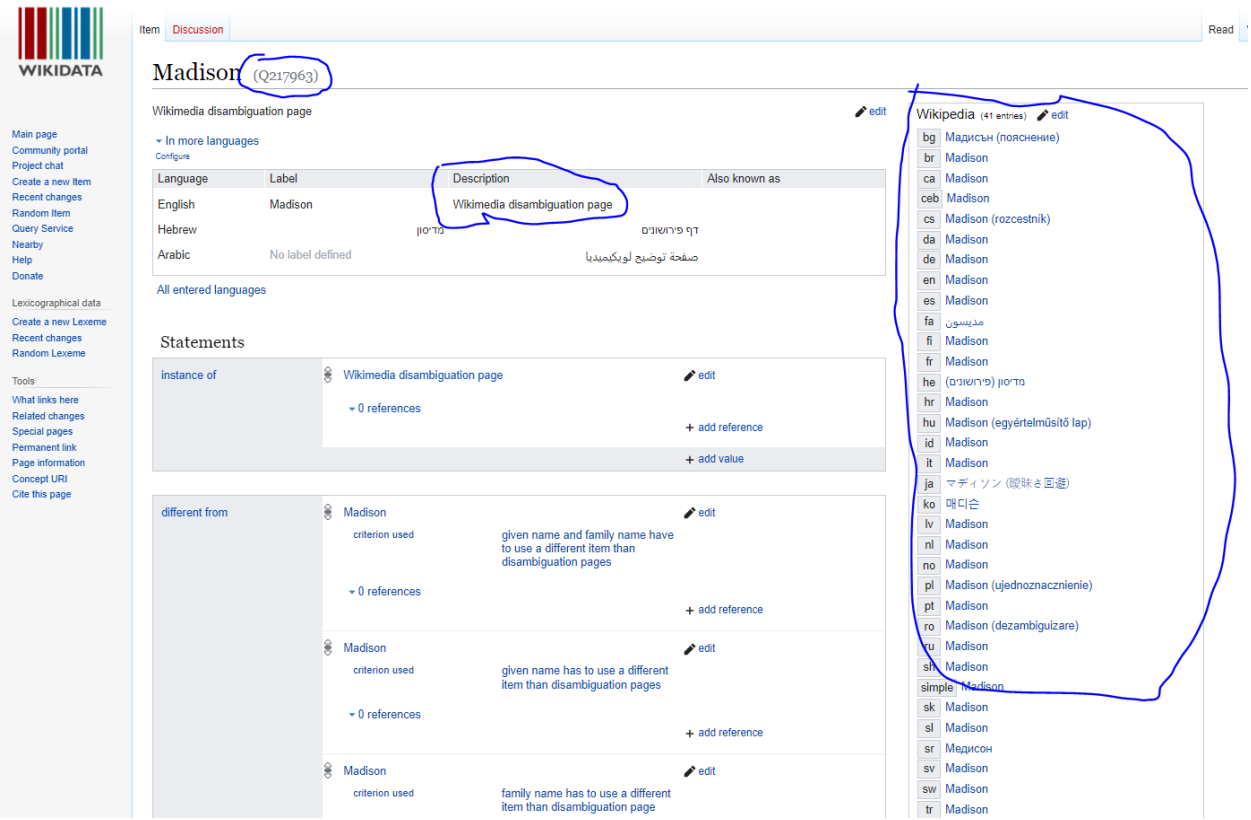
```
The Wikidata page ID for Abd al-Hafiz is Q4665356
```

ניתן לראות שנכנס השם Abd al-Hafiz ויוצא כפלט הערך המקביל לו מ-WikiData בשם Q4665356.

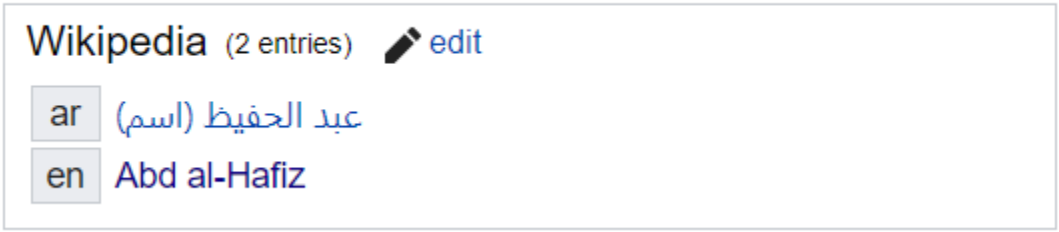
5. בהינתן שמצאתם דף ב-WikiData שמתאים לשם שברשותכם מ-Wikipedia. עליכם לגשת לדף הזה ב-WikiData. מצ"ב תמונה מהשם Abdl Hafiz



מצ"ב תמונה נוספת עבור השם Madison



6. עליכם לאסוף את ה-entries בדף זה תחת הכותרת Wikipedia ולשמור לרשימת tuples. דוגמה:



7. לאחר ששמרתם את כל השמות עליכם להעביר את רשימת ה- Tuples הזו ל-Pandas DataFrame (כמו בדוגמאות שראינו בביתה).

מבנה הקובץ צריך להיות כזה:

Label	WikiDate ID	English Description	Language	Wiki Short Lang	Entry
Abdul Hafiz	Q4665356	male given name (عبد الحفيظ)	English	en	Abd al-Hafiz
Abdul Hafiz	Q4665356	male given name (عبد الحفيظ)	Arabic	ar	عبد الحفيظ (اسم)
…		…	…	…	…
Aapo	Q16001311	male given name	English	en	Aapo
Aapo	Q16001311	male given name	Estonian	et	Aapo
Aapo	Q16001311	male given name	Finnish	fi	Aapo
Aapo	Q16001311	male given name	French	fr	Aapo
….		Surname	…	….	…

חשוב מאוד לשמור על שמות העמודות הללו!

8. לבסוף עליכם לשמור את הקובץ באמצעות פקודה to_csv. חשוב מאוד לשמור עם הפקודה encoding. חשוב מאוד לפתוח את הקובץ ולראות שהמידע כתוב כמו שצריך ולא בג'יבריש.

<u>דוגמה למחברת שתוכלו להעזר בה וראיתם בכיתה</u>

Example for HW1 - Data Mining Names.ipynb

https://colab.research.google.com/drive/1zTJzyaMejynsxSVyVXk1MiQZC2AX2054?usp=sharing

הקובץ נדרש להכיל את השם בשפות השונות הוא נדרש להכיל את העמודות הבאות:

1. שם (Label) - עמודה השומרת את השם מהאתר של ויקיפדיה. מדובר על השם הפרטי או שם המשפחה שאנחנו מחפשים אחריו.
2. WikiData ID -עמודה זו שומרת את ה-ID שמצאתם.
3. English Description – עמודה זו שומרת את התיאור הכללי באנגלית עבור השם הנתון.
4. Language – עמודה ששומרת את השפה המקבילה בה כתוב השם (Arabic ,English).

5. Wiki Short Lang – עמודה ששומרת את הקיצורים של השפה כפי שמופיעים בדף en עבור אנגלית, ar עבור ערבית, וכ"ו.

6. Entry – נועד לשמור את השם בשפה המסויימת.

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 (אפשר אחד שמייצג את השם הפרטי ואחד את המשפחה ואפשר אחד מאוחד. לבחירתכם), מחברת Jupyter Notebook או Google Collab.
3. את קובץ ה-ZIP עליכם להעלות לתיבת ההגשה.
4. הגשה בזוגות בלבד! רק אחד נדרש להעלות את העבודה.
5. תאריך ההגשה המעודכן ומיקום ההגשה יופיעו באתר הקורס.