

# 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 שבאלו, 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:Given_names)

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

ולסדר את השמות בקבצי CSV קל ונוח לשימוש.

#### הסבר על התהליך:

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](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.

## Pages in category "Given names"

The following 200 pages are in this category, out of approximately 19,425 total. [This list may not reflect recent changes.](#)

(previous page) (next page)

- Given name
- Template:R from given name

\*

- List of most popular given names
- Onomancy

### A

- A'Quonesia*
- A'Shawn*
- Aad
- Aadil
- Aadu (name)
- Aage
- Aagje
- Aagot
- Aajonus*
- Aake
- Aaliyah (given name)
- Aaly
- Aamir (given name)
- Aapo
- Aare (given name)
- Aarika*
- Aarne
- Aami (given name)
- Aarón

- Abdul Hamid
- Abd al-Haqq
- Abd al-Jabbar (name)
- Abd al-Jalil
- Abd al-Jamil
- Abdul Karim
- Abd al-Khaliq
- Abdul Latif
- Abdul Majid
- Abd al-Mannan
- Abd al-Mun'im
- Abd al-Muttalib (name)
- Abd al-Nur
- Abd al-Qayyum
- Abd al-Ra'uf
- Abd al-Rabb
- 'Abd al-Rahim
- Abd al-Rahman
- Abd al-Raqib
- Abdul Rashid (name)
- Abd al-Raziq
- Abd al-Razzaq
- Abd al-Sabur
- Abd al-Salam (name)
- Abdul Samad
- Abd al-Sattar
- Abd al-Shakur
- Abd al-Uzza

- Abena
- Aber (name)
- Abera
- Aberra
- Aberu*
- Abeti
- Abey (name)
- Abhijit (name)
- Abhinav
- Abhishek (name)
- Abhisit*
- Abia (name)
- Abiathar (name)
- Abid
- Abida
- Abidin
- Abiel
- Abigail (name)
- Abihail
- Abijah
- Abilio
- Abily
- Abimbola
- Abiram
- Abish (name)
- Abla (name)
- Ablabius
- Ablade

- 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.
- 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.

```
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 Abdul Hafiz

The screenshot displays the Wikidata page for the item **Abdul Hafiz** (ID: Q4665356). The page includes a sidebar with navigation links, a main content area with a table of labels and descriptions, and a right-hand panel with links to related Wikipedia and Wikisource pages. The 'Wikipedia' section shows two entries: 'عبد الحفيظ (اسم)' in Arabic and 'Abd al-Hafiz' in English, both circled in blue. The 'Statements' section at the bottom shows the item is an instance of 'male given name'.

Language	Label	Description	Also known as
English	Abdul Hafiz	male given name (عبد الحفيظ)	Abdel-Hafiz Abd al-Hafiz Abdullah Abdel-Hafiz Abd al-Hafiz Abdel-Hafiz Abd Al-Hafiz Abdul-Hafiz Abd al-Hafiz عبد الحفيظ Abdelhafiz
Hebrew	No label defined	שם פרטי של גבר	عبد الحفيظ
Arabic		عبد الحفيظ	اسم شخصي مذكر (عبد الحفيظ) عبد الحفيظ

מצ"ב תמונה נוספת עבור השם Madison

Item **Discussion** Read

## Madison (Q217963)

Wikimedia disambiguation page

✎ edit

Wikipedia (41 entries) ✎ edit

Language	Label	Description	Also known as
English	Madison	Wikimedia disambiguation page	
Hebrew		מדסון	דף פירושים
Arabic	No label defined		صفحة توضيح لويكيبيديا

All entered languages

### Statements

**instance of** Wikimedia disambiguation page ✎ edit

↗ 0 references + add reference

+ add value

**different from**

<div> <div>✎</div> <div>Madison</div> </div> <div> <div>✎</div> <div>edit</div> </div>	<div> <div>✎</div> <div>edit</div> </div>
<div> <div>✎</div> <div>Madison</div> </div> <div> <div>✎</div> <div>edit</div> </div>	<div> <div>✎</div> <div>edit</div> </div>
<div> <div>✎</div> <div>Madison</div> </div> <div> <div>✎</div> <div>edit</div> </div>	<div> <div>✎</div> <div>edit</div> </div>

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

Wikipedia (2 entries) ✎ edit

ar عبد الحفيظ (اسم)

en Abd al-Hafiz

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

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.

### דוגמה למחברת שתוכלו להעזר בה וראיתם בכיתה

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. WikiDate 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 (English, Arabic).

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.