

Group 3

Andrew Aziz - Ebram Thabet - Mohamed Alashkar - Omar Elwaliely

Documentation for web scraping the AUC Digital library

“<https://digitalcollections.aucegypt.edu/digital/collection/p15795coll20/search>”

Firstly, the user should run the Python script named "Webscrap_Selenium". This script will open the library website:

["https://digitalcollections.aucegypt.edu/digital/collection/p15795coll20/search"](https://digitalcollections.aucegypt.edu/digital/collection/p15795coll20/search). It is very important **not** to close this website while the script is running.

For each book on this website, a new folder will be created in the directory where the Python script is being executed. The folder name will match the book's name. All images **containing faces** from the book will be downloaded into this folder as ".jpg" files. The image filenames will follow the format: "BookName_PageName".

If two images have the same name, the first image will be saved as "BookName_PageName", and the second will be saved as "BookName_PageName (1)", with subsequent images following this numbering pattern.

If a folder is named "Unknown_Book_Name" or an image is named "Unknown_Book_Name_Unknown_Page_Name", it means the code could not find the book name or page name. The user may need to rerun the code to resolve this issue.

Books will be processed and downloaded in the order they appear on the digital library website. If the user stops the script and reruns it, the script will start from the beginning, iterating through the books in order. However, if it finds a folder with the same name as a book, it will skip that book and continue to the next one. This process will repeat until it reaches a book for which no corresponding folder exists. Essentially, the script resumes the downloading process from the last book it was working on before being stopped.

This code relies on the **HTML** structure of the AUC Digital Library website. If there are any changes to the website's HTML, the code may stop working or crash.

This Google Drive folder contains the **expected output** that the user should receive after running the whole code (Use AUC email to be able to access it): ["https://drive.google.com/drive/folders/1NAi29f37pdWhlwIJSzQ6YNC1V4s1YS56?usp=sharing"](https://drive.google.com/drive/folders/1NAi29f37pdWhlwIJSzQ6YNC1V4s1YS56?usp=sharing)

For the **OCR** (Optical Character Recognition) functionality, we use the library “Pytesseract” to extract text from images in each book. The OCR process specifically targets names written in capital letters or those preceded by titles such as 'Mr.', 'Mrs.', or 'Dr.'.

To use the OCR code, the user needs to specify the path to the folder containing the images by setting a variable named “folder_path”. Once the path is set, the OCR code can be executed to perform the extraction and matching.

When the user runs the OCR code, it matches the extracted names with the correct face encodings in some images (you can find these images below) from the first book in the AUC digital library, titled 'Al Hawdaj of the Campus Caravan 1958'. This process correctly associates names with their corresponding face encodings and saves this data in a CSV file named 'faces_data.csv'.

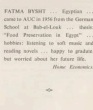
Our GitHub repository also includes a CSV file named “Webscraped_Book.csv”, which contains the names and face encodings for the below images from the book 'Al Hawdaj of the Campus Caravan 1958'. Below is a list of images that the current OCR can process and match with the corresponding face encodings:





FARIDA TAKAZI ... Palestinian ... a graduate of the American College for Girls ... came to AUC in 199 ... thesis: "Finishing Cotton Fabric for Clothing" ... hobbies: reading and trips ... was a member of the Arab Cultural Club and the Science Club ... feels sorry to leave her friends but happy to meet new ones.

Rene Esmont



FATMA HYSHT ... Egyptian ...
came to AUC in 1956 from the German
School at Bab-el-Luck ... thinks
"Food Preservation in Egypt" ...
hobbies: listening to soft music, using
reading novels ... happy to graduate
but worried about her future life.
Home Economics



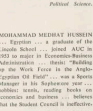
...



00
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99



MAHMOUD BAHLOUL... Bachelor
... came to AUC in 1955 from the
Preparatory Section in AUC ... thesis:
"Toward the Unity of Syria, Egypt and
Jordan" ... a member of the Arab
Cultural Club ... hobbies: reading
political books related to the Arab
states ... believes that AUC paves the
way for the student to work indepen-



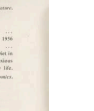
MOHAMMAD MEHDIAT HUSSEIN
... Egyptian ... a graduate of the
Lincoln School ... joined AUC in
1953 to major in Economics-Business
Administration ... thesis: "Building
up the Work Force in the Anglo-
Egyptian Oil Field" ... was a Sports
Manager in his Sophomore year ...
hobbies: scenic, reading books on
economics and business ... believes
that the Student Council is ineffective.



It ...
sets the
theistic
right of
one so
... state
happy
and the
ing in



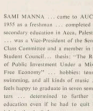
to
gr...
in the
north-
south
man?"
golf,
e four
red to
d diff-
or has
prob-
ed in



*Diet in
evolutionary
life,
economics.*



SALIMA SHAM ... Jordanian came to AUC in 1955 from C.M.S. Amman ... thesis: "Helping Problem Children: Help Themselves Through Play Therapy" ... hobbies: swimming and shooting ... was a member of Tearing Club and in her Senior 3 the Social Manager of the Arab-Café Club ... feels happy for obtaining B.A. degree, which will pave the way for her contribution in raising the standard of the Jordanian women.



KAMI MANNA ... came to AU in 1955 as a freshman ... completed secondary education in Azad, Pakistan ... was a Vice-President of the Student Class Committee and a member in Student Council ... thesis: "The Role of Public Investment Under a Mixed Free Economy" ... hobbies: collecting, swimming, and all kinds of music ... feels happy to graduate in seven semesters ... determined to further education even if he had to quit



ehonest ...
from Tripoli
utilization of
al Reference
umber in the
the Science
ding science
as an assis



... Egyptian
coming from
ed ... was
Stinkys Club
ice President
rice year ...
pears's Plays"
n, and acting
an M.A. in