

Group 3

Ebram Thabet - Omar Elwaliely - Mohamed Alashkar - Andrew Aziz

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.

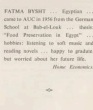
When the user runs the OCR code, it matches the extracted names with the correct face encodings in some images (you can find these images below) from the first book in the AUC digital library, titled 'Al Hawdaj of the Campus Caravan 1958'. This process correctly associates names with their corresponding 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 TARAFI ... Palestinian ... a graduate of the American College for Girls ... came to AUC in 1995 ... these: "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.



FATMA HYSHT ... Egyptian ...
came to AUC in 1956 from the German
School at Bab-el-Lawk ... teaches
"Food Preservation in Egypt" ...
hobbies: listening to soft music and
reading novels ... happy to graduate
but worried about her future life.
Rose E. Korman



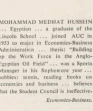
100



95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914



MAHMOUD RAHLOUL... Bachelor
... came to AUC in 1955 from the
Preparatory Section in AUB ... 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 gives the
way for the student to work indepen-
dently and feel responsible.

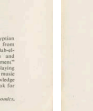


MOHAMMAD MEDHAT HUSSEIN
... Egyptian ... a graduate of the
Lincoln School ... joined AUC in
1953 to major in Economics-Business
Administration ... thesis: "Building
up the Work Force in the Anglo-
Egyptian Oil Field" ... was a Sports
Manager in his Sophomore year ...
hobbies: tennis, reading books on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

Economics-Business



all ...
with the
theatre;
right of
one so
... So
happy
at the
ing in
nature,



typical
from
sub-
n and
ness."
playing
music
wedge
look for
ments.



ne to
get ...
in the
winter-
months
most?"
golf,
or four
red in
differ-
or has
prob-
and in
coming

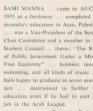


1956
Diet in
reins
e life,
omics.



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

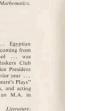
Nawal



SAMI MANNA ... came to AUG 1935 as a freshman ... completed secondary education in Aziz, Palestine ... was a Vice-President of the New Class Committee and a member in Student Council ... thesis: "The Role of Public Investment Under a Mixed Economy" ... hobbies: tennis, swimming, and all kinds of music ... feels happy to graduate in seven semesters ... determined to further education even if he had to quit job in the Arab League.



ehonist ...
from Tripoli
utilization of
al Reference
number in the
the Science
ding science
g to music
and an African



... Egyptian
coming from
eel ... was
drinkers Club
vice President
first year ...
pease's Plays"
n, and acting
an M.A. in