Querying for ACM Articles

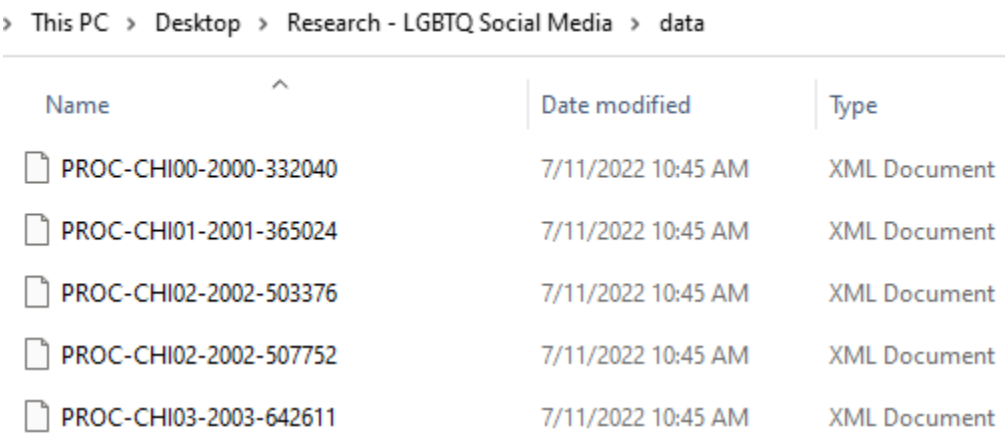
by Forum Modi

This document serves as a guide on how to use *XML_scraping.py* and *queries.py* when trying to obtain an excel sheet of articles using keywords, conference venues, and year ranges in the ACM database. These scripts will only work with ACM's XML files which can be obtained through Craig Rodkin (rodkin@hq.acm.org), the current ACM Publications Operations Manager.

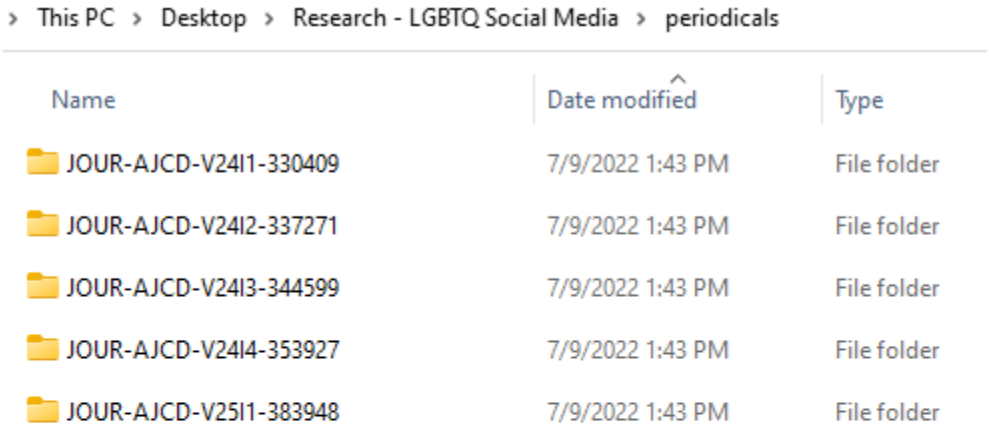
XML_scraping.py: Takes a folder of ACM's XML files and inserts the articles into a database

SKIP IF YOU ALREADY HAVE A DATABASE FILE

1. You will need to start with this script: Enter *python .\XML_scraping.py*
2. Enter the folder name with your XML files
3. Enter 0 if the files are proceedings or Enter 1 if the files are periodicals. (MAKE SURE TO ENTER THIS CORRECTLY!!)
4. You will be prompted to enter whether your XML files are in nested folders
    a. Open up the folder with your XML files
    b. If they look like this (not in nested folders), Enter 0



    c. If they look like this (nested folder), Enter 1



5. Press Enter to fill database
    a. It may take some time for the database to fill based upon the amount of files you have. I would recommend letting this run and working on something else for a while.
    b. Should output a .db file with the same name of folder that will be used with *queries.py*

```
> python .\XML_scraping.py
<---Hello, welcome to the ACM database XML parsing script!--->

<---Please enter the folder name that includes your XML files--->
periodicals

<---Do you have proceedings or periodicals?--->
Selection Menu
0 - proceedings
1 - periodicals
1

<---Are your XML files in nested folders?--->
Selection Menu
0 - No(default)
1 - Yes
1
Press Enter to Start!
```

queries.py: Takes database file (named after folder) filled with ACM articles and outputs an excel file of article data with many user options such as searching for specific conferences, year ranges, and keywords

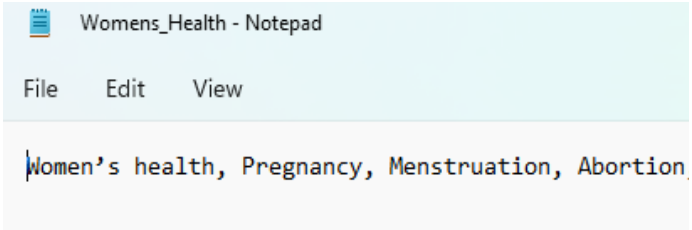
START HERE IF YOU ALREADY HAVE DATABASE FILE

1. Enter *python .\queries.py*
2. Enter database name (same as folder name in XML_Scraping.py) <u>without the .db ending</u>
    a. <u>Make sure it is in the same directory as queries.py</u>
3. Enter the name of what you want the excel file to be called <u>without the .xlsx ending</u>

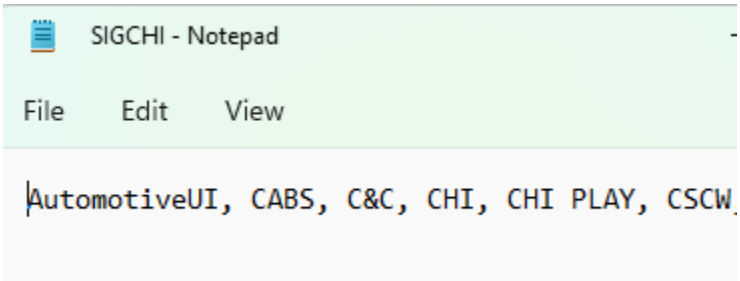
```
> python .\queries.py
<---Welcome to the database parsing script!--->

What is the database name (do not include .db)?
periodicals
Trying to enter database...
...Connected to database!

What would you like to call the excel file?
periodicals_example
```

4. Enter a numeric menu option

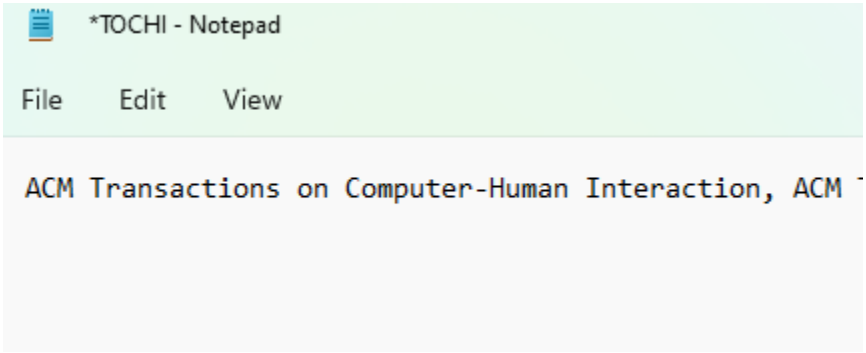
a. **Enter 1 - Keyword Search**: Search for articles with keywords in the abstract, title, full text, and keyword list
    i. Enter the name of a text file <u>without the .txt ending</u> with keywords separated by commas. Keep in mind, searches include plurals



b. **Enter 2 - Year Range:** Enter starting and ending year range
c. **Enter 3 - Conference Search**: Search for articles from specific conference venues
    i. Enter 0 for proceedings, Enter 1 for periodicals
        1. *Proceedings*: Enter the name of a text file <u>without the .txt ending</u> with **<u>publication venue acronyms</u>** separated by commas.
            a. There is a text file with all SIGCHI conferences: SIGCHI.txt



        2. *Periodicals:* Enter the name of a text file <u>without the .txt ending</u> with **<u>publication venue full names</u>** separated by commas.
            a. There is a text file with TOCHI's full name: TOCHI.txt



d. **Enter 4 - Start Search:** Outputs excel file that meets searching criteria
    i. This may also take awhile depending on how large the database is and how intensive the search is. I would recommend working on something else while it runs.

```
<--Query Menu--->
Please enter the corresponding menu option!
1- Keyword Search
2- Year Range
3- Conference Search
4- Start Search
4
Press Enter to start the search!
Please be patient... starting the search
Creating periodicals_example...
...finished creating periodicals_example
```

5. You can input any combination of the three menu options before starting to search.

```
<---Search by Keywords--->
Please enter a text file with each keyword seperated by a comma.
Ex: hci, binary trees, algorithm

What is the textfile name?
Womens_Health
Keyword Search Set!

<--Query Menu--->
Enter 4 to start search or set another menu option.
1- Keyword Search
2- Year Range
3- Conference Search
4- Start Search
2
<--- Search by Year -->

What is the starting year?
2007
What is the ending year?
2017
Year Range Set!
Keyword Search Set!

<--Query Menu--->
Enter 4 to start search or set another menu option.
1- Keyword Search
2- Year Range
3- Conference Search
4- Start Search
3
<--- Search by Conference -->

Are you searching for proceedings or periodicals?
Enter the corresponding menu option
1 - Proceedings
2- Periodicals
2

Please enter the name of text file with each conference venue seperated by a comma.
Ex: ACM Transactions on Mathematical Software, ACM Transactions on Computer-Human Interaction

What is the textfile name?
TOCHI
Year Range Set!
Keyword Search Set!
Conference Search Set!
```

Excel File Output: There are several fields to the excel file output, most are straight forward outside of the additional columns added when doing a keyword search.

- *Score:* A score from 0 to 5, based on how relevant the search term was
    - Keyword in Title: 2 points
    - Keyword i Full Text: 1.5 points
    - Keyword abstract: 1 point
    - Keyword in keyword list: .5 points
- *Included Keywords*: list of keywords present in article
- *Fields With Keys*: list of present keywords along with where they were found in article
- *Empty Fields*: fields that were not in the articles metadata (may affect ability to find in query)

	article_ids	title	dois	urls	blication_y	cation_ac	ication_v	article_typ	keywords	abstracts	authors	or_institu	author_ids	page_num	page_start	page_ends	full_text	publishers	sponsors	score	ded_keyw	ds_with_k	mpty_fields
0	1620516	A cognitiv	10.1145/1	http://dl.a	2009	Automoti	Proceedin	regular_ar	cognitive	<p>Intuiti	Sandrine I	University	P1680830;	8	35	42	INTRODU(	ACM	N/A	1.5	Conceptic	Conceptic	N/A
1	1969792	Making us	10.1145/1	http://dl.a	2010	Automoti	Proceedin	regular_ar	automotiv	<p>Intera	Dagmar K	University	P2627872;	4	110	116	INTRODU(	ACM	Carnegie I	1.5	Fertility,	Fertility: f	N/A
2	2381438	An angry	10.1145/2	http://dl.a	2011	Automoti	Proceedin	regular_ar	affective	<p>Most e	Myounghe	Sonificati	P3847970;	4	137	142	INTRODU(	ACM	ICT Center	1.5	Conceptic	Conceptic	N/A
3	2381449	Sublimina	10.1145/2	http://dl.a	2011	Automoti	Proceedin	regular_ar	N/A	<p>Follow	A. Riener;	Johannes	P3848008;	2	203	206	further st	ACM	ICT Center	1.5	Conceptic	Conceptic	N/A
4	2390291	Developm	10.1145/2	http://dl.a	2012	Automoti	Proceedin	regular_ar	automotiv	<p>Desigr	Hao Tan;	Body (Hur	P3871395;	8	201	208	INTRODU(	ACM	N/A	1.5	Conceptic	Conceptic	N/A
5	2390297	Physical a	10.1145/2	http://dl.a	2012	Automoti	Proceedin	regular_ar	Car2X com	<p>Millior	Monika M	Saarland U	P3871414;	8	249	256	INTRODU(	ACM	N/A	1.5	Birth,	Birth: full_	N/A
6	2516558	ADAS HMI	10.1145/2	http://dl.a	2013	Automoti	Proceedin	regular_ar	ADAS; HM	<p>We pr	Sabine Lar	Renault, C	P4296040;	8	74	81	INTRODU(	ACM	Eindhover	1.5	Conceptic	Conceptic	N/A
7	2516542	Standard	10.1145/2	http://dl.a	2013	Automoti	Proceedin	regular_ar	SAE J2944	<p>This pa	Paul Gree	University	P4296090;	8	184	191	measure/	ACM	Eindhover	1.5	Conceptic	Conceptic	N/A
8	2667341	Informati	10.1145/2	http://dl.a	2014	Automoti	Proceedin	tutorial	Adaptive	<p>Adapti	Andreas H	Center for	P4728919;	8	1	8	INTRODU(	ACM	N/A	1.5	Conceptic	Conceptic	N/A