NLP Project – First Cut

# Overview

The python module is “scholarly 0.2” is used to create the following tables. However, some modifications were required to the original script (the last update was on 2015-08-05) to collect the correct data. Here’s the link for the module: <https://pypi.python.org/pypi/scholarly/0.2>.

MySQL is used to store the database. Link: <https://dev.mysql.com/doc/connector-python/en/>

# Database Schema:

## Table 1: Author Detail (Author\_detail)

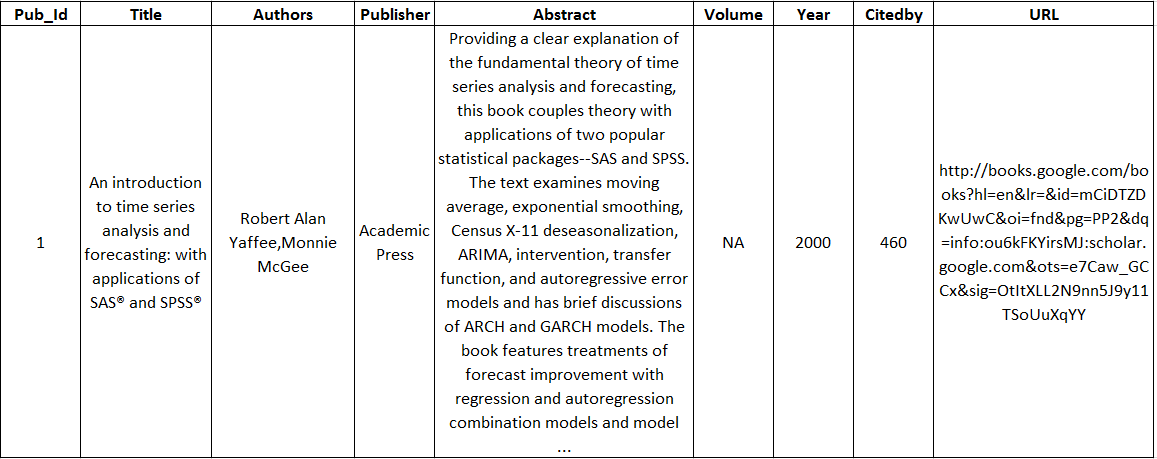
Primary key: ID

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Name** | **Affiliation** | **Email** | **Interests** | **Citedby** |
| 967CPlsAAAAJ | Monnie McGee | Southern Methodist University | @smu.edu | Data Analysis,  Flow Cytometry Data, Gene Expression Microarray Data, DNA sequence data | 1249 |
| … | … | … | … | … | … |

Code details:

Create an author\_list = [] – includes all the authors whose profiles needs to be scraped. For example, scrape the following url for table 1: <https://scholar.google.com/citations?view_op=search_authors&mauthors=monnie+mcgee&hl=en&oi=ao>

## Table 2: Publication details for a given author (Publishing\_Detail)



### Code details:

Scrape the following url for table 2:

<https://scholar.google.com/citations?user=fOZJofkAAAAJ&hl=en>

Extracts all the publications made by the author and reported on google scholar.

## Points

1. The two tables are not linked together yet. I’ve been having some trouble setting the keys for the two tables in python. Hopefully, this will be solved soon.
2. I have also written a script to extract similar information in table2 for articles that cite publications in table 2. For example, if a publication has been 120 times, we can create a new table which include all those 120 other publications.
3. I am not sure currently on how to categorize these articles into their main areas of research like machine learning, statistics, operational research etc. Any thoughts?

Any suggestions?