Objective:

The objective is to extract information from the site.

There are following pieces of information that we want to extract.

a. Author details
b. Author coverage
c. Article coverage

Site:

http://seekingalpha.com/leader-board/opinion-leaders

Detailed task:

The task is to extract information from the above-mentioned site in following forms

Author Information

Get author information in following form:

`blogger_id` varchar(50) NOT NULL, - assign UUID for first time
`name` varchar(30) NOT NULL, - assign name
`short_name` varchar(25) NOT NULL, - short name based on URL
`style` varchar(15) DEFAULT NULL, - "For ex. Investing Ideas"
`category` varchar(25) DEFAULT NULL, - For ex. Long Ideas
`individual` tinyint(1) DEFAULT NULL, - Y/N
`firm_name` varchar(100) DEFAULT NULL, - If it's a firm than use firm name
`description` varchar(255) NOT NULL, - description of an individual
`since` varchar(4) NOT NULL, - publishing since
`picture` longblob, - picture in base64
`followers` int - # of followers

Get coverage information in following form:

`blogger_id` varchar(50) NOT NULL, - take it from above
`ticker` varchar(15) NOT NULL, - tickers the blogger covers

Get articles:

`blogger_id` varchar(50) NOT NULL, - take from above
`article_url` varchar(255) NOT NULL, - article url
`story_id` varchar(50) NOT NULL, - ignore for now