**Crawler**

**Objective:**

To implement a crawler to crawl web pages from the internet.

**Source Code:** [**crawler.py**](https://drive.google.com/open?id=0B3n4oQctrRw6aGo4LTExNDFhMUE)

**Install the libraries:**

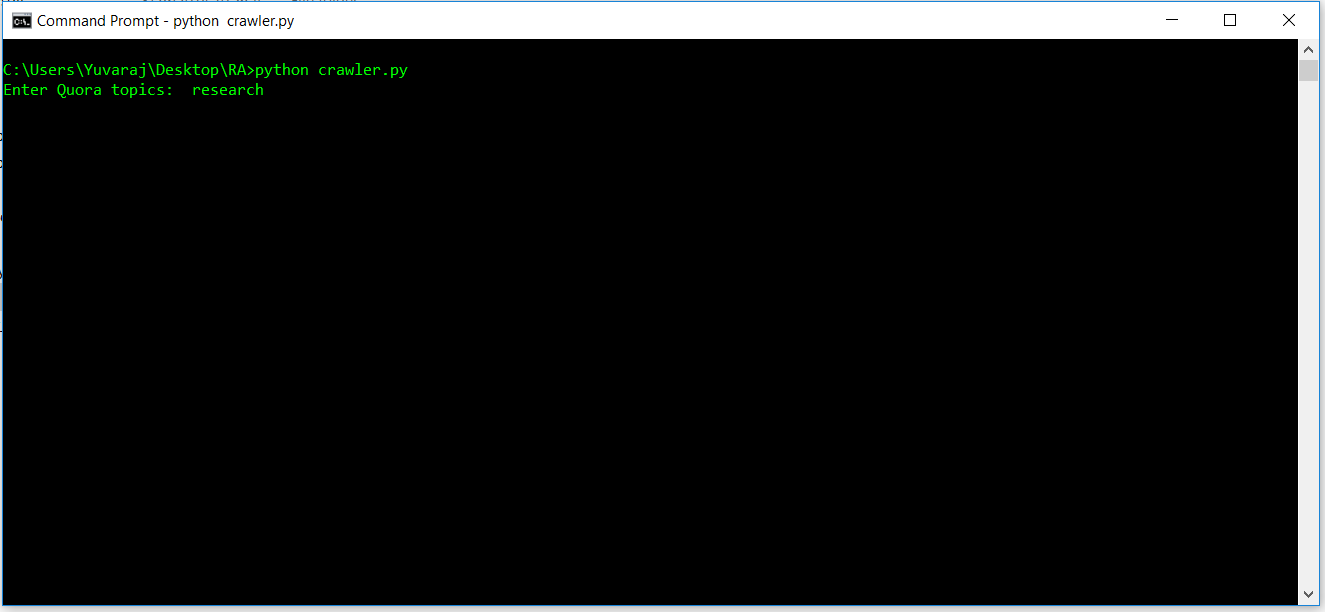
pip install beautifulsoup4

pip install requests

**Running the program:**

python crawler.py

Now enter a quora topic as input



Now the python script runs and crawls the web pages related to that topic **research** from quora. And then generated a output file with the questions and the top answers for each question along with the upvotes.

**Description:**

1. This program gets user input a quora topic as user input and crawls [quora.com](https://www.quora.com/) .
2. If the specified topic exists it crawls the questions related to the topic from the topics page.
3. Then it retrieves the hyperlinks of those questions and crawls the answers to those questions with the number of upvotes for the answers.
4. Finally it writes all the crawled data into a file.

If you search for any topic in quora the URL is of the following pattern below :

Format: <https://www.quora.com/topic/><topic name>

For eg: Machine Learning topics page has the URL below.

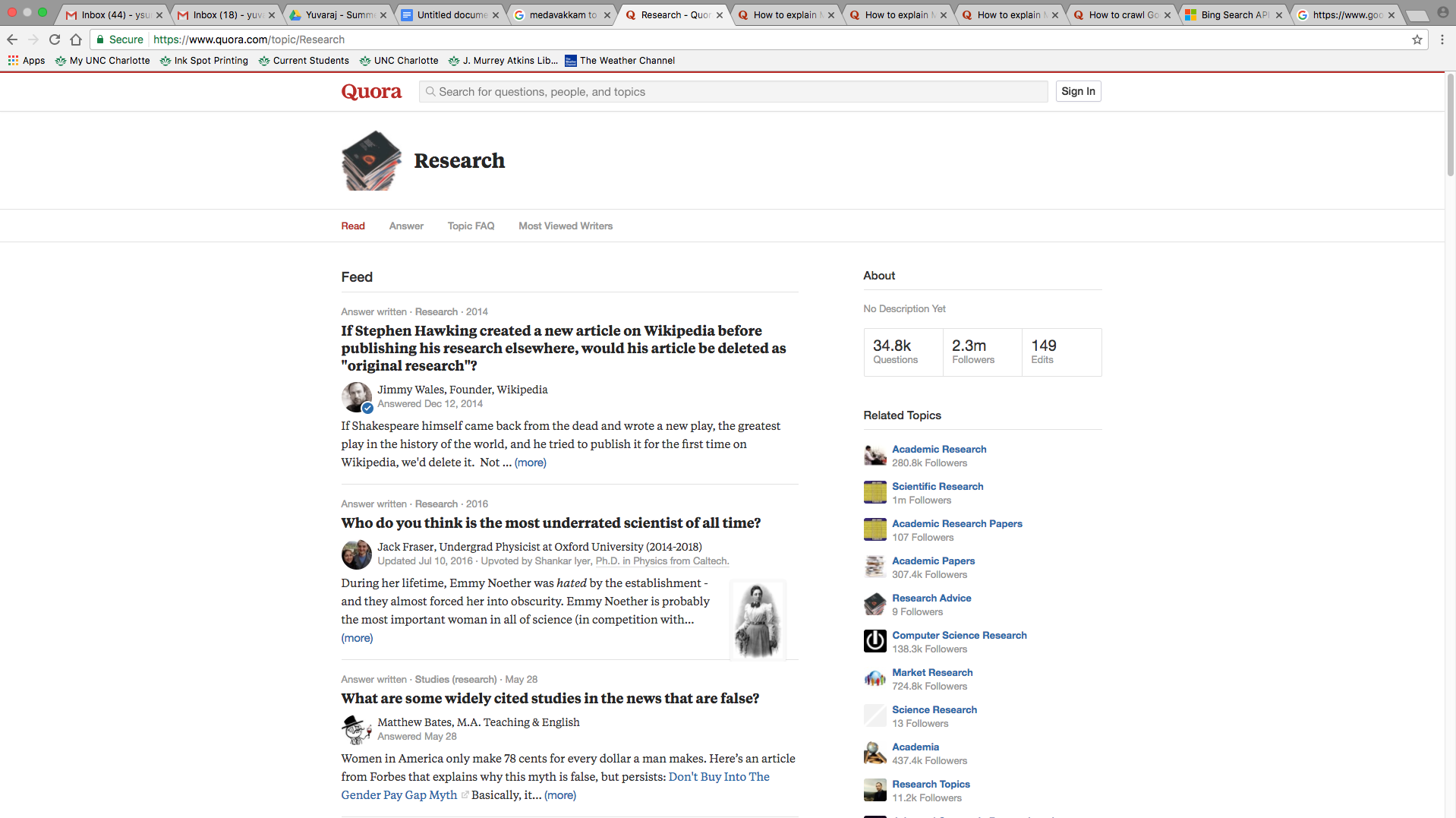
https://www.quora.com/topic/Research

So we form the URL for topics using the below lines of code. And send request using the requests package in python. This retrieves the entire source code of the page

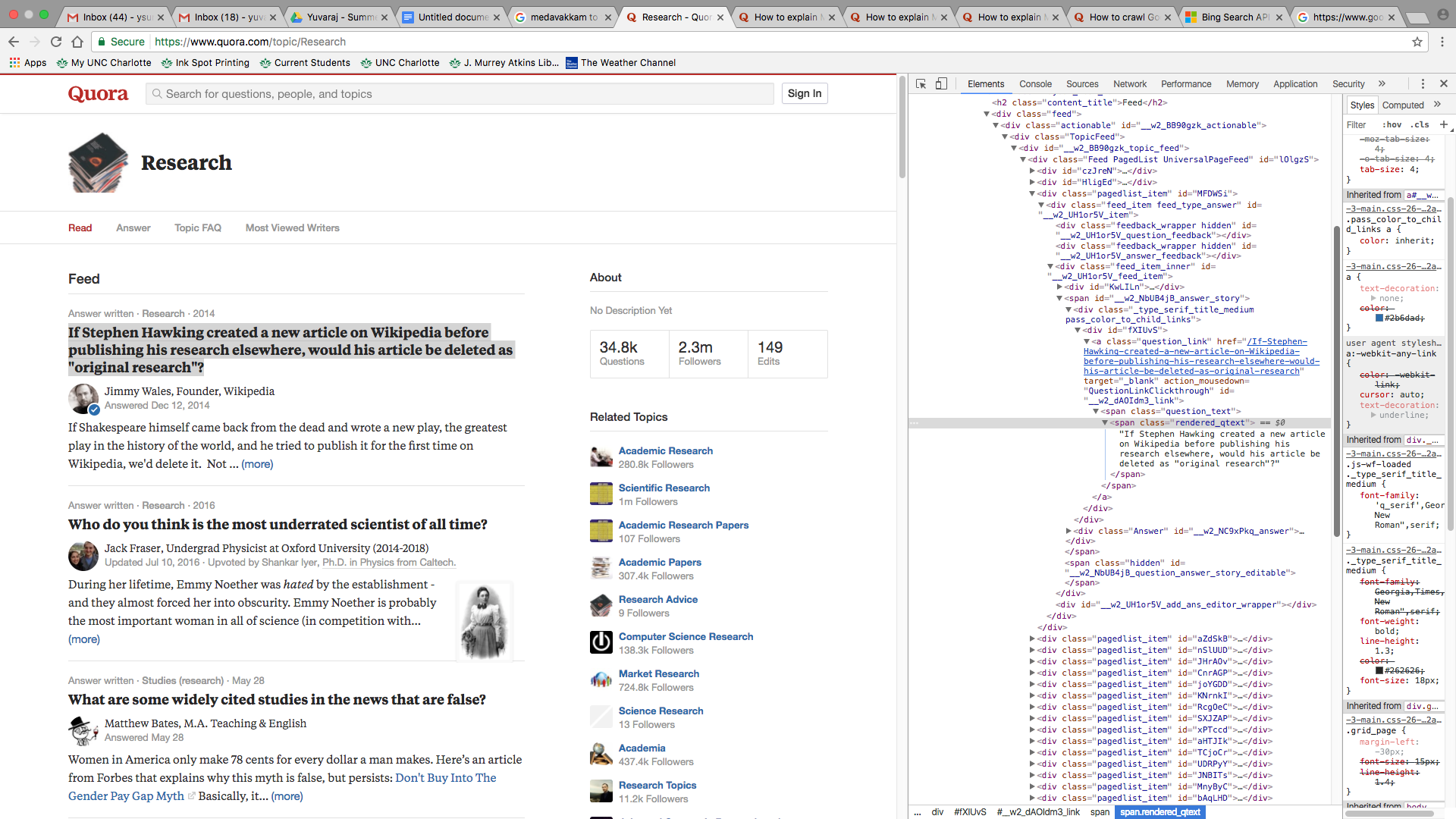
url='https://www.quora.com/topic/'+topic;  
 source\_code=requests.get(url)

Now we use beautifulsoup to parse the html document and then extract the data need from the web page.

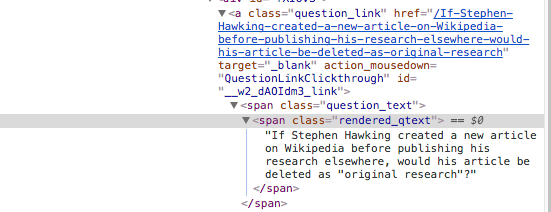
Go into any of the pages in topic pages in quora. Right Click on any question and select inspect



Topics page



Now on the right in the chrome console you could see that the links for each question is inside the **a** tag under the class **“question\_link”** and the question text is available under the span tag under the class **“rendered\_qtext”**



Now the lines below in python script is used to extract the information related to questions and then the question URL’s

plain\_text=source\_code.text  
 soup=BeautifulSoup(plain\_text,'html.parser')  
 for question in soup.find\_all('a',{'class':'question\_link'}):

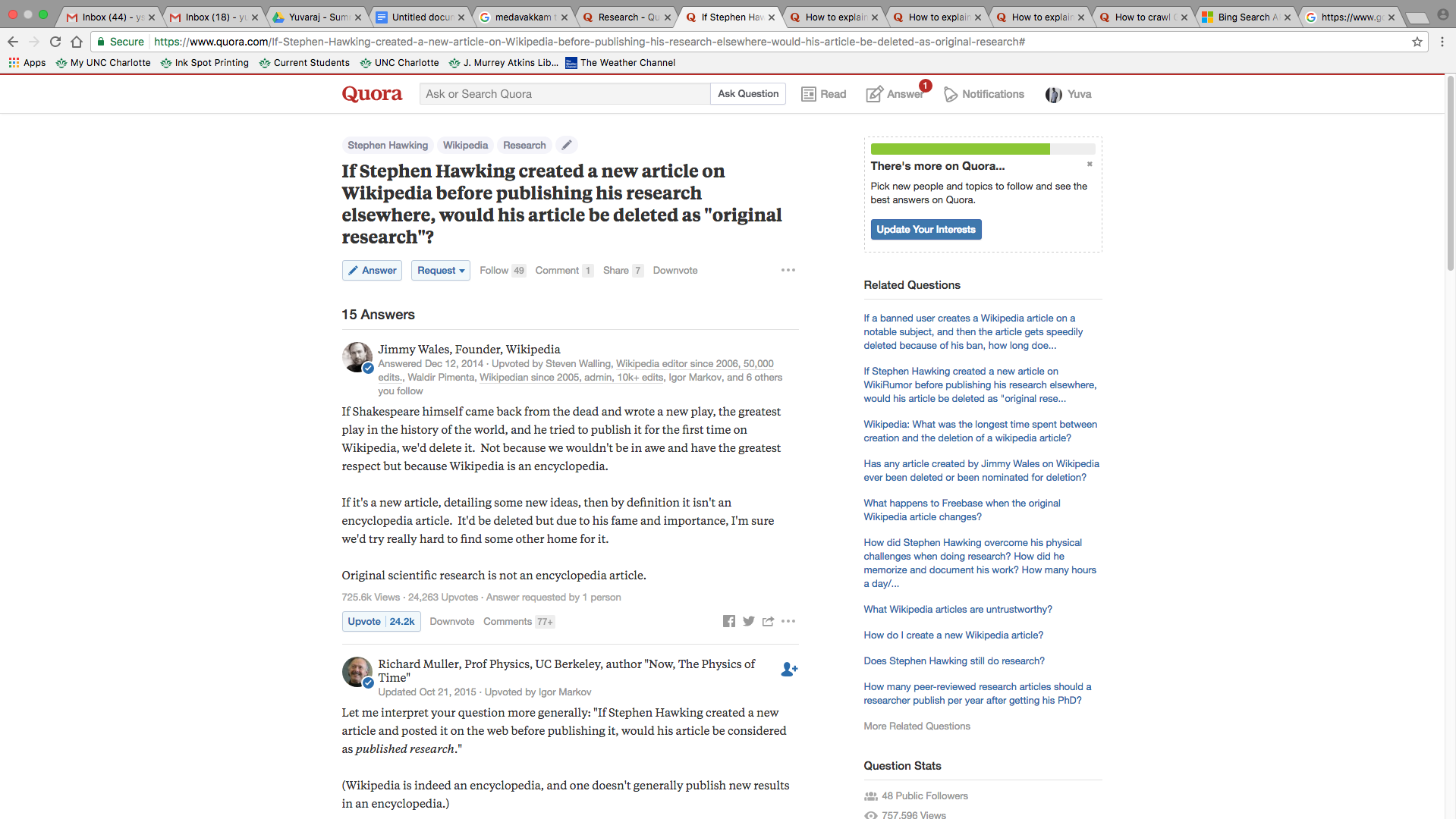
Now I have the hyperlinks of the questions and the questions text. Now I navigate into each question page using the hyper links and then retrieve the answers for those questions.

Now we form the URL of each question and then use requests to download the source code of the questions page. Then we extract the answers for each source code from the questions page.

url='https://www.quora.com'+question.get('href')  
print(url)  
source\_code=requests.get(url)  
plain\_text=source\_code.text  
ansSoup=BeautifulSoup(plain\_text,'html.parser')

for answer in ansSoup.find\_all

('div',{'class':'ExpandedQTextExpandedAnswer'}):



**Questions Page**

**Sample Output File:** [**research.txt**](https://drive.google.com/open?id=0B3n4oQctrRw6X09oRXJyYjBlVWs)

This python script can be used to crawl data related to any topic in quora.