

Steps to create dataset for Programming Task 07:

Steps for crawling Wikipedia to get documents and images:

Step 1: Open the “**download.py**” python file. Replace the **topic** variable with the topic of your choice.

Step 2: Pass the text parameters in Json format to the get request to retrieve the text from the introductory page of the selected topic. (This has already been done)

Step 3: The introductory text is downloaded in **topic.txt** file in the current directory from which “**download.py**” was called.

Step 4: Pass the image parameters in Json format to the get request to retrieve the images from the page of the selected topic. (This has already been done)

Step 5: The images are downloaded in the current directory from which “**download.py**” was called.

Please Note:

The “**download.py**” file retrieves a single topic at a time. Since you need to retrieve a minimum of 1000 topics, you need to write a few lines of code to read the topics from **Wikipedia_topics** and gives it as an input to the “**download.py**” file.