Files Downloader Tool

**TABLE OF CONTENTS**

[PREREQUISITES](#h.jdbeyu3w0gu5)

[SOURCE CODE LOCATION](https://docs.google.com/a/synerzip.com/document/d/1rJFjRx8I-yCSbD9Dnmhld-VNsoaWre_fTJMu05TTxdg/edit#heading=h.qlyearuookp)

[CODE ORGANIZATION/ARCHITECTURE](https://docs.google.com/a/synerzip.com/document/d/1rJFjRx8I-yCSbD9Dnmhld-VNsoaWre_fTJMu05TTxdg/edit#heading=h.piuivkvboshl)

[HOW TO SETUP](https://docs.google.com/a/synerzip.com/document/d/1rJFjRx8I-yCSbD9Dnmhld-VNsoaWre_fTJMu05TTxdg/edit#heading=h.hnrfkimhoy9)

[HOW TO RUN](https://docs.google.com/a/synerzip.com/document/d/1rJFjRx8I-yCSbD9Dnmhld-VNsoaWre_fTJMu05TTxdg/edit#heading=h.vbzlltnlq6jq)

[TEST DATA USED](#h.bv3dbt4s5yw6)

[LINK TO REPORT](#h.dlhc03chk1pw)

[NOTES](https://docs.google.com/a/synerzip.com/document/d/1rJFjRx8I-yCSbD9Dnmhld-VNsoaWre_fTJMu05TTxdg/edit#heading=h.623q1iw10tkd)

# **PREREQUISITES**

It runs on Windows,Ubuntu and Mac :

* Python 2.7
* Java 1.7
* Psutil module of python

# **SOURCE CODE LOCATION**

scratch/TETools : [PuneTETools/FileDownloader](https://quickoffice-internal.googlesource.com/html-office/+/scratch/miscTETools/PuneTETools/FileDownloader/#)

# **CODE ORGANIZATION/ARCHITECTURE**

# **HOW TO SETUP**

* Install java 1.7 ,python and psutil module of python.
* Clone the scratch TETools branch.
* Create directory “te-tools/File\_Downloader/builds”
* Open the File\_downloader.py file from te-tools/File\_Downloader.
* Configure the jar parameter in File\_downloader.py :

**-- Parameter Details :**

1. **-download** :To download the File on system it should be true always.
2. **-search\_engine** :To download files from search engine like yahoo,bing and google.
   1. e.g -search\_engine=google i.e default.
3. **-timeFrame** :To download the files for specific time frame like last day,last week,last month,last year.Its an optional parameter.
   1. e.g -timeFrame=PAST ALL i.e default.
4. **-language** :To downlaod the files of specific language.Its an optional parameter.
   1. e.g -language=en .
5. **-location** :To download the files of specific loadcation.Its an optional parameter.
   1. e.g location=in .
6. **-domain** :To download the files from specific domain.Its an optional parameter.
   1. e.g domain=microsoft.com.

* For language and location parameter have to give codes instead of name the code from [link](https://docs.google.com/a/synerzip.com/spreadsheets/d/1RAuXo2BNdeCzbZs7bJqlLmOteCuhJNZRYK0ZuWG8ylI/edit#gid=1649766924).
* Removing duplicate files and rename using hash code md5.py:
  + inputFilePath: path of downloaded files.
  + duplicateFilePath: path of any dir where you want to move duplicate files.

# **HOW TO RUN**

* Open te-tools/File\_Downloader/Keyword\_List.txt.
* Enter List of keyword in that file and save it.
* Run Python File\_downloader.py <File-Type> <File-Count>

1. File-Type: Formate of files have to download like xls,ppt etc
2. File-Count:Number of file to download per keyword.

* File are Downloaded at location: te-tools/File\_Downloader/Build/Download/
* Downloaded File name and url mapping CSV also created in te-tools/File\_Downloader/Build/Download/URLCSV.csv
* python md5.py --inputFilePath <inputFilePath> --duplicateFilePath <duplicateFilePath>

# **TEST DATA USED**

# **LINK TO REPORT**

# **NOTES**

1. Known issue: Sometimes Google may block IPs as repeated automated searches and downloads may get perceived as a DoS attack. In that case specify the search engine as Yahoo or Bing.