The Domain Specific Search Engine is specific towards "Mobile Phone Related Document Searches". It is designed into three basic parts:

1. Corpus collection: The file named web\_scraper.py scrapes the data from the website <https://www.theverge.com/mobile/archives/> and download the web pages. The web pages are then traversed, extracting the necessary text.
2. Creating Inverted Index: This task is handled by the file named create\_inverted\_index.py The corpus is traversed and tf-idf values are stored in the hash tables. Data Structure: The basic data structure used to store the inverted index is hash table of hash tables. Inner hash table reflects the document number and the outer hash table reflects the words After fully forming the data structure, it is saved as a .pickle file for further search queries.
3. Search User is asked to give a search string .pickle file saved above is used to access the inverted index of the corpus Top documents are ranked accordingly

Running time:

* For creating inverted index for 100 documents: 3.1 secs
* For search query: 0.02 secs

Following two files carry out the task for the given domain specific search engine:

* create\_inverted\_index.py:
  + For each file in corpus:
    - Words are tokenized
    - Stop Words are removed
    - Lemmatization
    - Document frequency is updated for each word
  + tf-idf is calculated and stored in the data structure
  + Data Structure is stored in a .pickle file
* ask\_query.py:
  + query is taken as input through command line
  + .pickle file is read and data structure is stored locally
  + words are tokenized
  + stop words are handled
  + inverted index is created just for the query
  + score is calculated for every document
  + scores are sorted and top documents are displayed

Following website is used for the corpus: <https://www.theverge.com/mobile/archives/>

Python dependencies required:

* + Support for python3
  + nltk
  + time
  + math
  + sys
  + string
  + operator

Given all the dependencies being met, the given search engine can be run by following commands in the unix terminal: For creating inverted index: python3 create\_inverted\_index.py

For running the query: python3 ask\_query.py <your query here>