Reveogle

**Search Product Reviews**

**Tools and Technologies** used in this project :   
 **OS** : Linux  
 **Crawler tool**: Scrapy (Python)  
 **Index and Search** : Lucene  
 **Web** : php, apache2.  
  
This assumes that you are on a **Linux** system and have installed **Scrapy, php, Apache2** . You can install Scrapy from here<http://doc.scrapy.org/en/latest/intro/install.html>.

## Directory Structure:

* **Part\_2**
  + **amazon\_review**s
    - **amazon\_reviews**
      * **spiders**
        + reviews.py
        + data.csv
  + SimpleFileIndexer.java
  + Searcher.java
  + revSearcher.java
  + index.php
  + **images**
  + **screenshots**
  + **index**

## How to create :

* Backend stuff:

(All commands are mentioned in “command.sh” file )

0. Open Terminal.

1. Start apache server :- “sudo service apache2 start”
2. Put entire folder in : “/var/www/” and go to “127.0.0.1/Part\_2/index.php” in browser.  
    Scraper script (reviews.py) is inside “amazon\_reviews/amazon\_reviews/spiders”.
3. Run the command “scrapy crawl amazon -o data.csv -t csv” .   
    This will start crawling Amazon website and will extract reviews from the website provided in crawler’s file into data.csv file. This will keep running until terminated by the user. **Note:**
   1. Terminate this after a while and the reviews would have been populated by then in the data.csv file.
   2. The data.csv file which has the reviews scraped from amazon, has commas in various fields of it and this caused some issues with this code. So we have to remove them before going further. To do this:
      1. Open csv file using libre office or Excel.
      2. Select “Search and Replace”
      3. In search type ‘,’ without quotes.
      4. Then in replace type “ “ (space, without quotes ) and select find and replace all.
   3. Now we have sanitised the csv file making it ready to use.
4. “SimpleFileIndexer.java” will Index this csv file with lucene. Run :
   1. javac -classpath .:lucene-core-4.0.0.jar:lucene-analyzers-common-4.0.0.jar:lucene-queryparser-4.0.0.jar SimpleFileIndexer.java
   2. java -classpath .:lucene-core-4.0.0.jar:lucene-analyzers-common-4.0.0.jar:lucene-queryparser-4.0.0.jar SimpleFileIndexer
5. Compile lucene-indexsearcher file - Searcher.java. Run :
   1. javac -classpath .:lucene-core-4.0.0.jar:lucene-analyzers-common-4.0.0.jar:lucene-queryparser-4.0.0.jar Searcher.java
6. Go to browser : “127.0.0.1/Part\_2/index.php” .

## Usage :

* Front end stuff.

1. Enter the product you want to see reviews for in the Search Box and click select.
2. A list of names of products matching the entered search term will be displayed.
3. Select one, and click Submit.
4. Reviews for the selected product will be displayed.

## Future improvements

1. Proper Designer can design a good appealing User Interface.
2. The scraper itself can be modified to remove the issue with commas in the csv.
3. We have currently implemented this to scrape data from Amazon.com only. This can be extended to other websites with minor changes in the crawler script based on which website are we targeting.

-- This is developed by Hardhik for a Freelance project.