Web Mining

Lab Assignment

Name: **Kritika Mishra**
Registration Number: **16BCI0041**

Question: Use scrapy or Beautiful Soup to crawl any one of the E-commerce website of your choice and perform the same.
The following information needs to be extracted from the page:
(Choose any one product: e.g laptop, Smartphone … etc)

1. Product Name
2. Product price
3. Product discount
4. Product image

Code:

```
import scrapy
from scrapy.crawler import CrawlerProcess
from scrapy.utils.log import configure_logging
from scrapy.utils.project import get_project_settings
import os
import shutil
import glob
curfilePath = os.path.abspath(__file__)
curDir = os.path.abspath(os.path.join(curfilePath, os.pardir))
tmpDir = os.path.abspath(os.path.join(curDir,'tmp/'))
# remove old crawling data
try:
    shutil.rmtree(tmpDir)
except:
    pass
# Get the search keyword from the user
print "Enter Search Keyword (product or brand name to search):",
pruduct = raw_input()
# configure logging
configure_logging({'LOG_FORMAT': '%(levelname)s: %(message)s'})
# get the project settings
s=get_project_settings()
# Change the depth limit here
# s['DEPTH_LIMIT'] = 2
process = CrawlerProcess(s)
# Add spiders to crawl
```

```
process.crawl('amazon',product=pruduct)
process.crawl('ebay',product=pruduct)
process.crawl('shopclues',product=pruduct)
process.crawl('olx',product=pruduct)
process.start()
# Add results to results.csv file after crawling is complete
interesting_files = glob.glob(tmpDir+'/*.csv')
header_saved = False
with open('results.csv','wb') as fout:
    for filename in interesting_files:
        if os.path.getsize(filename) > 0:
            with open(filename) as fin:
                header = next(fin)
                if not header_saved:
                    fout.write(header)
                    header_saved = True
                for line in fin:
                    fout.write(line)
print 'Crawling Completed'
```