Webscraping and its applications using Linux

A.Sasank(411564) Y.Sai harsha(411581)

April 5, 2018

1 Introduction

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

2 Steps followed in scraping a website

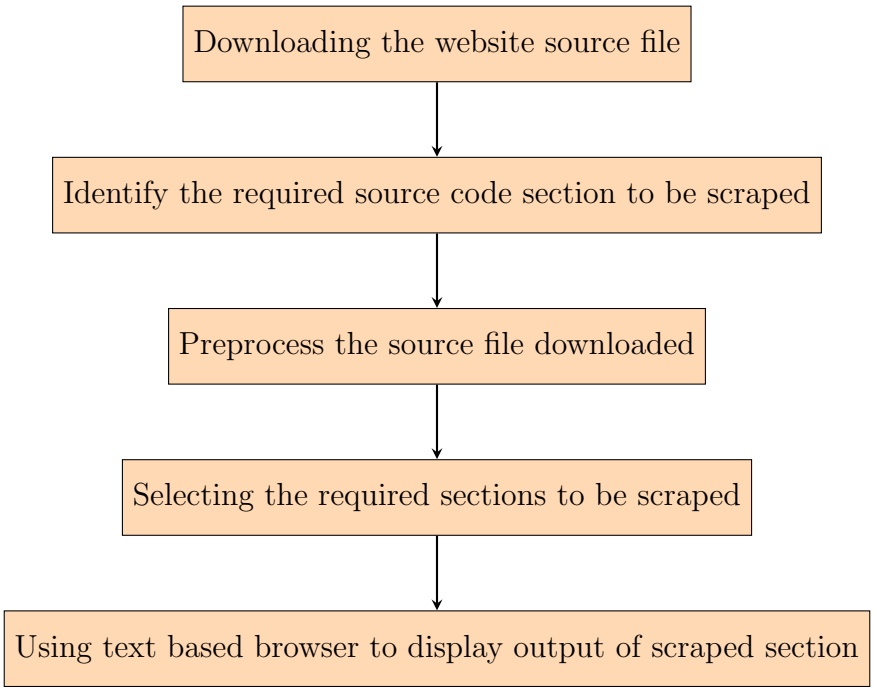

Table 1: Steps and Tools used

S.no	Step	Tools used
1	Downloading the website source file	curl, wget
2	Identify the required source code section to be scraped	inspect tool of browser
3	Preprocess the source file downloaded	hxnormalize
4	Selecting the required sections to be scraped	hxselect
5	Using text based browser to display output of scraped section	lynx, w3m

Tools like sed, awk, etc are also used for transforming the data to required form.

3 Description

In this project we developed following applications using webscraping:

- Stock Market Data extraction
  It has four options.

  1. Search
     Search feature allows to view stock codes of all the companies begining with a particular alphabet. It also allows to view stock code by using full company name.
  2. Display historical stock data
     This feature is used to display details like Opening stock , closing stock , volume , high and low values for a given date. These values are extracted from yahoo finance website.
  3. Generate a csv file
     The historical stock values which are extracted above , are then processed using various tools like sed , awk to store them as csv file, so that they can be used for future analysis purpose.
  4. Shows statistics of stock and plot them.
     This module is used to find various statistics from extracted historical data such as average opening and closing stock, max and minimum closing stock and dates when they occur.

- Criket score extraction
  This module displays the scores and match details from live cricket matches as well as recently played matches. This data is extracted using web scraping from ESPN Cric info website.

- Word dictionary
  This tool is used to show Phonetics, meaning and other usages of the word like verb, adverb, adjective as in a dictionary. This data is scraped from thefreedictionary.com website.

4 Conclusion

Web scraping has many applications such as trending topic extracting and analyzing in social media, e-commerce, finance, marketing, banking and recruitment. In this project, we used webscraping for extracting stock market data. It can also be used to develop useful tools such as cricket score monitor and word dictionary which make the targeted access from a website much simpler.