**Web Scraping using Java**

A Synopsis Submittedin Partial Fulfillment of the Requirements for

**PROJECT SUBMISSION**

By

Bharat Mathpal

ooxWord://word/media/image3.jpeg**Synopsis**

**1. Introduction**

Web scraping is the act of fetching data from a third-party website bydownloading and parsing the HTML code to extract the data you want.It can be done manually, but generally, this term refers to the automatedprocess of downloading the HTML content of a page, parsing/extractingthe data, and saving it into a database for further analysis or use.

Web scraping allows you to acquire non-tabular or poorly structured datafrom websites and convert it into a usable, structured format, such as a.csv file or spreadsheet.

Scraping is about more than just acquiring data: it can also help youarchive data and track changes to data online. If you’ve ever copy andpasted information from a website, you’ve performed the same functionas any web scraper, only on a microscopic, manual scale. Web scrapinguses intelligent automation to retrieve hundreds, millions, or even billionsof data points from the internet’s seemingly endless frontier

It is closely related to the practice of web indexing, which is what searchengines like Google do when mass-analyzing the Web to build theirindices. But contrary to web indexing, which typically parses the entirecontent of a web page to make it searchable, web scraping targets specificinformation on the pages visited.

Web scraping is also increasingly being used by scholars to create data setsfor text mining projects; these might be collections of journal articles ordigitized texts. The practice of data journalism, in particular, relies on theability of investigative journalists to harvest data that is not alwayspresented or published in a form that allows analysis.

Furthermore, the data scrapped can be useful in tons of areas includingfinancial data analysis, social media analysis, brand monitoring, andcompetition and customer analysis.