# Read Me

This is a draft of a paper, periodically updated over time. It might not conform purely to the norms of an academic paper in the beginning but this work will be polished and edited numerous times during the summer 2021 semester and more if required and as such, does not reflect the final product.

# Abstract

The process of collecting data from the internet is known as web scraping. In a time when the internet is rich with data and there isn't enough time to look through it all, web scraping has become much more necessary and feasible to utilize in many applications. Beautiful Soup and the Python library requests are effective tools for the purpose. This project intends to extract text content from websites using web scraping to make it easier to detect a pattern in the information store or find relevant queries. The data is then stored into a database. As, User can get necessary information more easily and query more effectively. This project is focused on scraping Bengali web pages and attempting to determine the most efficient method of doing so.

# Introduction

The amount of data kept online has only grown over time. The use of the internet and cloud servers has increased dramatically in recent years. As a result, searching for valuable data from available sources has become extremely challenging. Data is an essential part of any research, either it can be academic, marketing or scientific (SysNucleus, n.d.). People might want to collect and analyse data from multiple websites. The different websites which belongs to the specific category displays information in different formats. Even with a single website you may not be able to see all the data at once. The data may be spanned across multiple pages under various sections. Most websites do not allow to save a copy of the data, displayed in their web sites to your local storage (Penman et al., 2009). The only option is to manually copy and paste the data shown by the website to a local file in your computer. This is a very tedious job which can take lot of time. Web Scraping is the technique which people can extract data from multiple websites to a single spreadsheet or database so that it becomes easy to analyse or even visualize the data. (A Comparative Study on Web Scraping SCM de S Sirisuriya)

The aim of this study is to offers a review on web scraping techniques and software which can be used to extract data from web sites focusing the most on bangla websites and text contents.

# Overview of Web Scraping

Web Scraping is a great technique of extracting unstructured data from the websites and transforming that data into structured data that can be stored and analysed in a database. Web Scraping is also known as web data extraction, web data scraping, web harvesting or screen scraping. Web scraping is a form of data mining. The overall goal of the web scraping process is to extract information from a websites and transform it into an understandable structure like spreadsheets, database or a comma-separated values (CSV) file