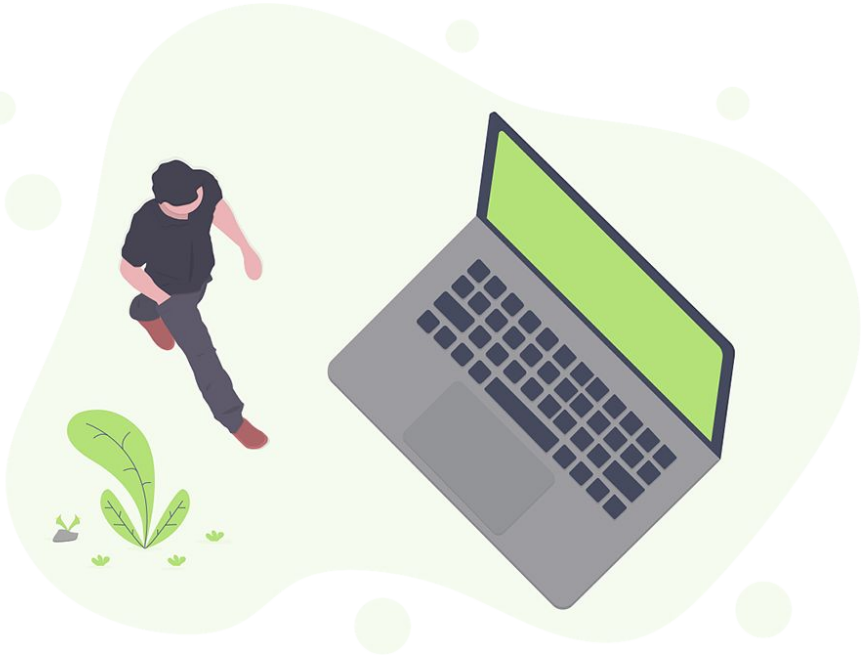
Wiki Data Scraper



Problem Statement

This is an application to build datasets from Wikipedia pages. Most common problem for data scientists

is collecting data, building datasets. Since Wikipedia has a collaborative data from the public it would

be a lot of help for the data scientists and programmers if they had a way to collect data from here. This

is where our application comes in, it creates an unstructured database from the page and gives it to the

programmer for further modifications.

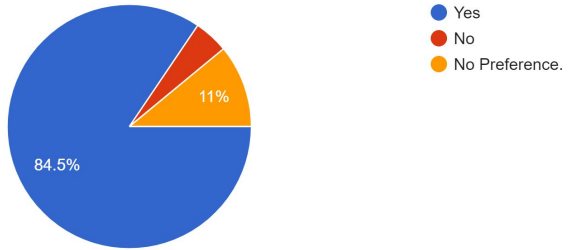
Literature Survey

Based on the survey conducted by us, this is the result.

Literature Survey

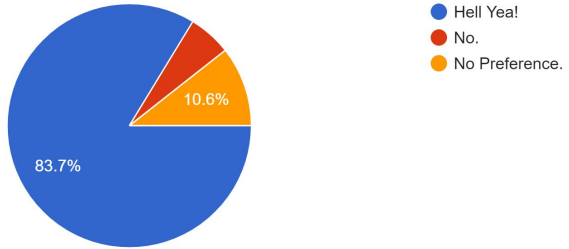
As a beginner / noobie would you prefer to work on a simple, small dataset as opposed to a large humongous, complex dataset?

264 responses



As a beginner / noobie would you prefer a tool to automate your collection of data points?

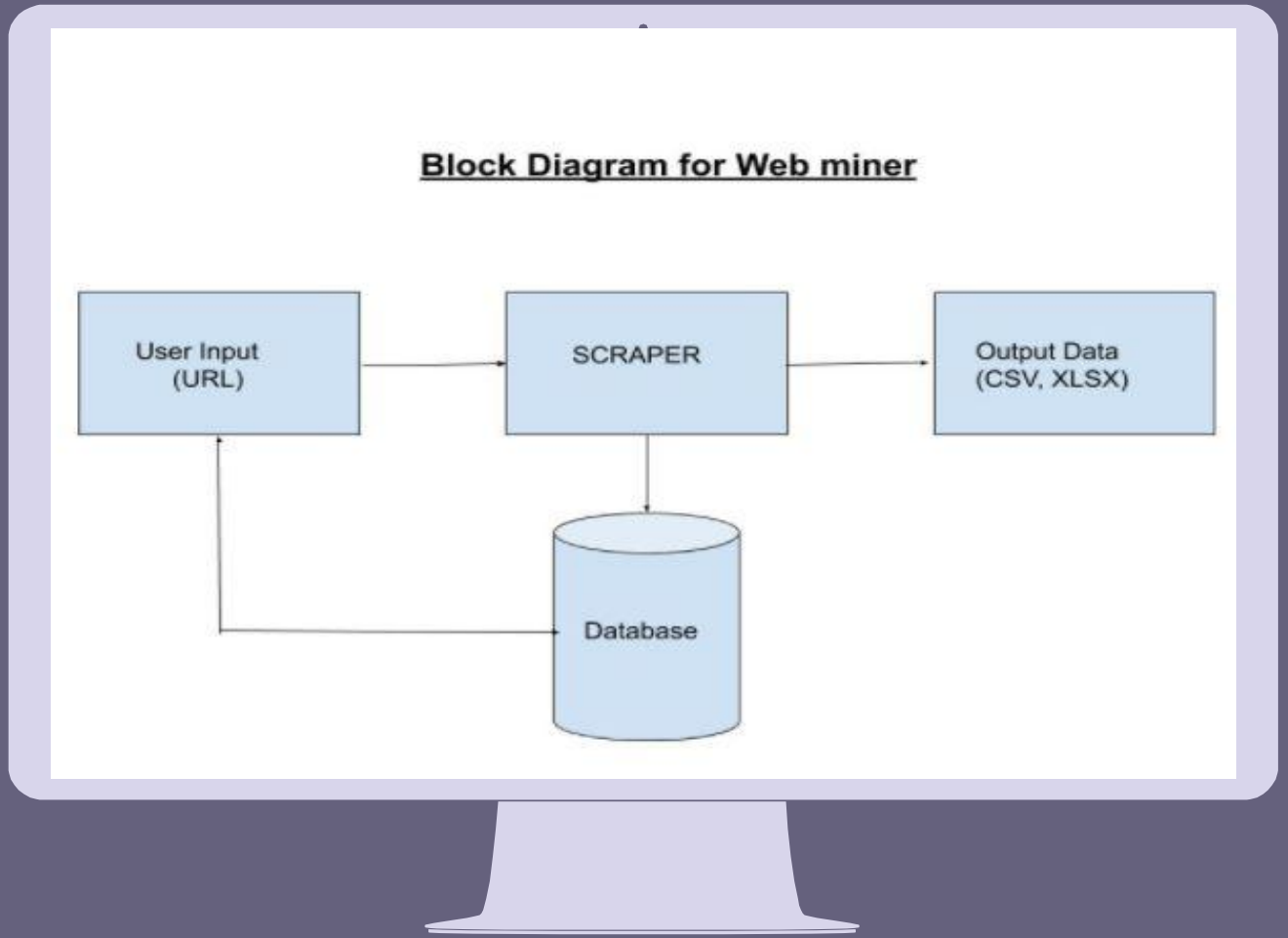
264 responses



Methodology

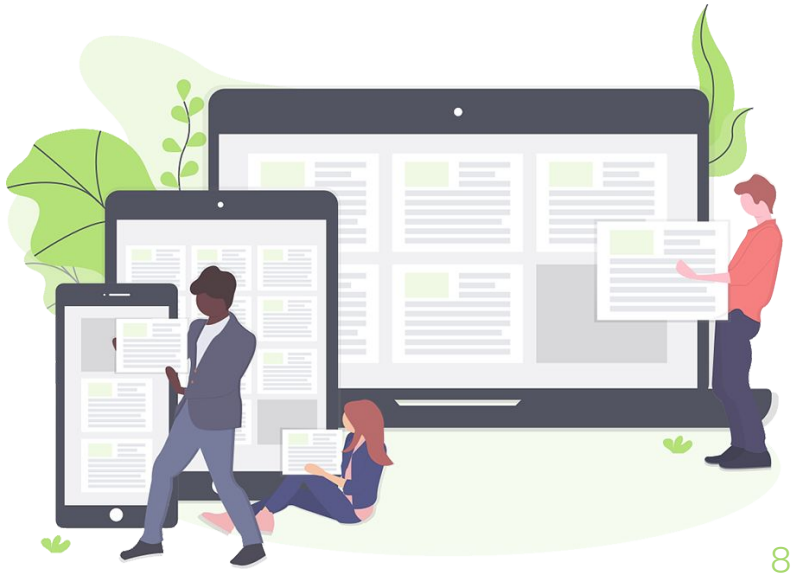
- There will be a database to collect the input queries provided by the user. (URL)
- If the URL contains any form of data collection, that will be scraped and converted to a csv / xlsx file which can be used by the developer to work with.
- The generated output along with the URL will also be stored in the database.

Flow Chart


Block Diagram for Web miner
User Input
(URL)
SCRAPER
Output Data
(CSV, XLSX)
Database

Project Objectives

- To facilitate **noobie data workers** in data collection.
- To **decrease collection time** of data points.
- To **help students** create new data sets in a hassle free manner.



Software Requirement Specification

- HTML, CSS, Bootstrap (Front End).
- Python, PHP (Back End).
- Web Browser: Microsoft Edge, Mozilla, Google Chrome. (Any Modern Browser)
- SQLite(Database).
- Operating System: Windows 10, Linux, Mac.

References

- https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac

- https://towardsdatascience.com/scrape-websites-using-python-in-5-minutes-931cd9f44443