# Foot Scraper

|  |  |
| --- | --- |
| Devs | Tyrone Lagore & Johnny Simmonds |
| Last Revision | February 8, 2019 |

## Introduction

Football (soccer) statistics are available from many websites, but there is no one place where someone can go to get all the aggregated statistics that they are interested in. Foot scraper is one component of the pipeline that will aggregate the data from these many other sources and combine them in a way that can be analyzed for further usage.

## Technologies

### Language

We have decided to use python as our language of choice because of the many modern website scraping utilities that are available for this language. Libraries such as beautifulsoup4, scrapy, and requests offer a way to pull and clean data efficiently. Furthermore, data manipulation libraries like pandas and numpy will allow for data analysis at later stages. It is not necessary that we use the same language for the scraper and for the data analysis, but keeping the language the same will help alleviate any learning curves that are encountered.

### Data Storage

Data stores are being investigated, but it is likely that some form of NoSQL data storage will be used for this project. Because of the naturally unstructured nature of scraped data, it will alleviate the extra effort that would be needed in order to clean the data into neatly formatted columns. Instead we can organize our scraped data into collections that are related, but may not have the same data, data types, or data names. It will still require some level of cleaning to ensure that data that IS the same data, but otherwise renamed.

## Design

It is important that we create a framework that can easily add new sources of data that have differing structures. This will allow us to add new sites with new data for aggregation without forcing a new component to be added for every single site that we add to our data collectors.