Scraping tool

## Abstract

These days market is hungry for informations. So, main idea of this tool is to fetch and structure unstructured data from different sources. Web pages and API-s are considered as a common sources. Data on the internet is [unstructured](https://en.wikipedia.org/wiki/Unstructured_data) and this tool will structure it and deliver it in one of selected format : .xlsx, .csv, .json. After this process ends it’ll be easy to visualize data.

## Market and customers

Our main targets will be :

* Web shop owners
* Managers
* Journalists
* Data scientists
* Different kind of agencies
* Marketers
* Developers
* Everyone who wants to compare some data from different sources
* Lead generation

## Problems

* Different templates for websites (paging, find tags of interest...)
* Protections of web site (number of requests, captcha, logged in, location restriction...)
* Complex user interaction (lot of options, workflow)
* Multi Concurrency (architecture which will satisfy lot of requests)
* Lot of bandwidth (moder pages are really heavy)
* It is niche
* Low regulation for [open](https://en.wikipedia.org/wiki/Open_data)/closed data
* Solution may be complicated so user needs training

## 

## Concurrents

[Octoparse](http://www.octoparse.com/) - Main problem of this tools is restriction for Microsoft operating systems.

[Webscraper](http://webscraper.io/) - Very simple solution.

[Mozenda](http://www.mozenda.com/) - They support scraping data from various type of documents(e.g. .pdf). And support custom solutions.

[Data-miner](https://data-miner.io/pricing) - Possible to hire a professional.

[Fminer](http://fminer.com) - Desktop solution that needs to be installed.

[Import.io](http://import.io) - Only solution which doesn’t need an installation. Very pure design and little buggy.

## Scalability

Business:

This product needs to meet requirements for professional users as well as newbie users. Newbie users needs to have out of the box solutions and possibility to solve easy tasks fast.

Intuitive design and hidden options for newbie users will ensure viral spreading. Scalability for advanced users won’t be so easy because support will be required for consuming tasks.

Technology:

Application will be stateless so horizontal scalability will be satisfied.