AIMultiple Coding Challenge

Congratulations, you made it through the initial interviews. Now it's time to show off your coding skills. The following task should be self-explanatory, but if you have any questions don't hesitate to contact me at cem@aimultiple.com

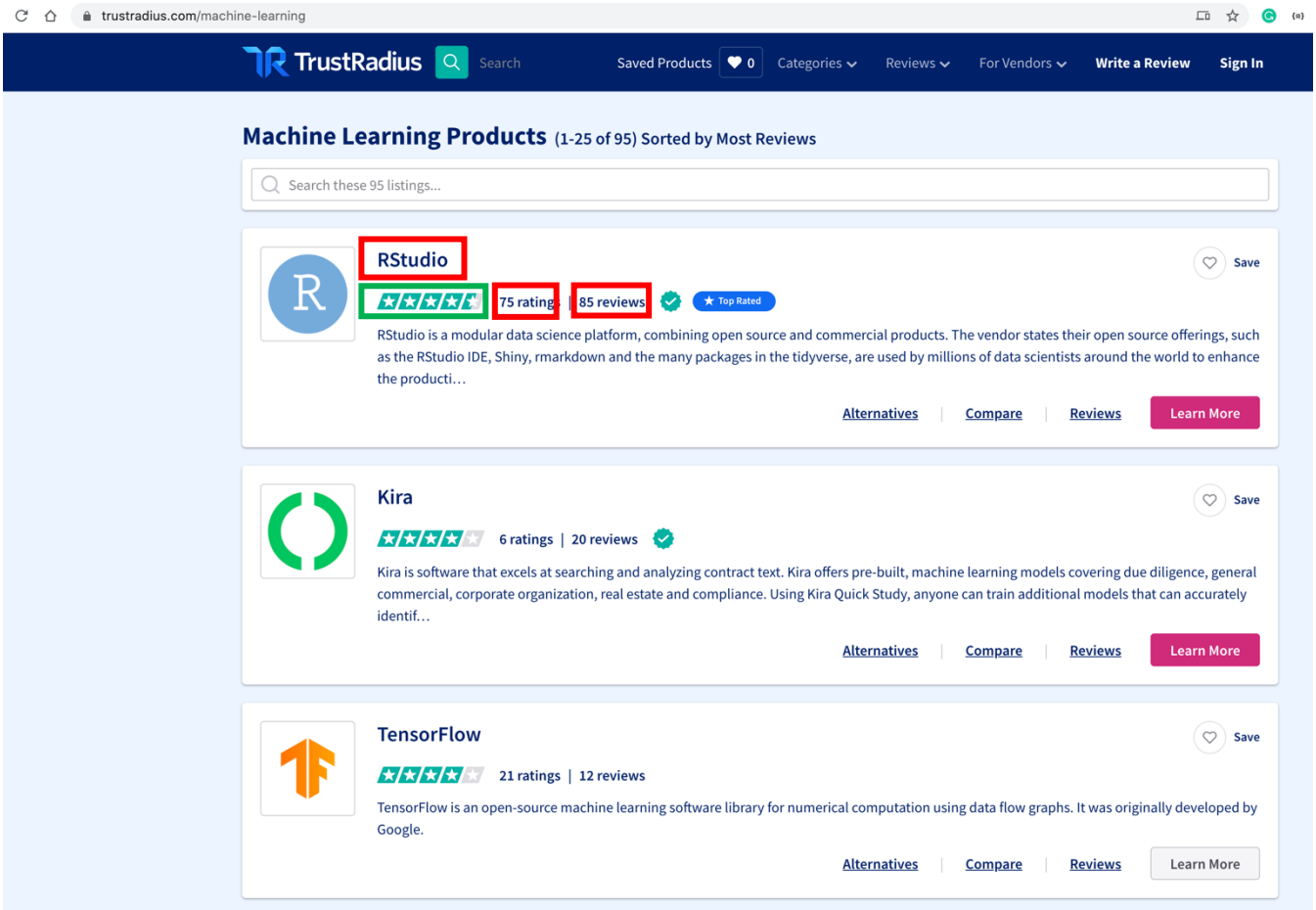
Your Task: Web Crawling

We crawl certain websites periodically to aggregate data. Then, we analyze them and then present them to our users in the most suitable way. Therefore, we would like you to crawl a website for us and create a dataset as an output.

Please push the source code to a Github repository for review.

Spec

For this assignment we would like you to crawl name, number of ratings and reviews of each product categorized under machine learning solutions on trustradius.com. The url you will work on is: https://www.trustradius.com/machine-learning. Once you land on the url you will see a list of products by scrolling a little bit down. Your task is to collect relevant information about each product and put them in an output file in the desired format. For example, you should crawl the information marked with red lines below and put them in a csv file:



Output will include

- a csv file with the following columns: productName, ratings, reviews. For an example output file with just one row of information, see output.csv. (your output csv will have as many products as the category has on the first page)
- your source code including clear explanations of what you did as comments and a README file. (If you are familiar with Jupyter notebooks you can just create a notebook and explain your code with markdowns, otherwise feel free to use comments in your code)

Technical spec

We believe there is no one-size-fits-all technology. Good engineering is about using the right tool for the right job, and constantly learning about them. Therefore, feel free to mention in your README how much experience you have with the technical stack you choose, we will take note of that when reviewing your challenge.

Here are some technologies we are more familiar with:

- Python
- SQL
- GCP

For one page you will not get blocked while using Selenium (we prefer to use this library but you are free to choose another one if you are more comfortable with it). Normally we would write our crawler in Python using the Selenium library and deploy it on a virtual machine that we created on GCP so that we don't get blocked while trying to access a website so frequently.

How we review

We will take into consideration your experience level.

We value quality over feature-completeness. It is fine to leave things aside provided you call them out in your project's README. The goal of this code sample is to help us identify what you consider production-ready code. You should consider this code ready for final review with your colleague, i.e. this would be the last step before deploying to production.

The aspects of your code we will assess include:

- Business understanding: This assessment is about extracting and preprocessing data to drive useful insights later on. Therefore, we will be checking the accuracy of the data crawled.
- Clarity: Does the README clearly and concisely explains the problem and solution? Are technical tradeoffs explained?
- Correctness: does the application do what was asked? If there is anything missing, does the README explain why it is missing?

- Code quality: is the code simple, easy to understand, and maintainable? Are there any code smells or other red flags? Is the coding style consistent with the language's guidelines? Is it consistent throughout the codebase?
- Does your README contain information on how to run it?

Extra Mile Challenge

- Crawl number of stars presented on the product card as well and add number of stars as a column named as stars in your output csv. (stars are marked with green lines in the example page above and for RStudio in the example number of stars should be 4.5 in the output csv)

Note: Extra mile challenge is in your advantage if you can do it, but it will not be assessed and it will not put you in any kind of disadvantageous situation in case you can’t complete it.