1nfluencersmarketing.com

Web Scrape Test

**Theoretical Test**

1. Describe which are the main challenges to implement a scraper in social media platforms as Instagram, You Tube and Tik Tock. Give at least 4 examples of challenges, explain why you believe that are an issue and present which methods you would use to handle it. If do you believe there is some challenges that cannot be solved, please provide the reason to support your point of view.
2. What do you understand by Data Mining? How this concept can be used to the mission to develop platform the support business to find the best Influencer to increase the ROI of their marketing actions.
3. Web scraping tasks. If you already use some of them, please describe the solution you have created.
4. Explain what HTTP is and how it works. Also, provide a brief description of methods, headers, and cookies in the HTTP context.
5. What do you understand by Proxies Servers and why their important for the web scraping process?
6. Differentiate Machine Learning, artificial intelligence, and data science.
7. In the Machine Learning context, what is Feature Engineering? Why this is important for Machine Learning Processing and how the Data Engineering cam interfere on that?
8. Differentiate the Data Scientist from the Data Engineering?
9. Imagine that you have been hired as a web scraper developer. On your first day of work you are asked to start a process to develop a data pipeline to create and update a influencers Database from Instagram. As a first step in this process, you are asked to create a simplified logic flowchart that shows the steps in the process of creating this pipeline.

**Practical Test**

Build a simple web scraper to collect data from 50 random profiles with over 1k followers on Twitter. The information that should be collected is:

* Username
* Displayed Name
* Description
* Number of Followers
* Number of Following
* Birthday
* Data that joined Twitter
* Website
* Within the last 10 posts
  + Number of favorites
  + Number of retweets
  + Number of replies

With this data in hand, generate a report. It should have at least the following information:

* Number of tweets per day
* Number of tweets per week
* Number of tweets per month
* Mean of tweets per day
* Mean of tweets per week
* Mean of tweets per month
* Median of tweets per day
* Median of tweets per week
* Median of tweets per month

Finally, explain what metrics are more representative (would better represent the user profile) and why.

**The answers to the theoretical questions and the practical test should be delivered in a GitHub repository. Is mandatory that you use Python. Is not recommend you use special tools develop by third parts as APIs or Scrapers**.