**Theoretical Challenge: 1nfluencermarketing**

**Author:** Fernando Limoeiro

**Question 1:**

**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.

**Answer:**

**1- Legality**: not all social media platforms allows access to user data via web scraping. Luckily most social media platforms provide (under user register) APIs to access this data via safe means. Yet another concern with legality is regarding user privacy. Regulations vary from country to country, demanding careful studies of the rules of a client’s country.

**2- Unstructured Data:** most of social media network data is unstructured by nature. Be it because of the origin (i.e.: user interaction, user events, ad interaction, etc), or the dimensions of data (ie.: timeseries or relational data), unstructured data might pose several constraints to a robust solution, creating bottlenecks from start (ie.: database update concurrence issues) to finish (ie.: high data retrieval time). In order to solve this problem 2 main processes might be applicable:

2.1- Accurate data modeling: In acquisition time (ETLs) social network data must be correctly parsed and formatted. For instance, if data can be stored in a graph format, there are solutions ready to be applied. The same goes for timeseries data. So, creating the right data model for the right kind of data is of utmost importance.

2.2- Efficient acquisition and parsing protocols: if data is correctly modeled, then the workload goes to the data processing pipelines. To achieve optimal performance, pipelines must be fine tunned to specific data formats, and extensively tested and tuned again. In this way, we can build more robust pipelines for each kind of data.

**3- Frequent structural changes**: most platforms change constantly and so does their code and the structure of their html documents. This poses a huge problem since it demands the data processing pipelines that depend on such websites to be constantly refactored. Therefore, this might increase maintenance time and, on a management point-of-view, allocate resources to a repetitive task. Unfortunately, there is no easy solution to this, since it depends exclusively on the platform developers. But problems might be mitigated by always staying up to date to software challenges and updates by these companies. Coupled with a robust testing pipeline.

**4- Captcha:** some websites provide access by solving a captcha (to stop bots from lurking in). Not all captchas are easily solvable and generally they contain images, making automation difficult and resource intensive. Given the complexity of the captcha solving, the most practical solution is to adapt the acquisition pipeline for the manual handling of this barrier.

**Question 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.**

**Answer:** Data Mining is a group of tools and techniques to intelligently extract data from large corpuses of information. Often times this data is not easily accessible, demanding the development of custom solutions. More often, the volume of data produced is enormous, demanding solutions in big data to be applied. One way of applying data mining to influencer marketing is by segmenting a client (company/business) followers base, based on “mappable” topics that might be found at a influencers previous social data. This is a way to match the clients interests to what a influencer commonly approaches as subjects in its work (performances, posting, etc). Besides, data mining can be used to gather several metrics from the business, influencer and client base, offering a whole new level of analysis.

**Question 3:**

**Web scraping tasks. If you already use some of them, please describe the**

**solution you have created.**

**Answer:** Needed to scrape NCBI’s website for thousands of papers to find some values to my research (at INMETRO). Applied selenium + beautifulsoup, and automated the scraping of papers abstracts. Then processed each of these abstracts with specific regex patterns to extract only desired data.

**Question 4:**

**Explain what HTTP is and how it works. Also, provide a brief description of**

**methods, headers, and cookies in the HTTP context.**

**Answer:** HTTP is a Hypertext Transfer Protocol, which basically means that it is a protocol that makes possible the access to a server by means of a human readable address string. It works by handling both the connection and data request to a website’s server.

**Methods**: those correspond to the actions of a general CRUD operation, controlling and parameterizing the following actions between client and server (website): POST, GET, PUT, PATCH, and DELETE.

**Headers:** are fields on HTML responses and requests that store metadata about the operation. **Cookies:** cookies are also another instance of user metadata from website interaction. It holds mostly information that enables websites to run more efficiently and manage resources to be loaded or requested.

**Question 5:**

**What do you understand by Proxies Servers and why their important for the web**

**scraping process?**

**Answer:**  A proxy server is a gateway between the client (users) and the internet. It can act as a router in some cases, and bypass several restrictions, as well as provide another layer of security (if used correctly). Proxies are important to web scraping because several websites and platforms employ several network filters, to dampen bot activity. In order to have a scalable web scraping pipeline, one must use several proxy servers to bypass these restrictions.

**Question 6:**

**Differentiate Machine Learning, artificial intelligence, and data science.**

**Answer:** Data Science is a research field, even though it is commonly referred to as a set of tools and techniques to analyze and extract insight from data. Data Science groups Statistics (and its subfields) and Computer Science, leveraging the power of advanced computers to perform deep and insightful analyses to data, from several sources. Artificial Intelligence is another research field that is also erroneously referred to as a toolset, or “algorithm set”. Indeed it encompases several algorithms, but an Artificially Intelligent system is composed of several algorithms (both Machine Learning or specialist systems) and solutions. Machine Learning is a field of Statistical Learning dedicated to developing algorithms that perform tasks without being explicitly programmed to do so. By being a subarea of Statistical Learning, it brings several notions and concepts from it, but leveraging the power of training algorithms and hardware paralellization to perform more efficient learning procedures.

**Question 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?**

**Answer:** Feature engineering is the process of understanding, analyzing and transforming data descriptors (or predictors) to conform to certain solution’s requirements and formats. It is also a process of fine tuning data representation and also comprehend the optimization of these predictors. For machine learning in general (including deep learning), Feature engineering is important to conform data to the sensibility of Machine Learning algorithms, and also elaborate data representations that might bring smooth convergence to training procedures, producing more accurate ML models. In some ML training pipelines the volume of data can reach enormous numbers, making the process cumbersome (computationally speaking) and resource intensive. Data Engineering helps tailoring better ways of developing such pipelines efficiently, optimizing the process to fit into available resources and optimizing training time, by providing effective and efficient solutions.

**Question 8:**

**Differentiate the Data Scientist from the Data Engineering?**

**Answer:**  The data scientist is mainly a modeling professional; it’s main responsibility is at analyzing data and creating models that might provide insights from it, while being specialized in ways of displaying this data to non-technical audiences. A Data Engineer is a more operations professional, specialized in creating infrastructure to bring most of the insights and models created by a Data Science team to production.