****

**Tools & Techniques for Data Science**

Supervisor: Sergiy Shevchenko

**Students:**

Mehrad Tavanamehr (Student ID: 202201051)

**Introduction**

After boomed of data’s job needs in the recent years, the lack of needed tools for more efficient cooperation in one project between different of members team showed up, why the cooperation is that important in this industry? The widespread different tasks in this industry in one side and different method of think and updating gradually in another side create the basic needs. The request was simple and old, sharing codes and control the changes on them, maybe it seems simple and no need any extra tools, but when we talk about a team with 10 members and more, or international team with members around world that make efficient communication hard or impossible and any of member working on code with their way of thinking in other hand the needs to other team members is essential, you need the efficient tools to control and sharing information.

With increasing request for these kinds of tools and needs to work with them in data industry, another challenge shows up, which tools are more efficient? We need to define efficiency in two different directions,

* Upgrading and support of tools

According to the speed of growing technology these days, the related tools need to be up to date and support previous versions of their product same as up-to-date version.

* User friendly

Inside of learning and working with different topics of data industry tasks, we need to learn and work with different tools as well, so make it easy to understand and work are essential things that the provider should be consider.

In addition of version controlling and sharing the latest updated of your code, some important part to learn is data gathering. Data gathering has different ways,

1. Internal generating data

For some kind of project that are new or no companies didn’t gather data

1. Direct access to generated data

If you are working as team member of company that needs analyze based on their current subscribers (in the first level of analyzing part) and you have direct access to generated information of sales, usage, and etc. probably you need to have knowledge of company’s business to gather correct information and used for involving task.

1. Data Scraping (Web Scraping)

Theses days, the majority needed data for different topic of projects can be found online and in different websites, here using your program language ability is needed and if you are working with python as program language, you need to know different libraries to write your own code for extracting information.

**Abstract**

The goal of this document is separated to two parts,

1. Understanding usages of different tools and technics in data scientist,

For this part we follow some simple definition of popular and useful tools in data industry and work with them in second part to be showed the usage in a project

1. Analyzing specific group of telecom company subscribers’ usage in one defined cycle,

One of telecom company have specific plan to give free data package to its subscribers end of each month as gift to increase the subscriber’s satisfaction and commercial of the quality of data services around the country.

The main goals of designed project are, first persuade current subscribers to use more of services and buying data packages and, second shows the quality of services in vendor to attract people that are not subscriber of this company to buy SIM and join. In summary, like all business around world the company trying to increase its income. But this program has big issue, that is give package randomly to the subscribers.

The company defined data packages with different amount of quota, now imaging below scenarios,

1. You are subscriber with huge usage of different this company services and when you try to take a gift package, you will get a small quota one.
2. You are subscriber with normal usage and you get gift packages, but because increasing usage of data services the company service faces a lack of bandwidth in one of the service layers.
3. You have SIM of this company and another company in same time and you just use gift package and no other usage on this company regularly.

In these scenarios, the telecom company not gain profit to give the gift and in the first and second one, the subscribers are devastated and their usage will gradually decrease and one day they left the company.

To avoid the explained problem, we make brief analyze on this group of subscribers’ usage to understand better their behavior in incoming cycle after took gift. In this project, our goal is a little different and we want to add some junk data to more analyzing and investigating.

In the below chart, you can see our steps briefly and in general,





**Tools**

In this part, we follow some open-source tools that are essential for working in different data teams,

* **GitHub**

It is most familiar website that providing repository to be share your code with other members of your team and providing a portfolio to be seen by companies and technical teams. You can create an account completely free in github.com

GitHub has different type of account that can be used by companies, but they need to pay for it to protect their codes and just people with specific credential, and the most amazing thing about this space is that, it can be control with your local system and no need to connect directly to site for put or push something from repository, it is done with creating and sharing the key from your system to website.

* **Git**

With similar name with GitHub create sense it is part of GitHub project, but it is not. Git is control version system that can be worked with different repositories and one of them is GitHub. This open-source control version tool is the famous because working with it is easy and in your local can have different type of it.

* **Docker**

Always has separate area to control of CPU and memory usage of local machine to avoiding any disaster of not end loop that make local machine down. Before that, we heard about VMware or other similar technology that create a virtual machine for us to be used. Docker is similar and different in this technology and it is the next generation of that for professional guys to use less space and less memory and CPU to run isolated area for using, another advantage of Docker is the repository it gives the it’s users with free login in hub.docker.com

With this repository you can share your all project on one package with your team members and just something they need to do is running your image. Docker is known as container management that can involve different containers separate from your local system, that you can find these images in docker website and after download and run, you can simply modify them for your own job and share it. Here, we use the PostgreSQL package for using database and connecting to python.

* **DBeaver**

It is database management application that can work with different types of databases (SQL, NoSQL, etc.). In this project we want use it to work with PostgreSQL database to create table and insert information and run some SQL program language to extract the need data and simply upload our result of investigation in different tables.

**Technics**

Some important general technic that any person that work in data industry should know and work with is web scraping to gathering needed information for task. It is important point that, data scraping and data mining are not similar and before we start to mining our needed data, we should gather them. Web scraping is technic that usage of it increases very fast and reason of that increasing data in web. Somehow, we should remember, data scraping from different web maybe is illegal because a lot of sellers store their subscriber’s information to increase the subscriber’s sufficiency, but it is important point saving subscribers information from web scraping is essential.



In the project, we used the two libraries in python for scraping of junk information to add our real one.

**Project**

* **Phase1: Data Gathering**

As in abstract part is mentioned, we gather our information from two ways

1. Gathering from direct access to data producer company

Working with real data always has difficulties from extracting and exporting data till working with extracted data and find the expected output of data, but it gives the open hand to work and understanding of data analyze and data prediction topics more than other datasets. According to our expected output and make some analyzes and in the end predicting next subscribers’ behavior, we will use below table of features and near 2.5Million rows of information

|  |  |  |
| --- | --- | --- |
| **Fields** | **Type** | **Description** |
| SERIAL\_NUMBER | Character | Unique subscribers code |
| PAYMENT\_TYPE | Character | Type of SIM that subscriber use (Postpaid/Prepaid) |
| PROVINCE\_NAME | Character | the region that subscriber belongs to |
| GENDER | Character | Gender of subscriber |
| PAYMENT | Number | all payment that subscriber is done in the next cycle of activation of gift package |
| SMS | Number | Usage of specific services based on amount |
| DATA | Number | Usage of specific services based on amount |
| DATA\_PACKAGE | Number | Usage of buying data package in the next cycle of activation of gift package |
| VOICE | Number | Usage of specific services based on amount |

1. Web Scarping

To increase the feature for analyzing parts and make better advantage in data analyze part, we added some fields. This job is done to using web scraping in project and facing with possible problem in working

|  |  |  |
| --- | --- | --- |
| **Fields** | **Type** | **Description** |
| SERIAL\_NUMBER | Character | Unique subscribers code |
| GENDER | Character | Gender of subscriber |
| Nationality | Character | Main nationality of subscriber |
| Age | Number | The current age of subscriber |

* **Phase2: Using Docker to create database container and uploading gathered data**

There are different ways to put your data in container and publish it, here, we use the PostgreSQL container and uploading the gathers data from all possible way and put them in different tables in database to store and work with them. We create two tables, “data\_source\_mondaydata” and “junk\_data\_Monday” and join them and create a view “Course2\_ToolTechniquesforDataScience” to work in pandas.

**create** **or** **replace** **view** Course2\_ToolTechniquesforDataScience **as**

**select** dsm.serial\_number, dsm.payment\_type, dsm.province\_name, jdm.gender, jdm.**age**, jdm.nationality,

dsm.payment, dsm.sms, dsm."data", dsm.data\_package, dsm.voice

**from** data\_source\_mondaydata dsm **inner** **join** junk\_data\_monday jdm

**on** dsm.serial\_number = jdm.serial\_number;

we updated some information to create sense from them,

-- according to sms service the amount of not delivered messages will be back to subscriber's account and when the cycle changed, if there are some remain amount from previous cycle, there are decreased from current cycle, we updated them to zero

**update** data\_source\_mondaydata **set** sms=0 **where** sms<0;

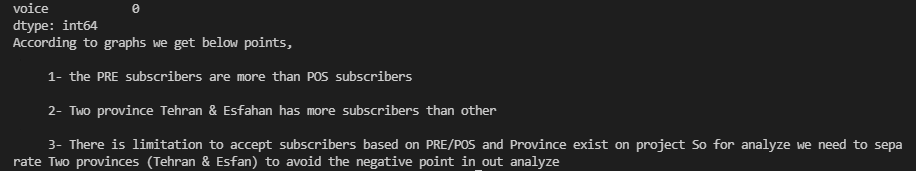
**commit**;

To make a job more efficient and possible to follow in the future there are a specific repository is created for insert new information to tables and the view make us sure our data in use is always up to date. To insert new data, you should run container like below,

docker run --rm mtavanamehr/insert\_into\_table <*your connection IP*> <*your WEBDriver path*> <*your csv path*> <*your log path*>

* **Phase3: read from database and do analyzing**

To read from PostgreSQL container, you need to put your connection IP as input argument so when you need to run the analyze container, you should go with  
docker run --rm mtavanamehr/analyze\_course2 <*your connection IP*>



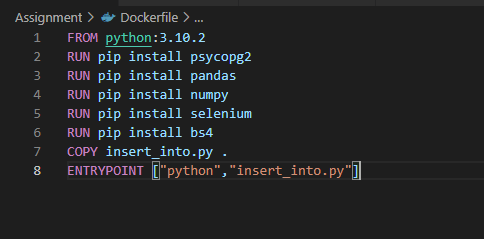
**Docker**

In this project, our focus on below items,

1. Create Docker container

docker build -t <*name*> .

you need to have DockerFile that contain some important configuration, and because we want to pass the argument in run time we use ‘ENTRYPOINT’ with our python file name



1. Share Docker container

To share your container, you need to create a tag for it first and then push it to your account repository that is connected to your system.

docker tag <*name*> <*your user in docker hub*>/<*name*>

docker push <*your user in docker hub*>/<*name*>

The repositories address that container are uploaded on,

<https://hub.docker.com/repository/docker/mtavanamehr/postgres>

<https://hub.docker.com/repository/docker/mtavanamehr/insert_into_table>

<https://hub.docker.com/repository/docker/mtavanamehr/analyze_course2>

**Resources**

<https://selenium-python.readthedocs.io/navigating.html>

<https://randomuser.me/documentation#nationalities>

<https://realpython.com/python-versions-docker/>

<https://realpython.com/python-command-line-arguments/>

<https://stackoverflow.com/questions/46245844/pass-arguments-to-python-argparse-within-docker-container>