Homework #2

The Big Data Theory

September 21, 2017

Dear Participants,

Here is homework #2 that I again ask you to do. Please take this homework seriously, because these will be crucial in solving the problems at the workshop. What I'm writing down below is for Windows machines, I assume Linux users are able to adjust the steps to their needs.

2.1 Install Jupyter Notebooks

Although for dataprocessing you'll be using an online platform called Databricks, it's necessary to install a Python interpreter on you computer as well. You'll need it not only for the project exercise but also for whole Tuesday.

Install Anaconda

Please download and install Anaconda Python version 3.6 and version 2.7 from the url below. (For Tuesday you'll need ver. 3.6. If you're a Hungarian student, and only want to do the project, but you don't come on Tuesday, installing ver. 2.7 is enough and vice versa.)

https://www.learnpython.org

I refer experienced Python users to the documentation below:

http://jupyter.readthedocs.io/en/latest/install.html

Change Jupyter Notebook path

Create a directory in your file system from where you want to run you notebooks. If you're using several drives, I'm warning you I could only get this stuff work on the same drive where my system is installed. (Regardless of the Anaconda folder.)

Start menu → AnacondaX folder (replace X by 2 and/or 3) → righ-click on *Jupyter Notebook* → *Properties* → in the *Target* field replace *%USERPROFILE%* by the working directory you created. E.g. in my case the *Target* field for Anaconda3 looks like:

```
... "D:/Anaconda3/Scripts/jupyter-notebook-script.py" C:\BigData
```

Run a notebook

It's best if you open a web browser, and then run *Jupyter Notebook* with the Python version you want. After a short waiting time, it'll open a new tab in your browser. What happened is that it has created a local server on your machine, and now you communicate with it through your browser. Ain't that cool?

Click *New*→ *Python X* (replace X by the version number).

In the notebook type e.g.

```
print "Hello world!"
```

for ver. 2.7 or

```
print("Hello world!")
```

for ver 3.6 and press *Shift+Enter* to run the code. Whammm, there you go!

Change Python version

Nothing much to do here, just run *Jupyter Notebook* from the Anaconda folder with the appropriate Python version.

2.2 Databricks

Now to create a userprofile on the data analysis platform that you'll use for the project.

Create profile

Create a Community Edition user profile for yourself on Databricks that you'll reach on the website:

https://databricks.com/try-databricks

Create cluster

There are some big servers behind Databricks, and you have to allocate some computational resources for yourself to be able to do anything. To do this, create a cluster by clicking *Clusters*(on the left) → *Create cluster*. Give it any name you like, leave anything else to default and click *Create cluster*.

Create a new python notebook

Click *Workspace* (on the left) → right-click on the white column that just appeared and *Create* → *Notebook*.

Write a *Hello World!* or whatever you'd like. The Python version running is 2.7, you can check it by executing

```
import sys
sys.version_info
```

(Just copy the whole code into a cell and press *Shift+Enter* to run.)

Note: Sometimes, especially if you have bad internet, you have to wait several minutes to get an answer.

Import a python notebook

If in the above menu you click *Import* instead of creating a notebook, you can upload a notebook that was given to you either as an URL or a file on your machine.

Upload data

This part is a bit tricky, watch out. On the left side click *Data* and click the + sign next to *Tables*. (*Create table* it says, if you move the cursor over it.)

Upload any file you want (I used the one I'll ask you to download under the next task, that'll be a tutorial to the data analysis library that we'll use; url:

https://github.com/graknlabs/sample-projects/blob/master/example-csv-migration-games/ign.csv),

and then **do not forget the path**, that looks something like this:

`/FileStore/tables/j27pmri11506025716765/ign.csv`

In case of *csv* files, such as the one I uploaded, there is also a possibility to look into the file content. To do this, click *Preview table*, enter a name (*ign* will do the job), select *First row is header* if it is (in case of this file, it is) and click *Create table*.

Whenever you want to access a file through the Python scripts you'll write, you have to use the path

`/dbfs/FileStore/tables/j27pmri11506025716765/ign.csv`

(Note the `/dbfs` in front that I added.)

From now on you can run a script in an IPython notebook such as

```
import pandas
reviews=pd.read_csv("/dbfs/FileStore/tables/j27pmri11506025716765/ign.csv")
```

This takes us already to the next section.

2.3 Pandas tutorial

In this section we'll look at the Python library that was developed for effective data analysis.

In order to solve this exercise, please import the notebook

`pandas_intro_Databricks.ipynb`

into your Databricks workspace, and go through it yourself. It follows the tutorials

https://www.dataquest.io/blog/pandas-python-tutorial/

https://pandas.pydata.org/pandas-docs/stable/10min.html

If you want more, I invite you to search for more yourself. (An option would be e.g. this:

https://www.learnpython.org/en/Pandas_Basics).

For a reference of Pandas, see

https://pandas.pydata.org/pandas-docs/stable/api.html

2.4 Geopandas tutorial in Databricks

Throughout the project you'll have to work with geographical data for which we'll use GeoPandas, that is a Python library just for this purpose, and builds on Pandas.

In order to solve this exercise, please import

`geopandas_intro_Databricks.ipynb`

into your Databricks workspace, and go through the notebook. Don't worry, if you don't understand everything, but I want you to get the big picture.

That's it for now, because I didn't have time for more, but brace yourselves, at least one more is coming.

Greetings,
József Mák
EESTEC LC Budapest