Homework #3

The Big Data Theory

October 2, 2017

Dear Participants,

One of your last homeworks is being uploaded. There is going to be one more extra, but it will be only for Wednesday. This homework is quite lengthy one, but for your own sake, I suggest you to go through it, because the project exercise will be quite similar.

3.1 QGIS

First I invite you to download and install a geographical software called QGIS from here:

http://www.qgis.org/en/site/forusers/download.html

If you're done installing, run *QGIS Desktop 2.18.13*. Go to

http://www.openstreetmap.org/search?query=Ames%2C%20Iowa#map=12/42.0259/-93.6265,

(this is Ames, Iowa, USA), click *Export* on the top, then *Export* again on the left. (If it doesn't work, click *Overpass API* instead of the second export.)

In QGIS go to *Layer* → *Add Layer* → *Add Vector Layer*, browse for the file you downloaded, *Open*, *Select All*, *Ok* and you've got yourself the map of Ames on you screen.

Now go to *Layer* → *Add Layer* → *Add Delimited Text Layer*, browse for *restaurants_Ames_GPS_QGIS.csv*. Click *Point Coordinates* if necessary, and in the drop-down menus for the *X*- and *Y-fields* set *lng* and *lat* respectively. (These are lateral and longitudinal GPS coordinates of the restaurants in Ames downloaded by the Google API that you'll use later.) Click *Ok*, and in the pop-up window set *EPSG:4326* as the Coordinate Reference System (CRS) and click *Ok*. Now you see dots corresponding to restaurants on the map.

Do the same with *mean_sale_prices_GPS.csv* and right-click on *mean_sale_prices_GPS* in the bottom left window. Select *Properties* ang go to *Style*. On the top drop-down menu click *Single symbol* and select *Graduated* instead, click the *Column* drop-down menu, slect *SalePrice*. Set the color map as you see fit and click *Classify* then *Ok*. Now you've also got the points of Ames' neighborhoods colored by the average price of house prices in the neighborhood.

3.2 Machine learning model

Upload *AmesHousing.csv* and import the IPython notebooks I've given you for this exercise into your Databricks account, open *Main.ipynb* and start reading. (Don't forget to create a cluster and where Databricks stores your files.) You'll see how to download data with Google's API for geographical locations and how to set a machine learning model that tries to guess the average prices of houses in Ames' neighbourhoods based on the neighborhood location and restaurants nearby.

I know this was long, but I hope you had fun. See you on the workshop.

Greetings,
József Mák
EESTEC LC Budapest