Homework #8
Due Apr. 11, 2022, 11:59pm

Problem 8.1

In the class data directory on the SciServer you will find a file `covid-19-us.csv`. This contains cumulative data on the COVID-19 cases for the whole US from the day the first infection was detected. The data is from the New York Times data collection on GITHUB (see https://github.com/nytimes/covid-19-data).

The problems presented here will serve as a warm-up exercise in learning the basics of Python and the basic data input/output and graphing techniques *(hint: read the csv file into a pandas DataFrame*).

a) Read the file and create a plot of the cumulative cases as a function of time
b) On a separate plot, create a count of the cumulative deaths as a function of time
c) Repeat both plots when you show the numbers of cases on a logarithmic scale
d) On the cumulative charts visually find the segments when there was an exponential growth in the number of cases, indicated by a linear increase of the log(counts). Estimate the slope visually, and overplot a straight line with approximately matching the trend of the data. Interpret the slope in terms of the number of days that the cases double.

In order to be able to use the shared data, you should have the following imports at the top of your iPython notebook:

```
import pandas
import os
```

followed by

```
path = '../../../AS_171_205_2021/AS_171_205_2021-Course/data/'
files = [f for f in os.listdir(path) if f.endswith(".csv")]
files
```

You should see the following output:

```
['covid-19-us.csv', 'covid-19-us-states.csv',
    'covid-19-us-counties.csv']
```

If you do not see this directory, you need to go to the SciServer Compute page with your Containers, and you need to create a new container. When you do that, make sure that you select the User Volume

```
AS_171_205_2021-Course, Storage Volume created by AS_171_205_2021
```

Finally, you can read the first data file with

```
df =pandas.read_csv(path+files[0])
```

There is an example notebook for database access below list of the homeworks.

Problem 8.2

Consider the data set describing wine quality vs different attributes, described at:
http://archive.ics.uci.edu/ml/datasets/Wine+Quality
The data itself is at `winequality.zip`, in the data directory [1].

(a) Use Principal Component Analysis to find a projection in which the excellent and poor wines can be best separated. Perform the analysis for both the red and white wines. If possible, try to determine the three most important factors influencing the decision.

(b) Repeat the analysis using t-SNE and UMAP

Problem 8.3

The nuts and bolts problem is defined as follows. You are given a collection of n bolts of different diameters and n corresponding nuts. You are allowed a test operation in which you try a nut and bolt together from which you can determine whether the nut is too large, too small, or an exact match for the bolt. You cannot directly compare two nuts or directly compare two bolts. Each test operation takes constant time, and the problem is to match with each bolt to its nut.

Design an algorithm to correctly pair up the n nuts and bolts in expected O(n log n) time. Show that any algorithm for the nuts and bolts problem must use the test operation Ω(n lg n) times in the worst case by measuring the execution times for 100 realizations with n nuts and bolts randomly scrambled, for each value of n=16,64,256 and 1024