ZHAW: Azure Databricks Access

ZHAW provides every group a cluster on Azure Databricks that is shared among the members of the group. **You can sign in here** (with your ZHAW login):

https://adb-8447367562806996.16.azuredatabricks.net/?o=8447367562806996#

The cluster that is given to your group is configured with one driver and four worker nodes. You can start the cluster either by navigating to the "Clusters" Tab or by attaching it to a Python notebook. Be aware that the cluster **shuts down after 30 minutes of idle time**.

As you do not have the permission to change the cluster configuration, **libraries have to be installed on a "per-notebook" basis** (unlike the steps outlined in the Machine Learning bonus material provided through the lectures). You can do this by adding this line of code at the beginning of your notebook to install, for example, the library numpy:

```
dbutils.library.installPyPI('numpy')
```

You can read more about this in Databricks' blog post here: https://databricks.com/blog/2019/01/08/introducing-databricks-library-utilities-for-notebooks.html

In case this procedure does not work and you need to explicitly install a library, please send an email to Kurt.Stockinger@zhaw.ch and yasamin.eslahi@zhaw.ch. We will forward your request to Remo Maurer, our **system administrator, who has access rights to install libraries**.

Moreover, all clusters have **access to a common shared file system**. This means that if you upload a file to the cluster, all other groups can see this file as well. In order to avoid overwriting of files, please use your **group name as suffix of a file**. For instance, if you upload the file "data1.csv" and you are in group 03, proceed as follows:

- First, **rename** the file from "data1.csv" to "data1**_group03**.csv" on your local computer.
- Then **upload** the file "data1_group03.csv" to the cluster.

Wise usage of cluster resources: The Azure Databricks cluster provided by ZHAW is hosted by Microsoft on a **"pay-as-you-go" license**, i.e. the more we use the cluster, the higher our bill. However, for the current setup there is no specific resource budget per group. In order words, you could in principle use the cluster for an (almost) unlimited amount of time and thus generate an (almost) unlimited amount of costs to ZHAW. In order to avoid this situation, please **test your code first on the Databricks Community Edition and use the Azure Databricks cluster for large-scale experiments**.

If you use the cluster for Spark Streaming, please make sure to **stop the producer after running more than, e.g. 5 hours per experiment and shut down the cluster** afterwards.