Tutorial 7

Epidemiology part 2

Step 0: If you have not done so already: *claim a virtual machine from infomdssLabA*

Log in to the Azure portal (https://portal.azure.com/),

Go to **Dashboard → infomdssLabA → infomdssLabA** and claim a virtual machine.

Login to your lab B VM (the default password is the same as in tutorial 1) and **change your password**.

```
$ passwd # choose your new password wisely
```

These actions are explained in tutorial 1.

Step 1: Prepare the Data

In the assignments of the previous tutorial, you were asked to generate a `epi.csv` file, by combining the `CHD`, `MENO` and `T2D` datasets.

You might want to use the `epi.csv` in the next tutorial, you can download the preprocessed set, but you are encouraged to use your own.

You also need to make a comparisement with data provided by the papers where these data-files are described. The top hits from the papers are also provided.

```
$ # Make and move to epidemiology (only execute if it does not exists)
$ mkdir /home/labuser/epidemiology && cd "$_"
$
$ # Download epi.csv and top hits from the papers
$ wget https://transfer.sh/Jl0tF/epi.csv.zip
$ wget https://transfer.sh/DeUjv/papers_top_hits.zip
$
$ # Get and execute the checksum
$ wget
https://raw.githubusercontent.com/Infomdss2018/infomdss/master/tutorial_7/checksum/epi.csv.zip.sha1
$
```

```
$ wget
https://raw.githubusercontent.com/Infomdss2018/infomdss/master/tutorial_7/checksum/papers_top_hits.zip.sha1
$
$ sha1sum -c epi.csv.zip.sha1
$ sha1sum -c papers_top_hits.zip.sha1
$
$ # Unzip epi.csv data and give permissions
$ unzip epi.csv.zip -d data/
$ unzip papers_top_hits.zip -d data/
$
$ # give permission to the data folder
$ chmod 777 data/*
```

Step 2: Open Notebook

Now all data is downloaded and prepared, we can start with the notebooks. You can open your notebook by navigating to https://vmlabaXXX.westeurope.cloudapp.azure.com:8000, where you have to replace XXX with the number of your VM.

You should get a login screen where you can fill in the username and password of your VM (e.g. “labuser” and your updated password).

After login you can choose if you want to continue this assignment in R or python, this tutorial is written in python, so you can select **new → python 3 spark-local** and start with the assignments.

Step 3: load the data-files into your notebook

Just like the previous tutorial, but the epi-set also needs some headers:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("epidemiology").getOrCreate()
sc = spark.sparkContext

meno_paper_RDD = sc.textFile("/home/labuser/epidemiology/data/MENO_paper.csv")
card_paper_RDD = sc.textFile("/home/labuser/epidemiology/data/CHD_paper.csv")
trans_paper_RDD = sc.textFile("/home/labuser/epidemiology/data/T2D_paper.csv")

epi_path = "/home/labuser/epidemiology/data/epi.csv"
epi_df = spark.read.csv(epi_path).toDF('Chr:Position', 'markername', 'chr', 'bp_hg19', 'effect_allele',
                                       'noneffect_allele', 'effect_allele_freq', 'median_info', 'model', 'beta',
                                       'se_dgc', 'p_dgc', 'het_pvalue', 'n_studies', 'allele1', 'allele2', 'HapMap_eaf',
                                       'effect', 'stderr', 'p', 'Allele1', 'Allele2', 'Effect', 'StdErr', 'P-value',
                                       'TotalSampleSize')
```

Assignments:

Sometimes merging cannot be conducted as names of the SNPs change over time or studies did not completely include the same SNPs (for instance because imputation failed for certain SNPs). Some of these problems can be solved, but this is beyond the scope of this assignment.

1. Provide the top hits for each of the outcomes (CAD, T2D, and age at menopause)

2. Check the overlap with the top hits mentioned in the papers or online.

3. Is there overlap between the top hits?

   For instance, what are the 50 hits with the lowest p-values per trait? How large is this overlap? Which SNPs are those and what is the ranking p-value for each of the traits for those SNPs (you can think of a table with SNPid, ranking per trait and p-value per trait, so 7 columns).

Stop your virtual machine at the end of the workshop