



EGT 305: BIG DATA PROCESSING & APPLICATION

BOSS

LAB 1:

INTRODUCTION TO PYSPARK ON COLAB

- HANDS ON WITH PYSPARK
- HOW TO RUN PYSPARK ON COLAB
- TRYING OUT SOME CODE
- INTRODUCTION TO SOME
FUNCTIONS



LAB 1

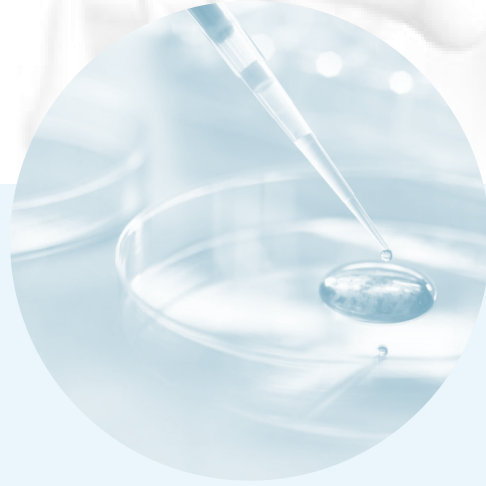
```

[1] 1
2 !apt-get install openjdk-8-jdk-headless -qq > /dev/null
3 !wget -q https://downloads.apache.org/spark/spark-3.5.2/spark-3.5.2-bin-hadoop3-scala2.13.tgz #If there is an error please check the spark version in the link above
4 !tar xf spark-3.5.2-bin-hadoop3-scala2.13.tgz#Please change the version according to the downloaded version
5 !pip install -q findspark

tar: spark-3.5.2-bin-hadoop3-scala2.13.tgz#Please: Cannot open: No such file or directory
tar: Error is not recoverable: exiting now

[2] 1 import os
2 os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
3 os.environ["SPARK_HOME"] = "/content/spark-3.5.2-bin-hadoop3-scala2.13" #Please change the version according to the downloaded version
4
5
  
```

- Download file from Brightspace
- Transfer file into Google Colab
 - Start running the code



THANK YOU



SLEEPYHEAD
SLEPPYMEL@GMAIL.COM