BU.330.740 Large Scale Computing on the Cloud

**Lab 2. Python Spark on Google Colab**

Learning Goal: Practice Python Spark on Google CoLab.

Required Skills: understand basic Spark syntax

1. Prepare:download an e-book *Pride and Prejudice* by Jane Austen in txt format from

<http://www.gutenberg.org/ebooks/1342> and save in your local folder. Or simply reuse from Lab 1.

1. Open a new notebook in Google CoLab. PySpark should already be installed in Google Colab. If not, install the PySpark package.

!pip install pyspark py4j

1. Create a SparkContext.

from pyspark import SparkConf

from pyspark.context import SparkContext

sc = SparkContext.getOrCreate(SparkConf().setMaster("local[\*]"))

1. Mount your Google Drive. Then you can access your drive. Upload the ebook to your drive.

from google.colab import drive

drive.mount('/content/drive')

A screenshot of a computer

Description automatically generated

1. Implement the code on page 42 of the lecture notes. Don’t forget to replace the file paths!

You should choose the correct path on your drive by clicking the three dots next to your file and selecting “Copy path”. Once copied, paste it into your code.

A screenshot of a computer

Description automatically generated

1. Now you can execute the code and then check results from your output folder.
2. Prepare the second exercise by downloading spam and ham examples.
3. Create a SparkSession.

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("SpamDetector").getOrCreate()

1. Mount your Google Drive. Then upload the files to your drive.

from google.colab import drive

drive.mount('/content/drive')

1. Implement the code on pages 46-48 of the lecture notes. Don’t forget to replace the file paths!