A picture containing diagram

Description automatically generated

**Course**: EECS 4415 Big Data Systems

**Topic:** Data Exploration and Analysis with Spark

**Notes**

All questions regarding the course projects should be directed to TA (contact e-mail information posted on eClass).

This project is to be done individually. Solutions you hand in for the project must be your own work. All code/solutions must be written independently. Students plagiarizing the solutions will get penalized. We use software to check the originality of your code.

Do not share the instructions and your solutions at any stage on external platforms. such as GitHub and Course Hero etc.

**Description**

You will be using Spark to complete three tasks in the Google Colab Notebooks by filling #YOUR CODE HERE and YOUR TEXT HERE boxes:

* Word Count
* Pattern Discovery
* High Dimensional-Data Analysis

Colab is a free cloud service from Google, hosting Jupyter notebooks with free access to GPU

Make a copy for Colab Notebooks after clicking on links below to save changes (otherwise changes will not be saved).

File -> Save a copy in Drive

It will appear in your Google Drive:

Google Drive -> My Drive -> Colab Notebooks

Use File->Save to save the changes.

Preliminary: make sure that you complete the [Spark Tutorial](https://colab.research.google.com/drive/1AXxMFFiYPU7mHIOOcVpZZ-enaDAvrKH2?usp=sharing) (Task 0)

**Task 1** (Week 1)

[Word Count](https://colab.research.google.com/drive/1-39cmxPVQ0TwE0lYn1CPXwZU1Y5peSH6?usp=sharing)

**Task 2** (Week 2-3)

[Pattern Discovery](https://colab.research.google.com/drive/1CXDtmKjW5U0TBhCNo-xtr7EGfE79VA8k?usp=sharing)

**Task 3** (Week 4-5)

[High Dimensional-Data Analysis](https://colab.research.google.com/drive/1kUZhfdMdrM2GEYPoc-Vc3emm_hCjuXoo?usp=sharing)

**Submission**

Submit one zipped folder in .zip that includes the following files :

* .ipynb: File-> Download -> Download .ipynb
* PDF: File -> Print -> Destination -> Save as PDF

That is there should be 6 files in total (each of the 3 tasks should have 2 files). Ensure that both files: .ipynb and PDF include all executed code, along with the printouts of the requested results and the accompanying text descriptions.

Ensure that your solutions are well-organized and well-written.