**

## COMP814 – Text Mining

Google Colab and Python intro lab

# Objective

1. To get a basic understanding of the Google Colab python environment.
2. To get a quick introduction of Python and Data Science related data processing.

# Notes

1. We will be using Google Colab environment to write python code.
2. If you have used some other python environment (such as pycharm or Jupyter notebook) for python programming and prefer to continue using it, you are free to do it.
3. If don’t have access to Google Colab, then you should use Jupyter Notebook. You can get access to online version [here](https://cocalc.com/).
4. You can also use Anaconda installation on your own machine. You can get installation files and instructions from [here](https://www.geeksforgeeks.org/how-to-install-jupyter-notebook-in-windows/).
5. If you are using Jupyter notebook, The instructions and information on Google Colab is still relevant but not exactly applicable. You should do similar steps on Jupyter Notebook. You will also need to do basic environment usage steps from the corresponding website. The learning steps on python in step 7 (later in this document) can be done on any python environment.

# Intro to Python

Python is an open-source (free) programming language that is used in web programming, data science, artificial intelligence, and many scientific applications. Learning Python allows the programmer to focus on solving problems, rather than focusing on syntax. Its relative size and simplified syntax gives it an edge over languages like Java and C++, yet the abundance of libraries gives it the power needed to accomplish great things with small amounts of code writing.

Python is an interpreter based language which means you the code is run one line at a time by a python interpreter as opposed to a language such as C++ which converts the whole program into an executable before you can run it.

A strong advantage of Python is that the code is highly transferrable as long as you have the appropriate libraries installed. This, combined with the availability of an abundance of Data Science related libraries makes it a very popular language for Data Science projects.

To be able to use Python on your own machine you will need to install python, however with Google Colab, you don’t need to install it as you write code and run it on Google’s cloud engine.

# What is Google Colab?

## Why you should consider using Google Colab

Apart from being a browser-based environment that requires a simple Google sign-in, Colab has several useful features that make it helpful for the data science community. The following are some of the advantages:

* Pre-installed data science libraries
* Easy sharing and collaboration
* Seamless integration with GitHub
* Working with data from various sources
* Automatic storage and version control
* Access to hardware accelerators such as GPUs and TPUs
* Colab is a cloud-based Jupyter notebook environment from Google Research. With its simple and easy-to-use interface, Colab helps you get started with your data science journey with almost no setup.
* Colab lets you kickstart your python work worrying about configuring your environment. It facilitates writing and execution of Python code from your browser.
* Colab notebooks allow you to combine executable code and **rich text** in a single document, along with **images**, **HTML**, **LaTeX** and more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab notebooks with other students, co-workers or friends, allowing them to comment on your notebooks or even edit them.
* When you start a project in Colab, you are allocated a **virtual machine** which is unallocated when you finish, so any initial setups will not be saved. This means any initial setups you will do will need to be done using code so that every time you run a Python script, it will first setup the environment before running the actual code. All this will become clear after you do your first lab.

## Getting started with Google Colab

1. You need a **google email account to use Google Colab**. If you don’t already have a google email, you will need to create one.
2. Follow instructions [here](https://colab.research.google.com/notebooks/basic_features_overview.ipynb#scrollTo=KR921S_OQSHG) to do your first tutorial on Colab.
3. Note that you can run the python code by clicking on the play button top left corner of the cells. Complete the steps in the tutorial and ensure that you understand the following about the environment.
   * Code cells versus text cells.
   * Adding/removing cells.
   * Merging cells.
   * Running systems commands using “!”.
   * Running code in cells.
   * Running code in a set of cells.
   * Clearing cell outputs.
4. Follow the instructions [here](https://www.tutorialspoint.com/google_colab/google_colab_quick_guide.htm) to write and run your first Colab program.
5. Follow the instructions [here](https://www.tutorialspoint.com/google_colab/google_colab_executing_external_python_files.htm) to learn how to run external python files and to mount your google drive to access your files from the virtual machine. After you mount your google drive you should be able to click on the folder icon on the left side to see the list of file and folders on your google drive as shown in the image below.

Graphical user interface, text, application, email

Description automatically generated

1. Follow the instructions [here](https://www.tutorialspoint.com/google_colab/google_colab_installing_libraries.htm) to learn how to install external libraries in Colabs. Note that you will need to include the library installation instructions as part of the script as this will need to be done every time you create a virtual machine to run code. Also note that some popular data science libraries come pre-installed on all virtual machines so you should, as a default, run assuming the library is installed and only include the installation instructions if there is an error saying the module can’t be found.
2. Follow the instructions [here](https://www.pythontutorial.net/python-basics/) to learn about python basics on the following topics. To learn each of the following you should copy and run the code in Colab cells. You should use separate cells organised with descriptions for each of the topics **as part of a single script file named, Lab1.ipynb.** You should then submit your work for lab 1 by sharing the notebook. To do this you should share the notebook (on the top left corner) and copy the link. This link should then be submitted for lab 1 on Canvas. **This is how you should submit the work for each of the future labs.** (if you are using a Jupyter Notebook environment then you should submit the .ipynb file instead)
   * Variables
   * Strings
   * Numbers
   * Constants
   * Type conversion
   * Comparison and Logical operators
   * If…else statement
   * For loops
   * Python functions
   * Lists
   * Tuples

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Lab 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*