Overview

[Please read over before proceeding]

**Introduction**

In this Google Drive Folder we have provided the python code and data used in this paper. We hope that this folder will serve several functions.

1. Provide an open and transparent repository of our code, data and output
2. Provide a free online work space (computational environment) where researchers can execute python code without having to download and install python on their machine (using the free *Goole Colab* tools). These tools also provide free access to high performance GPUs and TPUs that researchers may not otherwise have access to.
3. Provide code that researchers can copy to their own Google Drive and utilize for their own research projects (by adapting our code).
4. Provide an infrastructure to keep our data and models updated even after the paper is (hopefully) published:
   1. Provide updated classification of AI patents if additional training data is collected (by us or other research teams).
   2. Provide tools and implementations of new AI tools as they become available (for instance Transformer models have only become popular in the last two years.

*This drive has been set up on a new and anonymous Google account, such that the identity of the authors is not revealed. Additionally, access to the drive is shared with anyone that has access to the link and therefore the authors do not have information about the identity of the reviewers.*

# Overview

This folder is structured as follows:

## Google Colab Notebooks Comparing Different Text Classification Models:

* + **Notebook A:** Bag of Words Models
  + **Notebook B:** Embeddings Based Models
  + **Notebook C:** CNN Based Models
  + **Notebook D:** Transformer Based Models

## Google Colab Notebook Comparing Model Outputs and Generating Figures for Paper:

* + **Notebook X:** Summarizing Classification Outcomes

## Folders with Supporting Information:

* + **Output:** Folder where classification outputs are stored
  + **Embeddings:** Folders where we save text embedding data (used in Notebook C). For Notebooks B and D, we download these in the notebook directly.
  + **Data:** Notebook where we store the training data that we then use to perform text classification.

# How to Use:

* When you access this notebook, a folder titled “*USPTO AI Patent Classification”*  will appear on their Google Drive under the directory “*Shared with Me*”. This is view-only and cannot be edited directly.
* We recommend that you right click and select “Add Shortcut to Drive” so that this appears on your own google drive. Then make a copy of the folder and the files in your own drive (you can copy individual files, or may have to use extensions to copy the whole folder - [link](https://chrome.google.com/webstore/detail/copy-url-to-google-drive/hhkdailooaapiplkadgdkkllbnkjpbel)).
* At this point you should be able to view and run the notebooks on your own computer, and the path for files the notebooks should work on your machine (please check, as there might be some unforeseen issues). If you would like to change the notebook, or implement on your own application, we would suggest making a copy of the notebook files before doing so.

# Google Colab - What is it and How to Use it:

Python is a free and widely used programming language, but in order to use it, researchers typically have to set up an environment on their own machine (e.g. Anaconda Python). Google has launched a free service called Colab which provides Jupyter Notebook-esque files to be stored in drive and to be executed remotely.

This means that when you open these files, a remote machine operated by Google launches python on the back end, allowing you to write and execute python code in your browser without having to install Python on your local machine. You can access data hosted on your Google Drive and use it for data analysis in Python. Additionally, this tool provides free GPUs and TPUs (subject to availability for free tier on Google servers). These are special processors that are quite expensive but provide the computing power necessary to run high end machine learning and deep learning models.

We have chosen to use Colab because it allows us to share our code and implement it using a GPU online. It also means that other researchers who want to use these tools can do so online, for free, and can use the code we have provided as a starting point to incorporate their own modifications.

Since this is provided by Google, there are extensive online resources on how this tool can be used ([Intro to Colab - By Google](https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index#scrollTo=P-H6Lw1vyNNd)). For all of the notebooks provided, we have set the runtime needed to execute the models (e.g. GPU runtime for Transformer Models). However, if you would like to change the runtime type you can follow these steps.

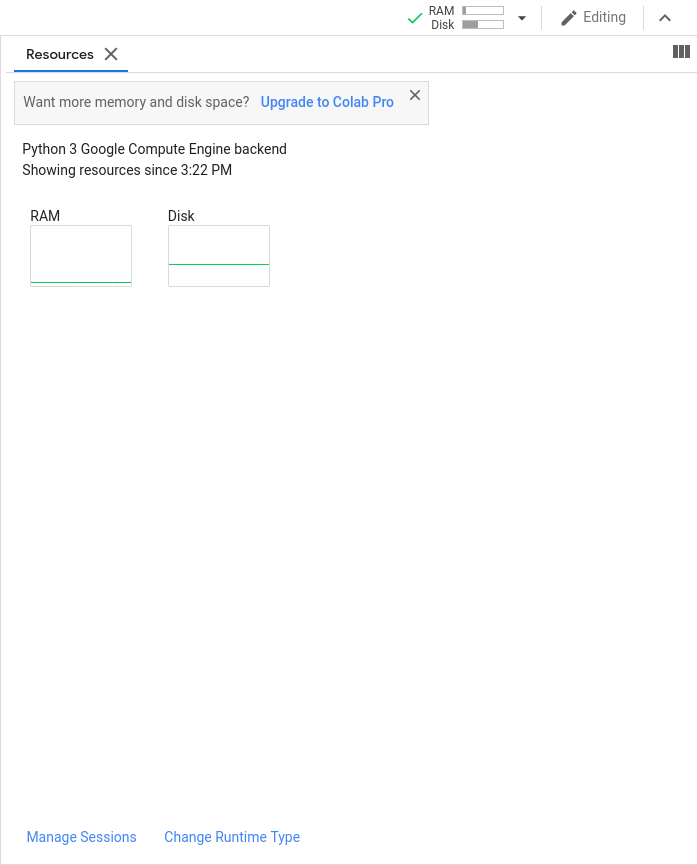
* On the upper right corner of the notebook, there is an icon that says “connect”



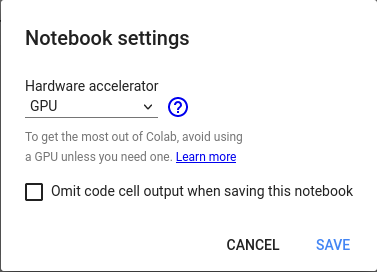
* Once you click on “Connect” it will connect your notebook to a remote Python machine and you are ready to run code. Once it is ready, Colab will display something like this which indicates how the computer is running.



* The default will usually be CPU (typical computer processor). To change to GPU or TPU, click on the bar graph displaying RAM and Disk usage (shown above) and it will open the following sidebar.



* Select “Change Runtime Type” and select GPU or TPU (We only use GPUs in the code in this file, not TPUs). Select GPU from the dropdown menu



* Click “Save” and rerun the code in the notebook.