CSCI 4820/5820 (Barbosa)
Project 7: NLP Application Using Hugging Face Pipelines
(**Optional**: Replaces lowest project grade)

Due: **See course calendar**. May not be turned in late.

Assignment ID: **proj7**

File(s) to be submitted: **proj7.ipynb, proj7.pdf, proj7doc.pdf**
data.zip (if applicable) – see below

Note: Your submitted code must run on the **TAMU FASTER** system – Ensure you test your code there before submission.

****** Clearly identify which semester's SIF is used on TAMU FASTER ******

Objective(s): This assignment will give you hands-on experience in constructing an NLP application using pipelines in the Hugging Face ecosystem.

Project description: Develop an end-to-end NLP application that primarily uses hugging face pipelines. You must use a **minimum of three pipelines** and your code must solve a relevant NLP need. Text pre, post processing and connective/glue code may be used in assembling the application. You may not develop an application that is based solely on the code demonstrated in class and posted in D2L (individual aspects/ideas from that code may be used but the purpose of the application may not be the same).

Note: Due to limited number of large capacity GPUs, it is preferred your code runs on a **T4 GPU** on TAMU FASTER.

With what you have learned in the course, develop an application that solves a real-world NLP problem:

- Immediately after the name/project markdown down cell, add a markdown cell that clearly describes:
  - The problem you are solving – describe why this is a real problem that is worth solving.
  - The dataset used – due to the limitations on file size submissions in D2L, it is preferred that you use a Hugging Face dataset. You may use an external dataset, but you must ensure that it does not exceed D2L size limitations. If you use an external dataset, submit it in a file named ***data.zip***.
  - How you solved the problem – provide details on your solution to include:
    - The pipelines used, why you chose them, and the function they perform in your application.
    - The preprocessing, input, output, and postprocessing of each stage of the application
- There are a number of NLP pipelines available on Hugging Face but there's no central list on their site. Refer to external lists, like the one at the website below, for available NLP pipelines:

  https://lazyprogrammer.me/list-of-hugging-face-pipelines-for-nlp/

- The application idea and its implementation must be yours. However, you may use external sources to generate candidate applications and solve small coding problems. Clearly cite/attribute any use of external code or use of AI tools in the file ***proj7doc.pdf***.
- Place a Markdown cell at the top of the source file(s) with the following identifying information: your name, CSCI <course number>-<section number>, Project #, due date.
- Generate a pdf file of the notebook (from the terminal):

```
$ jupyter nbconvert --to html proj7.ipynb
$ wkhtmltopdf proj7.html proj7.pdf
```

- Submit required files via the D2L dropbox for this project.
- **Test your code on TAMU FASTER - The quality of your application and the pipelines used matter in grading.**