Weekly Mentor + Team Meeting: Every Saturday
5:00-6:00 PM Our github repo: acmucsd-projects/wi23-ai-team-1 (github.com)

Team Meeting Time: 2/11/23 at 5pm in CSE B230
Attendees: **Aniket Gupta** **Arnav Modi** **Jimmy Ying** **Steven Shi** **Vincent Tu** **Vivian Liu**

What have we done so far

- Finished brainstorming project: toxic comment classification
- We're ahead of Team 2 muahahahaha

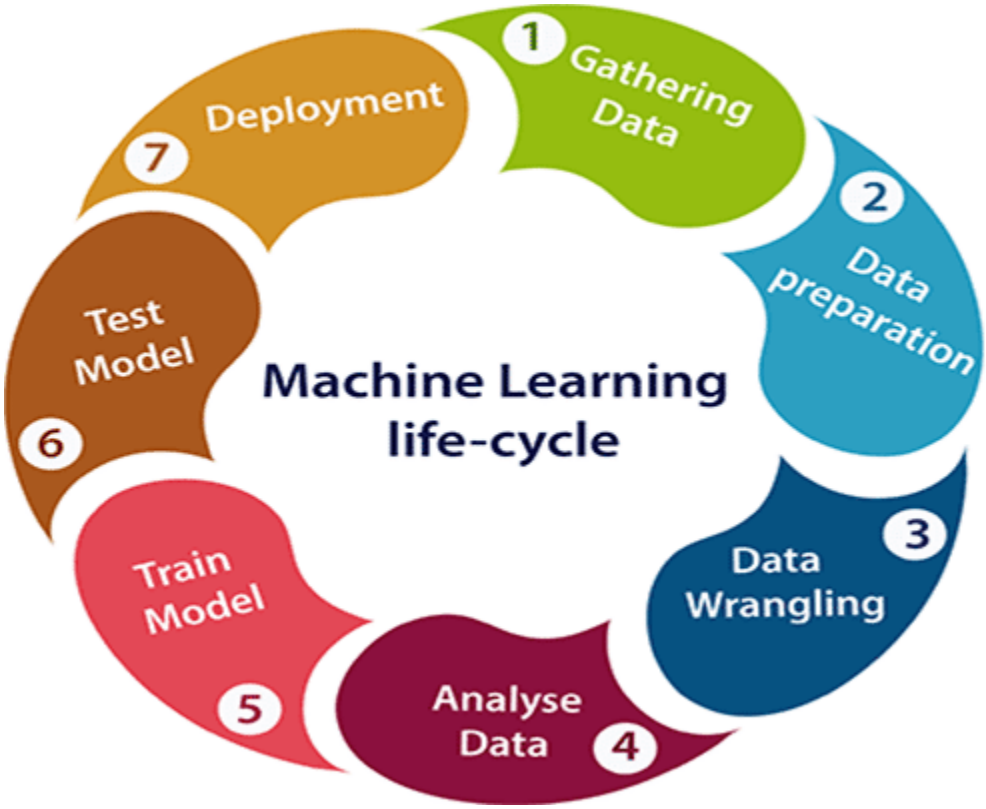
What is the point of this meeting, what are we going to discuss

- Goals for today: logistics, timeline, approach, dividing up work
- Colab or Kaggle Notebook
    - Easier to load data into kaggle
    - Come preinstalled with packages, but might run into problems if versions don't match across environments
        - Can bypass with requirements.txt (list of all the packages and versions)
    - Stick to Kaggle Notebook for now because it's easy to switch from Kaggle to Colab
    - Don't use local laptop (can write code but don't train anything locally)
- To train a good model for this NLP project, you need at least 24 GB. Ideally 32 or 48. At least 12-16 GB GPU memory
- Dataset is around 50MB
- Colab and Kaggle Notebook aren't connected to the repo
    - After a work session, download locally to add it to the repo
- Resources: Vincent, Kaggle (since competition's over and there will be lots of resources)
- Preprocessing
    - Can't feed words into a model, need to turn dataset into numbers
    - Tokenizing: encode words into numbers through vocabulary table (basically a dictionary)
    - Embedding: type of layer in a deep neural network
        - For each number that corresponds to a word, you map it to a vector of n dimensions (n >= 1000) to train the model
    - Tensor: generalization of a matrix so there's more than 2 dimensions, aka multidimensional array
    - All arrays fed into the model have to be the same length

What we did:

1. Make a Kaggle account using your **personal email**
2. Clone GitHub repo to local machine (if you haven't already)
3. Git pull to get the most recent changes
4. Download dataset and put it in your local repo's input folder (extract all and include only the CSVs in input folder). Do not commit the dataset

Timeline:



- **Week till 2/18**
    - Learn pandas, learn numpy, learn matplotlib (don't watch all of them; watch as many as you can)
        - Numpy:
            - https://www.youtube.com/watch?v=QUT1VHiLmmI&ab_channel=freeCodeCamp.org
        - Pandas:
            - https://www.youtube.com/watch?v=vmEHCJofslg&ab_channel=KeithGalli
        - Matplotlib:
            - https://www.youtube.com/watch?v=DAQNHzOcO5A&ab_channel=KeithGalli
    - Work on data wrangling (cleaning the dataset; don't do preprocessing just yet)

- **Week till 2/25**
    - Data exploration (and possibly data preparation/preprocessing)
    - *Check out Competition page "Code" and "Discussion" sections for how to explore, preprocess, and clean the dataset*