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: 2/25/23 at 5pm in CSE B230
Attendees: Aniket Gupta, Jeffrey Lee, Jimmy Ying, Steven Shi, Vincent Tu, Vivian Liu

What have we done so far

- We have cleaned the data
- We have explored the data via plots
- We have begun planning out possible model structures using tensorflow

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

- Don't read and rewrite the data in the clean function
- Everything before modeling can be done with a notebook
- Modeling should be done with multiple python files

There is a fast track way to plug in a numpy array into the model, but we will do it the more tedious, complete way, which apparently exists

We ran through the text classification rnn tutorial, mentioned in discord

Notes on pytorch w/ Vincent's old project

- Put data set in a class that inherits from base dataset
- __init__ constructor
- __len__ (how many instances, images, comments, etc.)
- __getItem__ (get an instance selected via an index, perform some kind of augmentation/transformation on it, returns data and label)
- are only functions needed in pytorch dataset, lots of legroom
- Pass in df for constructor, mostly just attribute setting
- Augmentations, tldr: apply transformation to data, ex. Feed rotated image
- This is to improve model performance and generalize, it knows how to handle more data
- We may or may not have augmentations, text augmentation is kinda sus

Generally, no augmentations to text

- Text is sequential, the order really matters

Try to get data working in tensorflow, if that doesn't work, use python

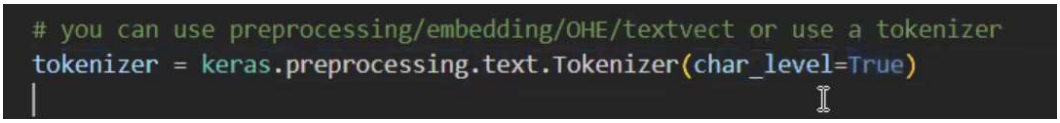
PyTorch	Tensorflow
•	•

Cleaning = remove unintentional errors and problems with the dataset

Preprocessing: take cleaned data, and convert to compatible format

- Same general steps, but different apis

What will we do going forward

- Convert our pandas df, preprocessing →
- Preprocessing & model
- **Decide between pytorch and tensorflow, make a dataset in both**
- Consider augmentation and feature engineering, probably won't be too useful, but fun to know
- Tokenize / text vectorization before preprocessing
  - 

  - 

- **Tokenizer maps keys (word) to values (integer)**