**Bill Yerkes**

**CS5542 Big Data Apps and Analytics**

**In Class Programming –4**

**17th September 2020**

**Submit ICP Feedback in Class. :** [**Lnik to Feed back Form**](https://docs.google.com/forms/d/e/1FAIpQLSesllFh5_STnj7RbHyQainRG_2EIKw1csp8ObP5FWjpVnGVOg/viewform)

**NLP:**

**Use the same data (that we obtained by in source code in ICP3** Data = pd.read\_csv('https://raw.githubusercontent.com/dD2405/Twitter\_Sentiment\_Analysis/master/train.csv')**) and perform the sentiment analysis task on this data using one of the Deep Learning Classifier (Keras Sequantial model) for text.**

ICP Requirements:

1. Data cleaning and preprocessing (at minimum have the following: Removing unnecessary columns or data, Removing Twitter Handles( @user ), Removing punctuation, numbers, special characters, Removing stop words, Tokenization, and Stemming, TFIDF vectors, POS tagging, checking for missing values , train/test split of data). (40 points)
2. Deep Learning Model building, adding right combination of layers, and successfully executing the model to make prediction. (50 points)
3. Code quality, Pdf Report quality, video explanation (10 points)

Submission Guidelines:

Same as ICP 2.

ICP Report:

**What I learned in the ICP:**

I am in the beginning phases of learning about Deep Learning. I watched the class video several times and found other videos and websites to build on what we went over in class. It is safe to say I am not an expert after one lab. I learned that the data has to be in a certain shape to perform each step. I learned more about other libraries which can help with the process and make coding these task easier.

**Description of what task I was performing:**

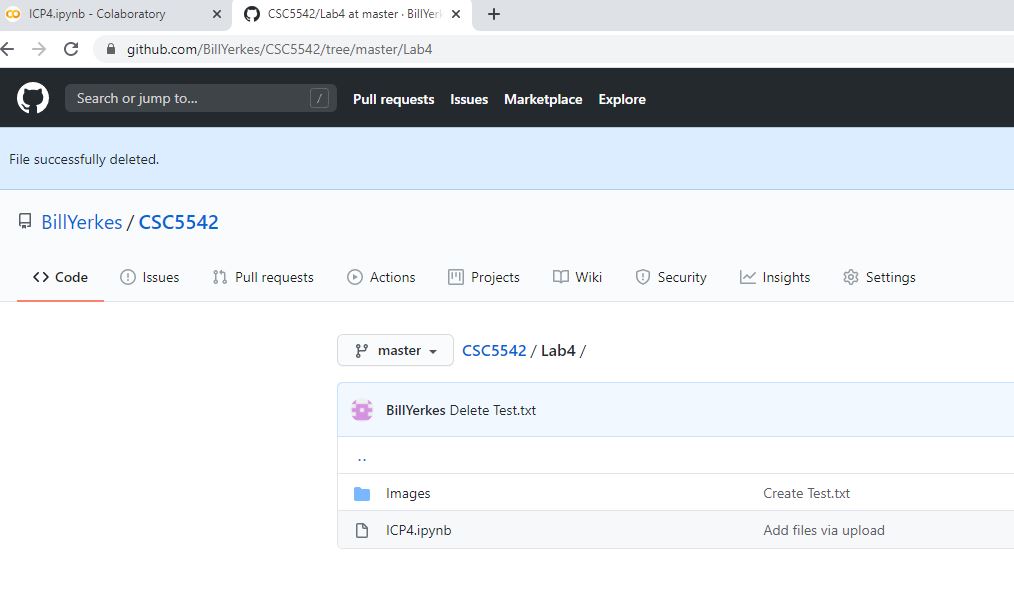
Use the given input file and perform tasks for cleaning analyzing the data using Deep Learning.

**Challenges I faced:**

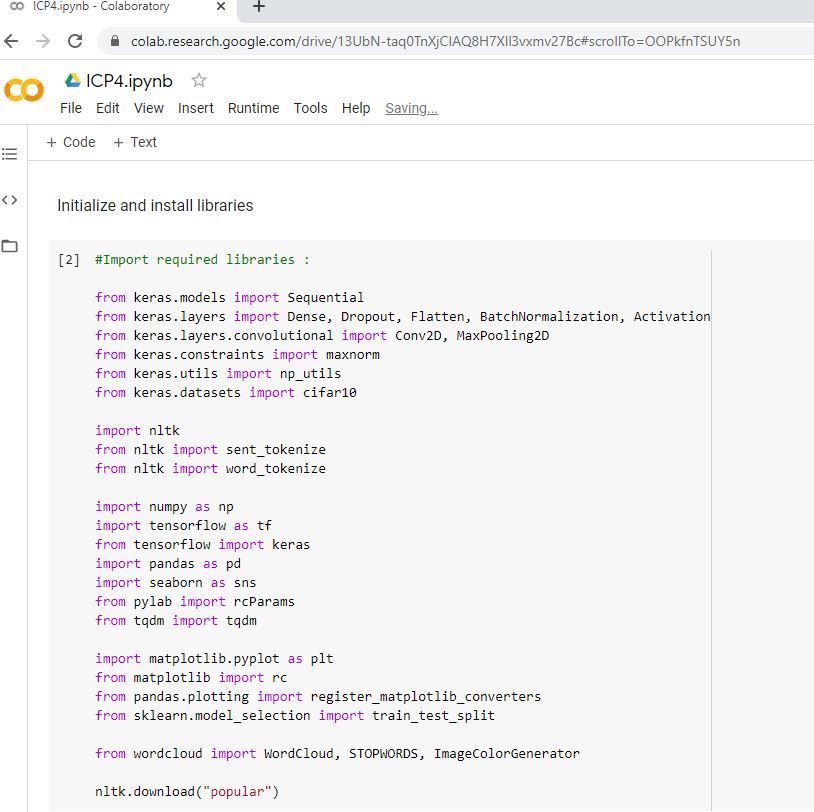
I had to figure out how to convert the sentences/words into numbers/vectors so that I could perform Deep Learning on the data. Figured out how change the shape of the Data Frame.

**Screen Shots**

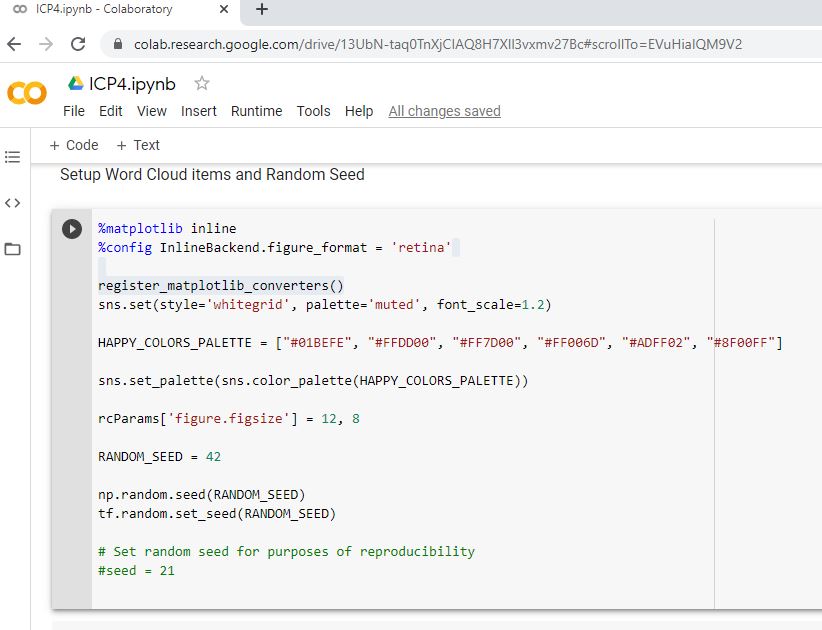
[GitHub Repository](https://github.com/BillYerkes/CSC5542/tree/master/Lab4)



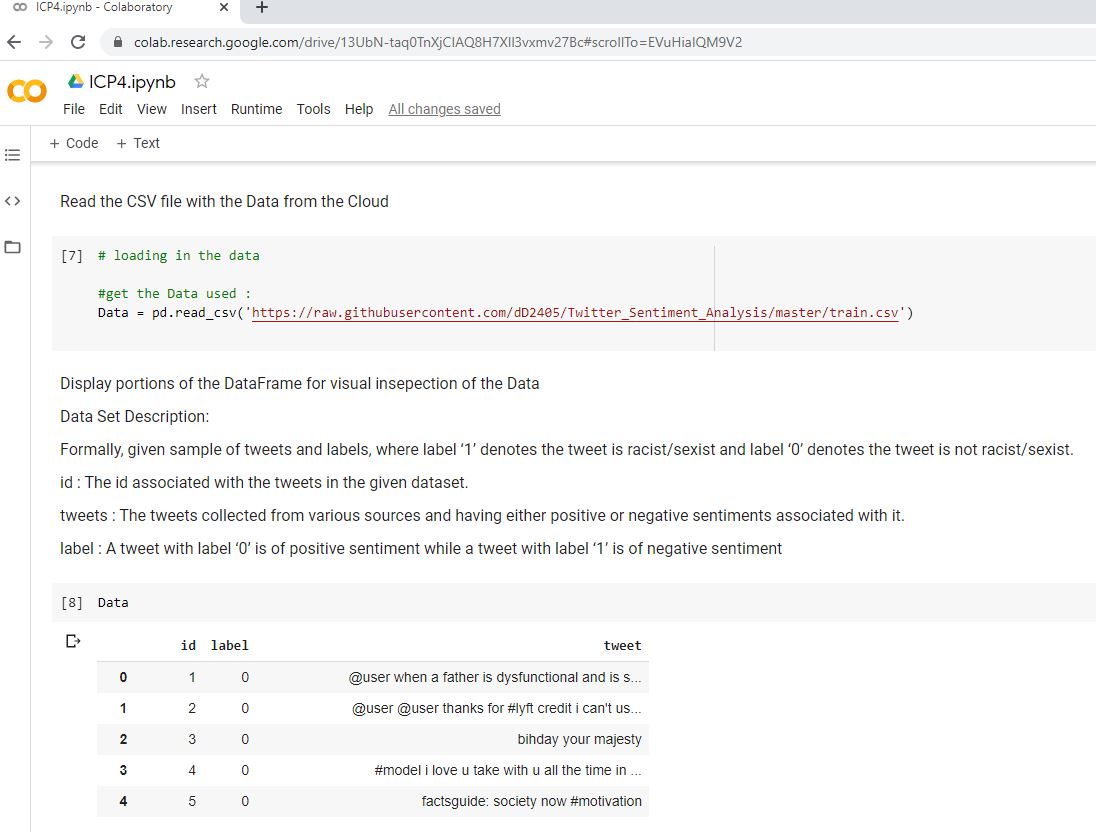
Initialize and Install Libraries



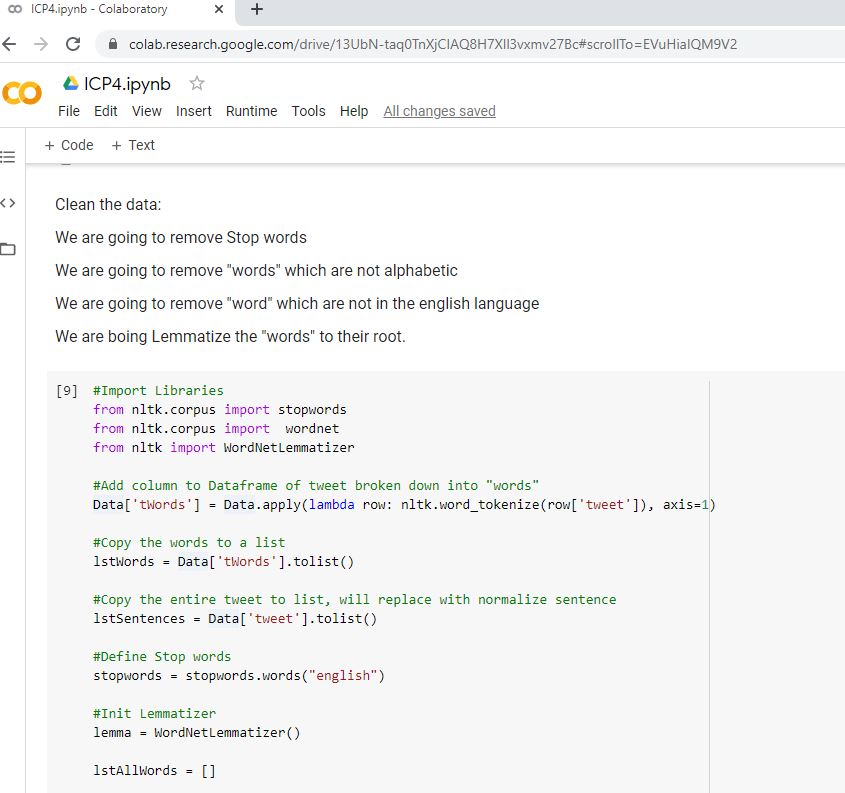
Setup Word Cloud items and Random Seed



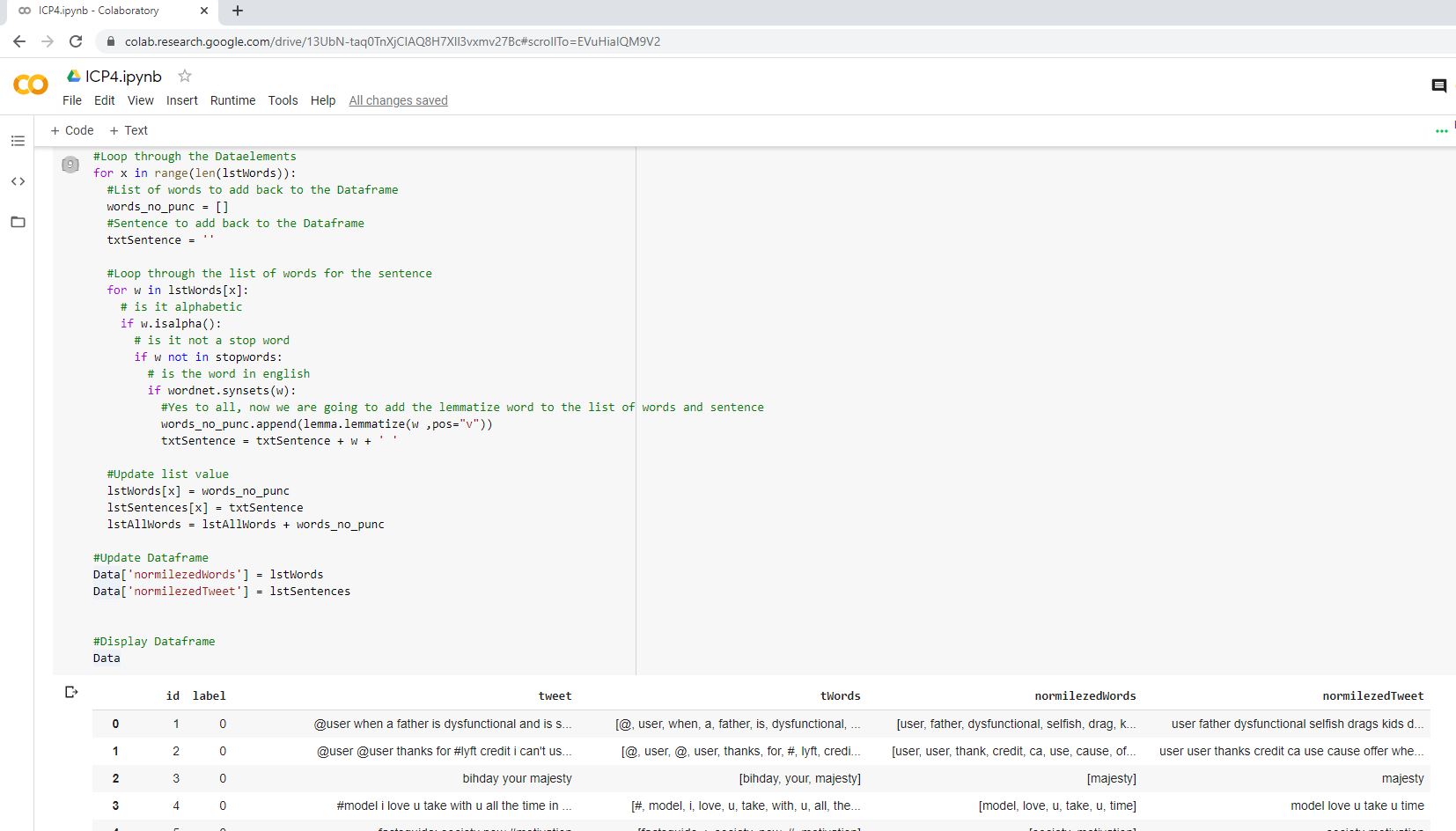
Read the Data



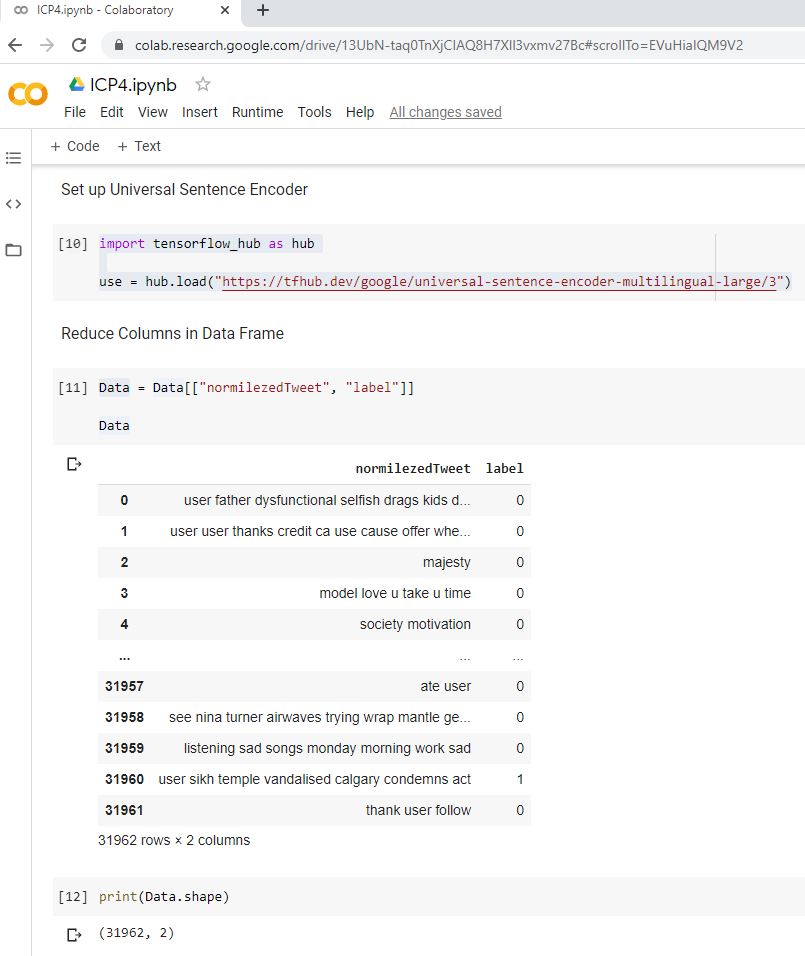
Clean the Data



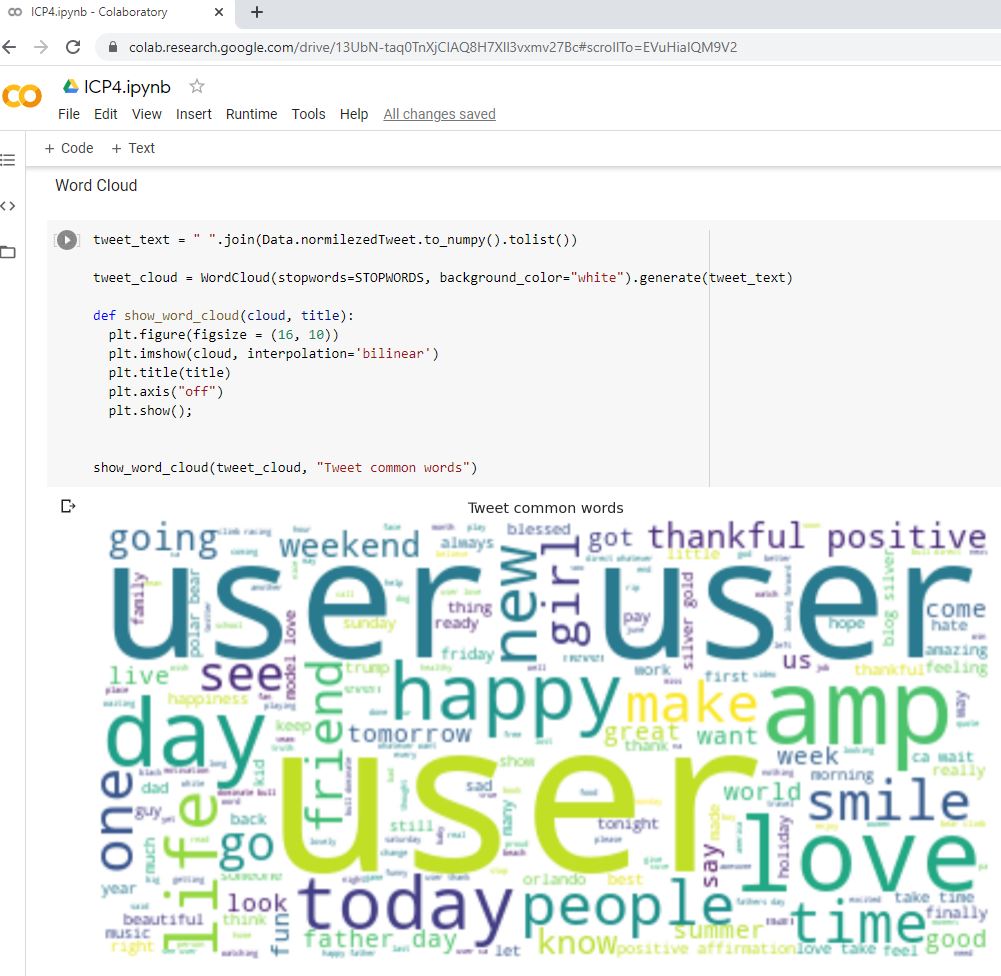
Clean Data Continued



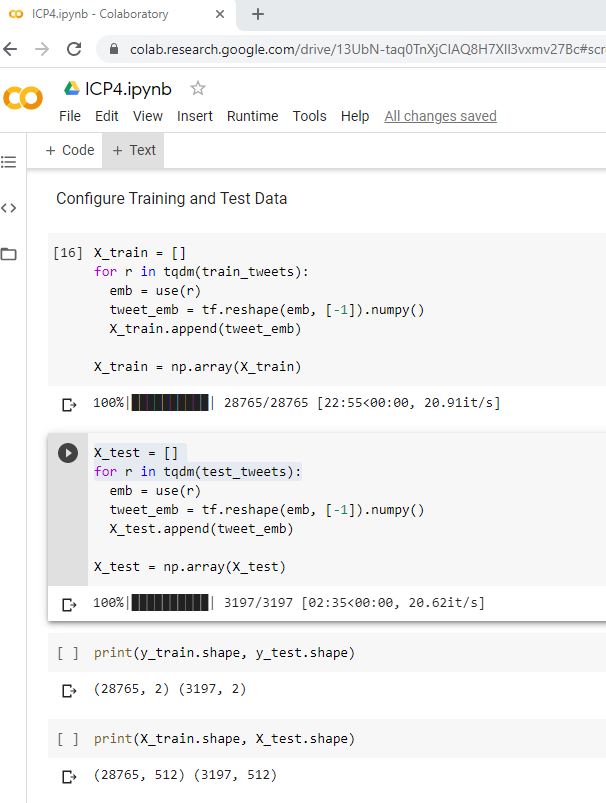
Setup Universal Sentence Encoder

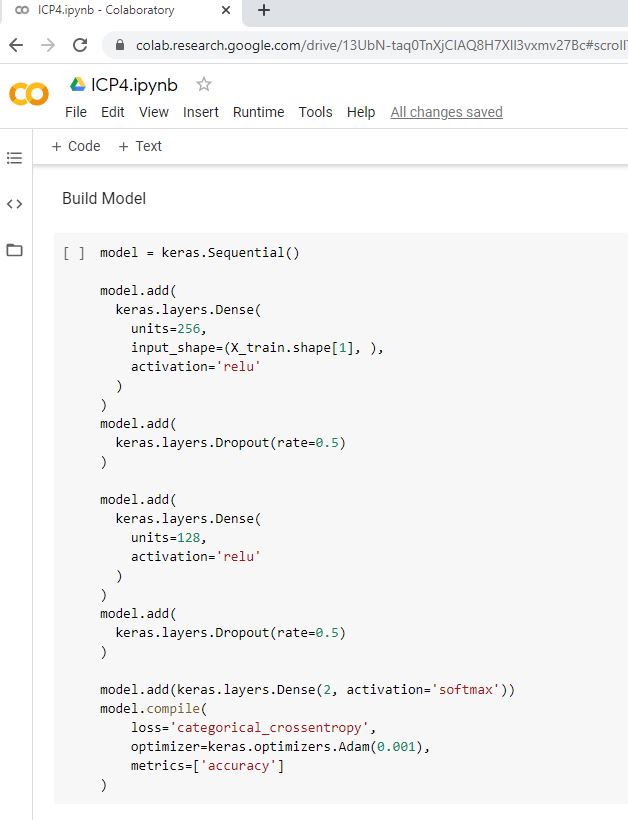


Word Cloud

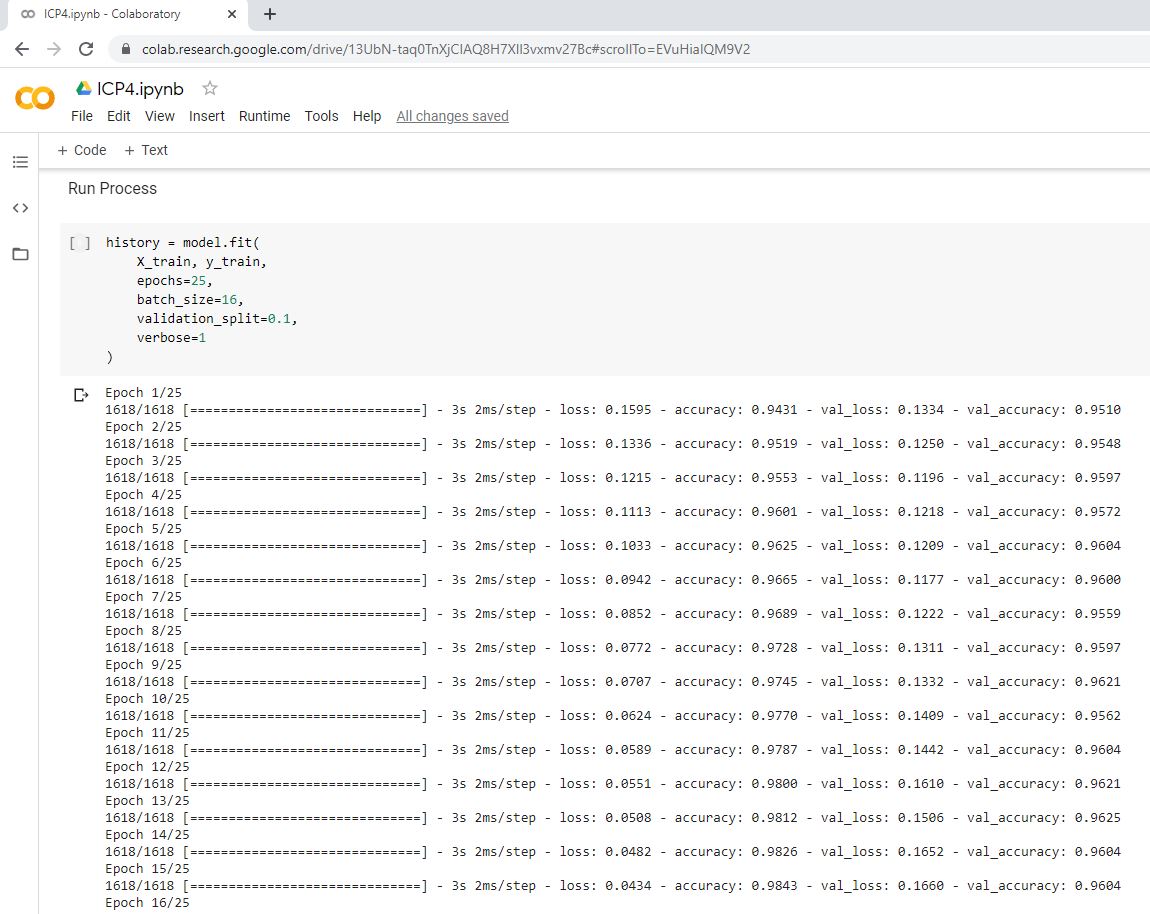


Configure Training and Test Data

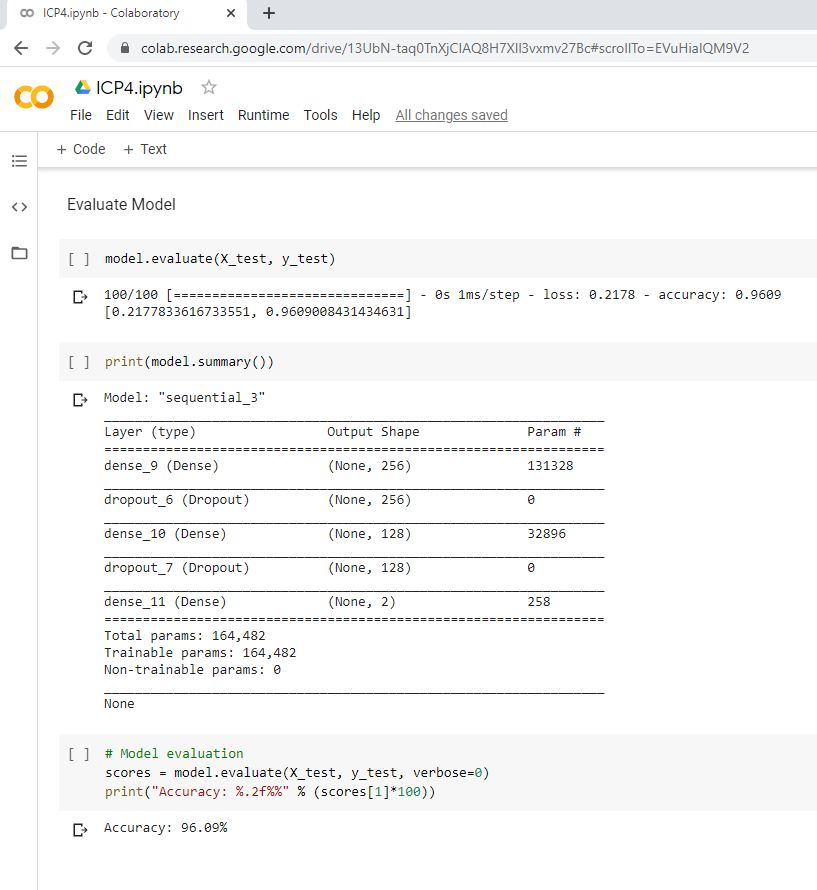


Build Model

Run Process



Evaluate Model



[**Video Link**](https://www.youtube.com/watch?v=V6TzODFhBGM)

**Any in site about the data or the ICP in general**

Data consisted of Tweet text, with the text being categorized as racist or not racist. I am loving CoLab more and more. The number of Python libraries for this area of computer science is amazing. The code this time ran faster then expect and I did not run into memory issues this time.