**CS5542 Big Data Apps and Analytics**

**In Class Programming –4 Report**

**(Jongkook Son)**

# Project Overview:

**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.**

# Requirements/Task(s):

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)

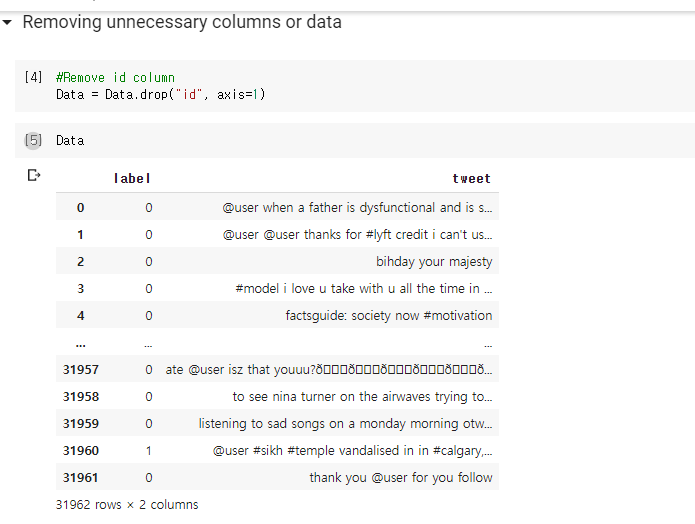
**What I learned in ICP:**

I could have learned the basics of deep learning model and data preprocessing with keras. First of all, I could have learned how to make deep learning model with keras. In this ICP4, I made a sequential model. Also, I could have learned basic structure of the model. I keep thinking about which layer should I add and which activation fuction should I use to maximize accuracy for model. Still I am confused and do not have intuitive about this concept. I can build some basic structure of model. Finally, using the module in keras I learned how to tokenize, which was very meaningful to me.

**ICP description what was the task you were performing and Screen shots that shows the successful execution of each required step of your code**

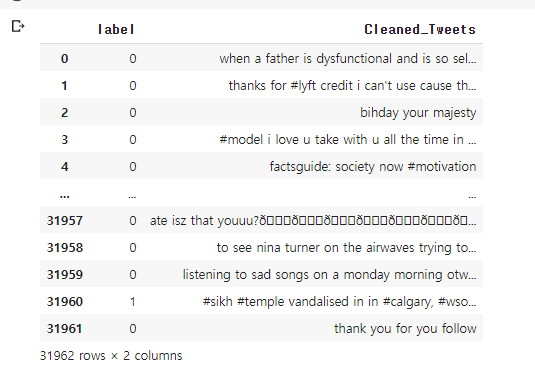
1. Data cleaning and preprocessing

**Remove unnecessary column in this case id column is not necessary**

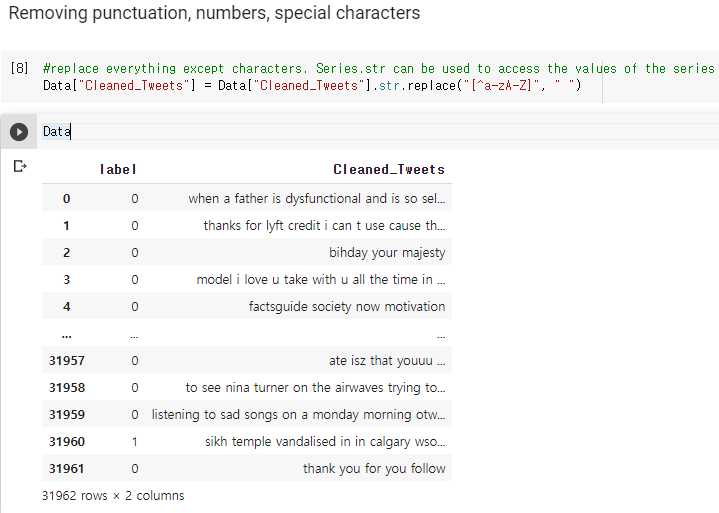


**Remove Twitter handler by using re module**

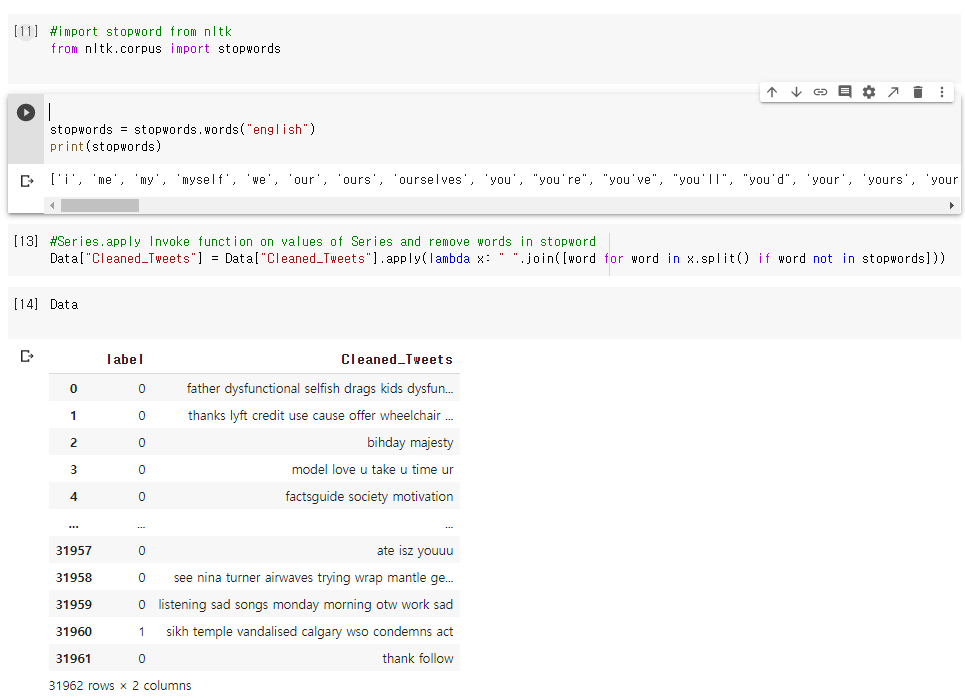




**Remove punctuation, numbers, special characters simply using Series.str**

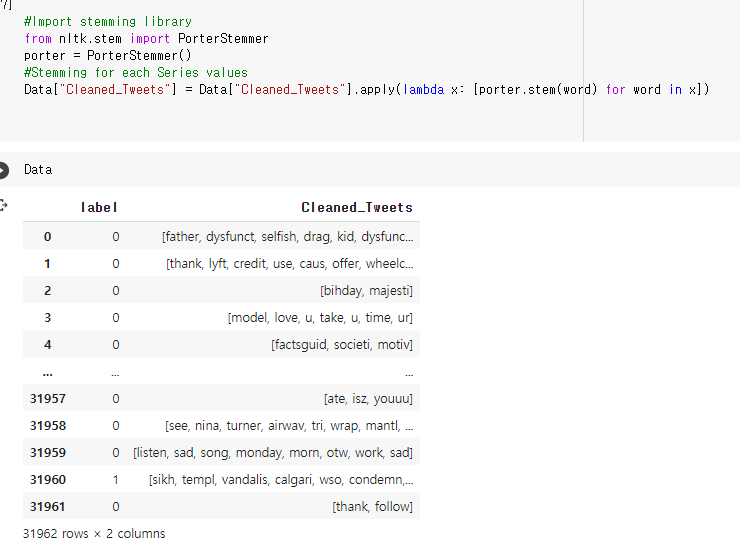


**Remove stopwords which are loaded from nltk**

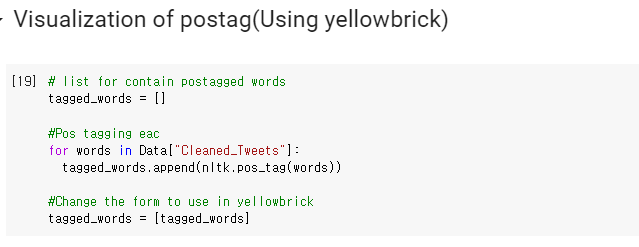


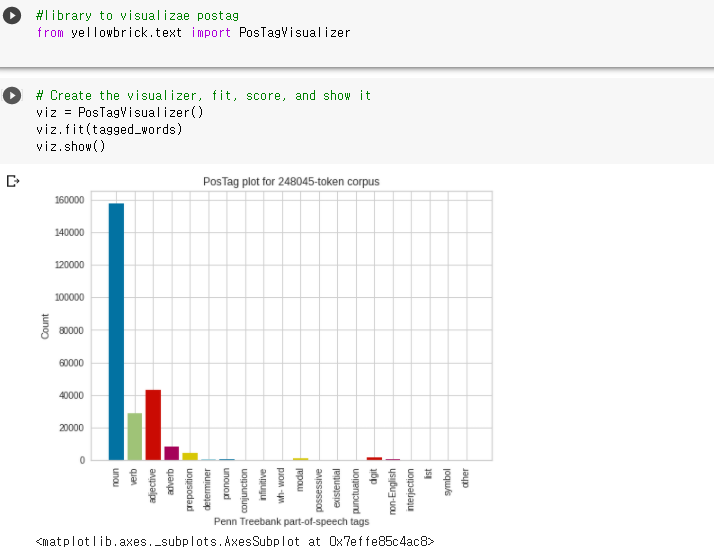
**Tokenization, and Stemming**





**POS tagging**

****

****

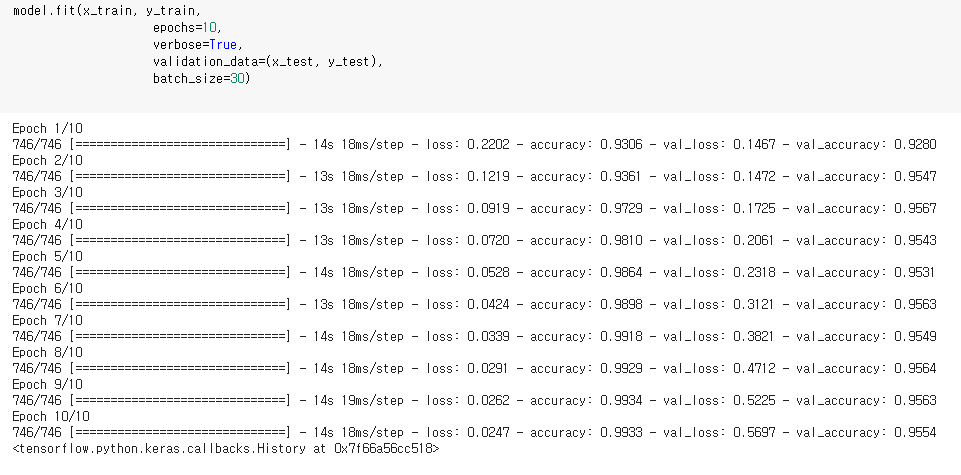
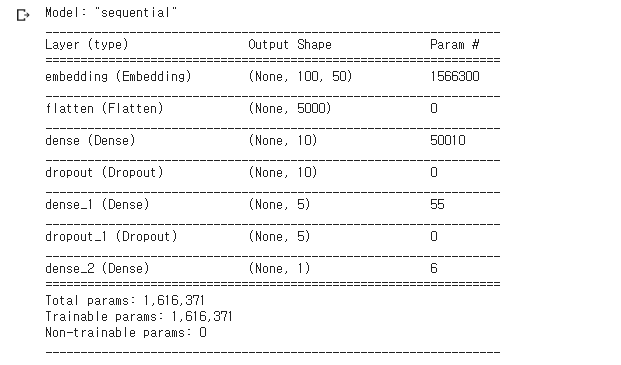
**Tokenization by keras**

****

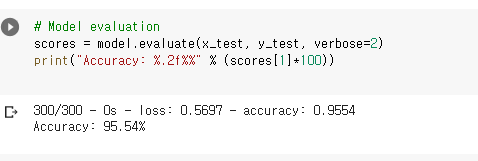
**Train/Test split for data**

****

**Simple deep learning model**

****

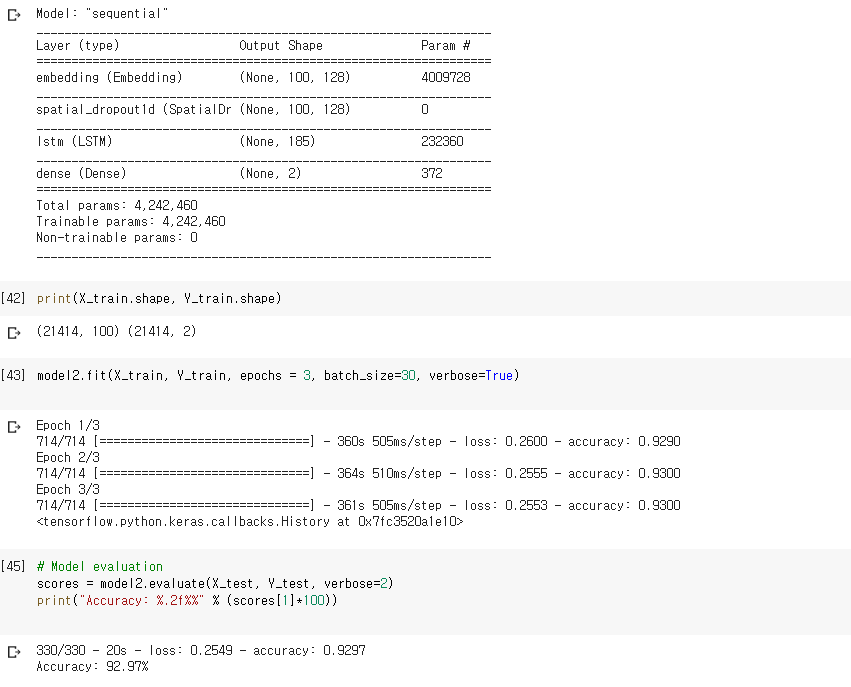
**Model evaluation**

****

**RNN model with LSTM layer**

****

**Model summary and evaluation**

****

# Challenges that I faced:

The most difficult challenge that I faced was that I was not used to building model deep learning. But I overcame this problem by exploring many materials that explain model building. Keras and tensorflow documentation was espescially helpful for me.

# Video link

# <https://www.youtube.com/watch?v=_PJs6zGnpr4>

# 