Environment details:

● Python3: programming language

● MySQL: for Database

● Jupyter Notebook: For python model preprocessing testing

● PySpark: For feathing data from DB and preprocessing it.

● NLTK: For NLP

# Repository:

<https://github.com/RAVIKUMARBALIJA/model-training-service.git>

# What goes in as an input?

Connects to mysql database in the container and fetches the news data and stores into pyspark dataframe.

# How the input is being processed?

All the news- text data gets cleaned by applying various cleaning techniques like stopwords removals, special symbol removals and stemming and then data will be reconstructed for the training.

# What comes out as an output?

A DataFrame/file that has been constructed which could be used as input for vectorization/model training.

# Challenges encountered and the way you tackled them.

* Python to mysql driver issues.
* Issues in setting up preprocessor container.
* Issues in connecting to db within mysql container from another container.
* Data loading into mysql db during db initialization.

Screenshots below while running the containers.









