Clarification and Expectations for Homework 3:

CLASS – USE THIS and not the item posted in class. The item posted in class is incorrect.

\*\* Please read this carefully ☺

For HW 3, you will use Python and CountVectorizer to convert two different datasets into two nice, clean, and labeled dataframes.

So you will end up with two data frames – one for each dataset.

**Dataset 1:** (Topic: Restaurants and restaurant reviews)

<https://drive.google.com/file/d/11H6AbWxKsPLY3yt__OrmK0rjjYShKhig/view?usp=sharing>

**Dataset 2:** (Topic: Movie reviews)

<https://drive.google.com/file/d/17nGHPsk4RXfvRoq-ndizTzc0_0PkzAqm/view?usp=sharing>

Both of these datasets are csv files. However, Dataset 1 is easier to clean and prepare. Dataset 2 is more of a challenge.

Your goal for both is to use **Python and CountVectorizer** and to prepare and format the datasets into dataframe. The first column in each dataframe should be called “LABEL” and will be the label of the data for that row.

**\*\*\* Because I am giving you two datasets to prepare, I am WAIVING the FORMAT – this time only. You DO NOT need to have an Intro or a Conclusions.**

In Analysis you will talk about the two datasets, how they are similar, how they are different, and the challenges you faced in preparing and formatting each into labeled data frames.

In Results you will include images of each of the dataframes (partial because they are huge) for each data set.

Show the BEFORE (the raw data) and the AFTER (the clean and labeled dataframe). Do this for each dataset.

Keep in mind that because the dataset and dataframe are BIG – you will only need to create images that show portions of each.

Best

DrG