Student Information

Name:

Student ID:

GitHub ID:

Instructions

1. First: do the **take home** exercises in the DM2024-Lab1-Master. You may need to copy some cells from the Lab notebook to this notebook. **This part is worth 20% of your grade.**

2. Second: follow the same process from the DM2024-Lab1-Master on **the new dataset**. You don't need to explain all details as we did (some **minimal comments** explaining your code are useful though). **This part is worth 30% of your grade.**

    - Download the the new dataset. The dataset contains a `sentiment` and `comment` columns, with the sentiment labels being: 'nostalgia' and 'not nostalgia'. Read the specificiations of the dataset for background details.
    - You are allowed to use and modify the `helper` functions in the folder of the first lab session (notice they may need modification) or create your own.
3. Third: please attempt the following tasks on **the new dataset**. **This part is worth 30% of your grade.**

    - Generate meaningful **new data visualizations**. Refer to online resources and the Data Mining textbook for inspiration and ideas.
    - Generate **TF-IDF features** from the tokens of each text. This will generating a document matrix, however, the weights will be computed differently (using the TF-IDF value of each word per document as opposed to the word frequency). Refer to this Scikit-learn guide .
    - Implement a simple **Naive Bayes classifier** that automatically classifies the records into their categories. Use both the TF-IDF features and word frequency features to build two seperate classifiers. Note that for the TF-IDF features you might need to use other type of NB classifier different than the one in the Master Notebook. Comment on the differences. Refer to this article.
4. Fourth: In the lab, we applied each step really quickly just to illustrate how to work with your dataset. There are somethings that are not ideal or the most efficient/meaningful. Each dataset can be handled differently as well. What are those inefficent parts you noticed? How can you improve the Data preprocessing for these specific datasets? **This part is worth 10% of your grade.**

5. Fifth: It's hard for us to follow if your code is messy, so please **tidy up your notebook** and **add minimal comments where needed**. **This part is worth 10% of your grade.**

You can submit your homework following these guidelines: Git Intro & How to hand your homework. Make sure to commit and save your changes to your repository **BEFORE the deadline (October 27th 11:59 pm, Sunday)**.

In [2]:
```
### Begin Assignment Here
```