Dear all,

for the final project, you are expected to submit a project which uses a methodology that we have seen in class.

A valid project is one that predicts customer reviews using the reviews of the text and Naive Bayes as a classifier.

**Grade scale**

Description of the data 5%

Data cleaning - text preprocessing 35%

Algorithm implementation and validation 25%

Description of the findings 35%

**Checklist to have in mind**

* Describe your data. What are the variables you are working with?
* Are there missing values? Are there wrongly encoded data? Describe how you approached fixing these issues (if applicable).
* in your specific case, does it make sense to remove numeric characters, punctuation, stopwords, lemmatize, or stem the tokens? Would it make more sense to use single tokens, bigrams, or n-ngrams? Argue why you are making certain choices rather than others.
* What are the features of the algorithm that you are using? Why did you choose this algorithm for your purpose? How does it perform when contrasted to other algorithms/pre-processing methodology?
* Describe your findings. What do they say? How to interpret the results? How would do your results change if you improve pre-processing?

For the final submission, are expected to submit a (.ipynb or .py) file that replicates your analysis and contains a description of your data, data cleaning/text-preprocessing, data visualization, algorithm(s), and findings.

**Rules!**

* Projects need to be pre-approved. Two weeks before the submission deadline you’ll send us a couple of lines about the main idea for the final project and the data you intend to use [link to the data].
* Feel free to google but try to understand the code that you are using.
* Everything that goes into your code should be reproducible. Reproducibility means we should be able to get the same results, from your code (.ipynb or .py). If you decide to collect/scrape your own data, you should also include the code for this step. Note that this is not required and you do not get bonus points for providing your own data.

Also note that if you decide to send a .py we expect a .pdf or .docx that documents your findings, including plots, comments, and descriptions.

Please let us know if there are further questions/comments.