Natural Language Processing in Data Science
Final Project Proposal
Jenny Tsai (G49487749)
November, 2022

- **What problem did you select and why did you select it?**
  By exploring and analyzing the customer reviews on Amazon electronic products, the project aims to find the words and model(s) that can best predict ratings. The reason why this problem is chosen is because the product ratings tend to have a great impact on sales, so the project results can be used in several ways, such product improvement (by comparing the key words in positive reviews vs. negative reviews), marketing (craft the advertisement using the words in positive reviews), and product outlook (by predicting ratings based on current reviews).

- **What database/dataset will you use?**
  The Amazon electronic products customer reviews dataset downloaded from Kaggle.com will be used in this project. The dataset including reviews from 2014 to 2018, with a total of 5000 reviews and 24 features including product brand, category, rating, review helpfulness, etc.

- **What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?**
  A classical, or rule-based model (e.g., logistic regression) will be first used in combination with SHAP to explain feature impact (Interpretability). If the results are not ideal, a customized neural network (NN) model will be implemented (e.g., LSTM) to see if the performance can be significantly improved.

- **What packages are you planning to use? Why?**
  For text pre-preprocessing, nltk and regular expression will be used as they are more efficient compared with hand-written codes. Wordcloud, sklearn, and SHAP will be the main tools used for EDA and modeling. If a NN is pursued, PyTorch will be used.

- **What NLP tasks will you work on?**
  Beside basic text-preprocessing, some EDA type of tasks in NLP such as frequency distribution, word cloud will be included in this project. The second part will be NLP modeling fitting, including logistic regression and possibly other NN models.

- **How will you judge the performance of the model? What metrics will you use?**
  F-1 score and accuracy score will be used to judge the model performance.

- P**rovide a rough schedule for completing the project.**

Week	**Tasks to be completed**
10/31 – 11/6	Find dataset and topic, complete project proposal
11/7 – 11/13	Data preprocessing
11/14 – 11/20	Model fitting
11/21 – 11/27	Draft presentation and report
11/28 – 12/4	Break - prepare for Exam 2
12/5 – 12/12	Finalize presentation and report