INFO 6105 – Final Project
Amazon Fine Food Review
Team: Flash
Natural Language Processing

- *Bhavani Shankar Telaprolu – 001083390*
- *Parag Laxmichand Shah – 001063214*
- *Dongzhe Wu – 001304924*
- *Chinwoo Haan – 001082279*

Project Description:

- The primary goal of this project is to find out whether the reviews present in the dataset are positive/negative
- The dataset used is from Kaggle at https://www.kaggle.com/snap/amazon-fine-food-reviews
- The data in this dataset is in two formats:
  - 1. Comma Separated values (CSV)
  - 2. SQLite
- We are using SQLite Data to perform NLP operations on this data using k-NN model
- Review score < 3 is considered as negative & Review score >3 is considered as positive

Workflow:

- 1. Sort data based on time
- 2. Convert reviews of "Amazon Fine Food Review" dataset into vectors using Bag of words
- 3. Split data into train and test
- 4. Find best hyperparameter by k-fold cross validation
- 5. Apply k-NN model on the train data
- 6. Find accuracy of the model
- 7. Print confusion matrix and plot error plot

Data Cleaning:

- The given dataset has a lot of redundant data which is eliminated using duplicate function
- The dataset contains “HN & HD”, helpful numerator and denominator which states the number of users find this data useful
- Eliminated useless data by checking the condition **HN<=HD**

Text Pre-processing:

- Used **Bag of Words model** for preprocessing the text data present in the Amazon reviews
- Regular expressions are used to remove any punctuations or any other special characters
- Converted all the text to lowercase
- **Stemming** is done by converting the word to it's base word
    - Eg: helpful,helping,helps → help
- **Lemmatizing** is done by grouping the words that are considered as one
    - Eg: San Jose → SanJose
- **Removal of stop words** is done by looping through the reviews
    - Eg: The food is very good
    - Common words like "The,food,is" are removed
- Finally, the review is reduced to lower dimensions

Featurization of Bag of words & k-NN:

- After looping through the reviews and processing of text, the final words are converted into vectors with lower dimension
- Used Cross-validation concept to decide the hyperparameter "K"
- k-fold cross-validation is done
- Eliminated overfitting & underfitting
- Trained/tested using k-NN model
- Checked the accuracy & plotted the confusion matrix

Observations & Results:

- Applied Bag of words to convert text to vector
- Hyperparameter for the k-NN model is 8
- Got accuracy of 85.066667%