**CS181 - Mid-semester Project Proposal**

**Vuong Nguyen and Cam Smith**

1. **Context - why we decide to launch this project:**

We thought it was interesting and it would be good for marketing to know how their products are doing in ways other than just the 1-5 star system

*Tell a lil bit about why we choose this project*

1. **Datasets:**

*Describe the datasets (source, variables, file types). What is interesting about the data that makes you feel this project is gonna make the most out of the data.*

## **Amazon Fine Food Reviews(csv)**

## **Amazon reviews: Kindle Store Category(csv)**

# **Amazon Musical Instruments Reviews(json)**

In this project, we will use 3 datasets which contain Amazon customers’ reviews from 3 types of products: beauty products, kindle store, and musical instruments. Since the 3 dataset have both ratings and reviews from customers, analyzing them will give profound insights as to how customers feel about the service that Amazon provides for each type of product. In the program we will implement sentiment analysis to find out how positive/negative the reviews are. We will utilize nltk (Natural Language Toolkit) to calculate the negative, positive, and compound score of the reviews. Then, we will visualize the result using the compound score with the ratings (from 1 to 5) to see the correlation between ratings and the positivity/negativity of the reviews. In that way, we can objectively assess the customers’ experience with the product.

In order to build an influential marketing campaign and an effective plan for product/service improvement, we will use word cloud to look for the most common adjectives and the most common nouns in each product category to find out what people like about the product provided.

1. **What questions do we hope to answer after this project?**

After this project, we hope to answer:

1. Which type of product receives the most positive review using sentiment scores?
2. What about the products/services they like the most?(Finding the most common adjectives and nouns to see how people actually feel about the products)
3. **What functions and Python module/package do we intend to use?**

sentiment analysis: → nltk → returns score for negative/neutral/positive and pos tag to categorize the word → calc the frequency of the adj and the noun to find what customers like about the products and their actual emotion through reviews

* 1. function to tokenize then remove special characters and stop words
  2. function to combine the sentiment analysis result into the dictionary
  3. function to plot the sentiment analysis

word frequency: → textlib module

1. function for the word cloud
2. function calculating the frequency of 5 most encountered words in the reviews then plot the frequency as a line graph.

Also, there will be 2 more functions to open file that handle 2 different types of data files and function to merge the datasets

*Overview of project*

1. *Open file*

*2. Combine files into DoL*

*3. Tokenize review column, remove special characters/remove stop words*

*4. Find sentiment score and add column sentiment and puts scores in that column*

*5. Find most common adjectives by creating word cloud*

*6. Find average sentiment score from each product and see which has the most positive reviews*