Amazon Fine Food Review:

# Strategy:

# Data Loading:

* Load CSV file
* Connect with SQLite and load database.sqlite

# Data Cleaning:

* Removing duplicate reviews, having multiple reviews at same timestamp, with same UserID, same review text.
* Remove data with, “HelpfulnessNumerator <= HelpfulnessDenominator”
* For Supervised approach, handle the review with Score as 3, consider neutral as a separate class (which makes total classes = 3), or just ignore (making total classes = 2)

# Data Preprocessing:

* Remove html tags
* Special characters + punctuation removal.
* Convert all words to lowercase
* Remove all words with length < 2 (no meaningful word)
* Remove stopwords (make changes in set of stopwords if necessary)
* SnowBall Stemming.
* Do Lemmatization, if necessary.

# Feature Extraction:

* TF-IDF simple (n-grams), probably n = 3
* Word2Vec with average
* Word2Vec with TF-IDF

For Word2Vec, form the dictionary with the local data, meaning consider only the data in our review column.

# Model Application:

## Supervised:

Convert the score column to:

* 2 classes namely, positive and negative
* 3 classes namely, positive, negative and neutral

Deal with class imbalance problem, refer: <https://towardsdatascience.com/yet-another-twitter-sentiment-analysis-part-1-tackling-class-imbalance-4d7a7f717d44>

Train/Dev/Text split ratio --- 90 : 5 : 5

* Random split
* Based on timestamp

### Model:

* Logistic Regression
* RandomForest
* Support Vector Machines
* Neural Network (Multilayer Perceptron)
* K- nearest neighbors
* Naïve Bayes

There can be more models

## Unsupervised:

Ignore the score column, classification will be done solely on base of review and summary.

For this no need to consider neutral class, consider only positive and negative, therefore there will be binary classification.

### Model:

* K-means Clustering
* Hierarchical and Density based clustering
* Gaussian Mixture Model Clustering
* PCA
* t-SNE

<https://towardsdatascience.com/an-overview-of-different-unsupervised-learning-techniques-facb1e1f3a27>

there can be more models.

# Result:

Compare the results of each model supervised and unsupervised with Word2Vec Average, Word2Vec TF-IDF and simple TF-IDF. Based on some common metric.

# Conclusion:

Conclude based on the results obtained, and suggest the best suitable technique for sentimental review classification based on this dataset.