# CSE291: Data Mining Challenge

October 17, 2024

## Challenge Objective

The dataset contains details about restaurants and their reviews. You are asked to design data mining models to predict the restaurant type using the observed variables.

**This challenge ends on November 19 (11:59 PM PT).**

**Link:** https://www.kaggle.com/t/e022579e073a4846bd6e2bafc5639303

## Data and Baselines

- **train.csv**: It contains all the training data that you can use in this challenge. The first column "id" provides you the unique key to identify the restaurants. Different columns show different features/attributes of a restaurant, including free texts, numerical features, and categorical features. There are also many missing values so please conduct some exploratory data analysis (EDAs) first before you work on feature engineering.
- **test.csv**: It contains all the restaurants that you need to predict their types. The format is the same as the train.csv, except that the "label" column has been removed.
- **baseline.ipynb**: It contains a logistic regression model trained on document vectors computed by averaging word vectors of its constituent words from reviews. By running this notebook, you will be able to get predicted.csv as output. This would give a score of **around 0.68**.

You can download all the datasets and baselines from the course website.

## Evaluation Metrics

Your predictions will be evaluated against the ground truth restaurant types. The f1 score is adopted as the evaluation metric.

## Submission Format

You are asked to run your models locally and upload your final prediction file. It is a CSV file with headers of two columns: **Id** and **Predicted**. The first character must be capitalized. The first column corresponds to the id in the test.csv file and the second column contains the predicted restaurant type.

Once submitted, the system will evaluate on a fixed portion (50%, randomly chosen) of the test set and compute f1-score accordingly. Then your score will be displayed on the leaderboard. Please note that the leaderboard during the challenge is **NOT** final. The final leaderboard will be refreshed once the challenge ends. A new f1-score will be calculated based on the other 50% portion which has not been tested yet.

Every day, you can make at most 20 submissions. So please start early and make sure you have enough time to tweak your models and hyperparameters. You will be able to choose 2 submissions for the final evaluation and the system will pick the best score you have.

## General Rules

- No external data.
- No teaming.
- No Cheating.
- Have fun!