CSCI-4502 Data Mining Project Part 1

Title: Fake Google Reviews in Colorado Using Data Mining

Team Members: Amber Tin, Nick Sietsema, Rebecca Toland

Description: We will use the google local review dataset for 2021 in Colorado to find fake reviews and remove or flag them.

Prior Work: Similar work has been done on other datasets but not on ours.

Datasets:

- URL: https://datarepo.eng.ucsd.edu/mcauley_group/gdrive/googlelocal/

Proposed Work:

- Data cleaning
    - Fill in empty cells using attribute mean in same class
    - Handle noisy data through regression
    - Inconsistent data will indicate fake reviews. We anticipate that our algorithm will use regression to identify outliers in spelling and grammar mistakes, unusual spikes in reviews, and mean rating by individual customers.
- Data reduction
    - Remove irrelevant data (business replies)

List of Tool(s):

- Python 3
    - Pandas
    - Matplotlib
    - Numpy
- Github

Evaluation:

We will evaluate our project using accuracy in terms of how well it spots fake reviews, using a baseline of reviews that we have determined meet our fakeness threshold. Our data will have a training set and a test set; the training set is for Colorado and our test set will be Washington, which is a state with a similar demographic profile to Colorado. We will then test our algorithm against the dataset from Washington for 2021. Finally we will visualize our results looking at the outliers using a regression model.