CS 189 Spring 2018 Final Project Proposal (Undergraduate)

Group Members:
Michael Liang
Samuel Tay
Allen Chen
Shun Lin

Objective:
Our objective is to use the concepts we learned in class and in relevant research papers to detect which Amazon product reviews are most likely to be fake reviews.

Goal:
Our goal is to build a machine learning model that can identify fake reviews with high accuracy and provide better online shopping experience for customers.

Data Sets:
Amazon product co-purchasing network metadata dataset from SNAP, https://snap.stanford.edu/data/amazon-meta.html.
Amazon Commerce Review set Data Set, https://archive.ics.uci.edu/ml/datasets/Amazon+Commerce+reviews+set.
Amazon Review Data, http://jmcauley.ucsd.edu/data/amazon/.
Potentially useful public dataset from AWS, https://aws.amazon.com/datasets/.
Potentially useful opendata for machine learning task, https://deeplearning4j.org/opendata.

Algorithms Planning to Implement/Use:
This problem will be a classification problem (fake review vs real review) and thus we can apply concepts we learned in class such as Neural Network, logistic regression, SVM, least square VM, Generative models, and K-mean cluster (in our case it will be 2-mean cluster). If we have time we will try to implement different classification models to understand which approach is the best.

Citation:

Playing with 80 Million Amazon Product Review Ratings Using Apache Spark, http://minimaxir.com/2017/01/amazon-spark/.

Amazon's Fake Review Problem, http://www.brianbien.com/amazons-fake-review-problem/.

An empirical study on detecting fake reviews using machine learning techniques, http://ieeexplore.ieee.org/document/8102442/?reload=true.

Fraud Detection in Online Reviews using Machine Learning Techniques, http://www.ijceronline.com/papers/Vol5_issue5/J055052056.pdf.

AWS tutorial, https://aws.amazon.com/getting-started/tutorials/