Part Time Bandits: Project Proposal

Group:

- Alex Spence, ajs811 (Member responsible for uploading submissions)
- Sammie Kim, sk7327
- Antonio Robayo, amr1059
- Tim Connor, tfc276

Summary of Plans:

- **Which project you have chosen:**
  - Project 1: Fake Review Detection

- **Proposed Approach and Suggested Experiments**
  - **Data Cleaning:** Limited, if any, data cleaning is anticipated because the data is already presented in an organized format. The data will be loaded into pandas dataframe/numpy nd arrays using Python.

  - **Feature Reduction:** Dimensionality reduction will likely be left out of scope, though feature reduction may be applied to remove short words, infrequent words, and/or mispelled words. Hypothetically, classification could be used to segment products and/or users into groups (based-on features derived from the text or volume of reviews).

  - **Feature transformations:** Create sparse word frequency and tfidf dictionaries based on the "review" column. Additionally, non-textual data may be generated, such as the number/frequency of reviews per user, the number/frequency of reviews per item.

  - **Data modeling:**
    - Split the data into 80% training, 10% validation and 10% test set. If we build any features that revolve around a user's, or item's review history, time will be an important decision to consider when splitting the data.
    - Develop baseline models using Scikit-learn for Naive Bayes (baseline model), logistic regression (with l1 & l2 regularization), and support vector machine.
    - Develop raw code for the aforementioned models
    - Iterate through the different models, feature sets, and hyperparameters to select the configuration which, when fitted to the training data, best minimizes AUC and average precision on the validation set(s).
    - Compare performance of the scikit learn baseline models and the "from-scratch" implementations. Any difference in our validation metrics will be noted and explained.
    - Once an optimal configuration (feature transformations + model + hyperparameters) is found, we will retrain the model on the entire training set and evaluate it using AUC and average precision on the test set.

  - **Final Submission:** A four page report will be developed to describe the methodology and approach of the above. The report will close with evaluation of findings and possible next steps.