Andrew Burns
Alexander Vansteel
Jacob Zelasko
February 21st, 2017
CIS 365 01

Project 2 Writeup

Project 2, while having a bit bigger of a scope than writing agents that try to capture squares on a board, seemed to go a lot smoother than Project 1. This project explored the use of the SnowBall library for stemming words, KFold Cross Validation for training and testing, and GridSearchCV for the final training and testing of the data.

The SnowBall import is used in the tokenizer function to create the stemmer, reducing redundancies in the vocabulary table. Earlier iterations of the tokenizer included stripping the HTML markup from the 'review' table; however, no noticeable improvements were made. During a redesign, the code was removed and never added back in.

The initial attempt at KFold Cross Validation was made using the KFold import from sklearn. The initial code attempt for the loop can be found in the code from lines 60-64. This was able to successfully declare the KFold model, create the X and y training and test arrays with numpy. However, we were not able to successfully train, fit, and score the data due to errors with vocabulary size. At this point, we set aside the KFold framework to look at making other improvements for progress with the intention of coming back to KFold with superior results from the basic framework.

Following the suggestion of looking at GridSearchCV, we began creating the framework for implementation. The param_gs has many of the same parameter options and the TfidfVectorizer used in the pipeline. Due to the similarities, the stop_words and tokenizer are taken care of in the pipeline. Implementing the tokenizer and including the stop words in the parameters for the grid search as well as the vectorizer actually caused the search to crash and fail. Before implementing the grid search, changing the vectorizer to 'unicode' from 'ascii' for strip_accents improved the results, and after implementing the grid search, using 'ascii' in the grid search was a slight improvement over 'unicode'.

Upon looking into GridSearchCV, we noticed that the function is able to implement KFold Cross Validation, or Stratified KFold Cross Validation through one of the parameters in the function call. Runs of the program were made with the cross validation set to be 5, 10, and even a test at 8. The result had no difference between the different numbers.