Assignment 3: Feature Engineering and Practical Issues

For this assignment, you will conduct programming tasks in Jupyter Notebook and answer provided questions in a Google Doc or MS Word document. You will submit **2** documents to Blackboard:

1. A PDF file converted from your Jupyter Notebook.
2. A PDF file converted from your Assignment 3.1 Template Word Doc (document is linked in Blackboard assignment prompt).
    - This template has a copy of all short answer questions for this assignment.

Clothing Data

For this assignment, you will work with data from a popular clothing website, modcloth.com. This gives us a chance to work with feature engineering, feature extraction, and unbalanced, multiclass datasets. You will train several different classification models based on different types of data from this dataset.

As you go through the assignment, you will create several tables and figures. After you complete the programming section, use the tables and figures you generated to answer the following questions.

Written Questions

(short answer, 2-3 sentences):

Using the confusion matrix and report in the Categorical Data for Reviews section summarize the results you obtained using the balanced and unbalanced perceptron. Report it in terms of both f1 score and accuracy.

The unbalanced perceptron seems to perform better based on the confusion matrix, and has a slighter higher accuracy. The unbalanced perceptron also has higher f-scores for each category across the board.

Of all the methods used on the clothing ratings dataset, which of them performs best as a classifier: would you recommend this system to predict product ratings?

The categorical classifier with an unbalanced perceptron gives us about 33% accuracy with relatively higher F1 scores. We could also consider the balanced classifier for categorical data, which gives us incredibly high results for category = 5, if we wanted to reframe the problem as a binary classification challenge of 5 vs. !5.

Short answer, 1-2 sentences each.

In the clothing ratings, you will plot height and waist size (plots appear below). What transformation could be applied to the waist data, to make it look more like a normal distribution?

Scaling and centering the variable is likely to result in a normal distribution.

You will obtain a plot of accuracy vs. the number of words chosen for count vectorizer and tf-idf vectorizer (or refer to the plot in the last section of this document). How many features are used to obtain the best accuracy on this dataset?

1000 words provided the highest accuracy with parsimony accounted for, accuracy did not improve significantly after k = 10000.

In the text classification section, you printed a list of the top ten words. What are they, and do you think these words make sense in the context of clothing reviews?

```
['cheap',
 'disappointed',
 'love',
 'perfect',
 'poor',
 'returned',
 'ripped',
 'terrible',
 'thin',
 'was']
```

All of these words can be reasonably related to clothing reviews – while love, cheap, disappointed, etc. can be generically applied to other goods, returned, ripped, thin, etc. are more likely to be specific to articles of clothing and the retail industry.