NLP: Text Classification

Business Problem
Building and producing products that are actually adopted by customers and solve real problems for them is a historically challenging task. Today, imagine that you have joined the machine learning team on the Amazon e-commerce site! Your webpage is full of reviews from customers for each of your products. Your product owners want to know about a negative review **immediately**. Ideally, they'd like to know why the review was negative.

Supervised & Unsupervised Machine Learning
Your research team just finished labeling a set of data. Your task is to find the best classification system, leveraging a wide array of machine learning models. In particular, because your data has already been cleaned, you will be asked to explore at least 2, if not 3, different modeling strategies to find the best way of describing your data. In particular you will focus on

- Blazing Text Supervised Classification
- BlazingText's Word2Vec Unsupervised Model to generate word embeddings
- Linear Learner for 1:1 model interpretation, using word embeddings as features
- Clustering negative reviews to extract topics and key words

Explore each model for their quality in prediction, and their run-time in training. Use the SageMaker Neo compiler to get the smallest version of your model possible,

Data Sets
The dataset you'll be working with comes directly from the Amazon review site. This is hosted on AWS through coursework via fast.ai https://course.fast.ai/datasets. Navigate to this page and click download for **Amazon Reviews: Polarity.** The Amazon reviews polarity dataset is constructed by taking review score 1 and 2 as negative, and 4 and 5 as positive. Samples of score 3 is ignored. In the dataset, class 1 is the negative and class 2 is the positive. Each class has 1,800,000 training samples and 200,000 testing samples.

Existing Research
Your research team just developed an innovative model that uses convolution to classify text. See this page for further details. http://xzh.me/docs/charconvnet.pdf

Sample Code
Code from your researchers is available here. https://github.com/zhangxiangxiao/Crepe
Download your data from the site, upload it to an s3 bucket via the AWS console, and then run this block of code on your SageMaker notebook instance to read the data into a pandas data frame.

```
import pandas as pd

!mkdir /Data
!aws s3 cp s3://nlp-workshop-reviews/amazon_review_polarity_csv.tgz /Data
!tar -xvzf Data/amazon_review_polarity_csv.tgz
df = pd.read_csv("amazon_review_polarity_csv/train.csv", names=["Label", "Title", "Rev
```