# Topic Modeling on Customer Review Data: NLP

**Business Problem**
Building and producing products that are actually adopted by customers and solve real problems for them is a historically challenging task. Today, image that you have joined the machine learning team on Amazon e-commerce site! Your webpage is full of reviews from customers for each of your products. Your research team just finished an extensive text classification example, so you are leveraging that existing work and extending it into unsupervised learning. Your task is to identify topics, key words, and insights from this data set to improve both your product offering and your customer experience.

**Unsupervised Machine Learning**
Here, you will engage a number of different unsupervised machine learning methods to extract the best information from your dataset. 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

- Topic Modeling
- BlazingText's Word2Vec Unsupervised Model
- Nearest Neighbors

Eventually you would like your end users to be able to define their own topics, and have your system identify the documents that most closely relate to those topics.

**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
```

**Unsupervised Machine Learning**
This is a tricky area of machine learning; model evaluation is quite different from classification. Oftentimes key metrics here are usability of the model, rather than a numerical indicator for its quality.

- Topic modeling
- Blazing Text

- Similarity Search