## **Overview**

The goal of this homework assignment is to demonstrate your ability to create a machine learning pipeline, train a model and optionally stand-it up for inference.

For this assignment you will be working with the *Apple Computers Twitter Sentiment* dataset. We’d like you to write a ML pipeline using any programing language you’re comfortable with that:

1. Reads the dataset into memory
2. Computes some set of engineered features - example features:
   1. Use regex functions to extract some specific terms
   2. Compute word embeddings (i.e. word2vec, GloVe, BERT)
   3. Tfidf
3. Trains any classifier to predict the `sentiment` class
4. Saves the model to some format that can be used for scoring
5. Stand-up an API service that has:
   1. Input: plain text
   2. Output: prediction of sentiment + the probability of that prediction
6. Deployment of model + api service should be dockerized

Even though at Carvana we dedicate a lot of effort to making our ML models as accurate as possible, for this assignment we are most interested in the process of how you piece all of these components together.

## **Expected Output**

The final result should be a Python script that we can run locally to replicate your work. We would like to be able to understand your approach to training the model as well as the steps you take to achieve the best accuracy of the model.

Once you complete the assignment, please send it back to Joey ([joey.meyer@carvana.com](mailto:joey.meyer@carvana.com)) with any instructions on how to run the script. If you advance to the on-site interview, we’ll ask you to give a short presentation about your approach and other techniques you would have tried if you had more time.

Good luck!