Unified Demonstration

PCF, PCC, and Greenplum with Machine Learning

Purpose

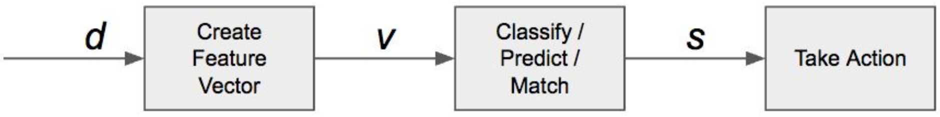
The purpose of this demonstration is to provide a jumping off point for conversations that address the broad spectrum of capabilities enabled by Pivotal through PCF, PCC, and GPDB or HDB.

Introduction

The terms machine learning and artificial intelligence encompass a broad set of technologies. The essence of machine learning is a set of algorithms that can learn from data and then make a prediction from new data. The data science process behind this typically involves a set of input data from which features are generated. An algorithm is applied to provide the prediction or classification. The result of the algorithm may then be used to determine what is shown to the user or to take action.

For some use cases, the model scoring or classification can be done as a batch minutes, hours, or even days after the data is collected. Other use cases require a near real time scoring of the model. For instance, someone has just placed an item in their shopping cart, what should be presented to them as other items to consider? If I have just placed an power drill in my cart, the system might identify a set of drill bits and safety goggles and suggest that I consider those items for purchase.

The streaming data processing pipeline for such use cases can be viewed as something like this:



Creating the feature vector may be as simple as transforming the data that arrives or may involve enrichment. For this discussion, we will focus on simple transformation. We will return to the need for enrichment during our discussion at the end of the presentation.

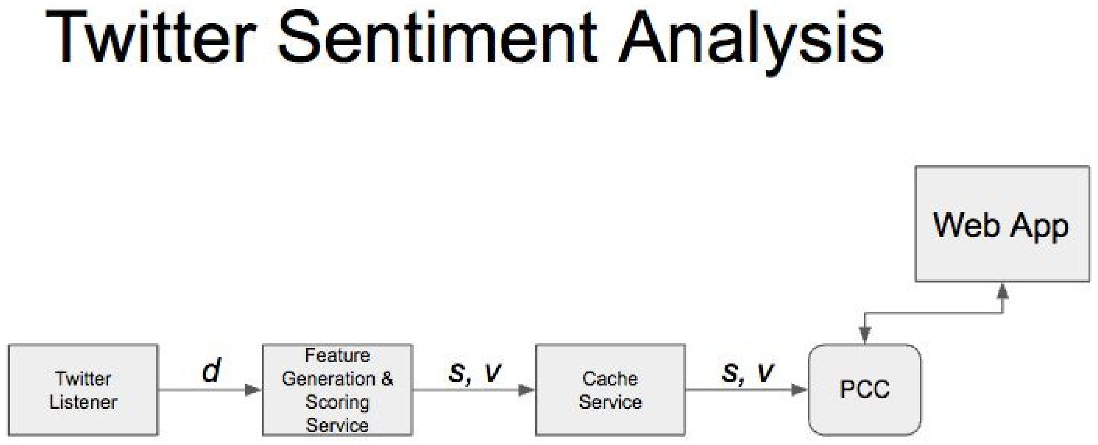
Twitter Sentiment Analysis

This demonstration will look at sentiment analysis of tweets. We will assume that the organization wants to analyze this in near real time and show the feed in a web UI as a timeline of sentiment vs. time of the tweet. To understand the amount of activity, they will present the rate of tweets as well. And, a selection of tweets will be shown so that users can see specific examples.

Perhaps they want to deal with highly negative tweets quickly, promote the positive tweets, or even adjust their marketing spend based upon what is being learned. Similar use cases might classify tweets based upon the subject matter to monitor events in real time. The general pattern can be extended to streaming data in general.

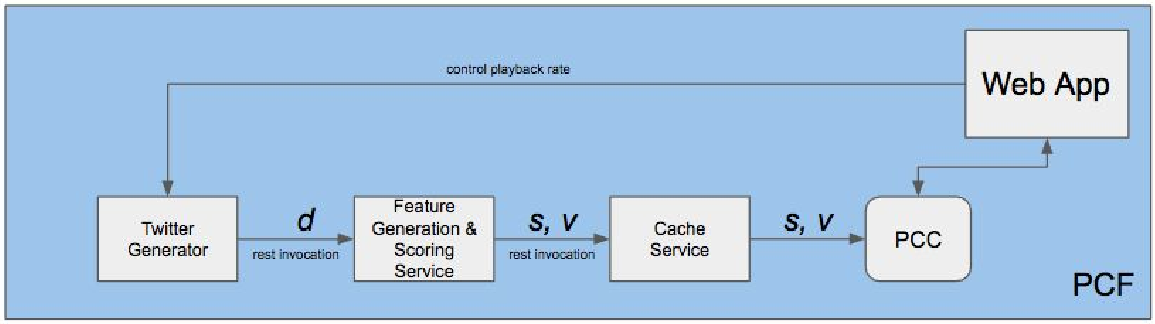
Each tweet will be scored to determine whether it is positive (+1), neutral (0.5), or negative (0.0).

The processing pipeline is like this:

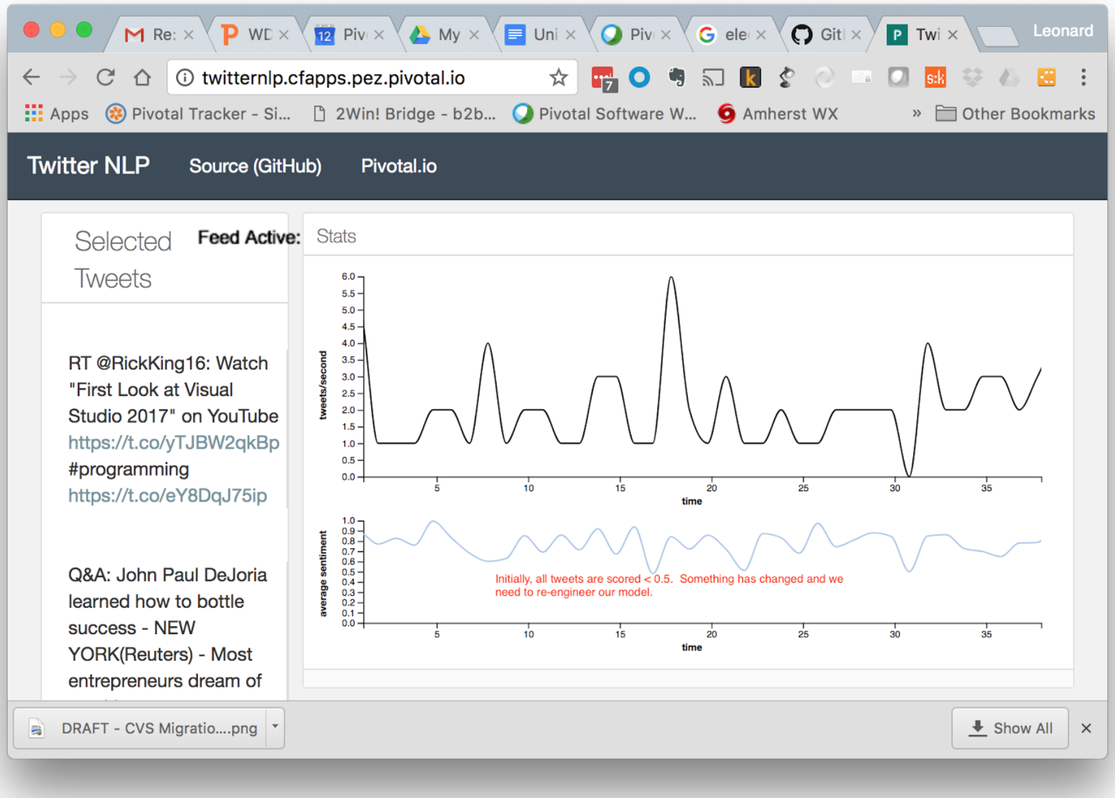

Twitter Sentiment Analysis
Web App
Twitter Listener
d
Feature Generation & Scoring Service
s, v
Cache Service
s, v
PCC

For the demonstration, we will simulate the twitter feed by playing back recorded tweets. Also, we will deploy the application in PCF:

Demo Twitter Sentiment Analysis



The UI looks like the following.



RT @RickKing16: Watch "First Look at Visual Studio 2017" on YouTube https://t.co/yTJBW2qkBp #programming https://t.co/eY8DqJ75ip

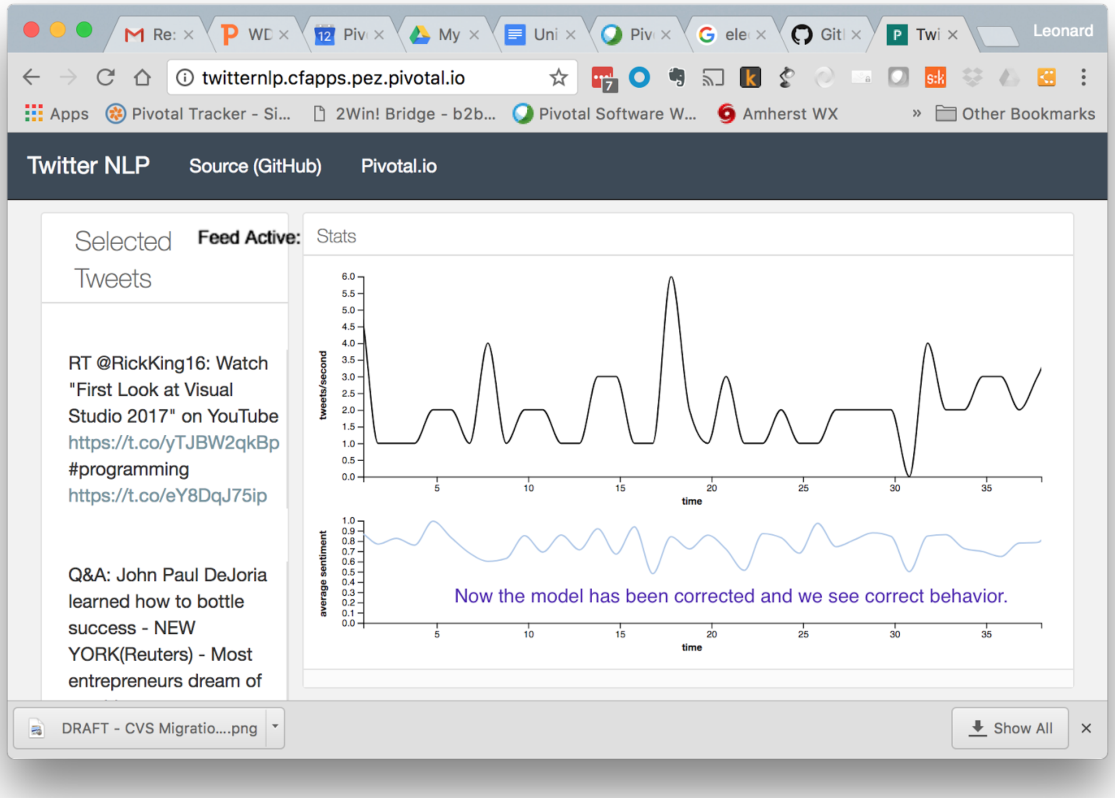
Q&A: John Paul DeJoria learned how to bottle success - NEW YORK(Reuters) - Most entrepreneurs dream of

Initially, all tweets are scored < 0.5. Something has changed and we need to re-engineer our model.

Note that there is a problem with the model. The sentiment scores are all negative. We need to engineer a new model.

… insert GPDB / Jupyter demo here….

Now, we push the updated model to the production environment. The technology allows us to move as quickly as we desire.

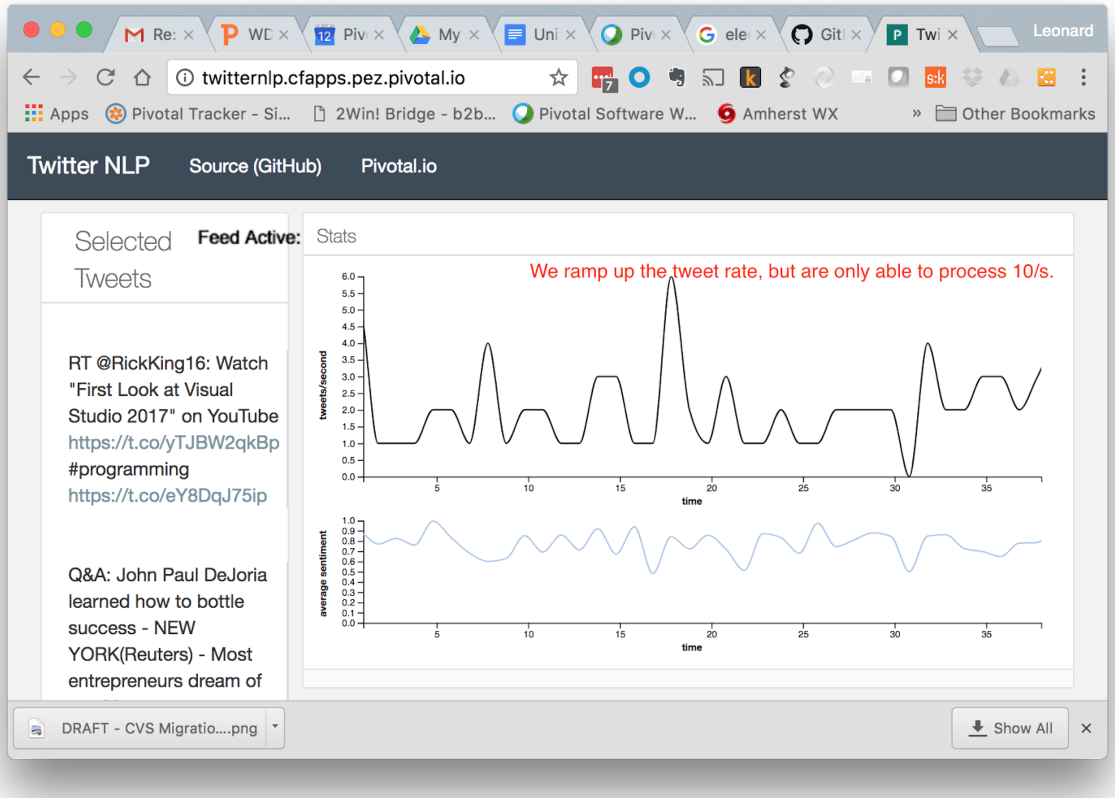


RT @RickKing16: Watch "First Look at Visual Studio 2017" on YouTube https://t.co/yTJBW2qkBp #programming https://t.co/eY8DqJ75ip

Q&A: John Paul DeJoria learned how to bottle success - NEW YORK(Reuters) - Most entrepreneurs dream of

Now the model has been corrected and we see correct behavior.

Now, let’s increase the rate at which the tweets are played back:

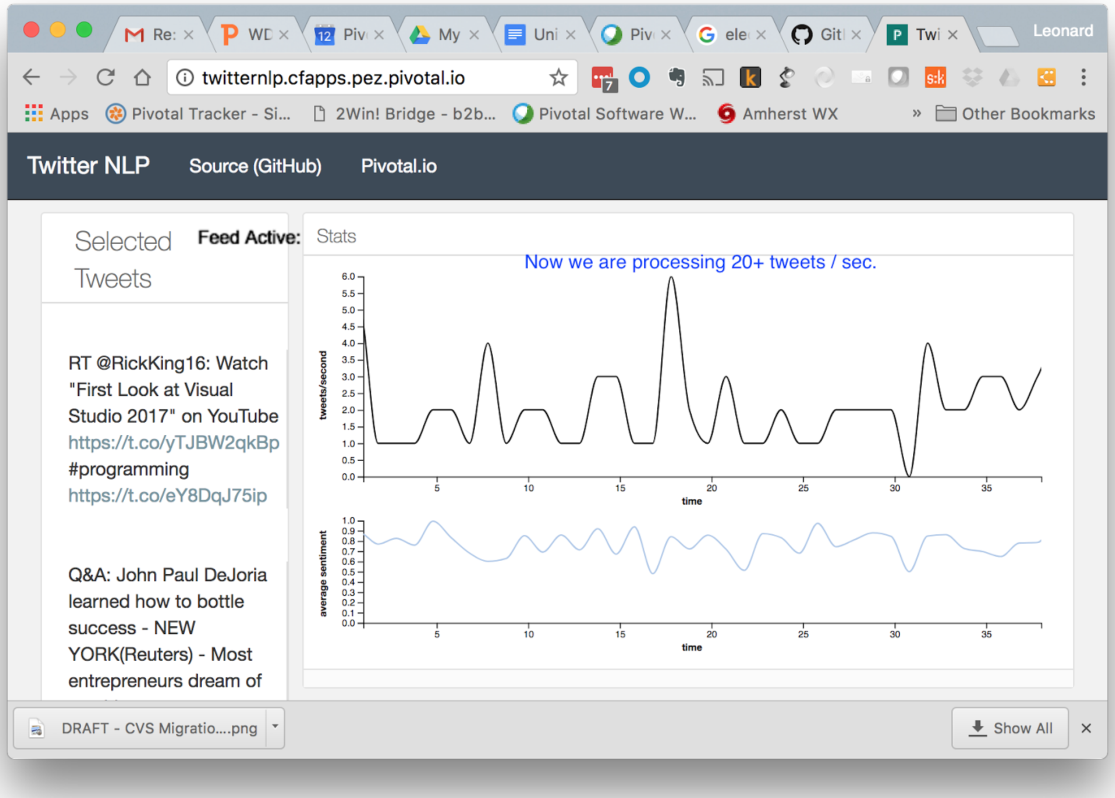


We ramp up the tweet rate, but are only able to process 10/s.

RT @RickKing16: Watch "First Look at Visual Studio 2017" on YouTube https://t.co/yTJBW2qkBp #programming https://t.co/eY8DqJ75ip

Q&A: John Paul DeJoria learned how to bottle success - NEW YORK(Reuters) - Most entrepreneurs dream of

We use “cf scale” to scale up the model scoring service:



RT @RickKing16: Watch "First Look at Visual Studio 2017" on YouTube https://t.co/yTJBW2qkBp #programming https://t.co/eY8DqJ75ip

Q&A: John Paul DeJoria learned how to bottle success - NEW YORK(Reuters) - Most entrepreneurs dream of

Summary

- Enables rapid deployment of operational data science models
- Provides operational efficiencies for deploying and scaling the micro-services architecture
- Enables the data scientists to be productive working with modern tools on data at scale