IBM Advanced Data Science Capstone Project

Sentiment Analysis of Amazon Customer Reviews

Harsh V Singh, Apr 2021

Architectural Decisions Document

This document provides a comprehensive architectural overview of the system, and intends to capture the significant architectural decisions which have been made while designing and implementing this project.

1. Data Source

The data that we are going to use in this project is sourced from a Kaggle user, Adam Bittlingmayer, who uploaded the data to the following Kaggle page -

Amazon Reviews for Sentiment Analysis

The dataset uploaded by Adam is modified for being used in a third-party tool called *fastText*. Since we would like to conduct our own analysis from scratch on the original data, we will use the link for the raw data provided in the description section of the page. This link points to a *Google Drive* folder which has the original training and test csv files for this dataset.

Xiang Zhang's Google Drive dir

The files are stored in **tar.gz** format on the drive with the following filename - **amazon_review_full_csv.tar.gz**.

a. Technology Choice

No technology is required at this stage.

b. Justification

We have manually identified the data source after reviewing various open-source datasets that were available. We reviewed the various datasets on platforms such as Kaggle and Qandl, and finally selected the Amazon reviews dataset as it pertains to a real-world, complex problem that can be addressed using data science and advanced machine learning. **Since the problem statement has been identified after picking the data source, we did not use any technology at this stage.**

2. Enterprise Data

Once we have identified the data source, we need to determine how to integrate this data into our project. Since the selected data source does not provide a recurring/ live stream of data, we only need to take a one-time dump from the external link and save the data to our local machine.

a. Technology Choice

Since we are dealing with a large dataset, we will use **Apache Spark** and **IBM Cloud Storage** to store our data in the form of parquet files.

b. Justification

We have selected **Apache Spark** in this case as this provides a lot of inbuilt features to parallelize the data exploration and feature extraction steps. The size of the dataset is quite large and we will be required to implement parallel processing in order to make the data exploration and feature extraction processes faster.

3. Streaming Analytics

a. Technology Choice

No technology is required at this stage.

b. Justification

This step is not applicable as our data source is static.

4. Data Integration

Data integration involves cleaning the raw data that we have collected and preparing it for the specific use case of our project.

a. Technology Choice

We have chosen Jupyter Notebooks, Pandas and Spark as the technologies for this capstone project.

b. Justification

Jupyter Notebooks allow us to quickly develop python projects with seamless integration of code and supporting documentation and analysis. This will be suitable for presenting the steps undertaken to prepare the data as well as to present it to various stakeholders. Pandas and Spark will be used as powerful tools for data manipulation.

5. Data Repository

There are a host of options available for storing and persisting data. We can choose between any one of them as the requirements of this project are fairly limited. Most of the data will be collected and processed as a one-time activity and there will not be much ongoing changes to it.

a. Technology Choice

We will use **IBM Cloud Storage** as the persistent data storage platform for this project.

b. Justification

IBM Cloud Storage provides an easy way to store and access data and can be scaled as per the requirements of the project.

6. Discovery and Exploration

Data discovery and exploration is one of the most integral steps of any data analysis project. In our specific case, we need to understand the characteristics of the Amazon Customer Review data in terms of natural language processing.

a. Technology Choice

We will be using Jupyter Notebooks as the primary technology and a number of Python packages such as **NLTK, scikit-learn, Keras, Spark, Matplotlib and WordCloud** for our data exploration.

b. Justification

The choice of packages and technology are driven by the need to analyze NLP data. NLTK provides a number of easy-to-implement methods for natural language processing. Similary, Keras and Spark also come with built-in functionalities that will be useful in transforming the data for machine learning.

7. Actionable Insights

Most of the decisions taken during the course of this project will be made in an iterative manner as we complete the various steps and get deeper understanding of the data as well as the results of our analyses. The actionable insights will come from building and deploying suitable machine learning algorithms that will be able to predict the sentiments of customers who provide product reviews on Amazon's ecommerce platform.

a. Technology Choice

We will use **Keras** as the primary technlogy for our model development. Keras provides an abstraction layer on top of TensorFlow, one of the most widely used deep learning frameworks. Within Keras, we will be developing MLP and LSTM neural network models. **TensorBoard** will be used to monitor the performance of the models during and after training.

b. Justification

The reason for choosing **Keras** is that it is an open source techology that provides a number of powerful features and functionalities to define, train and test deep learning algorithms. It is also able to handle large datasets easily.

We will be generating 2 different feature sets, one using **TF-IDF bag-of-words** while another which will convert the text into padded, vectorized word sequences. The former doesn't retain the ordering of words and hence we will train an **MLP neural network** while the latter feature set will be trained on an **LSTM neural network**.

We will compare the performance of these methods using **binary accuracy** metric, as we are building a binary classifier with balanced class distributions.

8. Applications / Data Products

Since this project only requires us to do a one-time analysis of existing data, we will not be creating a data product for its deployment. The results of our analysis will be presented to the stakeholders in an easy-to-replicate and distributable format.

a. Technology Choice

We will use Jupyter Notebooks and slide presentations to present the results of our project.

b. Justification

Jupyter Notebooks are very versatile and can be used to present detailed analysis and graphical results along with the code to replicate the various steps used.

9. Security, Information Governance and Systems Management

Security and governance are an important consideration for any enterprise project. In this specific project, since we are using open source data and we are not developing this project for any proprietary analysis, we will provide unrestricted access to the code as well as the results via Github.

a. Technology Choice

We will share the assets and analysis via Github.

b. Justification

Github is the world's most widely used platform for software development and version control. It provides all the functionality we need to control the access of our data and assets.

In [ ]: