The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document (ADD)

Advanced Data Science Capstone

Project by Ricardo Santos

1 Architectural Components Overview

1.1 Data Source

1.1.1 Technology Choice: external data source, CSV format, size: 233MB (csv)/79.42MB (zip), file publicly available at website.

1.1.2 Justification: availability (only format provided), and ease of use. Besides, csv file is a common format for tabular data.

1.2 Enterprise Data

1.2.1 Technology Choice: Not applicable.

1.2.2 Justification: Not Applicable.

1.3 Streaming analytics

1.3.1 Technology Choice: Not applicable.

1.3.2 Justification: Not applicable.

1.4 Data Integration

1.4.1 Technology Choice: due to relatively small size the data was extracted to a file storage service (Google Drive) but it could also be extracted to an object storage solution (IBM Cloud Object Storage, GCP, AWS, etc.)

1.4.2 Justification: Google Drive provides a free plan (up to 15GB file storage) and is easy to integrate into Google Colab projects, which gives access to GPU (NVIDIA Tesla K80 GPU). But again, other solutions (e.g. IBM Cloud Storage with Watson Studio) could also be used.

1.5 Data Repository

1.5.1 Technology Choice: due to relatively small size the data was extracted to a file storage service (Google Drive) but it could also be extracted to an object storage solution (IBM Cloud Object Storage, GCP, AWS, etc.)

1.5.2 Justification: Google Drive provides a free plan (up to 15GB file storage) and is easy to integrate into Google Colab projects, which gives access to GPU (NVIDIA Tesla K80 GPU).

1.6 Discovery and Exploration

1.6.1 Technology Choice: Jupyter Notebooks in Google Colab. The data quality and exploration were assessed using the following dependencies: pandas, numpy, seaborn, re, wordclouds, and matplotlib.

1.6.2 Justification: The Jupyter Notebook is an open-source, interactive web application, easy to use. The size of the dataset was a key factor in deciding which data exploration tools.

1.7 Actionable Insights

1.7.1 Technology Choice: Jupyter Notebooks in Google Colab. The following dependencies were used: pandas, tensorflow, and keras. Feature Engineering was created using pandas, Tensorflows (tokenization), and keras (padding). The variables were pre-processed before loading into the LSTM, CNN, and CNN-LSTM models.

1.7.2 Justification: The Jupyter Notebook is an open-source, interactive web application, easy to use. Sentiment Analysis requires certain NLP specific tasks (e.g. tokenization, padding) that call for specific methods. Tensorflow and Keras modules offer the necessary methods for the transformation as well for the modeling (ease of use).

1.8 Applications / Data Products

1.8.1 Technology Choice: The system of sentiment classification was based on a CNN model, which could be accessed using its model's weights through a Jupyter Notebook. For benchmarking, the NLTK's VADER was chosen and could also be accessed using Jupyter Notebook.

1.8.2 Justification: The chosen model has performed better (accuracy in the testing set) than its competitors and its weights were stored for ease of use.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice: Not applicable.

1.9.2 Justification: Not applicable.