ArangoML

Data Analytics

Luca Fioravanti Andrea De Angelis

University of Camerino

Academic year: 2020/21

Objectives

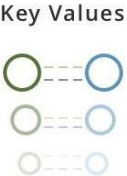
- Perform data analysis with graph database and machine learning.
- Understand the advantages that a graph database offers.

Technologies

Technologies

ArangoDB

ArangoDB is a native multi-model, open-source database with flexible data models for documents, graphs, and key-values.



Python

Python is a general-purpose programming language that can be used for a wide variety of applications.

Features:

- Easy to read and to code
- Robust standard library
- Object-oriented and procedure-oriented

Thanks to its popularity, Python has hundreds of different libraries and frameworks focused on Data Analytics and Machine Learning that make it a great choice for Data Scientists.

Technologies

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool written in Python.

Some interesting features are:

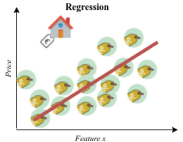
- Handling of (Big)data.
- Handling missing data.
- Alignment and indexing.
- Cleaning up data.
- Input and output tools.
- Multiple file formats supported.
- Merging and joining of datasets.
- Visualize and grouping datasets.
- Find unique data.

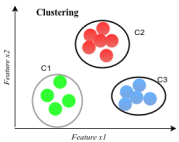
Technologies

Scikit-learn (Sklearn) is the most useful and robust library for machine learning written in Python and built upon NumPy, SciPy and Matplotlib.

Provides a selection of efficient tools for machine learning and statistical modeling including:

- Classification
- Regression
- Clustering





Technologies

Colab(Colaboratory) is a free Jupyter notebook environment that runs entirely in the cloud.

- Write and execute code in Python.
- Create/Upload/Share notebooks.
- Import/Save notebooks from/to Google Drive.
- Import/Publish notebooks from GitHub.
- Import external datasets e.g. from Kaggle.
- Integrate PyTorch, TensorFlow, Keras, OpenCV.
- Free Cloud service with free GPU.

Case study

Our case study focused on defining an application use case of ArangoML.

ArangoML allows Data Scientists to manage all information related to their ML pipeline in one place.

Use case:

- Which dataset influences which model?
- What was the performance of that model?

Arangopipe is a ArangoDB API component for tracing meta-data about machine learning projects.

Implementation

We created a notebook on Colab where we show how machine learning and ArangoDB can work together by creating a knowledge graph and providing missing information using a machine learning model that will be stored with arangopipe in ArangoDB.

Steps

- Selected a csv dataset.
- Preprocessed it.
- Imported it on ArangoDB, splitted into various collections for creating the graph.
- Defined a machine learning model to improve our graph from missing information.
- Updated our graph.

For our analysis we used an instance of ArangoDB running on the cloud.

Achieved results

By storing machine learning model metadata thorugh its pipeline (data ingestion, feature engineering, model training) we can easily keep track of our model updates as well as compare them.

Future implementation

Arangopipe provides a web interface that arranges in a clearer way all the collections created to keep track of your model metadata that we didn't investigate since its not available by the cloud.