Proposal for UBC Urban Data Lab

04.28.2020

Connor Lee, Claudia Nikel, Eva Nguyen & Alex Tamm
MDS Capstone 2020
UBCO

Background

The UBC Urban Data Lab (UDL) is located on the UBC Vancouver campus and was established to improve the access, management, and analytics capabilities regarding data generated on the campus. Its primary goals are the following [1]:

1. "Provide open access of UBC sustainability data to researchers, policymakers and operational staff; [1]"
2. "Support the monitoring and measurement of sustainability performance for buildings, transportation, biodiversity specifically as it relates to the policy commitments of UBC Sustainability Initiative (USI) and Campus and Community Planning (C&CP). [1]"

Urban Data Lab has access to a server and platform managed by UBC Energy and Water Services (EWS). EWS has a large open-source time series database on the power, energy, water, and gas use of each UBC building. This data is important because understanding which instruments are attached to each end-use will allow UDL to analyse the campus wide environmental impact of upgrading specific instruments. This is helpful for the university as targeting specific instruments for upgrading to more energy-efficient ones will help efficiently reduce energy consumption on campus. However, one problem with the data is UDL is unsure of which instruments are for which end-use. This is because most of the instruments were installed a long time ago before any documentation was created, resulting in uncertainty of which instrument belongs to which end-use. Another issue is instruments in specific buildings (i.e. the Pharmacy Building) have tags that are not understandable or are too detailed for practical use by building managers, making it difficult to decipher what the instrument is and where it is located. Figuring this information out will help UDL better understand the data and the instruments that need to be upgraded. Our job is to assist UDL by creating a tool to automate sorting and cleaning of the data, and figuring out the end-uses for the sensors. Overall, this is a small step in a larger end goal of assisted artificial intelligence (AI) that the university is working towards to help with proactive and preventative maintenance.

Aims & Objectives

The primary objective of this project is to deliver a program that collects live streaming sensor data from the UDL SkySpark mirrored database for the Pharmacy building. The program will then use this data to classify each sensor in accordance with the Natural Resources Canada (NRCan) Secondary End-Use Classification Guidelines [2]. In addition, the program will perform any data cleaning required to complete the end-use classification

task. These end-use classifications will be integrated into a Grafana dashboard for visualization. The Grafana dashboard will be embedded into the UDL website.

Data

UDL provides public access to two databases through InfluxDB, one is an ION database and the other is a SKYSPARK database. The ION database mirrors the UBC EWS ION database that updates power, energy, water, and gas use data for each UBC building every 5 seconds [3]. The UDL mirror of the UBC EWS ION database is updated every 2 kW change in electricity power [3]. The SkySpark database mirrors the UBC EWS SkySpark database, it provides data that was recorded by smart devices and meters in each of the UBC buildings [3]. This database is updated every 15 minutes with smart device and meter data, and every hour with weather data [4]. The database elements are named in conformance with the Project Haystack tag guidelines [5]. UDL provides public access to the mirrored ION and SkySpark databases in order to provide researchers and students access to UBC building data, without overloading the UBC EWS databases with queries.

Project Plan

In order to efficiently plan out the work to be completed and develop a framework to effectively monitor progress five themes with associated goals have been developed. Each week has been assigned a theme and associated task completion goals. The implementation of themes will help keep the team focused on tasks relevant to the given week, thereby reducing the risk of working on tasks out of sequence and in turn reduce the risk of missing project deadlines. The proposed themes, and associated dates and goals are shown in **Table 1** below and how these goals will be completed are outlined in the subsequent Methodology section.

Table 1: Proposed Timeline

THEMES	WEEKS	DATES	GOALS
Investigation and Data Prep	3	27 April - 14 May	Identify project objectives and key data features + understand data dictionaries + transform data for machine learning tasks.
Model	1	15 - 21 May	Develop a classification model to apply group tags to end-uses for the Pharmacy building.
Model	1	22 - 28 May	Validate and evaluate models.
Scale + Analysis	2	29 May - 11 June	Expand the model to other UBC buildings (if time permits) + complete user-acceptance testing of model + identify conclusions + create visualizations of results + complete user-acceptance testing of dashboards + UBC mid-term presentation
Wrap-Up	2	12 - 26 June	Final report + package final code + UDL final presentation + UBCO final presentation
Total Weeks	**9**		

Methodology

Figure 1 shows the flow of data under the proposed framework. The primary objective of this project is to develop the code that will complete the tasks contained within the yellow box, which we are proposing to run through Databricks hosted on AWS.

Figure 1: Proposed Flow of Data

In order to develop a viable end-use classification product the following steps need to be completed:

I. Data Collection and Cleaning
II. Feature Selection and Engineering
III. End-Use Classification
IV. Visualization

Each of these steps are discussed in greater detail below.

I. Data Collection and Cleaning

Data collection and cleaning consists of the following tasks:

- Collect data from the SkySpark database
- Develop a tool to identify any missing information
- Develop a tool to populate missing information
- Format the data such that it can be fed into the Feature Engineering and/or End-Use Classification algorithm

II. Feature Selection and Engineering

Feature selection and engineering consists of the following tasks:

- Identify which features are relevant/useful
- Identify if any feature engineering can be done to aid in classification
- Develop a tool to implement any feature engineering identified

III. End-Use Classification

End-use classification consists of the following tasks:

- Determine the best method of classification for our data
- Develop code in Spark for classification of sensors/equipment by end-use using the classification method and features determined above
    - Proposing to use Databricks hosted on AWS

IV. Visualization

Visualization will consist of developing a simple visualization in a Grafana dashboard for integration into the UDL website

Deliverables

Our team will collectively produce a detailed report of the analysis, including complete Python code and Grafana visualizations, as well as appropriate presentations to Urban

Data Lab and within accordance of the UBCO MDS requirements. The Python code will be able to do the following:

- Query and clean data required for classifying instrumentation by end-use
- Classify instruments by end-use
- Time permitting, scale the model such that it can be used to classify new instruments within the same building, or classify instruments in other buildings

Throughout the project regular input and advice will be sought from UDL in order to refine our project and produce the correct product.

Conclusion

Overall, the main goal of this project is to deliver a program that queries live streaming sensor data from UBC's Pharmacy building, cleans the data appropriately, classifies the end-use of each sensor based on NRCan guidelines and integrates everything into a Grafana dashboard. Time permitting, the program will be developed such that it can be scaled up to incorporate other UBC campus buildings. The data is provided publicly from UDL through two InfluxDB databases: ION and SKYSPARK. Data includes measurements from sensors and equipment on building power, energy, water and gas usage on campus. At the request of UDL, our team is proposing to produce a detailed report of the analysis, including complete Python code and an integrated Grafana dashboard. All of this is in mind of the end goal of helping UDL and UBC reach their vision of assisted AI for proactive & preventative maintenance. The project is broken up into five sections: Investigation, Data Prep, Model, Scale & Analysis and a final Wrap-Up. All of these to be completed within the 9 week framework and according to UBCO's outline.

References

[1] J. Wei, “ABOUT,” UBC Urban Data Lab. [Online]. Available: https://urbandatalab.io/about/. [Accessed: 01-May-2020].

[2] Natural Resources Canada (NRCan) Secondary End Use Classification Guidelines

[3] J. Wei, “UBC Building Energy Data,” UBC Urban Data Lab, 11-Feb-2020. [Online]. Available: https://urbandatalab.io/news/ubc-building-energy-data/. [Accessed: 01-May-2020].

[4] J. Wei, “Leveraging Energy Data for Sustainable Buildings,” UBC Urban Data Lab, 12-Feb-2020. [Online]. Available: https://urbandatalab.io/project/analyzing-ubc-building-energy-use/. [Accessed: 01-May-2020].

[5] “Project Haystack,” Tags – Project Haystack. [Online]. Available: https://project-haystack.org/tag. [Accessed: 01-May-2020].