Open Data Engineering & Systems Architecture

Master on Visual Tools to Empower Citizens

Universitat de Girona

2020-2021

Matt Osborn

Overview

This document lays out the *Data Engineering* component of the first trimester's final project. For a complete reference this document should be viewed alongside the *Statistical Programming* and *Dataviz Programming* project documents.

Objective

Create a data pipeline and living dashboard to provide up-to-date COVID predictions based on your predictive model, using currently available daily data.

This should: **(1)** provide up-to-date predictions based on current, daily data, and **(2)** track the historical performance of your predictions by comparing them to real data as it is released.

(1) Current predictions and predictive model

This should use the predictive model you created and refined in the statistics section of the project. If necessary, i.e. if it was previously operating on longer timeframe data, you should modify your model to operate on daily data.

(2) Tracking prediction accuracy

The data output and dashboard should track the performance of your predictive model over time by comparing your predictions with real data as it becomes available. You will not be marked on your analysis methodology, which can be a straightforward comparison, but you will be marked on the engineering involved in working with a dynamic dataset that changes over time.

Living dashboard

This should be an Observable notebook, and will be the same notebook you create to visualise the prediction model as part of the *Data visualization programming* component of the project. You will not be marked on the visualisations themselves for this part of the project, but they should use the data coming from your data pipeline, and they should update automatically each day with new data.

Descriptive overview

In addition to the two living visualisations described above, the observable notebook should contain a short overview of the data pipeline to describe how it functions. This overview should include:

1. A flow diagram visualising the pipeline. This should cover the entire pipeline, from the source data to the dashboard, referencing any data sources used (both static and daily), and it should make clear where parts of the pipeline live, i.e. the cloud services used, where data is being pulled from, where it is being uploaded to, etc.
2. A brief description of the different components
3. Any specific challenges or difficulties encountered and how they were addressed

Specific components that should be mentioned in this overview section:

4. Cloud services used
5. Scheduling approach
6. Configuration of error notifications
7. Any data wrangling that was required

Requirements

Some requirements to note:

- **Must update daily** using currently available, daily data.
- **Should be fully automated and hands off**. Your visualisations should update automatically each day.
- **Operate using cloud services** of your choice. i.e. AWS, Heroku.
- **Error notifications** must be set up to notify you of any critical errors (i.e. via email)

Deliverables

1. **Living dashboard** on Observable. Broadly, this should contain the following sections, as described above:
    a. Current predictions based on available, daily data
    b. Historical prediction accuracy tracker
    c. Descriptive overview of the data pipeline
2. **All code used.** GitHub should be used for collaboration and the repository shared with me as a deliverable.