Predicting influenza occurrences based on weather

FAIR Data Science

Contact person • Ondrej Hudcovic (e11919907@student.tuwien.ac.at)

Based on Common DSW Knowledge Model, 2.0.1 (dsw:root:2.0.1)

Generated on 20 Apr 2020

Data Management Plan created in Data Stewardship Wizard <https://ds-wizard.org>

Abstract

The goal of this experiment was to model the relationship between weather observations and the prevalence of new influenza infections. It included reading, preparing and transforming data. Subsequently, these data were visualized and used for building a prediction model - we tried to predict influenza infections based on weather conditions.

Section A: Data Collection

1. What data will you collect or create?

Instrument datasets The following instrument datasets will be acquired in the project:

- **influenza.csv** - number of influenza occurences in Vienna (weekly data, 2009 - 2018)
- **weather observations** - temperature, humidity, wind and similar data for Vienna (2012-2018)

  This dataset will be collected by experts in the project, with our own equipment.

  The equipment is very well described and known.

Data formats and types We will be using the following data formats and types:

- **CSV Dialect Description Format**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

2. How will the data be collected or created?

Instrument datasets

- **influenza.csv** - number of influenza occurences in Vienna (weekly data, 2009 - 2018)
- **weather observations** - temperature, humidity, wind and similar data for Vienna (2012-2018)

  No instruments for this dataset have been specified.

  We will not be using quality process for this dataset.

Storage and file conventions We will use a filesystem with files and folders with the following folder conventions:

- There will be a **folder for each sample/subject**. Each of those will use the following conventions:
  - **data** - these are the datasets used in the experiment
  - **images** - generated images from "visualization" part of the project
  - **src** - source code

Every "type" of data (be it source code, datasets, ...) has its own designated folder in the repository.

Moreover, we have made appointments about naming the files.

We will not be storing data in an "object store" system.

We will not use a relational database system to store project data.

We will not use a graph database for data in the project.

We will not be storing data in a triple store.

Section B: Documentation and Meta-data

3. What documentation and meta-data will accompany the data?

List of data to be published is given in Section E, Question 9. This also includes information about catalogs where the data can be found. Information about data types used is given in Section A, Question 1.

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

We will be documenting the data with W3C PROV provenance.

Section C: Ethics and Legal Compliance

4. How will you manage any ethical issues?

5. How will you manage copyright and Intellectual Property Rights (IPR) issues?

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open immediately.

Section D: Storage and Backup

6. How will the data be stored and backed up during the research?

Storage needs will be the same during the whole project.

All essential data is also stored elsewhere to prevent a total loss of data. All project data stored outside of the working area will be adequately backed up.

7. How will you manage access and security?

Project members will not store data or software on computers in the lab or external hard drives connected to those computers.They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services addressed via secure http (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

Only project members will have read access; only selected project members will be able to write data.

Section E: Selection and Preservation

8. Which data are of long-term value and should be retained, shared, and/or preserved?

We plan to publish the following datasets:

- **influenza.csv, weather observations** – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.

9. What is the longterm preservation plan for the dataset?

- **influenza.csv, weather observations** will be stored in a domain-specific repository: GitHub. This is the link to the repository of the project. We don't need to contact the repository because it is a routine for us. We will be adding a reference to the published data to at least one data catalogue.

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication.

Section F: Data Sharing

10. How will you share the data?

- **influenza.csv, weather observations** – freely available for any use (public domain or CC0).

Information about used repositories (i.e. where will potential users find out about the data) is provided in Section E, Question 9.

Embargo on the data is described in Section C, Question 5, and Section F, Question 11.

11. Are any restrictions on data sharing required?

Ethical and legal restrictions are documented under Section C. We have used the Data Stewardship Wizard, which made us aware of options to minimize the restrictions.

No data sharing agreement will be required.

Section G: Responsibilities and Resources

12. Who will be responsible for data management?

Ondrej Hudcovic is responsible for implementing the DMP, and ensuring it is reviewed and revised.

13. What resources will you require to deliver your plan?

To execute the DMP, no additional specialist expertise is required.

We do not require any hardware or software in addition to what is usually available in the institute.

Charges applied by data repositories (if any) are mentioned already in Section E, Question 9.