Data science

Data Collection

What data will you collect or create?

The following instrument datasets will be acquired in the project:

- influenza.csv - number of influenza occurences in Vienna (weekly data, 2009 - 2018). The uncompressed file takes 20kB on disk.
- weather observations - temperature, humidity, wind and similar data for Vienna (2009 -2018) containing multiple csv files (each csv for a year). Each of the uncompressed csv files takes 25kB on disk, therefore 250kB in total.

Both datasets are in 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.
Also, both datasets are us-ascii encoded.
The following data will be produced by the analysis:

- charts containing visualizations of various dependencies among features. Each chart is in Portable Network Graphics (.png) format and its size varies from 15kB to 57kB.
- prediction dataset. As well as the acquired datasets, this will be in CSV Dialect Description Format and encoded in us-ascii. The volume of this dataset is 25kB.

The scale of both acquired and produced data is up to 500kB in total, therefore there should be no additional costs or any other challenges in terms of long-term storage of those data.

How will the data be collected or created?

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
    - visualisation - 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.
Naming conventions:

- influenza - this is one csv file and is named influenza.csv
- weather observations - the data files are named in this convention: <year>-ZAMG_Jahrbuch.csv where the <year> may be in [2009, ..., 2018].

Random seed is set to specific value which enables reproducibility of the experiment in

the future.

Documentation and Metadata

What documentation and metadata will accompany the data?

We will be documenting the data with W3C PROV provenance. Also, the data is in CSV format us-ascii encoded, so any text processor should be able to read such files. The data will be described in the README file in the project directory.
We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

Ethics and Legal Compliance

How will you manage any ethical issues?

We obtained the data via publicly available sources and therefore we assume we won't increase the chance of possible de-anonymisation (possibly applicable to infleunza dataset). Any other data are not sensitive.

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

The data will be open and will be publicly-accessible immediately via Github platform. There won't be restrictions on reuse of the data. We won't claim ownership of the data.

Storage and Backup

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

Storage needs will be the same during the whole project. All project data and source code will be stored on Github repository as well as on a cloud provider highly-available storage solution. This will be considered as a backup if the primary storage, Github, wouldn't be publicly available.

How will you manage access and security?

Every person who worked or will work on this project 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 of them were informed of the risks doing so. Though, the impact of loosing such data wouldn't be that high.

All data centers where project data is stored carry sufficient certifications (Github and Google Cloud). All project web services addressed via secured protocol - HTTPS.

The data doesn't contain any personally identifiable information.

Only project members will have read access; only selected project members will be able to write data (predictions and visualisations).

Selection and Preservation

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.

Foreseeable research uses for this data might be for example comparison of average influenza occurences vs. COVID-19 occurences in Vienna with respect to the weather.

What is the long-term preservation plan for the dataset?

influenza.csv, weather observations will be stored in a domain-specific repository: GitHub.

Costs associated with storing the data are zero, mainly because GitHub doesn't charge for their services and the secondary storage - cloud, doesn't have any costs as well.

Data Sharing

How will you share the data?

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

Embargo on the data is described in question "How will you manage copyright and Intellectual Property Rights (IPR) issues?"

Are any restrictions on data sharing required?

No restrictions on data sharing are applied. There won't be an exclusive use of the data needed after publishing this DMP. No data sharing requirement will be needed.

Responsibilities and Resources

Who will be responsible for data management?

Alex Marksfeld is responsible for implementing the DMP, and ensuring it is reviewed and revised.

What resources will you require to deliver your plan?

To execute the DMP, no additional specialist expertise is required. We require only Python3 installed where a user can install other packages and therefore reproduce the experiment.