Curating data - Assignment 1

Collecting and building datasets

Lisa Golla

WH-QUESTIONS

In the course of the assignment I personally decided to choose *bottles* as objects to collect. Therefore, I took pictures of *analogue materials*. I decided to choose bottles, since it provides an easy way to distinguish between certain labels. Next to that it is an appropriate way to display how analogue objects, respectively certain information, become digital and how labels are applied to the objects. Especially, I chose different bottles made of glass, plastic and I also used cans. Here it is possible to distinguish between what the bottle is usually filled with, as well as whether the bottle contains drinkable liquid. For example, I also integrated Oil and Ketchup which are not drinkable whereas coke or water would be drinkable. Concerning the various information the containers provide, it seemed to be a sensible approach to use these bottles as a demonstration of how labels can be applied to certain objects in a dataset. In order to digitize the bottles I used my mobile phone to take pictures of the objects. Further, I created a Spreadsheet and created a relational database in a sense that I entered certain information in the table like 'Kind of bottle' where the labels glass, plastic and can could be entered. Also there is one column called 'Picture' which contains the file name of the corresponding picture where the information matches. Further, I created a csv file out of this table and I coded to match the picture files to the respective labels, whereas a dataset is created thereby. For that, I used the programming language python and coded in jupyter notebooks. The code and all the libraries I imported can be tracked here. I created a public github repo where the source code is available. When it came to the decision of my dataset format I decided to create a dataset with *python* because I'm already used to such datasets in terms of training *artificial neural networks* (ANNs). In the following when I'll explain more about my decisions you can definitely see some references to a Machine learning (ML) approach. Such pictures and labeled data can be used for training in a *supervised* manner in order to detect certain patterns which are corresponding to given labels. For example, you can try to make an ANN learn to decide if a bottle is a can, a glass bottle or a plastic bottle and track the accuracy afterwards. The data from which the ANN learns is a fundamental basis since the knowledge learned from the data is all the network can use when trying to label new, unseen pictures. This is why it is highly important to display a *variety* when choosing a dataset. Generally you don't want the model to overfit. Essentially, by producing training data with a wide range of resolutions, lighting conditions, and angles, you will be able to produce a more ***robust model.*** I also tried to consider this in my dataset.

Critical reflection & Report

As has been revealed in [4] data can be defined as "transmittable and storable computer information."[4]. There are certain methods through which analogue objects become translated into data, namely the logic of digitization is underlying. Everything can be virtually represented in digital formats. In the first place when shooting the pictures I was doing a so-called objectification since I produced digital images and thereby followed the system of mapping. Objectification does not mean that before doing this, the underlying data is no object. Rather, the data is formalized as objects through human agency in the first place

and afterwards recognized as objects by computers which is the objectification process. The second process of how the material got digital was the labeling, or adding of tags to objects, and finally coding them into the "digital milieu"[4]. This process is called dataification of objects. The explained way of representing objects is defined as knowledge representation(KR) which is a key topic in AI. KR is not just storing data into some database, but it also enables an intelligent machine to learn from that knowledge and experiences so that it can behave intelligently like a human, as well as the KR is responsible for representing information about the real world so that a computer can understand and can utilize this knowledge to solve the complex real world problems such as diagnosis a medical condition or communicating with humans in natural language. What can be inferred from the fact that one is choosing a certain KR is that knowledge is actively produced [2]. Data requires the active human-based part and at a certain level the collection of the data presupposes an interpretation. Data has to be "generated"[2], formed and is consequently based on an interpretative dimension. In other words, data cannot speak for itself, rather "We speak for them"[3]. Therefore, there does not exist an objectivity of data [2+3], the nature of data is subjective based. In order to critically reflect, it is important to look under data to consider their root assumptions[2], or in other words to detect the subjective interpretation, or bias which is underlying. Data also has an aggregative nature, whereas an ubiquitous structure of data aggregation can be a relational database as I used for this dataset. Relational databases are organizing data into separate tables, and each column of a table encodes information of a particular sort[2+1]. I chose this format because it is flexible and allows a small number of operations to be defined that apply across those operations [1]. Generally, it is a so-called "data modeling problem" [1] to decide on a format for the given objects. Also my fundament of interpretation and active engagement with the representation is obvious, while my decision also has to be revisited critically. So as to make sense of the represented data, and use it as a basis for argument, it is necessary to include a graphical representation [1]. Any interface is a data visualization, in my project I employed Spreadsheets and printed the dataset in terms of printing the pictures in combination with their label. Data visualization emphasizes the rhetorical dimension of data itself, because particular visualizations are effective on different levels, and also persuasive in certain ways [1]. Which leads me to the fact that my visual representation here, may also be persuasive with regard to the contextual domain of training a neural network I displayed. The usage and interpretation I chose for the bottles in that way is attributing the labels to the bottles and therefore reducing them to a certain degree and recognizing them in an application-specific way. Concerning my dataset I'm using an inductive approach since I provided already given examples from which an ANN for example could infer specific patterns. I chose this approach because the labeling demonstrates how data becomes labeled and how these data is ascribed and combined into a dataset. Another approach which is also promising is abduction [3]. This approach takes incomplete observations and is doing the best prediction out of it. Especially unsupervised ANNs and (Deep) Reinforcement Learning approaches are applying the abduction method and are achieving promising results. As a matter of fact, to stick to the data-driven science whose essence lies in a more exploring and flexible essence, it might be reasonable in general to employ an abduction approach since the approach will lead to more extensive models of entire complex systems [3].

Sources

[1] Paul Dourish. “No SQL: The Shifting Materialities of Database Technology.” *Computational Culture* 4 (9th November 2014). http://computationalculture.net/no-sql-the-shifting-materialities-of-database-technology/.

[2] Lisa Gitelman and Virginia Jackson. “Raw data is an Oxymoron” (2013). https://doi.org/10.7551/mitpress/9302.001.0001

[3] Rob Kitchin. “The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences”(2014) https://dx.doi.org/10.4135/9781473909472

[4] Hui, Yuk, and Bernard Stiegler. “On the Existence of Digital Objects”(2016) https://doi.org/10.5749/minnesota/9780816698905.001.0001