setup knowleadge engineering

f.r.peeters

May 2023

1 Data Transformation and Presentation

For the intermediate status report, you need to finish this chapter. For those questions that you have not implemented or executed yet, please write out a plan for how you'd like to approach the data transformation.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The authors state that the entries and descriptions in the database were derived from published papers, reports, data, and internet documents. published in the time frame 1908 to 2018, representing a variety of sources, including geologic and exploration studies described in State, Federal, and industry reports. The documentation states authors but lacks unique identifiers such as doi keys.
for the employment data the methods explain that the data is gathered by a general population survey, however the link to it is broken and thus not reachable. [1]

Which degree of interaction with the data was needed to prepare the data? (Discovery, Capture, Curation, Design, Creation)

i do not know the terms but for each dataset i took quite a few steps wich i will describe below
cobalt sites where only geocoded by appropriate lattitude and longitude, i

transformed this by using an onine api: **??** Which gave more categorical location data including the state Employment data was not provided in an csv and this

resulted in me copying and pasting the data in an usable format for python. The

railroad data was non conclusive in their terminology as to what is important to our client and therfore requires further research. [1]

Did you find contradictions in your data? If yes, how did you deal with them?

The Query dataset contained something that could be considdered a contradiction, The documentation stated to be restricted to amarican cobalt mines, whist in actuality it also contained 5 locations in puerto rico. due to the diffi-coult politiqual situation of this erea is not a state, but it is us territory. This has the conclusion that it is not present in any of the other dataset and as a result over 10% of the dataset can not be considered for a suitable location.

employment the singular nature of the sources combined with the obscure/non existent references to raw data, make finding possible contradictions hard as reasoning is not provided and the individual detests do not contradict themselves. the numbers for seem pllousible for a non expert on the domain.[1]

Did you find conflicts in your data? If yes, how did you deal with them?

The element that the main cobalt dataset and the transportation and crimes datasets have in common. Is the naming of American districts, the contradictions in this category where solvable by de-capitalizing and removing special characters. resulting in a category that could be used interchangably. The authors have also published a explanations for parts that could have been understood as per their words possible "contradictions" this however as per terminoligy in the lectures seems to closer fit conflicts as they are not inherrently wrong but do explain why choices where made. "The locations of mines, mineral occurrences (which includes deposits and prospects), and mineral regions are represented as points, and some of these point locations have corresponding "footprints" or polygonal outlines. The polygonal footprint may represent the approximate outline of a mineral occurrence or mineral region. The exceptions are areas disturbed by mining-related activity, or surface workings, which are derived from imagery rather than published reports. Surface working outlines have no corresponding point location, nor do they have links to other tables. Polygonal outlines, except for surface workings, may overlap. Overlapping surface workings are merged into a single outline. Surface workings do not distinguish different types of mine features, such as pits, tailings piles, dumps, etc. Two points may occupy the same location. This occurs when there is a deposit with a mine, and the location of either the mine or the deposit is unknown. For example, a report provides a map showing the location of a deposit. The report also provides production data for underground "Mine X" that is mining the deposit, but does not provide the location of "Mine X". In this case, a second point representing "Mine X" is placed at the point location of the deposit." [1]

How did you organize your data? Describe the categorization and/or classification of your data set. Feel free to include diagrams or other imagery.

the question centers arround location, and the axis of consideration are also center able around a metric in a certain location, therefore it made the most sense, to pick the smallest possible location indicator shared across all data-sets. Therefore the process of organizing the new question was one of enriching the location categories to include all parameters of interest

[1]

How did you integrate the different datasets? What was the process for deciding which data to map to which data? If you wrote code for this, please link to it here.

The code can be found in the zip and is called :Merge_data.ipynb

[1]

If you use data exchange in your project, describe your process and share your code.

data exchange? i dont know what this is [1]

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? (i.e., to what extent does this dataset achieve answering the knowledge analytics requests of the client?) If so, how? If not, what are the limitations?

The question of the client centers arround balencing unemployment rates and railroad access in earea's wher cobalt can be mined in large queanities. This dataset provides these quantities for each query. making it possible to answer the question of the client, regardles of what balance they require.

[1]

Any other comments?

[1]