

UNIVERSITY OF KOBLENZ - AG SOFTLANG

DATA SCIENCE

Assignment 2: Data Collection

TO BE SUBMITTED UNTIL THE 24.11.2021 (2 P.M)

November 10, 2021

Question 2.1: Collect Data from the Web

Be creative and collect data from the web which can be used to answer a research question that you choose. You are not expected to submit a fully-fledged study with all kind of methodological issues resolved. Considered this assignment as some sort of blueprint; however, the technical collection process should work. We need to be able to reproduce it. Further, the data should roughly correspond to the research question that you have picked. The data does not need to provide the full answers, but it needs to be relevant for answering the question.

You are not limited to any particular domain. You may cover questions related to social science, finance, economy, software engineering, medicine, epidemiology, ecology or psychology. Furthermore, you may also reuse data from existing studies, if you put it into another context.

Provide us with the following artifacts:

- Provide a description of the research question (in PDF form) that you may want to answer based on the data. You are not limited to questions on causality (see the intro lecture for the other question types).
- Provide a description of the data. A few sentences on what is contained in your collected data does suffice (in PDF form).
- Provide the code (R or Python) to automatically download/crawl the corresponding data sets. Please stick to relative/configurable paths in your scripts. Extra Points: We are aware that there are no limitations on developing advanced data collections methods. We will give extra points for outstanding solutions.
- Provide the data in terms of a CSV file. Do not submit huge files; take a sample or limit files to the first rows.