Supplementary material for GitTables: A Large-Scale Corpus of Relational Tables (Paper ID: 112)

Madelon Hulsebos[1,2], **Çağatay Demiralp**[1], and **Paul Groth**[2]

[1]Sigma Computing, San Francisco, CA 94105
[2]University of Amsterdam, Amsterdam, 1012 WX

1 Societal consequences

The success of prior large-scale data collection efforts across sciences and engineering in accelerating research illustrates the importance of easy access to realistic data at scale as well as evaluating research on shared datasets. The positive impact of large-scale data repositories in machine learning research and applications has been also immense in improving tasks, from image-based disease detection to automated translation. Similarly, GitTables has been developed to bring the value of data to advance research in for example data management. To date, the main interest in our earlier work on semantic column type detection models came from medical institutions who used these models to integrate data across different medical databases. One of our motivations for building GitTables is to improve the application of such learned table models in the medical domain and beyond.

Despite the positive applications that our work is intended to have, we recognize that it can have negative effects as well. We expect that the main potential risk of GitTables lies in privacy. Although the tables were extracted from CSV files that are already publicly available, it might occur that the underlying data files were posted unintended, contain sensitive information, or that the original files were removed for another reason. Furthermore, it is known that combining publicly available datasets may impact privacy of individuals. As GitTables makes public data files more accessible at scale, it might have a multiplicative negative impact associated with the undesired spread of such files or their contents.

Given the nature of the data, relational tables, we do not foresee that applications built on this corpus would have severe negative impact with regard to unfair treatment of individuals. However, just like with other large-scale corpora, such negative consequences can be unforeseen. As GitTables inherits the biases of its source, namely, Github, it is predominantly English and is data from those who use Github. It is important for users of the datasets to reflect on this when using this data for downstream tasks.

As authors of this paper and creators of GitTables, we bear responsibility for the negative effect that our work might have. In the case that GitTables is observed to have negative impacts, the authors will act appropriately to avoid further damage and spread of the data.

2 Corpus access

The table subsets of GitTables are hosted on Zenodo[1]. Zenodo attaches a DOI to each of the data assets and provides metadata for discovery through DataCite. It also ensures long-term persistence of the data as well as versioning of the datasets. This implies that no issue will arise for users of the corpus, researchers and other parties, who depend on a particular version of GitTables, as they will always have access to the respective version. This also ensures that our work is reproducible.

---

[1] `https://zenodo.org`

Preprint. Under review.

Table 1: Metadata included in the table Parquet files.

Table metadata	Column metadata
Table ID	Atomic data types (from Pandas)
URL GitHub CSV	Syntactic annotations (per ontology)
Table dimensions	Semantic annotations (per ontology)
Table topic annotation	Semantic annotation similarities (per ontology)

All tables in GitTables were extracted from public GitHub repositories, hence it is assumed that publication of this data is not restricted. However, the original CSV files may be licensed and restricted in use, hence products built on GitTables inherit these licenses. It is considered the responsibility of the user of the corpus to comply with these licenses, and make responsible use of this data. We publish GitTables itself under the Creative Commons Attributions 4.0 International license (CC BY 4.0).

3 Corpus usage

All material can be accessed through the website of GitTables `https://gittables.github.io`. We publish the subsets of the corpus used for the analysis in this paper along with the ontology data files used for the annotations. We will iteratively publish the corpus as more parts come in. We also share the code used to construct, annotate and analyze the corpus, along with documentation, on GitHub. Finally, we share the manual reviews of the annotations by T2Dv2 and our annotation pipeline used for the analysis in section 4 of the paper.

As discussed in section 4 of the paper, the topic used to query a table might be informative for domain-specific analyses or enhancing table semantics. We therefore keep this structure in tact and provide the table corpus in subsets, a subset of tables for each topic. We also include a "topic metadata" file (JSON) consisting of statistics like the distribution of table dimensions, the number of annotated tables per ontology, distribution of semantic types, etc.

The parsed tables are stored in the Apache Parquet file format which is widely used for efficiently storing and processing tabular data. Filenames are kept as found on GitHub but in case of filename duplication an identifier is added. The table and column metadata as specified in table 1 and structured as dictionaries (key,value pairs) are stored in the metadata of the file. Annotations (if any) are attached in this metadata as well.