CAP Babel Machine Upload Instructions

Metadata

1. **Name:** Your name.
2. **E-mail address:** The e-mail address where you want to receive the results.
3. **Institution name:** The name of the institution you are affiliated with.
4. **Dataset name:** The full name of the dataset. E.g. Presidential speeches (Argentina).
5. **Dataset language:** The language of the dataset. E.g. English. If the language is not in the dropdown list, please choose "Other".
6. **Domain:** Choose one of the 10 CAP domains from the dropdown list that fits the dataset. As noted in the page description, there are differences compared to the CAP website domains.
7. **Unit of observation:** The unit of observation's level. You choose it from a given list: quasi sentence, sentence, paragraph, full text, budget item.
8. **Period (from):** The starting year of the period investigated by the dataset. e.g., 1972.
9. **Period (to):** The ending year of the period investigated by the dataset. e.g., 2018.
10. **Level of dataset:** The level of territorial unit investigated by the dataset. You choose it from a given list: supranational/international, state, substate. We kindly ask you to classify international organisations as supranational/international level, and parties, social movements or national-level NGOs as substate level.
11. **Geographical unit:** The name of the geographical unit investigated by the dataset. If it is in substate level, we kindly ask you to include both the state and substate name. E.g. India, Nagaland
12. **Use case:** Your use case for the CAP Babel Machine. You can choose from a given list: research, commercial use, personal use, other.
13. **Description**

- An introduction (max. 1000 characters) of the database.
- List sources and important metadata for your database.
- Provide any other significant information that you deem important.

Dataset Requirements

The dataset must pass two rounds of validation checks for a successful prediction.

First Check Requirements

The first check takes place when you submit your dataset through the upload form. If the dataset fails this check, then your upload will not go through and the upload form will inform you of the error.

- It must have UTF-8 encoding.

- It must be in CSV format.
- The first row must be the header of the dataset.
- All spaces must be substituted with underscores in variable names.
- We kindly ask you not to include diacritical marks to your variable names and write them by Latin script.
- Mandatory variables must be included in a given order. (See below.)
- Additional variables can be included freely in a desired order, but we do not incorporate them in our classification process.

Mandatory Variables Mandatory variable names must be entered as-is. If you have a column that corresponds to the mandatory columns but is named differently (e.g., `row_id` instead of `id`), then it must be renamed accordingly.

Non-coded datasets:

- **id:** The unique identifier of the unit of observation. Can be a numeric or combination of text and numeric.
- **text:** Full text of the unit of observation.
- **year:** The year of the unit of observation's origin (based on Georgian calendar). Four character numeric variable.

Pre-coded datasets:

- **id:** The unique identifier of the unit of observation. Can be a numeric or combination of text and numeric.
- **text:** Full text of the unit of observation.
- **year:** The year of the unit of observation's origin (based on Georgian calendar). Four character numeric variable.
- **major_topic:** The major topic of the unit of observation based on the codebook of the Comparative Agendas Project.

Second Check Requirements

The second check takes place as the file is processed by our pipeline. If the dataset fails this check, the prediction will fail, and you will receive an email informing you of the failure. If this happens, please re-check your dataset and re-submit it.

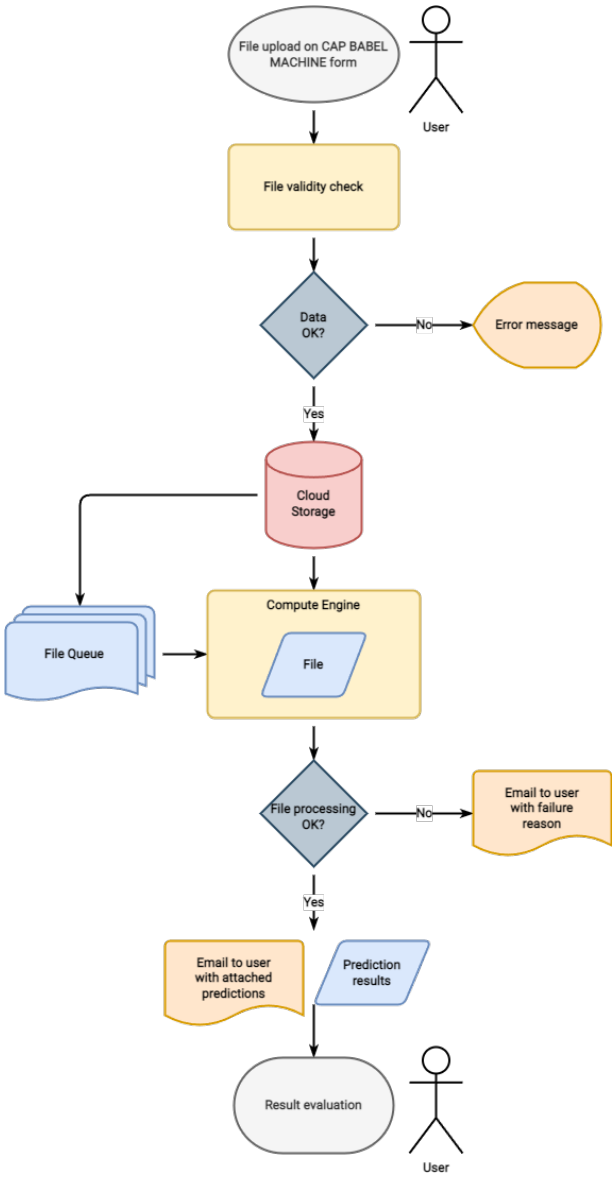
- Please make sure that the rows follow the amount of columns the file has.
- Please make sure that delimiters are used correctly and consistently. (e.g., don't mix comma and semicolon separators within one file).
- If the cell contains delimiter characters, then please make sure that it's quoted properly.

Codebook Requirements

- You may upload a codebook in PDF format.

- Explanation of variables must be in English, following the same order as in the dataset.
- The codebook must contain a short description of the dataset, including its content, the investigated period, the number of observations and the list of the preparers of the dataset.

CAP Babel Machine pipeline



Help and Support

For additional help and inquiries please don't hesitate to use our contact form or write to the following e-mail address: poltextlab@poltextlab.com.