Dataset upload rules

1. Select data type
    - Data type here refers to the datasource type. Currently there are eight different data types that can be uploaded to our system: 1) legal document, 2) speech, 3) media, 4) social media, 5) budget, 6) manifesto, 7) judicial decision, 8) other.
    - This is necessary for coding purposes.
2. General requirements for uploaded dataset
    - All files must be UTF-8 encoded
    - All files must be in CSV format
    - First row must be the header of the dataset
    - Variable names must be all lower case
    - 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 have included in a given order.
    - Additional variables can be included freely in a desired order but only after the mandatory variable names that have to be in the required order
    - Additional variables will be downloadable, but we do not incorporate them in our classification process
3. Mandatory variables
    - id: The unique identifier of the unit of observation. Can be a numeric or combination of text and numeric.
    - 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 Comparative Agendas Project. At this point, please, left it empty as dataset will be coded by CAP Babel Machine.
    - text: Full text of the unit of observation.
4. Codebook to upload
    - All datasets must be uploaded with 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.
5. Dataset name
    - The full name of the dataset to be displayed on the webpage. E.g. Presidential speeches (Argentina).
6. Dataset type
    - The type of data included in the dataset. You choose it from a given list: legal document, speech, media, social media, budget, manifesto, judicial decision, other.
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
    - The starting and ending year of the period investigated by the dataset. E.g. 1972-2018.
9. 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.
10. 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
11. Description
    - A short (max. 300 words) introduction of the database that will be visible on the webpage under the dataset.
    - List sources and important metadata for your database.
    - Provide any other significant information that you deem important.
12. Pre-validated dataset
    - You have the option to upload a dataset containing a given number of randomly selected units of observation including major topics previously coded by you.
13. Help
    - For additional help and inquiries please don't hesitate to write to the following e-mail address: poltextlab@poltextlab.com