Aditya Nadimpalli

Indian Election Data Extraction – File Contents and Explanation

Folder Contents

1. The folder Data_Extracton_Code contains:
   - Python files used to convert the raw data output from tabula into the final cleaned data output. The files state_conversion_1971.py to state_conversion_2014.py use the tabula output (Year_detailed_results_raw_output.csv) in the folders 1971 - 2014 to generate the output file Constituency_Totals.csv. **Order of execution of files should be from state_conversion_2014.py to state_conversion_1971.py as each program appends data to previously written data.**
   - The files candidate_conversion_1971.py to candidate_conversion_2014.py use the tabula output (Year_detailed_results_raw_output.csv) in the folders 1971 - 2014 to generate the output file Candidate_Details.csv. **Order of execution of files should be from candidate_conversion_2014.py to candidate_conversion_1971.py as each program appends data to previously written data.**

   - The file constituency_summary.py converts the tabula output Year_constituency_summary_raw_data.csv in the folders 1971-2014 into the output file Constituency_Summary.csv.
   - **Candidate_Details.csv, Constituency_Totals.csv and Constituency_Summary.csv are the final three data output files containing all the years' data**.

   The state_conversion and candidate_conversion python files follow similar format but vary from year to year due to the differences in data present and document layout. **The state_conversion_2014.py and candidate_conversion_2014.py files contain detailed comments explaining the code.**

2. The folders 1971 to 2014 contain the following files:
   - XXXX_Constiuency_Summary.pdf and XXXX_Detailed_Results.pdf, the pdfs containing the constituency data summary and detailed results respectively for the year XXXX.
   - The jruby files XXXX_detailed_results.rb and XXX_constituency_summary.rb contain the tabula extractor code used to get the data from the detailed results and constituency summary pdfs respectively. The corresponding output files are XXXX_detailed_results_raw_output.csv and XXXX_constituency_summary_raw_data.csv
   - These raw data files are used by the python programs in the Data_Extraction_Code folder to generate the final three output files containing all the years' data together.

    - Jruby progams can be executed at the command line as follows:
      >>jruby <filename>.rb

Work Flow for Data Extraction:

- Use jruby based tabula extractor code to roughly convert the data tables in a pdf document into csv format. The data in this raw csv file is not yet structured and readable.
- Tabula uses column spacing dimensions to convert tabular data into csv format. For each year's pdf documents, these dimensions need to be manually adjusted so that all the data in the document is correctly captured in the csv file.
- Use a python program to restructure the tabula output into a final readable data file.