**Documentation on the Order of Implementation of Scripts for Raw Data Cleaning Associated with NREGA Assets (State-wise)**

Process: *Cleaning NREGA State data files*

*Author: Ewan Nikhil Thomas*

Use the Python Scripts in the “src\data\” subdirectory in the following order to produce processed data files from raw NREGA assets data.

1. The process requires an external data dependency to clean file of Rajasthan. So, primarily run the Scrapy spiders “devanagiri” and “english” in the subdirectory, “.\devanagiri”. For more details, refer Unicode Dictionary Documentation.docx available in the reports sub directory.
2. Run the script “unicode\_english\_transliteration.py” to create necessary external data in the “data\external\devanagiri.csv” location. For more details, refer Unicode Dictionary Documentation.docx available in the reports sub directory.
3. Run the script “state\_interim\_data\_create.py” to create interim data. For more details, refer Data Wrangling Documentation.docx available in the reports sub directory.
4. Run the script “state\_interim\_to\_processed.py” to create processed and cleaned data ready for LGD mapping. For more details, refer Data Wrangling Documentation.docx available in the reports sub directory.
5. After creating the processed data, run script “JK\_to\_LADAKH.py” to extract the districts “KARGIL” and “LEH LADAKH” and create a separate state file called “LADAKH.csv”.
6. LGD mapping for this dataset can be conducted now by running the scripts “lgd\_mapping.py”. The mapping was originally run using the script “lgd\_iterator.py” which runs the fuzzywuzzy mapping multiple times to identify blocks for manual mapping. So in actual execution, just use the former.