**Manifest of intermediate files**

**1. Princeton folder**

|  |  |
| --- | --- |
| **File name** | **File function** |
| Princeton\_clean1\_deletenull.json | Clean all the records of which every attribute value is “null” from raw data. |
| princeton\_clean2\_removegeocode.py | Delete those records of which the geocode value is “Place depicted” (7462 records deleted) , “(not assigned)” (17 records deleted),“geography attribution” (9 records deleted). |
| Princeton\_clean2\_removegeocode.json | Outfile |
| princeton\_clean3\_removeduplicates.py | Delete the duplicate records. |
| princeton\_clean3\_removeduplicates.json | outfile |
| princeton\_clean4\_fixtypo.py | Fix continent value typos |
| princeton\_clean4\_removeduplicates\_fixtypo.json | Outfile |
| princeton\_clean5\_countryvaluefixed.json | Fix country values |
| princeton\_clean5\_countryvaluesfix.py | Outfile |
| princeton\_clean6\_split\_iflonglatitude.py | Split the file princeton\_clean5\_countryvaluesfix.json. |
| princeton\_clean7\_nonelist.json | outfile |
| princeton\_clean7\_withlonglatitude.json | outfile |
| princeton\_clean8\_fromlongilatinumbertostatecity\_finalversion.py | Get the country, state, city information values using reversegeocode package. |
| princeton\_clean9\_transferlocation.csv | Outfile |
| princeton\_clean9\_withlongitudelatitude.csv | outfile |
| princeton\_clean10\_withlongitudelatitude\_final.csv | Combine princeton\_clean9\_transferlocation.csv and princeton\_clean9\_withlongitudelatitude.csv |
| princeton\_clean11\_nonlist.csv | outfile |
| princeton\_clean11\_nonlist.py | Add state, city values to records in “princeton\_clean7\_nonelist.json” |
| princeton\_clean 12\_final version.csv | Final cleaned Princeton dataset |

**2. gm folder**

|  |  |
| --- | --- |
| **File name** | **File function** |
| gm\_geography\_clean1\_columndelete.csv | A file in which every record has only “ObjectID”, “term”, “ThesXrefType”,”LatitudeNumber”,”LongitudeNumber” columns. |
| GM\_GeographyTerm\_Clean.py | Delete the records of which the ThesXrefType value is either “place”, “0900”, or “culture area”. |
| GM\_Clean2\_CorrectGeographyTerm.csv | Outfile |
| GM\_CSV\_RemoveDuplicates.py | Remove duplicates |
| GM\_Clean3\_CorrectGeographyTerm\_withoutDuplicate.csv | Outfile |
| GM\_Clean4\_RemoveDuplicates\_ObjectID.py | Delete rows with repeated object ID. For rows with repeated object ID, only keep the first-occurred row. |
| GM\_clean4\_RemoveDuplicate\_ObjectID.csv | Outfile |
| GM\_clean5\_RemoveDuplicate\_ObjectID\_RemoveEmptyLine.csv | Split the file GM\_clean4\_RemoveDuplicate\_ObjectID.csv into two sub-files. In this file, each record has both longitude and latitude number. |
| GM\_clean6\_RemoveDuplicate\_NoLocationNumber.csv | Split the file GM\_clean4\_RemoveDuplicate\_ObjectID.csv into two sub-files. In this file, each record has neither longitude nor latitude number. |
| gm\_clean7\_fromlatitudelongitudetolocation.py | Transfer the longitude number and latitude number to country, state, city. |
| gm\_clean7\_fromlatitudelongitudetolocation.csv | Outfile |
| gm\_clean8\_codetocountryandcontinent.py | Transfer the ISO 3166-1 alpha-2 country code to country name and continent name. |
| gm\_clean8\_continent.csv | Outfile |
| gm\_clean8\_country.csv | Outfile |
| gm\_clean9\_withlongilanti\_final.csv | Hand combine three files: (1) gm\_clean7\_fromlatitudelongitudetolocation.csv (2) gm\_clean8\_continent.csv; (3) gm\_clean8\_country.csv |
| GM\_clean10\_RemoveDuplicate\_NoLocationNumber\_withlocationnumber.csv. | process the records with neither longitude number nor latitude number |
| gm\_clean11\_finalversion.csv | Combine file  “GM\_clean10\_RemoveDuplicate\_NoLocationNumber\_withlocationnumber.csv” with file “gm\_clean9\_withlongilanti\_final.csv”. |

3. Gm dataset manifest

|  |  |
| --- | --- |
| **File name** | **File function** |
| SAAM\_removeDuplicates.py | Remove duplicated lines |
| saam\_WebConGeography\_CleanDataRemoveDuplicates.csv | Outfile |
| saam\_clean2\_removegeocode.py | remove those rows in which ConGeoCode value is “(not assigned)”, “Associated Place”, “Associated Place (at time of NEA award)”, ““Last Known Residence”. Keep the records of which ConGeoCode values are “Place of Birth” and “Place of Death”. |
| saam\_clean2\_correctgeographyterm.csv | Outfile |
| saam\_clean3\_correctgeographyterm\_addmissingvalue.csv | I added “missing” to the cells to represent the city name is missing in this record. I added the country name according to the given city name and state name. If no enough information is given to have a country name, then I enter “missing” in the cells. If the state name is missing, I enter “missing” in the cell. |
| saam\_clean4\_addcontinent.csv | Hand add continent values according to the country name. |