MicroDataCleaning

政府統計の調査票データの読み込みなどに役立つプログラムを公開しています。ここで公開されているプログラム等を使用される場合、以下をよく読んでご使用ください。お問い合わせ等はIssueからお願いします。場合によっては対応が非常に遅くなる、もしくは対応できないこともあるかもしれませんがご容赦ください。

This repository provides desktop application and Python programs to make do-files for data cleaning from layout tables as well as the resulting do-files. Before using the application or programs, please read the followings. You can find the English version here.

Overview

glmiceは政府統計の調査票データ読み込み用do-fileを符号表から自動で作成するプログラムです。

- glmice.exe: Windows用デスクトップアプリケーション（main.pyを元にしています。）
- main.py: 符号表からdo-fileを生成するためのメインプログラム
- do-file/: 固定長データを読み込むためのdo-fileがあります。上のプログラムを用いて作成したものと他の研究者の方々からご提供いただいたものがあります。本プロジェクトに協力してくださった方のリストはここをご覧ください。

 demo

Requirement

- glmice.exe：デスクトップアプリケーションなのでPythonのインストールは必要ありません。（動作確認環境：Windows 10）
- main.py：Python3（動作確認環境：Windows 10, Python 3.7.0）Pythonはここからインストールできます。また、main.pyを実行するには`python-Levenshtein`が必要です。以下のコマンドでインストールできます。

```
pip install python-Levenshtein
```

Usage

glmice

1. dist.zipをreleaseからダウンロードして、任意の場所に展開します。展開後、dist/glmice/glmice.exeをダブルクリックしてください。
2. 符号表の情報とアウトプットのファイル名を入力してください。('Excel file', 'Excel sheet index', 'Output file', 'Data file')
3. 'Add'ボタンを押すと、画面左のリストに入力した情報が追加されます。リスト内の項目を選択して'Check'ボタンを押すことで、入力した内容を確認できます。
4. また、リスト内の項目を選択して、'Remove'ボタンを押すことで、選択した入力内容を取り消せます。

5. 読み込む符号表の情報を全て入力して、リストに追加した後、'Survey name'に調査名を入力します。入力しなくても問題ありませんが、調査名を入力することで、変数名を調査年のデータの間で統一するプログラムのパフォーマンスが改善する可能性があります。
6. 最後に'Run'ボタンを押してください。

ステップ１～３は必要な情報を入力したcsvファイルを用意することで省略できます。その際にはdist/docs/input_list_template.csvを使用してください。csvファイルに入力後、`File > Import`からcsvファイル読み込むことができます。

また、アプリケーションを起動する際、以下のメッセージが出るかもしれませんが、`詳細情報 > 実行`から実行してください。

> **Windows によって PC が保護されました**
> Windows Defender SmartScreen は認識されないアプリの起動を停止しました。このアプリを実行すると、PC に問題が起こる可能性があります。

Main program

main.pyは符号表からdo-fileを生成するためのメインプログラムです。
（現在、このプログラムはPyPIに登録されていないので、ソースコードをここ）からダウンロードして使用してください。

```
Main(infile_list, index_list, outfile_list, data_list,
        xls=False, reservation=0.2, SurveyName=None)
"""
    infile_list: list of input files.
    (i.e., Excel files for the layout tables)

    index_list: list of Excel sheet indices (Count from 0).

    outfile_list: list of output file names.
    If outfile_list = ['File1', 'File2'], the resulting output files are
    'File1_const.do', 'File1_const.do', and so on.

    data_list: list of the raw data (in most cases, .txt or .dat files).

    xls: whether to make 'cleaned' layout sheet.
    (This option may be useful if you are an R user.)

    reservation: reservation distance which is used to make a do-file to
    harmonize several data. If the Levenshtein distance of two variables
    (from different survey years) is more than this reservation distance,
    those variables are judged to be different variables even if they are
    the closest pair. (You may not want to change this parameter, and rather
    making synonyms will be more useful.)

    SurveyName: Name of the survey (in Japanese). Specifying the survey name
    could improve variable matching process when harmonizing several data.

    Method defined here:
```

```
    run(): Run the program and make do-files.
"""
```

Output files

- FILENAME_const.do: 固定長データの読み込みます。
- FILENAME_var.do: 変数ラベルの貼り付けます。
- FILENAME_val.do: 値ラベルの貼り付けます。
- FILENAME_validate.do: データが正しく読み込まれていそうか確認します。
- rename.do: 複数年データを接合するために変数名を変更します。
- master.do: 上記のdo-fileを実行、データの接合、データの保存をします。
  このファイル内の`DoFilePathTemp`と`DataFilePathTemp`（グローバルマクロ）を使用する前に変更する必要があります。（rename.doとmaster.do`outfile_list`のルートディレクトリに作成されます。）
- rename.xls: 複数年のデータに関して、変数名のマッチングを行った結果がまとめてあります。必ずrename.doを実行する前に、正しく変数名の変更がされているか確認してください。
- FILENAME_layout.xls: `xls=True`を指定したときに作成されます。各変数の開始位置と終了位置がまとめられており、繰り返しは展開されているため、Rを用いて固定長データを読み込む場合、こちらのファイルを使うことができます。

Example code

```
from main import Main

infile_list = [
    'RootPath/Layout/XXX.xls',
    'RootPath/Layout/YYY.xls',
    'RootPath/Layout/ZZZ.xlsx'
]

index_list = [0, 0, 0]

outfile_list = [
    'RootPath/do-file/XXX/XXX',
    'RootPath/do-file/YYY/YYY',
    'RootPath/do-file/ZZZ/ZZZ'
]

data_list = [
    'RootPath/data/raw/dataXXX.txt',
    'RootPath/data/raw/dataYYY.txt',
    'RootPath/data/raw/dataZZZ.txt'
]

main = Main(
    infile_list, sheet_index_list, outfile_list, data_list
)
main.run()
```

Before running master.do ...

1. dist/ado/CheckAppendValidity.adoとdist/ado/DestringAll.adoを適切なディレクトリに追加してください。Stataで`adopath`と入力するとadoファイルのパスが表示されます。
2. rename.xlsを参照しながら、rename.doの内容が正しいか確認してください。rename.xlsでは、前後の調査年の変数の項目名が変わっている場合、セルが黄色く塗られています。

Harmonize variable names across survey years

複数年の変数名の統一は符号表上の項目名と変数名を用いたfuzzyなマッチングによって行われていますが、必ずしも正しくマッチングできるわけではありません。調査ごとに項目名や変数名の類義語リストを用いることで、マッチングの質を高めることができます。（例：WageCensus.py）リスト内の各要素は最初の要素に変更され、その後、項目名や変数名のLevenshteiの距離が計測されます。

以下の手順で新しい類義語リストを作成・適用出来ます

1. 新しい類義語リストを作成して'python/module/Writers/VarNameThesaurus/Thesaurus/'にそれを保存します。ファイル中で定義するリストの変数名は`thesaurus_jargon`である必要があります。
2. '/python/module/Writers/VarNameThesaurus/StrDistMeasure.py'内の`StrDistMeasureFactory`クラスを修正します。例えば、1.で作成したファイル名が'NewThesaurus.py'で調査名が'xxxSurvey'の場合、以下を`StrDistMeasureFactory`クラスのif文に追加します。

```
elif SurveyName == 'xxxSurvey':
    from .Thesaurus.NewThesaurus import thesaurus_jargon
```

3. `SurveyName='xxxSurvey'`を指定して、main.pyを実行します。

ここでの変更はglmice.exeには反映されません。変更をアプリケーションに反映させるには

```
pip install pyinstaller
```

でPyInstallerをインストール後、python/build.batを実行してください。ディレクトリの構造を変えていなければ、ファイル内容を変更する必要はありません。また、build.bat内で配布用のzipファイルを作成するために7-zipを用いていますが、圧縮する必要がなければ、このソフトウェアは必要ありません。

各調査ごとの類義語リストは他の方にとっても有用である可能性が高いので、作成された場合はプルリクエストを出していただけますと幸いです。

Remarks

- アプリケーションやプログラム中では相対パスではなく絶対パスを使用した方が良いかもしれません。Windowsを使用されている方はパスの指定する際に、`\`（バックスラッシュもしくは円マーク）ではなく、`/`を用いてください。（例：`C:\Users\Takahiro\Desktop` -> `C:/Users/Takahiro/Desktop`）
- アウトプットファイルの保存先ディレクトリが存在しない場合は自動で生成されます。

- 生成されたdo-fileは必ずしも正しいとは限らないため、データ読み込み後の確認作業は必ず行ってください。もし、プログラムや生成されたファイルに問題がありましたら、Issueにてご報告いただけますと幸いです。（このページの`New issue`ボタンをクリックしてください。）
- _validate.doは各カテゴリ変数が想定されている値を取っているか確認します。
- rename.doは使用前に内容を必ず確認してください。すでに、調査ごとの類義語リストが用意されている場合以外では、正しく変数名の変更が行われていない箇所がある可能性が極めて高いです。各変数の変数名がどのように変更されるかについてはrename.xlsを参照してください。
- 変数の値ラベルは個別のデータには付与されますが、カテゴリ変数の定義が調査年によって異なる可能性があるため、複数年のデータを接合したものには付与されません。そのため、カテゴリ変数の整備はご自身で行う必要があります。
- プログラムでは、基本的に同じ調査に関してデータを接合することが想定されていますが、異なるデータ間で変数名を統一する目的にも使用可能です。ただし、その場合、符号表の項目名や変数名が大きく異なる可能性があり、あまりうまくいかないかもしれません。（未確認）

License

Author

Takahiro-Toriyabe