Thomas Lander

12/06/24

Digital Humanities Team #1 (Blue)

BirthParse Documentation

BirthParse is a Python parsing script that is made to work with the Alabama Authors master file ("Full AL Collection Master File (Updated Aug 13 2024).xlsm" in our case). This parser works very similarly to Clayton Onwere's "AuthorParse" ; it functions by first cleaning up the biography section for every author, removing all HTML tags for ease of use. Then, shortened excerpts of each biography for each author are extracted based on a set of configurable keywords (i.e. "Born", "born", "Born in" for birthplaces). These excerpts are then parsed to see if they include information relating to an Alabama city (located in "alabama_city_coordinates.json"), if no city is found it then checks for Alabama counties instead (located inside BirthParse.py). When running BirthParse.py, two separate csv files will be output: Formatted_AL_Authors.csv, containing all the data that was found by the parser (including LastName_FirstName FirstName_LastName, City or County, Longitude and Latitude) ; and Authors_Outside_Alabama.csv: containing all the authors for which data could not be found and the shortened biography so it will be easy to check by hand.

<u>Files Needed to Run BirthParse.py:</u>

- BirthParse.py
- alabama_city_coordinates.json
- Full AL Collection Master File (Updated Aug 13 2024).xlsm

Changes Needed in order to Run BirthParse.py:

- Change the following variables to your local system's path for the master file and output files. (master_filePath, output_csv_path_alabama, and output_csv_path_outside)

```
# Load the uploaded files
master_filePath = '/Users/thomaslander/Desktop/AL_Authors_QGIS/data/Full AL Collection Master File (Updated Aug 13 2024).xlsm'
```

(lines 5-6)

```
# Save Alabama authors to a CSV
output_csv_path_alabama = "/Users/thomaslander/Desktop/AL_Authors_QGIS/data/Formatted_AL_Authors.csv"
in_alabama_df[['Formatted_Output']].to_csv(output_csv_path_alabama, index=False, header=False)

# Format and save the output for authors outside of Alabama or missing coordinates
outside_alabama_df['Formatted_Output'] = outside_alabama_df.apply(
    lambda row: f'"{row["Last_First"]}","{row["First_Last"]}",{row["info"]},,,',
    axis=1
)

output_csv_path_outside = "/Users/thomaslander/Desktop/AL_Authors_QGIS/data/Authors_Outside_Alabama.csv"
outside_alabama_df[['Formatted_Output']].to_csv(output_csv_path_outside, index=False, header=False)
```

(lines 165-176)