politician-experience-project

This repo is dedicated to code related to Professor Malmendier's Politician Experience Project. These codes are for tracking the location history of the politicians.

Data source

The `BioguideProfiles` folder was downloaded from Biography Directory of the United States.

The `H102_votes.csv`, `HSall_members.csv`, `HSall_votes.csv` were all downloaded from Voteview.com

The `Lawmaker Location by County-Year, BG and CD (Best Version, June 2019).csv` is a past result of Politician Voting project.

Code

The first part (101-104) of the project is focused on extracting the birthplaces of the politicians.

- `101 Generating Bios File.ipynb`: takes in all the bioguide `json` files under `BioguideProfiles`, and outputs `results/bioguide.csv`, which contains all information from the json files.
- `102 Parse birth Places and School.ipynb`: takes in `results/bioguide.csv`, and outputs `results/bioguide_birth_places_schools.csv` which contains appended birth places and school columns.
- `103 Geocoding Birthplaces with Google Maps.ipynb`: takes in `results/bioguide_birth_places_schools.csv` and outputs `results/geocoded_birthplaces.csv`, which added the latitudes and longitudes of the birthplaces. Additionally, to compare the results from the old codes in 2018, this file also takes in `tables2018/current_legislators_birth_places_schools.csv` and `tables2018/historical_legislators_birth_places_schools.csv`, and outputs comparison table `result/comparison.csv`.
- `104 Extracting Places and Years.ipynb`: takes in `results/bioguide.csv` and `prompt.txt`, and outputs `results/sample_tables.txt`, which contains 3 sample tables of time and location generated by ChatGPT.

The second part (201) of the project uses ChatGPT to generate the location history.

- `201 Filtering bios.ipynb`: takes in `results/bioguide.csv`, `HSall_members.csv`, and `HSall_votes.csv`, and outputs `'results/icpsr_bioguide_id_crosswalk.csv`, `results/full_bios_after109th.csv` and `results/sample_bios_after109th.csv`, which limits the data to the politicians voting in any of the Congresses after the 109th Congress.
- `202 Generating location history with ChatGPT.ipynb`: takes in `results/sample_bios_after109th.csv` or `results/full_bios_after109th.csv` and any of the prompts, and outputs the files in the folder `gpt_bio_result`. (different GPT models and different prompts)

- `203 Converting timeline.ipynb`: takes in a gpt bio result, converts the text result to a table format, and outputs files `results/sample_timeline.csv` or `results/timeline_gpt4_new_full.csv`. (depending on the input gpt bio result)
- `204 Geocoding locations and getting fips`: takes in `results/timeline_gpt4_new_prompt_full.csv`, geocodes the locations, converts to fips, and outputs `results/gpt4_new_prompt_full_timeline_and_fips.csv` and `results/gpt4_new_prompt_fips`.

manual checks

The files in this folder are to check the results generated by ChatGPT.

Archives

- `finetune_dataprep.ipynb`: takes in `samplebios.`, and manually generates the tables of time and location for 10 random sample bios for finetuning the model. (not used)
- `Fixing high school.ipynb`: takes in `gpt_bio_results/gpt_bio_result_gpt4_new_full.csv`, filters the bios that has "high school", extracts the sentence, and outputs `results/high_school.csv`.
- `Comparing GPT and past geocoding result.ipynb`: takes in `results/sample_timeline.csv` and `Lawmaker Location by County-Year, BG and CD (Best Version, June 2019).csv`, and outputs `results/sample_timeline_fips.csv`, `results/sample_location_fips.csv` and `results/sample_compare_fips.csv`. It geocodes the locations with Google Maps API, and get the fips with censusgeocode API, then compare with the past result.

Conclusion

If you have any questions about the files in this folder, feel free to contact me at elaine220615@gmail.com