Description:

The features (Year, District, and Candidate names for all past and current elections) are found through the historical House of Representative election data on Wikipedia (https://en.wikipedia.org/wiki/United\_States\_House\_of\_Representatives\_elections,\_2012 . for example.). Because the data on the Wikipedia comes in the format of tables, we wrote a script to download the HTML source code of the page and then use Python library to parse the HTML content to get the candidates' names and zones. Because Wikipedia contains data for all past candidates, our dataset is very complete. In addition, we use web search and Machine Learning and manual inspection to get the gender for each candidate. There are many existing APIs(https://gender-api.com/) that do an excellent job to identify genders, and wikipedia itself provides some info that links to a candidate's gender.