# Notes

* ~~Determine size constraints of R datasets~~
* Don’t get crazy
* Go through .json again, see what might be missing
* Check through the original files
* Priority
* Bill level (meta, progress, rollcall, sponsor, subjects)
* People level (just people)
* Individual votes
* Text
* Amendment
* How to get R/Bash to compress files for me
* Single file that goes from original, then new\_original, then data\_r, then collapse.
* Include line of code
  + Embed .sh in .R; could wait till the end.
  + If takes too long, link to .sh from inside the .R.
* Clean up the .sh file
* Have a prompt so it’s either input State or \_all.

# Steps

1. Within */Volumes/Samsung\_T5/Data/Legiscan/Data/data\_json/each\_state*, I created an *All\_Sessions* folder. (It was easier than looping through each of the subdirectories within each of the states. Tried, but my code was convoluted and missing things.) Created a bash script (*things.sh*) within the *data\_json* folder that loops through each type of state meeting and copies the .json files to the state’s *All\_Session* folder.
2. Within the */Legiscan/Data* folder, created two separate folders for the decoded data, *data\_r* and *data\_txt*.
   1. *.txt* instead of *.csv* because the bill text and bill descriptions have commas and semicolons embedded in the text. Seemed to be better to change the delimitator to a tab (which is default for a *.txt* file opened in Excel) from any of those than to change the raw data. There’s probably a way around it, but I don’t know it [yet].
3. Within in each state’s *data\_r/data\_txt* are the following files:
   1. *state\_bill\_master* (contains *bill\_meta*, *bill\_progress*, *bill\_rc\_votes*) [aggregate]
   2. *state\_bill\_people\_master* (contains *bill\_people*, *bill\_people\_votes*) [individual]
   3. *state\_bill\_text*
      1. This last file, the *bill\_text* file, has been throwing problems. It may not be included in the file list for some (or none) of the states.
      2. This is also separate from the *state\_bill\_master* because I was afraid that the files would reach the maximum number of entries in Excel.
      3. An additional problem (seemingly-unrelated to the errors mentioned above) is that the decoded bill text contains tab formatting, which throws off the *.txt* tab deliminator.
4. Within the R script:
   1. Had to add an option to a line in Funghui’s code, so that the *parse\_people* command would deal with nested json files. I don’t actually know what the rest of the commands in the function do, but she showed me this (similar enough, anyway) with the *decode\_text* function.
   2. Command loop for each state, both input and output.
   3. Both roll-call and individual-case votes are parsed from *the bill\_vote\_json* files.
   4. *Bil\_progress* and *bill\_meta* are both parsed through the general bill json file.
   5. I’ve included commented-out code for saving each file (votes, meta, progress, etc) individually, but it seems to just take up a lot of unnecessary space to save them all separately as well as combined.
5. For those states completely finished, both master files and text files, I compressed *All\_Sessions* to clear up some room.

# Issues

1. *Bill\_text* for the following states seems to be problematic:
   1. AL, throws:
      1. PDF error: Invalid least number of objects reading page offset hints table
   2. AK, throws:
      1. PDF error: May not be a PDF file (continuing anyway)
      2. PDF error (2): Illegal character <21> in hex string
      3. PDF error: Couldn't find trailer dictionary
      4. PDF error: Couldn't read xref table
      5. Error in poppler\_pdf\_text(loadfile(pdf), opw, upw)
      6. PDF parsing failure.
2. Parsing Errors
   1. Ohio is being a jerk
      1. Error: Column `current\_body` can't be converted from character to logical
   2. CO is also a jerk.
      1. Error: Column `chamber` can't be converted from character to logical
   3. HI
3. Hawaii has no vote records.
4. Some states have Supplemental files.
   1. KY, 2013 1st Special
   2. KY, 2013 Regular
   3. KY, 2014 Regular
   4. KY, 2015 Regular
   5. KY, 2016 Regular
   6. KY, 2017 Regular
   7. KY, 2018 1st Special
   8. KY, 2018 Regular
   9. KY, 2019 1st Special
   10. KY, 2019 Regular
   11. TN, 2017-2018 General
   12. TN, 2019-2020 General
   13. MD, 2013 Regular
   14. MD, 2014 Regular
   15. MD, 2015 Regular
   16. MD, 2016 Regular
   17. MD, 2017 Regular
   18. MD, 2018 Regular
   19. MD, 2019 Regular
   20. RI, 2015 Regular
   21. RI, 2016 Regular
   22. RI, 2017 Regular
   23. RI, 2018 Regular
   24. RI, 2019 Regular
   25. WY, 2017 Regular
   26. WY, 2018 Budget
   27. WY, 2019 Regular
   28. LA, 2013 Regular
   29. LA, 2014 Regular
   30. LA, 2015 Regular
   31. LA, 2016 1st Special
   32. LA, 2016 2nd Special
   33. LA, 2016 Regular
   34. LA, 2017 1st Special
   35. LA, 2017 2nd Special
   36. LA, 2017 Regular
   37. LA, 2017 Regular
   38. LA, 2018 1st Special
   39. LA, 2018 2nd Special
   40. LA, 2018 3rd Special
   41. LA, 2018 Regular
   42. LA, 2019 Regular
   43. UT, 2017 1st Special
   44. UT, 2017 Regular
   45. UT, 2018 2nd Special
   46. UT, 2018 3rd Special
   47. UT, 2018 Regular
   48. UT, 2019 Regular
5. Some states have Amendment files.
   1. AL, 2017-2017 Regular
   2. AL, 2018-2018 Regular
   3. AL, 2019-2019 1st Special
   4. AL, 2019-2019 Regular
   5. AZ, 2013-2013, 1st Regular
   6. AZ, 2013-2013, 1st Special
   7. AZ, 2014-2014, 2nd Regular
   8. AZ, 2014-2014, 2nd Special
   9. AZ, 2015-2015, 1st Regular
   10. AZ, 2015-2015, 1st Special
   11. AZ, 2016-2016, 2nd Regular
   12. AZ, 2017-2017, 1st Regular
   13. AZ, 2018-2018, 1st Special
   14. AZ, 2018-2018, 2nd Regular
   15. AZ, 2019-2019, 1st Regular
   16. FL, 2013-2013 Regular
   17. FL, 2014-2014 1st Special
   18. FL, 2014-2014 Regular
   19. FL, 2015-2015 1st Special
   20. FL, 2015-2015 2nd Special
   21. FL, 2015-2015 3rd Special
   22. FL, 2015-2015 Regular
   23. FL, 2016-2016 Regular
   24. FL, 2017-2017 Regular
   25. FL, 2018-2018 Regular
   26. FL, 2019-2019 Regular
   27. FL, 2013-2013 Regular
   28. IL, 2015-2016, 99th
   29. IL, 2017-2018, 100th
   30. IL, 2019-2020, 101th
   31. KY, 2013 1st Special
   32. KY, 2013 Regular
   33. KY, 2014 Regular
   34. KY, 2015 Regular
   35. KY, 2016 Regular
   36. KY, 2017 Regular
   37. KY, 2018 Regular
   38. KY, 2019 1st Special
   39. KY, 2019 Regular
   40. TN, 2017-2018 General
   41. TN, 2019-2020 General
   42. ME, 2017-2018, 128th
   43. MD, 2013 Regular
   44. MD, 2014 Regular
   45. MD, 2015 Regular
   46. MD, 2016 Regular
   47. MD, 2017 Regular
   48. MD, 2018 Regular
   49. MD, 2019 Regular
   50. LA, 2013 Regular
   51. LA, 2014 Regular
   52. LA, 2015 Regular
   53. LA, 2016 1st Special
   54. LA, 2016 2nd Special
   55. LA, 2016 Regular
   56. LA, 2017 1st Special
   57. LA, 2017 2nd Special
   58. LA, 2017 Regular
   59. LA, 2017 Regular
   60. LA, 2018 1st Special
   61. LA, 2018 2nd Special
   62. LA, 2018 3rd Special
   63. LA, 2018 Regular
   64. LA, 2019 Regular
   65. MS, 2017 1st Special
   66. MS, 2017 Regular
   67. MS, 2018 1st Special
   68. MS, 2018 Regular
   69. MS, 2019 Regular