Replication files for Huang, Perry, and Spirling (2019)

Style estimation is implemented in our R package `stylest`, which can be installed from CRAN:

https://cran.r-project.org/web/packages/stylest/

Specs

This code was tested with:

- Mac OSX with 16 GB RAM
- RStudio Version 1.2.1335
- R version 3.6.0

and with the following R packages installed via CRAN:

- stylest 0.1.0
- corpus 0.10.0
- ggplot2 3.2.0
- Matrix 1.2-17
- dplyr 0.8.3
- reshape2 1.4.3
- stargazer 5.2.2
- plm 2.2-0
- lmtest 0.9-37
- jsonlite 1.6
- magrittr 1.5
- quanteda 1.5.1
- trend 1.1.1
- gridExtra 2.3
- RColorBrewer 1.1-2
- strucchange 1.5-2
- devtools 2.0.2

and with the following R packages installed from GitHub:

- quanteda.corpora 0.87, using `devtools::install_github("quanteda/quanteda.corpora")`

Directory contents:

`pa_replication` contains the following subfolders:

- `code`: contains code to output results, tables, figures
- `figures`: figures labeled as they appear in the paper
- `tables`: tables labeled as they appear in the paper
- `data`: metadata files, the second part of the raw data (see next section for details) and intermediate data (mostly `.csv` files) generated by the Workflow

`style_text` contains the bulk of the raw data. This folder should be at the same level as `pa_replication`.

The file `/code/master.R` will run all scripts necessary to generate the tables and figures for this paper. Total time: 2 hours on a 2014-era MacBook Pro.

Data files

The raw data is in two parts, and is loaded separately for each part.

1935-2013 is loaded, parsed, and filtered from raw XML files found in `style_text`, a directory at the same level as `pa_replication`.

2013-2018 is loaded from an .RData file `sessions_2014_2018_data.RData` in the directory `/data`.

We have made every effort to ensure that all local, relative file paths will work on any machine. If R says that a file cannot be found, it is likely because the relative location in your file system has not been recognized correctly.

Metadata files

These are included in `/data` and include MP-level metadata about names and IDs; session-level data; and party-level data. These are necessary for replicating the main results and can be identified by their filenames as called in the replication code.

Main results

Code files are in the subdirectory `/code`. "Workflow" should be run prior to the "Tables" and "Figures" sections.

Workflow

A. Cross validated vocabulary:

- (60 min) `01_select_vocab.R` generates `data/vocab_cutoff.csv`

B. Model fitting:

- (3 min) `02_member_accuracy.R` generates `data/member_accuracy.csv`, `data/term_influence.csv`
- (1 min) `03_unique_speakers.R` generates `data/unique_speakers.txt`

C. Influential terms:

- (1 min) `04_influential_terms.R` generates `tables/table7.txt` and `tables/table8.txt`

Figures in the paper

Fig 1

- (2 min) `obama-in-HOC.R` generates `data/obama_in_HOC_member_accuracy.csv` and `data/obama_in_HOC_term_influence.csv`

- (1 min) `viz-point-est-obama.R` uses these estimates to produce Figure 1

Fig 2

- (5 min each) `05_correct_classifications.R` and `05_correct_classifications_mw.R` generate point estimates;

- (1 min) `06_compare_classifications.R` uses these estimates to produce Figure 2

Fig 3

- (2 min) `overtime_viz_distributions.R` produces Figure 3 and also includes the code for the Cox-Stuart and Mann-Kendall tests mentioned in section 9.1 of the text

Fig 4

- (1 min) `interesting-v-experience.R` produces Figure 4

Fig 5

- (1 min) `federalist_predict_mwj.R` produces Figure 5

Fig 6

- (1 min) `validation_tokens.R` produces Figure 6

Tables in the paper

Table 1

- (1 min) `summary_distinctiveness_row.R` generates the values in the "Distinctiveness" row as `table1distinct.txt`
- (3 min) `summary.R` generates everything else as `table1.txt`

Table 2, 3, 4, 5

- (2 min) `validation_prepostBlair.R` generates `table2.txt`, `table3.txt`, `table4.txt`, `table5.txt`.
- *Note: we fill in some speakers' first names manually in the paper, as the tables generated by the code use data that lists some speakers as e.g. "Mr. Lastname" or "Mr. (Title)".*
- *Note: the "Newspaper Mentions" column in each table was manually tabulated by the authors.*

Table 6

- (3 min) `basic_regression_logodds.R` generates `table6.tex`

Table 7, 8 (Appendix C)

- (1 min) `04_influential_terms.R` generates `table7.txt` and `table8.txt`