Code for "Gender Bias in Rumors among Professionals: An Identity-based Interpretation"

Alice Wu

This file lists all programs provided in the "code" folder. I begin with the analysis .do files that produces the results in this paper, and then explain the programs for cleaning and merging raw data sets.

- **Final Analysis Files (Stata)**
  - **"static_analysis_2019.do"**: this .do file uses "full_sample2019_stata_final.dta", provides summary statistics and analyzes topic differences between Female and Male posts (Section 4). The results in the log-files (see the folder "results_static") are used to produce Table 1 and 2, Figure 2, 3 and 4.

  - **"dynamic_analysis_2019.do"**: this file uses "full_sample2019_stata_final.dta" and runs multinomial logistic regressions to estimate the effects of a gendered post on transition rates between topics (Section 5). The results in the log-files are used to produce Table 4, and Figure 5, 6, and 7.

- **Code for Cleaning & Merging Data (R & Python)**
  - **"clean_scraped_data.py"**: reads "..data/scraped_posts.txt", assigns unique thread and post IDs, and generates the first .csv file "EJR0_raw_text.csv"

  - **"clean_raw_text.R"**: cleans up "EJR0_raw_text.csv" such that each row corresponds to a different post. Outputs "EJR0_raw_text_cleaned.csv", which contains 2.2 million posts across 223,475 threads in total.

  - **"word_count.py"**: identifies the most frequent 10,000 words from the 2.2 million posts and save the dictionary into "vocab_Nov2017.pkl". Record the occurrences of each word in each post into "counts_Nov2017.npz", a 2.2 million by 10,000 sparse matrix in compressed format.

  - **"vocab.R"**: summarizes the 10,000 most frequent words as in "..data/vocab/cleaned_vocab.csv", and save the indices of words under each category (see the "../data/vocab/category" folder).

  - **"prep_for_analysis.py"**: counts the occurrences of words under each category in each of the 2.2 million posts, and saves it into "EJR_ALL_categories_2019.csv"; counts the number of female and male classifiers in each post and saves it into "EJR_ALL_gender_classifiers_2019.csv".

  - **"NBER-post-history-ID (by full).R"**: identifies posts that include an NBER author's full name, and saves the thread and post IDs in "author-history-ID-by-full.csv"

- **"NBER-post-history-ID (by part).R"**: under threads that include at least one post mentioning an author's full name, identifies posts that mention his or her first name or last name. The IDs are saved into "author-history-ID-by-part.csv".

- **"JMC-post-history-ID.R"**: first, identifies posts that mention a person's full name; under threads that include such posts, identifies posts that mention his or her first name, last name or initials. The IDs are saved into "JMC-history-ID.csv".

- **"merge_sources.R"**: this file merges threads with at least one Female/Male post from three sources – (1) gender classifiers such as she/he, (2) names of NBER authors, and (3) names of recent economics Ph.D. graduates / job market candidates. It outputs "full_sample2019.csv" that preserves all posts under such threads.

- **"job-rank.R"**: counts the number of keywords indicating different job ranks in each post in the full gender sample. The output file "full_sample_job_rank.csv" is merged into the final dataset through the R file below.

- **"gen_full_sample.R"**: starting from "full_sample2019.csv", this file finalizes the definition of gender at the post level, merge in counts by category ("EJR_ALL_categories_2019.csv") and thus define Professional versus Personal topics, and merge in keywords and information about the subjects – NBER authors or JMCs to define the job rank at the post level. The output "full_sample2019_stata.csv" is read in Stata and saved as "full_sample2019_stata_final.dta".