Description of Data and Programs for:

Title: Gendered Language on the Economics Job Market Rumors Forum
Author: Alice H. Wu
AEA Papers & Proceedings 108: 175-79, May 2018

Datasets:

1. "gendered_posts.csv":
   - a dataset of Female/Male posts identified from the four-year sample of EJMR data. Please see the codebook on page 2 for details.
2. "vocab10K.csv":
   - a list of the most frequent 10,000 words that emerge from 2.2 million posts from Oct 2013 to Oct 2017, and each's marginal probability on a post discussing a female from the Lasso models. Please see the codebook on page 3 for details.
3. "X_word_count.npz"
   - this file contains a matrix that records the number of occurrences of each word from the most frequent 10,000 words in each post. This matrix is called in the python programs for logistic/linear Lasso models.
4. "keys_to_X.csv"
   - this file contains unique identifiers for each post in each thread (title_id and post_id) in the *Same* order as the matrix of word counts saved in the .npz format. Useful for merging in the python programs below.
5. "trend_stats.csv"
   - monthly summary stats for Figure 1.

Programs:

1. "lasso" folder contains three Python programs for:
   a. "lasso-logit-full-sample.py": logistic Lasso on the full gender sample (Female and Male posts identified through the comprehensive list of gender classifiers; see Section II in the paper)
   b. "lasso-logit-pronoun-sample.py": logistic Lasso on the pronoun sample (Female and Male posts identified through pronouns; see Section III in the paper)
   c. "lasso-linear-pronoun-sample.py": linear Lasso on the pronoun sample (see Appendix Figures 2 & 3)
2. "tables-figures.R" starts from the final datasets and constructs all tables & figures included in the paper.

Codebook for the dataset of Female/Male posts: "gendered_posts.csv"

Variable	Description
title_id	uniquely identifies a thread
post_id	uniquely identifies a post in each thread (1 refers to the title)
topic	title of a thread (adding "https://www.econjobrumors.com/topic/" to the front generates the URL to the original thread)
raw_post	scraped content of a post
Using the Comprehensive List of Gender Classifiers	
fem_all	total number of female classifiers (from the comprehensive list) included in a post
male_all	total number of male classifiers (from the comprehensive list) included in a post
training	1 if a post is in the training sample for a Lasso-regularized logistic model, 0 if a post is in the test sample for selecting optimal p-score cutoff, NA if a post includes both female and male classifiers and needs to be re-classified based on the predicted probabilities ("ypred")
ypred	predictied probability of a post discussing a female(s) rather than a male(s)
female	1 if a post is Female, 0 if Male. (This is the final classification of gender after reassigning genders for the posts that include both female and male classifiers through the Lasso-logistic model)
Using the Gender Pronouns (Robustness Check)	
fem_pronoun	total number of female pronouns included in a post
male_pronoun	total number of male pronouns included in a post
training_pronoun	same definition as "training" but restricted to the pronoun sample (gendered posts including either female or male pronouns) in the robustness check
ypred_pronoun	same definition as "ypred" but restricted to the pronoun sample in the robustness check
female_pronoun	same definition as "female" but restricted to the pronoun sample in the robustness check
Time Stamps at the Thread level	
time_latest	time stamp for the Latest post in each thread, listed on the main pages of EJMR as of scraping in the end of Oct 2017
month_latest	number of months between a thread's latest update and Oct 2017 (0 if last updated in Oct 2017), identified among threads initiated or updated between Nov 2016 and Oct 2017.

Codebook for the dataset of the most frequent 10,000 words: "vocab10K.csv"

Variable	Description
index	unique ID for words, ranging from 1 to 10,000
word	most frequent 10,000 words (in lower case) identified from the full four-year sample, including posts that are neither Female nor Male.
female	1 if a word is a female classifier (refering to a female), 0 if a male classifier, NA if neither.
i_pronoun	1 if a gender classifier is a female or male pronoun (e.g., "he"/"she")
exclude	1 if a word is NOT used as a predictor in the Lasso models. This list includes all gender classifiers, plus names of celebrities who are not economists.
Results from the Lasso-logistic Model on the Full Sample (Section II)	
coef	estimated coefficient on each word predicting a Female post in the Lasso-regularized logistic model (see Section II in the paper)
ME	estimated average marginal effect of each word on the probability that a post is *Female* in the Lasso-regularized logistic model (see Section II in the paper and the online Appendix)
nFemale	number of *Female* posts in the full sample that includes each word
nMale	number of *Male* posts in the full sample that includes each word
Results from the Lasso-logistic Model on the Pronoun Sample (Section III)	
coef_pronoun	same definition as "coef", but restricted to the pronoun sample (see Section III)
ME_pronoun	same definition as "ME", but restricted to the pronoun sample (see Section III)
nFemale_pronoun	number of *Female* posts in the pronoun sample that includes each word
nMale_pronoun	number of *Male* posts in the pronoun sample that includes each word
Results from the Lasso-linear Model on the Pronoun Sample (Appendix)	
linear_coef	marginal probability of each word predicting a Female post in the pronoun sample (estimated coefficients in the linear Lasso model; see Appendix Figures 2 and 3)