Documentation for replicating *"Machine Learning for Public Administration Research with Application to Organizational Reputation"*

L. Jason Anastasopoulos (ljanastas@uga.edu)
Andrew B. Whitford (aw@uga.edu)

DATA FILE DESCRIPTIONS

1. "coded-tweet-data.csv" - data file containing the sample of tweets used for training the gradient boosted tree algorithm.
2. "agency_tweets_database.csv" - the full database of the agency tweet sample collected from Twitter.

CODE FILE DESCRIPTIONS

1. "textcleaner.R" - a function which is called within the main analyses ("xgboost-analysis-final.csv") which prepares the tweets for analysis before transforming them into a document term matrix.
2. "xgboost-analysis.final.R" - the R code which can be used to replicate all plots and trained machine learning algorithms using the gradient boosted tree method.

CODEBOOK FOR "coded-tweet-data.csv"

- **HITID -** the ID associated with the classification of a unique tweets.
- **Text -** the text of the tweet as seen by both workers and the expert coder.
- **Answer1 -** Response of the Mechanical Turk worker # 1.
- **Answer2 -** Response of the Mechanical Turk worker # 2.
- **Agreement** - Whether both coders agreed.
- **Answer** - The "recommended" final answer according to Amazon. If both coders agreed, the agreed upon response would be the recommended answer, if both coders disagreed, then "no_agreement" would be the recommended answer.
- **Date** - The date and time of the responses.
- **JasonCode** - Coding decision made by the expert coder where:

  1 = Performative Reputation
  2 = Moral Reputation
  3 = Procedural Reputation
  4 = Technical Reputation
  0 = None of the above

CODEBOOK FOR “agency_tweets_database.csv”

- **agency_id** - Twitter handle for the agency
- **tweet_text -** text of the agency tweet.
- **tweet_favorites -** how many times the tweet was favorited.
- **tweet_retweets -** how many times the tweet was retweeted.
- **tweet_created -** date and time that the tweet was created.