Requirements

1. Validated FAA data file - ***DATA\_FAA\_split\_validated.csv*** 
   1. This is the output from the text cleaning and merging workflow
   2. If a data set does not require cross validation, the cross validation script can be skipped but the file needs to be formatted to match its ouput
2. R script – ***topic\_modeling\_training.R***
   1. Ingests the file in step 1. and outputs 2 R objects
      1. LDA – topic model - ***topic\_model.RDS***
         1. This can ingest new reports, and assign probabilities to belonging to a topic
      2. Logistic regression predictive model - ***predictive\_model.RDS***
         1. This ingest the out put from the LDA model and predicts if a new report

Steps

1. Update the working directory and file path to ***DATA\_FAA\_split\_validated.csv***  at the top of ***topic\_modeling\_training.R***
2. Tuning parameter can be changed at the top of ***topic\_modeling\_training.R*** as well. It will change the number of topics that are uncovered and later used for predictive modeling.
3. Run the entire script
4. Output is the 2 models(Topic Model and Prediction Model) as R objects

Workflow details

* This first processes all of the report text and creates topics (the number selected by user).
* With the topics built, all of the reports are then run through topic model and assigned a probability of belonging to each topic category
* The topic probabilities are then feed into a predictive model, in this case, random forest, and our predictive model is trained
* The script out puts the 2 models as R objects for use with new data in ***topic\_modeling\_prediction.R***