Script Usage Guide – Auburn Big Data

Disclaimer: This script is rather straightforward to operate, **as long as you do so properly**. Any improper use of the script may cause breakages and/or invalid results. We have failsafes in place for that, but please refrain from making improper use of this script.

1) Overall Behavior & How to Run

In this section, we describe the basic behavior of our script and how to run it.

Basic Behavior

The script, when executed, pulls in the supplied FNDDS, IRI, and PPC data needed (*More on supplying data in Section 3*). Then, according to how it is configured, it will 1) format the data for use, 2) train the model, and 3) test it.

Note: If the script has already generated something (embedded FNDDS data, embedded IRI data, train/test splits, etc.), it **will not** generate it again unless you remove the file it is looking for. More on this in Section 3.

How to Run

To run this script, you **only** need to run "main.py". This script will kick off all other functionalities of the script.

2) Configuration

In this section, we outline the configuration file and what options are to be configured.

General Options

This section of the configuration file contains general options that apply to all functions within the script. The options are as follows:

- num_threads: This option tells the script how many parallel threads you wish to run for all operations. The default is 8, and it is recommended not to exceed the amount of cores (or virtual cores) your machine has. *More often than not, you can just leave this value alone*

Data Options

This section of the configuration file contains options that are specific to storing/retrieving data in the filesystem. **It is recommended that these options are not changed, as they are currently half-implemented**

Method-Specific Options

This section of the configuration file contains options that are specific to each method in the script. Currently, we have some general options and two available methods, with room for more options later. The current options are as follows:

- active_method: This option tells the script which method to use. **The recommended method for evaluation is "random_forest"**, but if you want to play around with the code, we have another method available, "ann".
- model_loc: This option tells the script where to write the models. This option is currently half-implemented, so *it is also recommended to not change this option*

- prep_data*:* This tells the script whether or not to prepare data for either training or inference. Your options here are **true** or **false** to turn it on or off, respectively. *For evaluation, this option will be "false"*
- train_model: This tells the script whether to train the model or not. Your options here are **true** or **false** to turn it on or off, respectively. *For evaluation, this option will be "false"*
- run_inference: This tells the script whether to run inference with the model or not. Your options here are **true** or **false** to turn it on or off, respectively. *For evaluation, this option will be "true"*
- train_size: This tells the script how what portion of data should go to the training set. *It is recommended that basic users refrain from changing this option*
- test_size: This tells the script how what portion of data should go to the training set. *It is recommended that basic users refrain from changing this option*
- inference_data: This tells the script what dataset to use to evaluate the model. *Leave this string empty to use the auto-generated testing set*
- method_list: This option contains a list of valid methods that can be used. *This list does not need to be modified, so we recommend leaving this alone*

Embedding Options

This section of the configuration file contains options that tell the script which embedding to use and handles more specific options related to each embedding. The options are as follows:

- active_embedding: This option tells the script which embedding to use. **The recommended method for evaluation is "glove"**, but if you want to play around with the code, we have another embedding available, "bert".
- embedding_list: This option contains a list of valid embeddings that can be used. *This list does not need to be modified, so we recommend leaving this alone*

3) Directory Structure

In this section, we outline the directory structure, where things are generally stored, and what files are important.

Configuration

This folder houses any configuration files that are created. The only important file here is *"default.json"*, as it is the default configuration for the script. *Without it, the script won't function, so take care not to delete this file*

Data

This folder contains *by far* the most things that can be adjusted. The following files are the most important, and removing these files will cause the script to re-generate them on its next run:

- data/raw/fndds_<year>.csv, data/raw/iri_<year>.csv, data/raw/ppc_<year>.csv
  - These files contain the raw information from the database. If you want to manually supply these files in these locations, the script will not overwrite them.
- data/embedded/fndds_combined_embedded.csv, data/embedded/iri_similarity_embedded.csv
  - These files, as well as some similarly-named, numbered files, contain the calculated embeddings and similarity scores. These files will be large, and are used to generate the training and testing data for the model.
- data/random_forest/train.csv, data/random_forest/test.csv, data/random_forest/full_set.csv

    - These files, as well as some similarly-named, numbered files, contain the training, testing, and entire datasets, respectively. These files will also be large, and are used to train/test the model whenever needed.
- data/random_forest/predictions.csv
    - This file is arguably the *most important* to know about, since this file contains the predictions generated by the model.

Models

This folder contains the models created by the script. Feel free to rename, copy/paste, and do whatever you'd like with the models, but *make sure there is a model named "rf_model.bin" in the "random_forest" folder if you plan on doing inference*.

Source Code

This folder contains the various other scripts called by the main script. *It is highly recommended not to tamper with this folder*.

Embeddings

This folder contains the word embeddings used by the script. Once again, *it is recommended to not tamper with this folder*.