This README describes the various files created for 2019S2 COMP90049 Project 2.

There are various files in these archives: other than this README, each one can be identified by its filename, in the format {set}-{type}.{filetype}

- {set} refers to either train, dev, or test:

    train: You should use this data when building a model

    dev: You should use this data when evaluating a model

    test: You should submit the outputs on this data; the labels (?) are not given

- {type} refers to either tweets, mostXX, or bestXX:

    tweets: This contains the raw text of the corresponding tweets, one tweet per line, in the following format:

        Tweet-ID,User-ID,Tweet-Text,Class

        where Tweet-ID is a unique value;

        and Class is one of NewYork, California, or Georgia.

    mostXX: For these files, we have pre-processed the corresponding tweets, and have recorded the term frequency for the top XX terms according to document frequency.

bestXX: For these files, we have pre-processed the corresponding tweets, and have recorded the term frequency for the terms with the greatest Mutual Information and Chi-Square values. The total number of features does not correspond to XX: rather, for each of the two feature selection methods, the top XX features were determined for each of the three classes, and then the resulting list was sorted, with duplicate entries removed.

It is worth noting that to process the raw text of the tweets, we folded case, and removed all characters that are not English alphabetic characters ([a-z]) or whitespace. This is significant, as many tweets contained hashtags, at-mentions, and hypertext links which were consequently garbled, and numerous tweets were not in English.

- {filetype} refers to arff:

arff: Each ARFF file contains a number of instances, proceeded by a a header. As an example, consider the first few lines of the train-most10.arff file:

@RELATION twitter-loc-most10 # This is simply a name for the dataset. Each of the attributes has a line

@ATTRIBUTE tweet-id NUMERIC # One line for the tweet ID, and

@ATTRIBUTE user-id STRING # One for the user ID, and

@ATTRIBUTE a NUMERIC # 10 for the (numeric) token frequencies.

@ATTRIBUTE i NUMERIC

@ATTRIBUTE im NUMERIC

@ATTRIBUTE lol NUMERIC

@ATTRIBUTE me NUMERIC

@ATTRIBUTE my NUMERIC

@ATTRIBUTE rt NUMERIC

@ATTRIBUTE the NUMERIC

@ATTRIBUTE to NUMERIC

@ATTRIBUTE u NUMERIC

@ATTRIBUTE class {NewYork,California,Georgia}

@DATA # This is the final line in the header; the instances follow

110,USER_ce270acf,0,0,0,0,0,0,0,0,0,0,NewYork

...

Note that when using development/test data, Weka will insist that the header is _exactly_ the same as the training data. Otherwise, it

will refuse to evaluate the model and/or predict the classes of the test instances.