Project 1: Who Tweeted That? 

Due:	17:00 Thurs 12th September 2019 AEST for Kaggle, Code archive.
Submission materials:	Predictions submitted through Kaggle Code archive
Assessment criteria:	Kaggle prediction accuracy;
Marks:	The Project will contribute 25% of your overall mark for the subject

Overview

The goal of the Project is to develop skills in researching and applying methods from subareas of Statistical Machine Learning (SML) that may be unfamiliar, by leveraging the fundamental knowledge built in lectures and workshops. The Project accomplishes this by challenging students with a difficult learning task: **predicting authors of test tweets from among a *very large* number of authors found in training tweets**. The Project also builds generic skills in problem solving, critical analysis, presentation/communication, and team work – all critical for practical SML.

Deliverables

1. The predicted labels of the test tweets submitted to the Kaggle in-class competition described below.
2. [Not marked[1]] A zipped archive of code (but please, no data). It need not be all code you wrote, but it should be able to run on lab machines to produce your final Kaggle predictions. You may use any freely available software you like. Submit this to LMS together with your final report.

Assessment Criteria

Kaggle performance: (7 marks), of which:

> Classification accuracy: (5 marks)
> Let *A* be the "categorization accuracy" % on Kaggle over the final private leaderboard (all of test data).
> Then this portion of your mark is taken to be (max(min(*A*,8),2)-2)*5/6.

Example 1: If you scored 2% (or less) you would receive 0 out of 5
Example 2: If you scored 5% you would receive 2.5 out of 5
Example 3: If you scored 8% (or more) you would receive 5 out of 5

Why Authorship Attribution?

I'm glad you asked! Authorship attribution has several applications. One is for identifying who wrote an ancient text to better understand its historical context; or similarly in literature, for books written under a pen name. Authorship attribution is closely related to plagiarism detection: did a student really submit this essay, or a professional essay writer? A related area is *stylometric attacks* in privacy research: online authors who might want to write text critical of their government or employer may go to pains to hide their IP address and name, but they should understand that their writing style can give their identity away. Demonstrating that authorship attribution is possible out in the open helps raise awareness to privacy risks of how data is collected, shared and released.[3] Finally, authorship attribution is an interesting application of so-called *xtreme classification* when framed as massively multiclass classification. Our hope is that you will look into some (not necessarily all) of these areas in developing your Project submissions.

Dataset

The data needed for the Project is available on Kaggle. Note, you must abide by rule (R1) below on data use.

> train_tweets.zip – contains a single text file representing the training set of over 328k tweets authored by roughly 10k Twitter users. The file is tab-delimited with two columns: (1) a random ID assigned to each user in the dataset; (2) transformed text of a tweet.
>
> test_tweets_unlabeled.txt – a text file representing the unlabelled test set of over 35k tweets authored by the same users as in the training dataset. These authors have been removed from the file. It is your task to identify them. An original file of all tweets combined was split with each tweet randomly assigned to test with probability 0.1 and train with probability 0.9 (subject to constraints that users appear in the training dataset). Thus, the test data is an approximately-stratified random sample.

The files are encoded as utf-8, with the tweets transformed somewhat:

- ✓ Whitespace (inc. tabs, newlines) have been converted to spaces, with repeat spaces stripped.
- ✓ Mentions like "@bipr" that would reveal useful user information have been converted to "@handle" so as to preserve some mention statistics.

Some generic pointers about Twitter data: tweets have short, restricted lengths; "RT" sometimes means retweet or in other words resharing of someone else's tweet; besides mentions being special, Twitter also permits hashtags such as #smlftw; images or other media that might have been tweeted have been removed. Other than these peculiarities of Twitter, tweets can be regarded as strings of text.