TweetLID corpus release V1.

Twitter Language Identification Workshop at SEPLN 2014

Last update: 2014/10/01

==========

License

==========

This resource is distributed research purposes, and includes both the content of the tweets as well as the manual annotation of the language each tweet is written in, following the guidelines of the TweetLID 2014 shared task.

The content of the tweets is distributed by adhering to the 6.b(i) clause of Twitter's TOS, which allows us to distribute a tweet collection of no more than 50,000 tweets: https://dev.twitter.com/overview/terms/agreement-and-policy#6._Be_a_Good_Partner_to_Twitter
The content of the tweets, hence, remains under Twitter's license (see I.B above).

The manual annotations we provide along with these tweets are released under the Creative Commons Attribution License (CC BY). The full details of this license can be found at http://creativecommons.org/licenses/by/3.0/legalcode

If you use this corpus please cite the following paper:

- Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza A., Fresno, V. (2014). Overview of tweetlid: Tweet language identification at sepln 2014. Proceedings of the TweetLID Worshop at SEPLN2014. Girona. pp. 1-11.

==========

Corpus files

==========

The corpus is composed of three annotated files:

--> '~~tweetLID-~~training-tweets.tsv': The file contains the training corpus for the TweetLID 2014 shared task. It includes 14,991 tweet IDs , user names and texts, as well as the language annotations for those tweets.

--> '~~tweetLID-~~test-tweets.tsv': The file contains the whole set of tweets composing the test corpus for the TweetLID 2014 shared task. It includes 19,993 tweet IDs, user names and texts, as well as the language annotations for those tweets.

~~--> 'tweetLID-testOfficial-7july.tsv': The file contains the final test corpus used during the TweetLID 2014 shared task. It includes 18,423 tweet IDs and user names, as well as the language annotations for those tweets, which were still available after the evaluation period. For the sake of~~

sparing some space, no tweet text is included in this file, but you can extract the actual tweets from the previous file.

NOTE: If you want to compare you results with those of tweetLID you should use this test set.

The format of the files is the following:

tweetId1<tab>user_screen_name1<tab>tweetText<tab>lang_annotation
tweetId2<tab>user_screen_name2<tab>tweetText<tab>lang_annotation
...

In addition, JSON formatted files are provided for train and test collections. They include the whole metadata information provided by Twitter for each status(tweet) in the collection. At the end of each element, a new field has been added, for the manual language annotation: "tweetlid_lang". These files are:

--> 'tweetLID-training-tweets.json' and 'tweetLID-test-tweets.json'.

==========

About the annotation of the corpus:

==========

Along with the tweet IDs and user names, the file also provides the manually annotated language(s) of the tweet. The
annotation uses the following names for languages:

* eu: Basque.
* ca: Catalan.
* gl: Galician.
* es: Spanish.
* en: English.
* pt: Portuguese.
* other: A different language from those listed above (e.g., French).
* und: Undeterminable, which means that the text of the tweet includes words that are widely used in any of the languages
  considered in the task, which makes it impossible to determine the language being used in that specific case.

In some cases, some tweets include more than a single language, annotated as follows:
* es/gl/pt: when a tweet is annotated with two or more languages separated by slashes, it means that the text of the tweet
  may have been written in any of those languages.
* es+eu: when a tweet is annotated with two more languages separated by plus signs, it means that the text of tweet
  contains parts in both languages.

Note that the corpus includes only tweets with at least one word (i.e., string fully made of a-z characters), and that
#hashtags and @user mentions have not been considered in the annotation of a tweet.

==========

--- Evaluation script

==========

This release includes the script used for evaluating the submissions during the TweetLID 2014 shared task.

--> tweetLID_eval.pl: This script takes as input two files, the reference we want to compare with (it defaults to 'tweetLID-testOfficial-7july.tsv' in the same directory) and a result data file. For help on how to use this script type:

% perl tweetLID_eval.pl --help