Language Identification

18 Nisan 17

Language identification is one of the most widely known text classification problems. The task is to predict the language of a given textual or spoken data.

The approaches to this problem are mostly based on statistical language models and the baseline model can be said to be those using the (word or character) n-gram weights of texts as features. Thus, it should not be expected that an identifier can recognize a language outside of its training set. Although the task is deemed as solved, there are some strong objections to it [1] since no system so far can recognize all the languages on earth where it is assumed that near 7000 languages exist [2].

We will not create a new model and use an existing method implemented in Python.

Available packages in Python

We list those whose performance scores are reliable.

1. LangID
    a. https://github.com/saffsd/langid.py
    b. No licence restriction
    c. 97 languages available
    d. Outputs confidence scores
    e. Performance:
    f. Allows training for more languages / re-training with more data
2. LangDetect
    a. https://github.com/Mimino666/langdetect
    b. Apache licence: no restriction
    c. 55 languages available
    d. Can output a confidence score (call detect_langs() instead of detect())
    e. Performance
        i. Accuracy said to be 99% (for only the available languages).
3. Google's language detector
    a. Directly from google API

i. https://cloud.google.com/translate/docs/detecting-language
ii. Has a quota and a price on exceeding; requires an API key.
iii. 160+ languages available
iv. Outputs a confidence score
v. Couldn't find a score but it is a very powerful identifier.

b. TextBlob
i. https://textblob.readthedocs.io/en/dev/quickstart.html#translation-and-language-detection
ii. Allows easier API calls but confidence scores are not available.

c. Detectlanguage.com
i. https://detectlanguage.com/
ii. 163 languages available
iii. Seems to be using google API but no explicit mention.
iv. Similar restrictions with google

Result

The most suitable package is LangID for its ease of use, confidence score, having pretrained texts in larger number of languages, licence and research-oriented background.

Resources

[1] Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pages 229–237, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
[2] Carnegie Mellon University, Robotics Academy, The FIRE Project. Number of languages. Available online at http://www.education.rec.ri.cmu.edu/fire/naclo/pages/Ling/Fact/num-languages.html. (Last visited on 2017-04-18).