Language Detection Proposal

DSC 680

Isabella Sturm

2022-02-10

Topic

The model(s) trained in this project will be able to detect the language of the input text and output the language used.

Business Problem

There are thousands of languages spoken around the world. While some languages, like English and Spanish, are common and somewhat easy to identify in written language, others are less common and/or much more difficult to determine the language.

It is useful to be able to identify the language of a text. For instance, being able to identify a foreign and unfamiliar language makes it possible to translate the text into a known language so the text can be read and understood.

Datasets

The data was downloaded from https://tatoeba.org/en/downloads. From this source, only the sentences (senteces.tar.bz2) was downloaded to get the tab delimited csv of sentences. There are three fields in this dataset: Sentence id, Lang, and Text. Lang is an ISO 639-3 language code, from a second dataset – downloaded from https://iso639-3.sil.org/code_tables/download_tables#termsofuse – the data will be used to map the language code to a more human-readable language (ie. eng will output English).

Tatoebo Sentences Data

Column	Description
Sentence id	Unique identifier
Lang	ISO-639-3 language code

Text	Sentence

ISO 639-3 Data

Column	Description
639-3	ISO 639-3 language code
639-2/639-5	ISO 639-2/639-5 language code
639-1	IS 639-1 language code
Language Name(s)	Language that correlates to the language names
Scope	Whether language is Individual or Collective
Language Type	Whether language is Living, Extinct, Genetic, Constructed

Methods

The Tatoeba dataset contains nearly 7 million sentences in over 300 languages. First, I will limit the number of unique languages to the top twenty-five. Then I will make sure the data is balanced by ensuring there are close to equal number of records for each language.

To prepare for modelling, I will need to use bag-of-words or ngrams then vectorize them as features and encode the target Lang. Then I can start training the model. I will start by training a Naïve Bayes model. If I have time, I will also try creating a Neural Network that will determine the language of the text.

Ethical Considerations

There are no ethical considerations that stand out to me for this project. Unless there are sentences in the dataset that have personal and identifying information, there are not ethical concerns with the data itself. Additionally, the datasets are free to use with attribution, so there is no major concern about obtaining the data directly from the sources indicated.

Challenges/Issues

I chose to do a natural language processing (NLP) project because it is an area I have struggled with throughout the program. It is possible I will continue to have issues and some delays working through this project as I gain a better understanding how to analyze and work with text data using NLP.

References

Collection of sentences and translations. Tatoeba. (n.d.). Retrieved February 13, 2022, from https://tatoeba.org/en/

Poirot, L. (2022, January 20). *25 most spoken languages in the world*. Far & Wide. Retrieved February 13, 2022, from https://www.farandwide.com/s/most-spoken-languages-eebb682e26eb4bc5

SIL International. *ISO 639-3* |. ISO639-3. Retrieved February 13, 2022, from https://iso639-3.sil.org/