Language Detection Draft

DSC 680

Isabella Sturm

2022-02-24

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.

Background

Data Explanation

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 limited the dataset using the top 25 languages based on the most spoken languages around the world by population – this was a list I found from www.farandwide.com. The dataset didn't have four of the languages and had limited samples for many of the languages from this list, so I also trained a model using the top 25 languages based which languages had the most samples in the dataset.

Once I had each dataset filtered by language, I then used random sampling to under sample majority languages so the number of records for each language were roughly equal. For the top 25 languages based on population spoken around the world, there were about 500 samples for each language; for top 25 languages based on most samples in the original dataset had about 50,000 samples per language.

Then, I split the data into train and test sets, and I used CountVectorizer to create bigrams and used the bigram features to train a Multinomial Naïve Bayes model and make predictions.

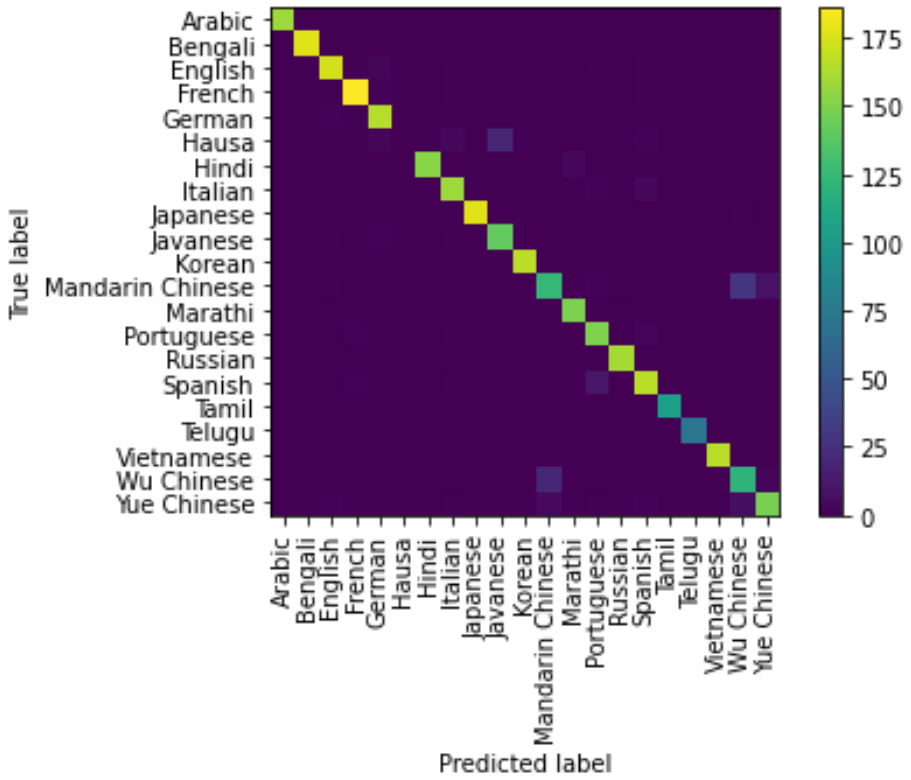
Analysis

After training the models, I used the test datasets to make predictions and used a classification report to numerically show how well my model preformed and plotted the confusion matrix as a more visual representation of the model results.

Both models detected the correct language 95% of the time. Here is the breakdown of the model prediction for the top 25 languages by population.

	precision	recall	f1-score	support
Arabic	1.00	1.00	1.00	159
Bengali	1.00	1.00	1.00	177
English	0.96	0.96	0.96	180
French	0.96	0.98	0.97	190
German	0.93	0.98	0.96	168
Hausa	0.00	0.00	0.00	30
Hindi	0.99	0.97	0.98	157
Italian	0.94	0.95	0.94	168
Japanese	1.00	0.99	1.00	179
Javanese	0.88	0.97	0.92	145
Korean	1.00	0.99	1.00	168
Mandarin Chinese	0.83	0.75	0.79	165
Marathi	0.97	0.99	0.98	150
Portuguese	0.90	0.96	0.93	156
Russian	0.99	1.00	1.00	161
Spanish	0.93	0.92	0.93	181
Tamil	1.00	1.00	1.00	105
Telugu	1.00	1.00	1.00	73
Vietnamese	1.00	1.00	1.00	166
Wu Chinese	0.77	0.84	0.80	144
Yue Chinese	0.93	0.89	0.91	166
accuracy			0.95	3188
macro avg	0.90	0.91	0.91	3188
weighted avg	0.94	0.95	0.94	3188

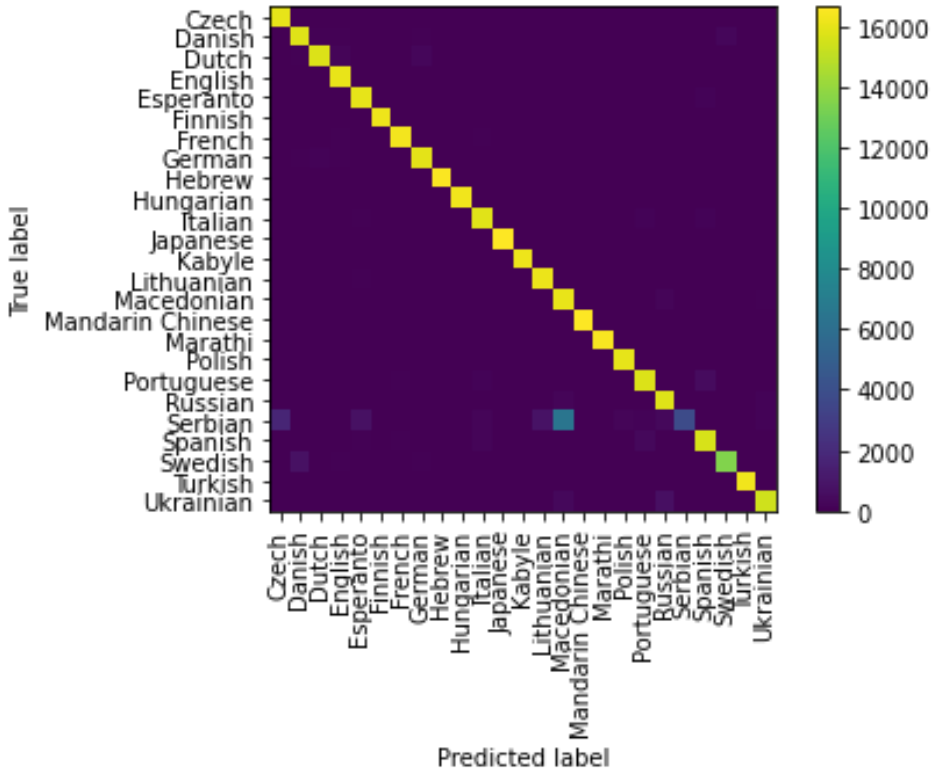
By looking at the confusion matrix below, it is no surprise that the most notable incorrect predictions were when trying to detect and differentiate between the three different Chinese languages in the dataset: Mandarain, Wu and Yue.



Below is the classification report and confusion matrix of the second model using the top 25 languages based on number of samples available in the original dataset.

	precision	recall	f1-score	support
Czech	0.90	0.98	0.94	16306
Danish	0.93	0.95	0.94	16612
Dutch	0.97	0.95	0.96	16566
English	0.95	0.98	0.96	16465
Esperanto	0.92	0.98	0.95	16401
Finnish	0.98	0.99	0.99	16486
French	0.98	0.98	0.98	16723
German	0.96	0.97	0.97	16384
Hebrew	1.00	1.00	1.00	16573
Hungarian	0.99	0.98	0.99	16557
Italian	0.95	0.96	0.95	16532
Japanese	1.00	1.00	1.00	16671
Kabyle	1.00	0.99	0.99	16368
Lithuanian	0.94	0.98	0.96	16425
Macedonian	0.70	0.98	0.81	16415
Mandarin Chinese	1.00	1.00	1.00	16694
Marathi	1.00	1.00	1.00	16473
Polish	0.98	0.99	0.98	16271
Portuguese	0.94	0.95	0.95	16568
Russian	0.92	0.97	0.95	16268
Serbian	1.00	0.25	0.40	15050
Spanish	0.93	0.95	0.94	16512

Swedish	0.97	0.92	0.94	14671
Turkish	0.99	0.99	0.99	16524
Ukrainian	0.97	0.94	0.96	16483
accuracy			0.95	408998
macro avg	0.95	0.95	0.94	408998
weighted avg	0.95	0.95	0.94	408998



Looking at the confusion matrix for this second model, again there are no surprises. In this case there is only one Chinese language so the model was able to predict Chinese with no issue, but instead occasionally mixed up Macedonian and Serbian. This is no surprise because both of these languages come from countries in the Balkans (Southeast Europe) so they likely have similar influences and therefore similarities in the languages.

Conclusion

Both models, even with differences in the sample sizes, has highly accurate models. Neither model is better for all scenarios. The ideal model to deploy would be one that uses the languages that are most likely to come up in a scenario where you would apply the model. For instance, if you are at war – say the U.S. at war with the Middle East – a useful model would be trained using samples from a variety of languages that include but are not limited to English, Arabic and at least some known popular minority languages so you can detect and then translate intercepted correspondence.

Since both models resulted in high accuracy even though different languages and different sample sizes were used, it seems likely that we would have a similar outcome if we selected languages with more business knowledge at hand and took similar steps to prepare and train the model.

Assumption

There was one major assumption made – only to top 25 languages are necessary for the language detection model. This number was somewhat arbitrary, but in both models I used some sort of additional knowledge to select the languages. In one case, I was training my model based on how many people would theoretically find the language detection useful – if more people are speaking a language then the language is more likely to come up in a scenario when you need to figure out what language you are translating from. In the second case, I am assuming that because there are more samples for language, that means the language is more likely to come up in a language detection/translation scenario.

Limitations

One limitation of this dataset is that the data comes from entries submitted by users around the world. In the case of my first model, some languages that are spoken by millions of people have little to no samples. This could be attributed to either the lack of resources for people who speak those languages to make their submissions or in general, to make a submission you need to know about the Tatoeba website and mission to do so.

A limitation of the project is that I am only focusing on a small set of languages. This means that the application of the model will only be useful for detecting languages used to predict the model.

Challenges

Aside from some classwork that touched on natural language processing (NLP), this is the first project performing NLP and I knew there would be some challenges. One challenge I came across was trying to use CountVectorizer on my dataset. After my computer restarted on its own while trying to vectorize the dataset, I discovered I did not need to this and created a pipeline for a multinomial Naïve Bayes model that used bigrams. The other challenge I came across was balancing my dataset. I wanted to use the best method possible, but after some research, with text data, I found there is no one best solution and ended up under sampling my data to get even classes.

Future Uses/Additional Applications

A future extension of this project would be to use the audio files from the Tatoeba dataset to create a model that can detect spoken language. Additionally, in both this current project and the audio detection suggestion, a future project would be to include translations of the written or spoken sentences.

Recommendations

One recommendation for this project would be to train a model using all the languages. One limitation here is that some languages have very few samples, so we would either need to find another dataset to join with this one, or somehow randomly create new samples. Additionally, we could limit the languages by requiring there to be at least 500 samples for instance. Training a model with more languages would be ideal for making an application that is able to differentiate between different dialects or languages that have similar influences. As an example, in the case of East Asian languages, if you are missing languages, you might not be able to differentiate Korean from Japanese or Chinese.

Implementation Plan

Ethical Assessment

On the surface, language detection does not seem like it would have many ethical implications. However, with natural language processing, there are more than I realized. With written language detection, one ethical concern is that in the wrong hands it could be used to

train a computer to mimic a person's language and going further, how they speak in written language. This opens doors for faking identities in an online or virtual space.

If we go further and use spoken language, this adds an additional layer of realism to something that could be faked by a computer. People could be led to believe a faked online identity is real by speaking to it over a text chat, but the reality of this faked identity becomes even more real when it can also speak the language. These fake identities are scary and can lead to potentially very dangerous situation.

However, in the right hands, this type of model can be used to connect or even protect in certain scenarios people across the world.

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/