CS505 HW5
Yuhao He

English and Spanish Prediction:
Data Cleaning:

1. For each tweet, used *tweet preprocessor* to remove URLs, SIMLELYs et cetera.
2. Change words to lowercase
3. Remove punctuation and stop word
4. Get stem for words

Word Embedding:
I choose TFIDF Vectorizer provided by *sklearn.* The reason for that is it can help us eliminate the inference caused by some common words, which occurs a lot but cannot provide us much useful information.

Model Selection:
I tried Logistic Regression and SVM for the emoji predictions of the English and Spanish. The best result I got shown in below:

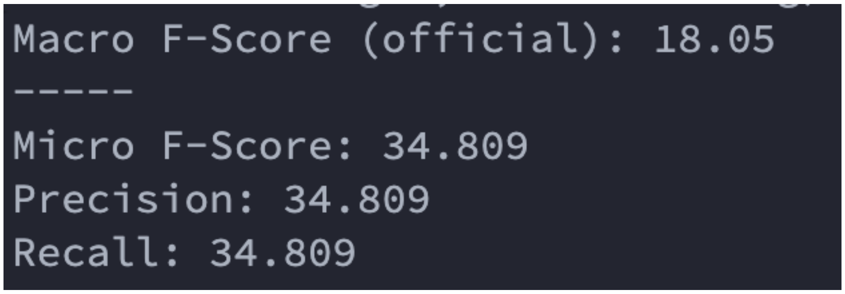
	Logistic Regression	SVM
English	21.748	12.142
Spanish	20.744	12.498

In the Jupyter notebook, you might found that I have tried a range of C to determine which one could give us a better score, in SVM, I didn't do that, because SVM is not very efficient with large number of observations, and it took me a long time to wait for it returns a result. I also tried a different ngram range for the TFIDF Vectorization, but it seems that the default parameter returns the best score.

Improve Prediction by Translating Spanish to English

To translate Spanish to English, the first thing I do is to remove labels of emojis that only occur in one language. Then I reorder the label for Spanish so the same emoji now share the same index in different language. After that I perform the same data cleaning schema as above.

The result I got show below, which didn't improve the predict scores.



```
Macro F-Score (official): 18.05
-----
Micro F-Score: 34.809
Precision: 34.809
Recall: 34.809
```