Translating a Classifier

Patrick Martin

February 2, 2018

1 Overview

1.1 The Data

In order to work on translating a classifier, I need similar data in different languages. The first thing I'm going to use is reddit, starting with March 2017 (and adding others if I need more data), using langdetect (https://pypi.python.org/pypi/langdetect) on comments that have at least 15 unique words. I'm going to try to pull 10k each of Spanish, French, and Italian. Unfortunately this is either slow or sparse, and is taking a while. Hopefully it will finish by tomorrow or something.

Alternatively/additionally, if I can find different-language wikipediae, I can use page category as the classes.

1.1.1 LID data

We can make sure the LID worked reasonably well by checking the subreddits represented

Subreddit	Count	Subreddit	Count	Subreddit	Count
argentina	4,632	france	7,772	italy	7,862
mexico	1,840	Quebec	1,052	oknotizie	525
podemos[1]	1,622	montreal	197	ItalyInformatica	351
chile	592	ParisComments	116	italy_SS	327
vzla	280	French	44	italygames	86
Argaming	87	Lyon	35	perlediritaly	69
Spanish	80	FiascoQc	34	lisolachece	62
uruguay	75	SquaredCircle_FR	32	Romania	61
PuertoRico	50	effondrement	27	ItaliaPersonalFinance	54
Colombia	44	melenchon	23	ItalyMotori	40

1.1.2 Subreddit corpora

Using the data from the LID stuff, we can also just create the corpora by using all the posts from some subreddits. I propose[2]

[1] Spanish political party

[2] We'll fix this later

Spanish	French	Italian
argentina	france	italy
mexico	Quebec	
chile	montreal	
vzla	Lyon	
uruguay		
Colombia		