Differentiating between romance languages given a block of text
Brynne Henderson
Level of Interest: 10/10

Description of project
There are many languages in our modern world that share ties to the Roman alphabet and lettering system. Most Americans can recognize at least a few of these languages on sight, but hardly anyone could recognize the majority. This machine learning model would be able to, given a block of text, determine which Roman-based language the text belongs to.

Features of the dataset
There are many different features that this dataset could have. I hypothesize the most critical features would stem from how the individual words are structured and from the commonality of specific letters, rather than from sentence structure or grammar. This hypothesis leads me to focus on features such as the average length of a word, the number of vowels in the block of text, the most common single-letter word in the block of text, and so on. Although many of these features are continuous, a few would be nominal. An example table with data, features, and classification is as follows:

Input	**Avg Word Length**	**Num Vowels**	**One-Letter Word**	**Classification**
A sentence is defined as "a group of words, usually containing a verb, that expresses a thought".	5.411	31	'a'	English
Las oraciones simples consisten en un sujeto unido a un predicado.	5.909	25	'a'	Spanish
Die afkappingsteken word hier gebruik omdat die woord vroeër syn was, maar dié spelling word selde nog in hedendaagse Afrikaans gebruik.	6.381	47	?	Afrikaans

How to gather data
Data can be gathered from articles on Wikipedia, which has a wide variety of articles in all the languages in question. This can be done using web-scraping libraries or tools such as Beautiful Soup. The data would already be labeled correctly, but would need to be cleaned. Some sentences may not contain any one-letter words, and this missing data would need to be taken into account when engineering the model. The need for normalization is also likely.