Veronica Espinoza
Diana Kuwano
LING 165
Final Project

English Language Recognition

Problem Statement

Can we build a text classifier that uses bigrams to effectively determine words as being English or foreign?

Inspiration came from the Spam Lab. We wanted to try working on a text classification problem with a different type of data. In particular we wanted to create binary classifier that would be able to determine if a given word is English or not. We trained our English data against Spanish in a supervised learning setting.

Literature Survey

Different approaches to solving text classification problems

- Naïve Bayes
- Maximum Entropy

Here are two different approaches to solving this problem. The first being Naïve Bayes text classification which would allow us to calculate the probability of a word being English using a simple but effective method of assuming that the features, in our case bigrams, were conditionally independent of each other. An alternative to this being maximum entropy and in this case, although being a probabilistic method as well, we would have to used weighted features. Our choice in the matter was truthfully due to simplicity, we had one feature type and one category thus we chose to work with Naïve Bayes classification.

Data Preparation

Collected data from The New York Times
English words are labeled ‘0’ and foreign words are labeled ‘1’
File format: label, tab space, word
File types: train and test files

We sorted through the data and got rid of punctuation, capitalized proper nouns, and duplicates We labeled the English words with ‘0’ and the Spanish words with ‘1’

We shuffled all of the English words and all of the Spanish words in 2 separate files
We selected 900 from each for the train file, 100 from each for one test file, and 100 from each for another
We shuffled the order within each of the three files
1 train file: with 900 English words and 900 Spanish words
2 test files: each with 100 English words and 100 Spanish words
All of the words are unique

Implementation

Naïve Bayes classifier
Separate words into two groups: English (0) and foreign (1)
Create dictionaries for the English and foreign groups
Apply add-one smoothing
Change the counts to probabilities
Use the probabilities to classify a word as English or foreign
Test the data

Our code is based off of Lab 1
The code looks at the train file, which has 900 English and 900 Spanish words
Then, it looks at the labels for each word and separates the words into two groups: English (0) and foreign (1)
The keys of the dictionaries are the bigrams and the values are the counts
First, the program needs to identify and count the bigrams within each group
We also need to pad the words, so we add a word boundary symbol at the beginning and end of each word
Then, the program looks at the label
If the bigram is already in that dictionary, then the program increments the frequency
If not, the program adds the bigram to the dictionary
We want to add a dummy bigram in case the program encounters a bigram that isn't found in our train file
Then we apply add-one smoothing by adding 1 to each frequency
We do this by dividing the frequency of each bigram by the total tokens of bigrams
We defined a function that calculates the log probability that a word is English or foreign
It calculates the scores by adding the log probabilities for each bigram in the word

So we get an English score and a foreign score

Whichever is greater determines what label the program assigns to the word

The program looks at the test file, which has 100 English words and 100 Spanish words

We need to pad the words and identify the bigrams for each word

Then, the program can calculate the log probabilities by looking at each bigram in each word and assign a label based on which score is higher, English or foreign

Results

causes	true	episodes		resulte
square	began	case		federales
molestation	includes	base		reptiles
candidate	lobbies	arcade		apropiarse
arrival	divide	are		tus
exposes	alone	politicians		noche
causes	global	comes		dure
able	race	invisible		seres
compares	carpenter	donor		orbe
area	canes	does		relativamente
value	miles	same		suroeste
defendants	nonnegotiable	if		
local	decade			

These are the words our classifier identified incorrectly. Interestingly , even though the test data had the same number of English and Spanish words, more English words classified as Spanish as opposed to the other way around. This is likely a result of the fact that the Spanish data shares quite a number of bigrams with English words.

Bigram Evaluation

Test 1

Precision: 0.934

Recall: 0.84

Test 2

Precision: 0.939

Recall: 0.77

Because we had two test files, 200 words each 100 of English and 100 of Spanish the results here are shown for both. Our classifier did a pretty decent job selecting actual English words from the one predicted to be English in that precision for both tests were both above 90%. However performance lacked in terms of recall as the classifier calculated a lower percentage for predicting English words out of those that were actually English which is reflected in the previous chart showing all of the English words that it misclassified.

Final n-gram Evaluation

2-gram (Test 1)

Accuracy: 0.895

Precision: 0.934

Recall: 0.85

F-1 score: 0.445

2-gram (Test 2)

Accuracy: 0.86

Precision: 0.939

Recall: 0.77

F-1 score: 0.423

3-gram (Test 1)
Accuracy: 0.905
Precision: 0.965
Recall: 0.84
Fscore: 0.449

3-gram (Test 2)
Accuracy: 0.905
Precision: 0.976
Recall: 0.83
Fscore: 0.448

4-gram (Test 1)
Accuracy: 0.92
Precision: 0.966
Recall: 0.87
Fscore: 0.457

4-gram (Test 2)
Accuracy: 0.915
Precision: 0.956
Recall: 0.87
Fscore: 0.455

5-gram (Test 1)
Accuracy: 0.92
Precision: 0.956
Recall: 0.88
F-1 score: 0.458

5-gram (Test 2)
Accuracy: 0.925
Precision: 0.956

Recall: 0.89

F-1 score: 0.461

Sources

https://www.nytimes.com/2017/05/04/us/university-of-kentucky-stolen-test.html?rref=collection%2Fsectioncollection%2Fus&action=click&contentCollection=us&region=stream&module=stream_unit&version=latest&contentPlacement=4&pgtype=sectionfront&_r=1

https://www.nytimes.com/2017/05/08/us/penn-state-prosecutors-fraternity-hazing-deaths.html?action=click&contentCollection=us&region=rank&module=package&version=highlights&contentPlacement=2&pgtype=sectionfront

https://www.nytimes.com/2017/05/08/us/legal-immigrants-who-oppose-illegal-immigration.html?rref=collection%2Fsectioncollection%2Fus&action=click&contentCollection=us&region=rank&module=package&version=highlights&contentPlacement=1&pgtype=sectionfront

https://www.nytimes.com/2017/05/09/magazine/how-homeownership-became-the-engine-of-american-inequality.html?rref=collection%2Fsectioncollection%2Fus&action=click&contentCollection=us&region=rank&module=package&version=highlights&contentPlacement=2&pgtype=sectionfront

https://www.nytimes.com/2017/05/09/climate/arctic-nations-to-meet-amid-unsettled-us-stance-on-climate-change.html?rref=collection%2Fsectioncollection%2Fus&action=click&contentCollection=us&region=stream&module=stream_unit&version=latest&contentPlacement=1&pgtype=sectionfront

https://www.nytimes.com/es/2017/05/04/gustavo-dudamel-condena-la-represion-en-venezuela-ya-basta-de-desatender-el-justo-clamor-de-un-pueblo-sofocado/

https://www.nytimes.com/es/2017/05/09/el-caso-contra-el-chapo-guzman-10-000-documentos-1500-grabaciones-y-decenas-de-testigos/?action=click&contentCollection=noticias&region=rank&module=package&version=highlights&contentPlacement=1&pgtype=collection

https://www.nytimes.com/es/2017/05/09/un-altar-dedicado-a-la-ciencia-abre-sus-puertas-en-miami/?action=click&contentCollection=noticias&region=rank&module=package&version=highlights&contentPlacement=2&pgtype=collection

https://www.nytimes.com/es/2017/05/09/otro-obstaculo-para-el-muro-de-trump-los-duenos-de-las-tierras-en-texas/?action=click&contentCollection=noticias&region=rank&module=package&version=highlights&contentPlacement=2&pgtype=collection

https://www.nytimes.com/es/2017/05/05/amamos-demasiado-a-los-celulares-y-no-hablamos-de-eso/?ref=nyt-es&mcid=nyt-es&subid=contentband&mccr=noticias&action=click&contentCollection=noticias&region=rank&module=package&version=highlights&contentPlacement&pgtype=collection

https://www.nytimes.com/es/2017/05/09/lo-mejor-que-puedes-comer-antes-de-ejercitarte-segun-un-estudio-nada/?rref=collection%2Fsectioncollection%2Farchive&action=click&contentCollection=noticias&region=stream&module=stream_unit&version=latest&contentPlacement=2&pgtype=collection

Manning, C. D., Raghavan, P., & Schutze, H. (2009). An Introduction to Information Retrieval. Cambridge University Press.