English Language Recognition

VERONICA ESPINOZA

DIANA KUWANO

Problem Statement

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

Literature Survey

Different approaches to solving text classification problems

- Naïve Bayes
- Maximum Entropy

Data Preparation

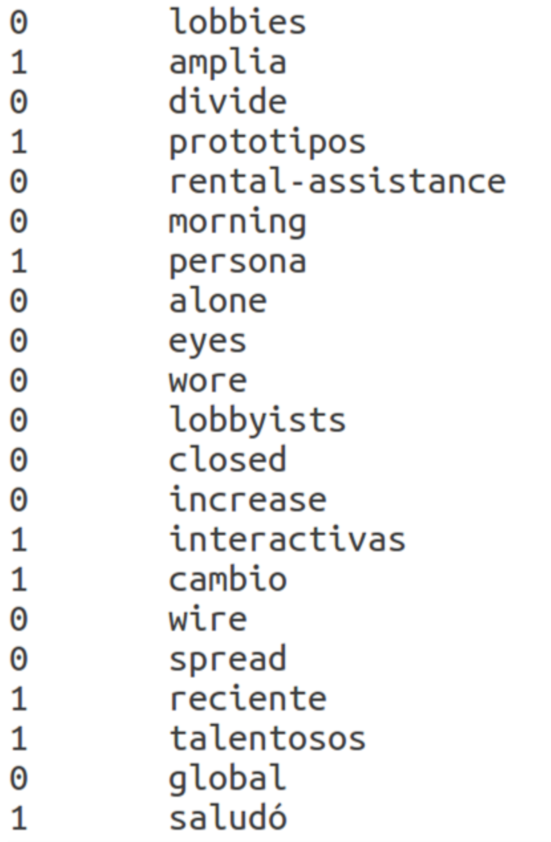
Collected data from The New York Times

English words are labeled ‘0’ and foreign words are labeled ‘1’

File format: label, tab space, word

File types: train and test files

Data Preparation



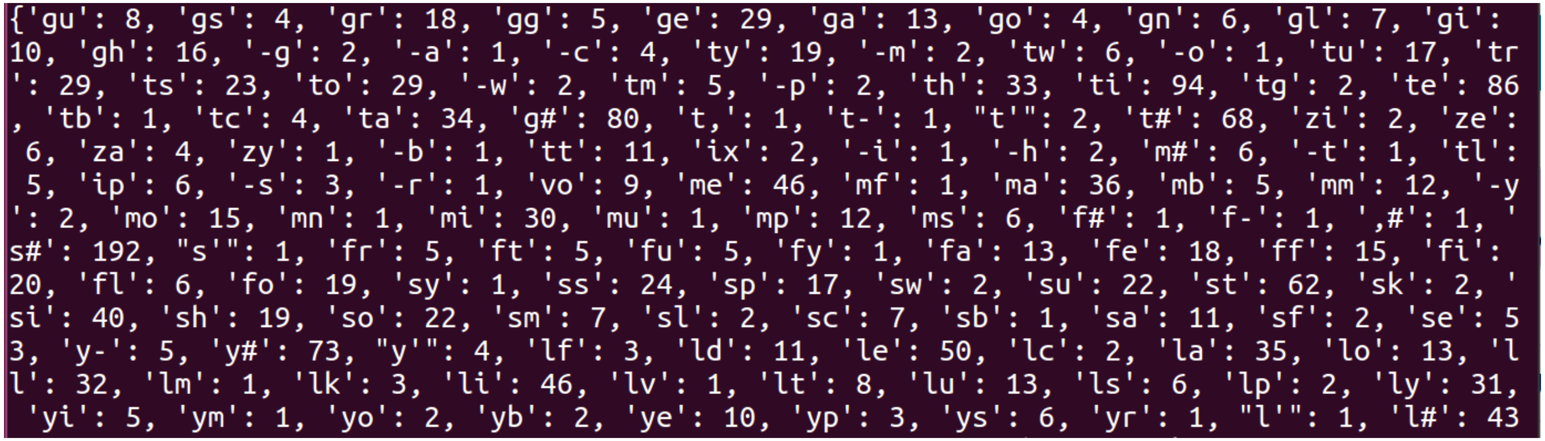
```
0       lobbies
1       amplia
0       divide
1       prototipos
0       rental-assistance
0       morning
1       persona
0       alone
0       eyes
0       wore
0       lobbyists
0       closed
0       increase
1       interactivas
1       cambio
0       wire
0       spread
1       reciente
1       talentosos
0       global
1       saludó
```

Implementation

Naïve Bayes classifier

Separate words into two groups: English (0) and foreign (1)

Create dictionaries for the English and foreign groups

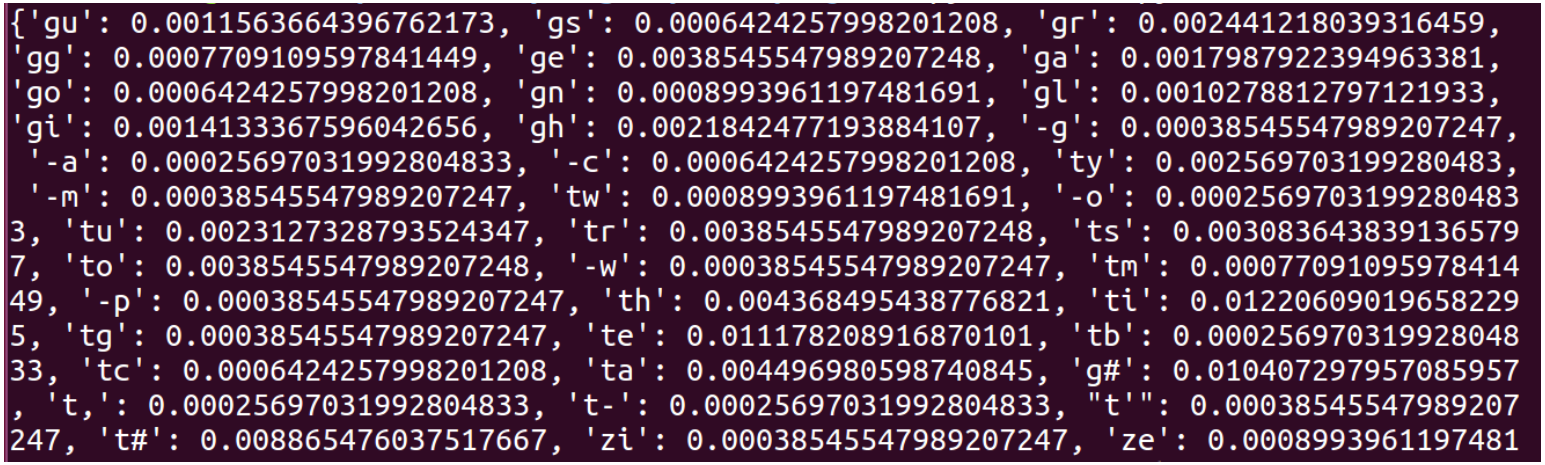


```
{'gu': 8, 'gs': 4, 'gr': 18, 'gg': 5, 'ge': 29, 'ga': 13, 'go': 4, 'gn': 6, 'gl': 7, 'gi':
10, 'gh': 16, '-g': 2, '-a': 1, '-c': 4, 'ty': 19, '-m': 2, 'tw': 6, '-o': 1, 'tu': 17, 'tr
': 29, 'ts': 23, 'to': 29, '-w': 2, 'tm': 5, '-p': 2, 'th': 33, 'ti': 94, 'tg': 2, 'te': 86
, 'tb': 1, 'tc': 4, 'ta': 34, 'g#': 80, 't,': 1, 't-': 1, "t'": 2, 't#': 68, 'zi': 2, 'ze':
 6, 'za': 4, 'zy': 1, '-b': 1, 'tt': 11, 'ix': 2, '-i': 1, '-h': 2, 'm#': 6, '-t': 1, 'tl':
 5, 'ip': 6, '-s': 3, '-r': 1, 'vo': 9, 'me': 46, 'mf': 1, 'ma': 36, 'mb': 5, 'mm': 12, '-y
': 2, 'mo': 15, 'mn': 1, 'mi': 30, 'mu': 1, 'mp': 12, 'ms': 6, 'f#': 1, 'f-': 1, ',#': 1, '
s#': 192, "s'": 1, 'fr': 5, 'ft': 5, 'fu': 5, 'fy': 1, 'fa': 13, 'fe': 18, 'ff': 15, 'fi':
20, 'fl': 6, 'fo': 19, 'sy': 1, 'ss': 24, 'sp': 17, 'sw': 2, 'su': 22, 'st': 62, 'sk': 2, '
si': 40, 'sh': 19, 'so': 22, 'sm': 7, 'sl': 2, 'sc': 7, 'sb': 1, 'sa': 11, 'sf': 2, 'se': 5
3, 'y-': 5, 'y#': 73, "y'": 4, 'lf': 3, 'ld': 11, 'le': 50, 'lc': 2, 'la': 35, 'lo': 13, 'l
l': 32, 'lm': 1, 'lk': 3, 'li': 46, 'lv': 1, 'lt': 8, 'lu': 13, 'ls': 6, 'lp': 2, 'ly': 31,
 'yi': 5, 'ym': 1, 'yo': 2, 'yb': 2, 'ye': 10, 'yp': 3, 'ys': 6, 'yr': 1, "l'": 1, 'l#': 43
```

Implementation

Apply add-one smoothing

Change the counts to probabilities



```
{'gu': 0.0011563664396762173, 'gs': 0.0006424257998201208, 'gr': 0.002441218039316459,
'gg': 0.0007709109597841449, 'ge': 0.0038545547989207248, 'ga': 0.0017987922394963381,
'go': 0.0006424257998201208, 'gn': 0.0008993961197481691, 'gl': 0.0010278812797121933,
'gi': 0.0014133367596042656, 'gh': 0.002184247719388410 7, '-g': 0.00038545547989207247,
 '-a': 0.00025697031992804833, '-c': 0.0006424257998201208, 'ty': 0.002569703199280483,
 '-m': 0.00038545547989207247, 'tw': 0.0008993961197481691, '-o': 0.00025697031992804833
3, 'tu': 0.0023127328793524347, 'tr': 0.0038545547989207248, 'ts': 0.0030836438391365797
7, 'to': 0.0038545547989207248, '-w': 0.00038545547989207247, 'tm': 0.000770910959784149
49, '-p': 0.00038545547989207247, 'th': 0.004368495438776821, 'ti': 0.0122060901965822 9
5, 'tg': 0.00038545547989207247, 'te': 0.011178208916870101, 'tb': 0.0002569703199280 48
33, 'tc': 0.0006424257998201208, 'ta': 0.004496980598740845, 'g#': 0.010407297957085957
, 't,': 0.00025697031992804833, 't-': 0.00025697031992804833, "t'": 0.00038545547989207
247, 't#': 0.008865476037517667, 'zi': 0.00038545547989207247, 'ze': 0.0008993961197481
```

Implementation

Use the probabilities to classify a word as English or foreign

Test the data

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			

Evaluation

Test 1

Precision: 0.934

Recall: 0.84

Test 2

Precision: 0.939

Recall: 0.77

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.