Damegender

David Arroyo Menéndez

[2019-08-21]

Outline

I have a string, I want the sex

All is simple in the beginning

```
$ python3 main.py David
David's gender is male
probability: 1.0
363559  males for David from INE.es
0 females for David from INE.es

$ python3 main.py Isabel
Isabel's gender is female
probability: 1.0
0 males for Isabel from INE.es
271166  females for Isabel from INE.es
```

Perhaps there are non binary probabilities . . .

All is possible if one name is found in different countries

```
$ python3 main.py Andrea
Andrea's gender is female
probability: 0.9808615955404946
2084  males for Andrea from INE.es
106807  females for Andrea from INE.es

$ python3 main.py Alex
Alex's gender is male
probability: 0.9966257742642983
41351  males for Alex from INE.es
140  females for Alex from INE.es
```

My string has different sex in different countries

. . .

Genderguesser (old sexmachine) did work for us

```
$ python3 nameincountries.py Andrea
grep -i " Andrea " files/names/nam_dict.txt > files/grep.
males: ['Italy']
females: ['Albania', 'Austria', 'Belgium', 'Bosnia and He
both: []

$ python3 nameincountries.py Alex
grep -i " Alex " files/names/nam_dict.txt > files/grep.tm
males: ['Azerbaijan', 'Denmark', 'East Frisia', 'France',
females: []
both: []
```

Now, string is using nicknames . . .

We can find a name called "silla". What is the gender of this string?

```
$ python3 main.py silla
silla gender predicted is female
0 males for silla from INE.es
0 females for silla from INE.es
```

The string is not in the dataset. But with damegender we can predict a gender using artificial intelligence. The classification such as with spam is only to reduce time or earn money for humans. It is not exact!!

With this command, we could count males and females in git, mailing lists, etc.

Now, you could count males and females with mails and git:

```
$ python3 mail2gender.py http://mail-archives.apache.org/r
The number of males sending mails is 5
The number of females sending mails is 1

$ python3 git2gender.py https://github.com/chaoss/grimoir
The number of males sending commits is 17
The number of females sending commits is 13
```

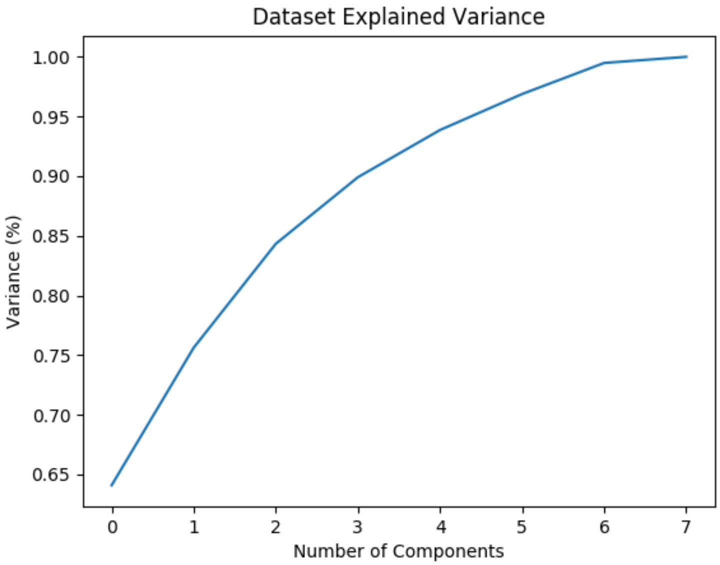
What features in a string is determining the sex?

```
$ python3 infofeatures.py
---------------------------------------------------------
Females with letter/s a: 0.7657420999768214
Males with letter/s a: 0.6717175543601788
---------------------------------------------------------
Females with last letter a: 0.4705246078961601
Males with last letter a: 0.1691037199787826
---------------------------------------------------------
Females with last letter o: 0.01730665224456464
Males with last letter o: 0.1075839078718084
---------------------------------------------------------
Females with last letter consonant: 0.2735841767750908
Males with last letter consonant: 0.4873854079854534
---------------------------------------------------------
Females with last letter vocal: 0.7262612995441552
```

A previous step to Machine Learning. PCA or not PCA (Principal Component Analysis)

```
$ python3 pca-components.py --csv='files/features_list_no
```

PCA or not PCA (Principal Component Analysis)



PCA or not PCA (II)

```
$ python3 pca-features.py --categorical="both" --componen
$ firefox files/pca.html &
```

PCA or not PCA (III)

first_letter	last_letter	last_letter_a	first_letter_vocal	last_letter_vocal	last_letter_consonant	target component
-0.2080025204	-0.3208958517	0.2352509625	0.2113242731	***0.6095269139***	***-0.6095269139***	-0.1035071139
-0.6037951881	***0.5174873789***	-0.4252467151	0.4278794455	0.0388287435	-0.0388287435	-0.0265942125
0.1049343046	0.1158117877	-0.2867605971	-0.3473950734	0.0901034539	-0.0901034539	-0.8697264971
0.2026467275	0.3142402839	***0.630802294***	***0.5325769702***	-0.1291229841	0.1291229841	-0.3811720011

In this analysis, we can observe 4 components.
The first component is about if the last letter is vocal or consonant.
If the last letter is vocal we can find a male and if the last letter is a consonant we can find a female.
The second component is about the first letter. The last letter is determing females and the first letter is determing males.
The third component is not giving relevant information.
The fourth component is giving tha $\text{last}_{\text{lettera}}$ and the $\text{first}_{\text{lettervocal}}$ is for females.
So, we have our scientific intutions to compose the machine learning model

Measuring tools and machine learning algorithms

APIs

	Accuracy
Genderapi	0.968768696648212
Namsor	0.753957037874505
Genderize	0.71537591859807
Gender Guesser	0.690220463538722

Machine Learning Algorithms

Support Vector Machines accuracy	0.7049180327868853
NLTK bayes	0.6677501413227812
Bernoulli Naive Bayes	0.5962408140192199
Gaussian Naive Bayes	0.5960994912379876
Stochastic Gradient Descendent accuracy	0.5873374788015828
Multinomial Naive Bayes	0.5876201243640475

License