Predicting age and gender of artists based on songs and lyrics

Marta Kuziora, EPFL 2018

Project overview

1. Introduction
2. Data downloading
3. Dataset preparation
4. Data analysis
5. Age and gender prediction

Data downloading

Spotify®

MusicBrainz

musiXmatch®
not only words

WIKIDATA

Dataset

Musicbrainz artist:
- id
- name
- year
- area (country)
- albums

Spotify artist:
- popularity
- followers
- genres
- albums

Spotify songs features:
- danceability
- loudness
- violence
- instrumentalness
- ...

Lyrics
- bag of words
- 5000 most popular

Dataset preparation

Dataset merging

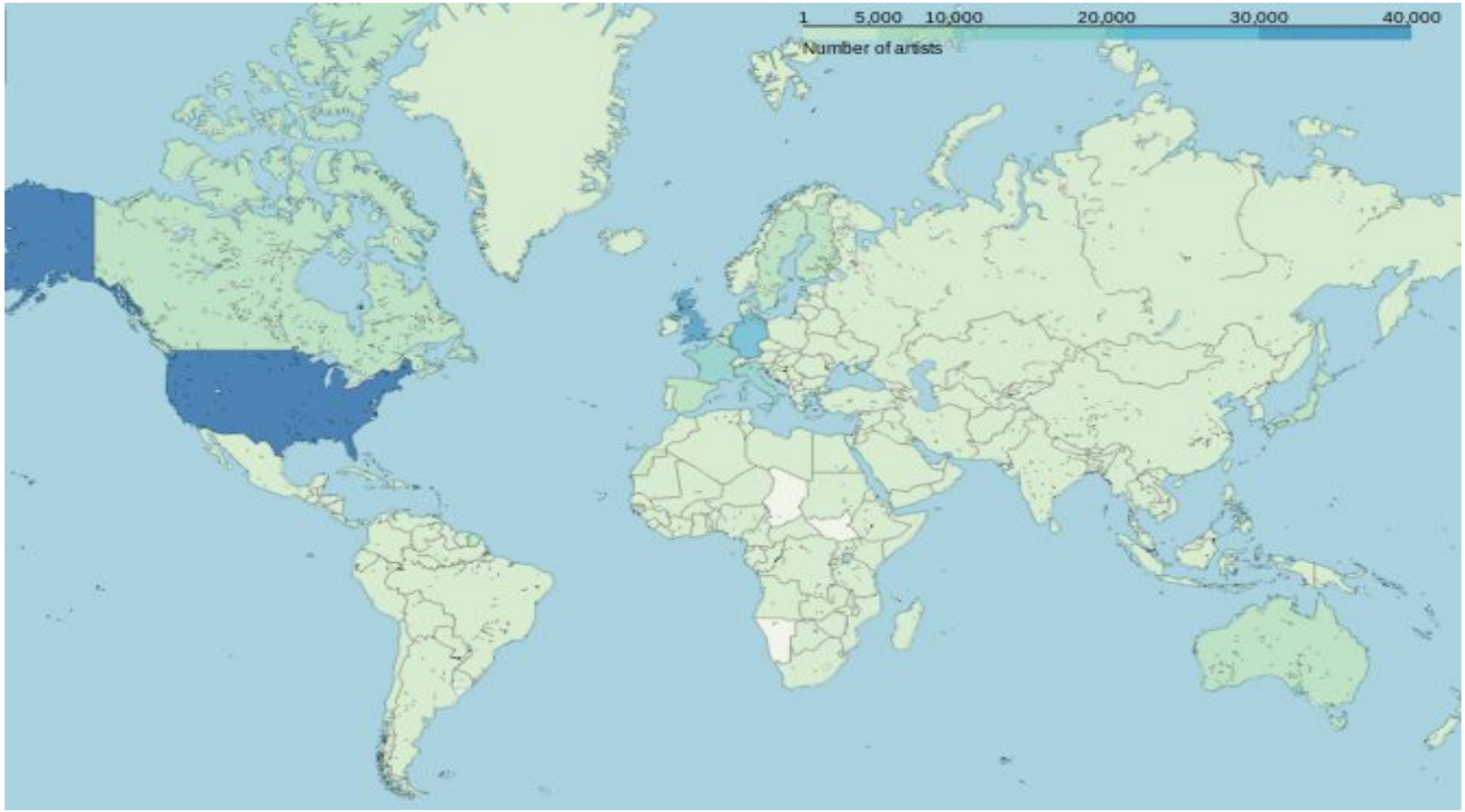
1. Using direct ids
2. Using edit distance on names and albums

Data preprocessing

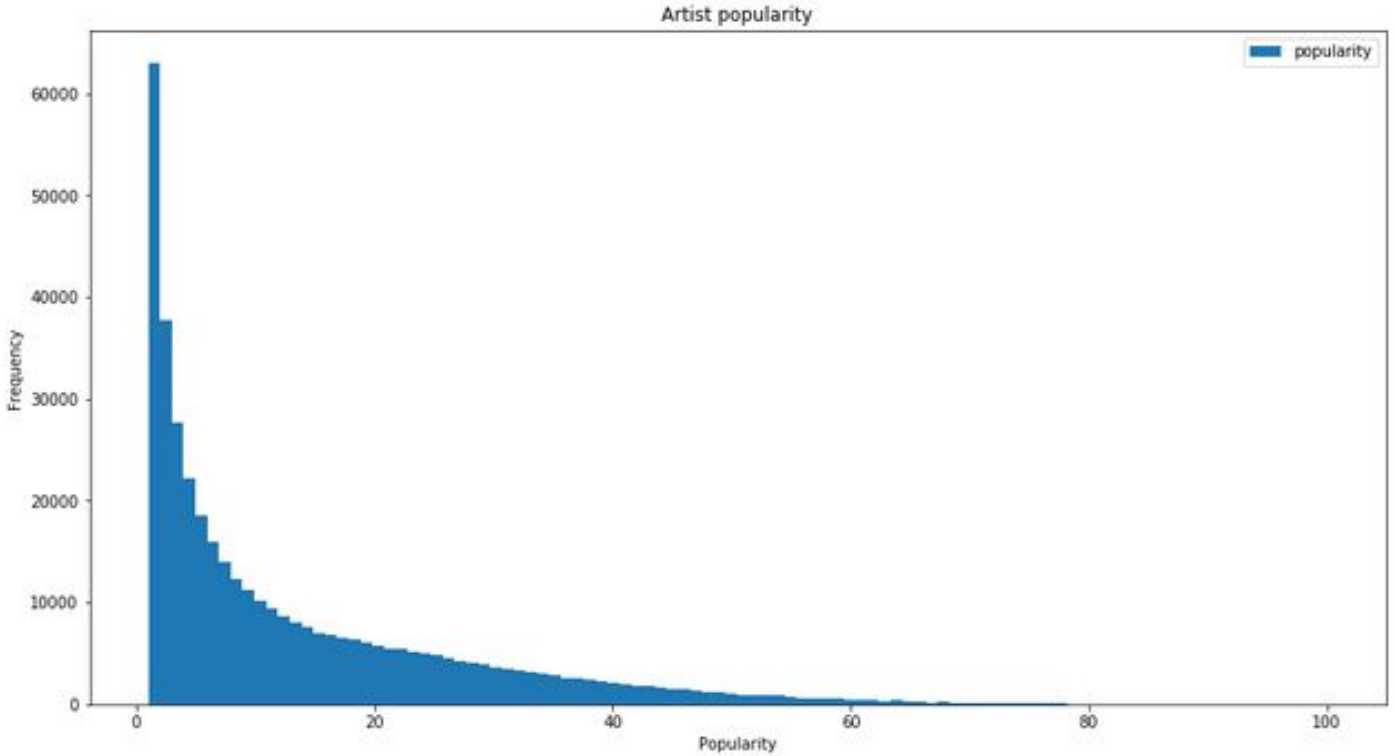
1. Lyrics - tf-idf
2. Songs features - averaged over all songs of one artist
3. Genres - reduced from 1600 to 600

	followers	popularity	danceability	energy	key	loudness	mode	speechiness	instrumentalness	liveness	valence	gender	age	lyrics	genres
0	3752.0	61.0	0.5459	0.440100	6.4	-9.4249	0.8	0.14020	3.180000e-07	0.15880	0.65750	2.0	38.0		(0, 417)\t1.0
1	1361.0	29.0	0.6975	0.554100	5.8	-11.2578	0.4	0.04146	7.196700e-01	0.09388	0.69470	2.0	58.0		(0, 201)\t1.0
80	18.0	22.0	0.2168	0.197430	5.5	-17.0261	0.7	0.04539	7.174316e-01	0.22717	0.08244	2.0	58.0		
93	9.0	14.0	0.2470	0.076460	7.1	-22.4985	0.7	0.04495	6.515000e-01	0.11944	0.11742	2.0	51.0		
98	40.0	20.0	0.3998	0.054605	6.8	-27.5582	0.6	0.04613	9.059000e-01	0.09894	0.21079	1.0	91.0		

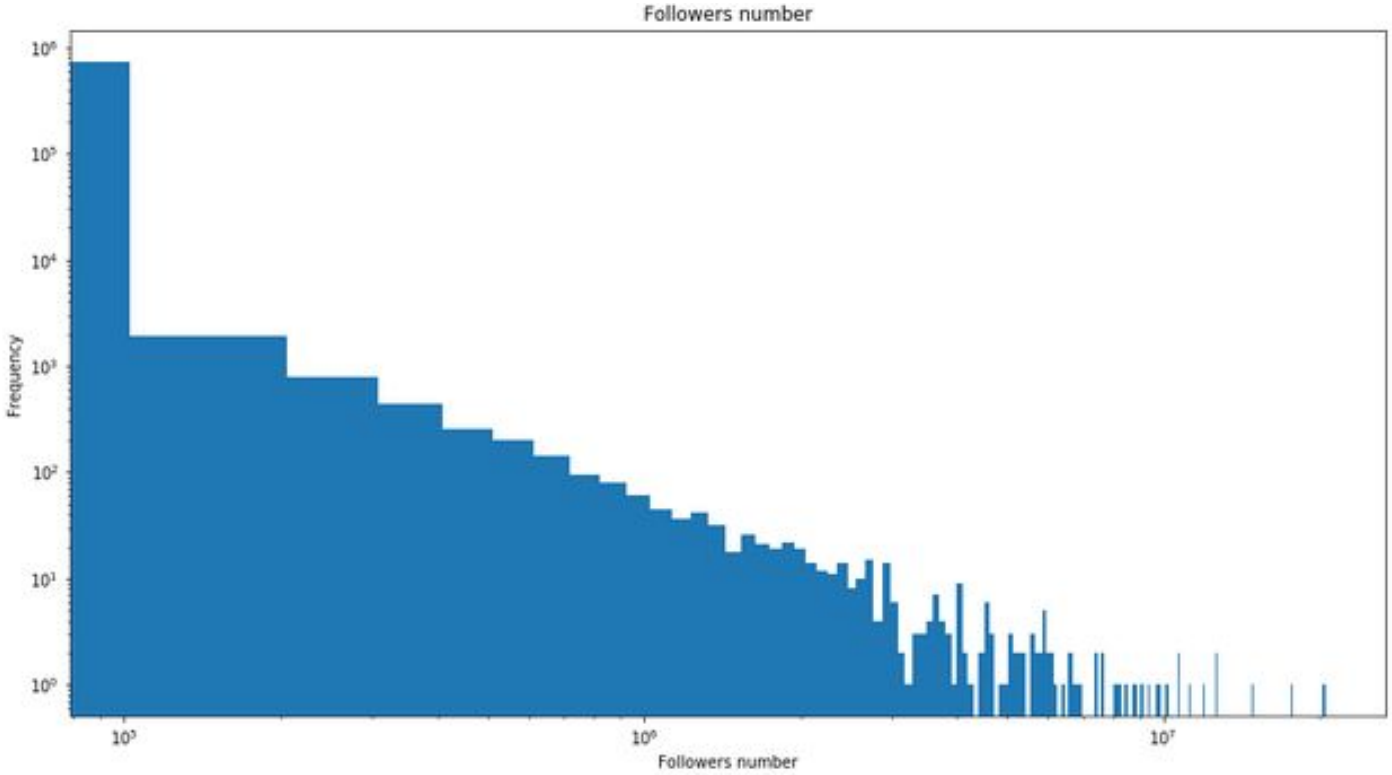
Artists by country



Data analysis - artist popularity

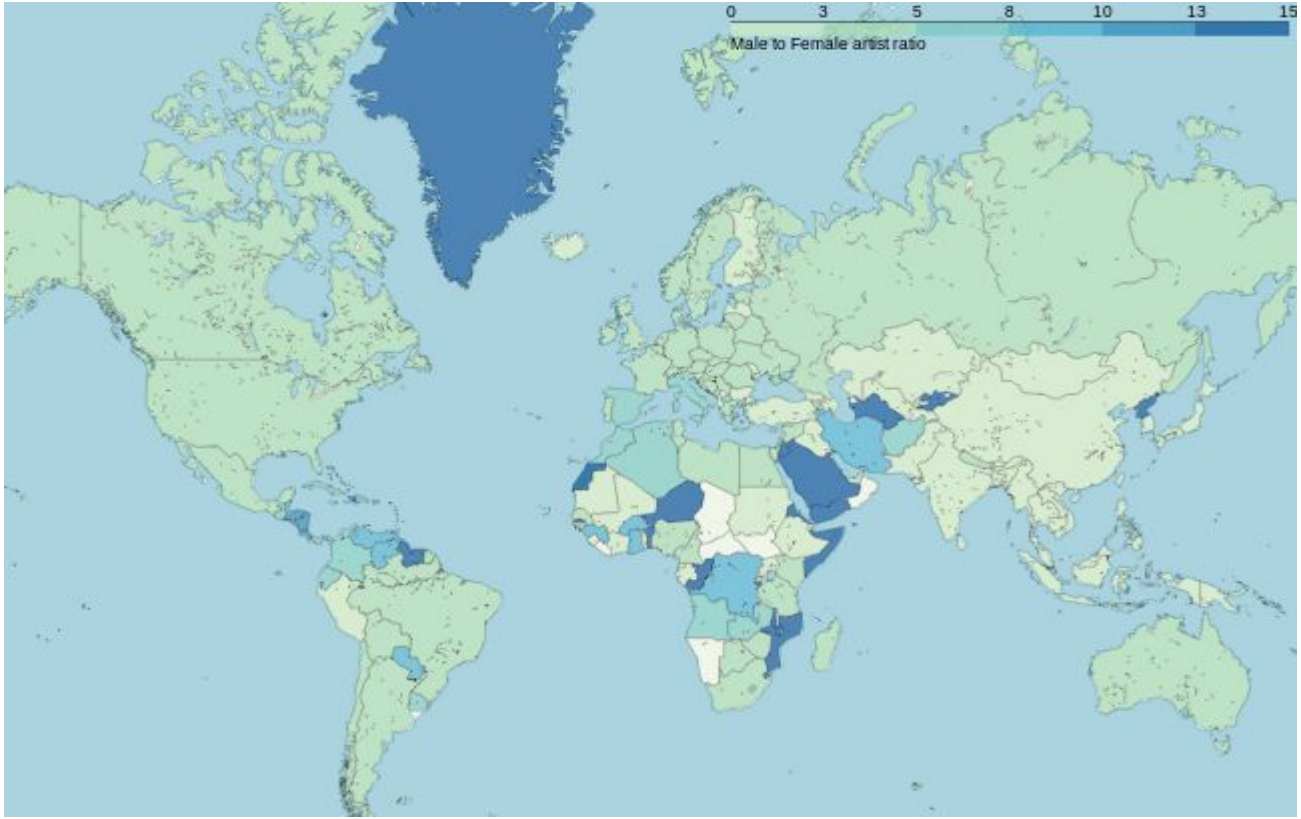


Data analysis - followers

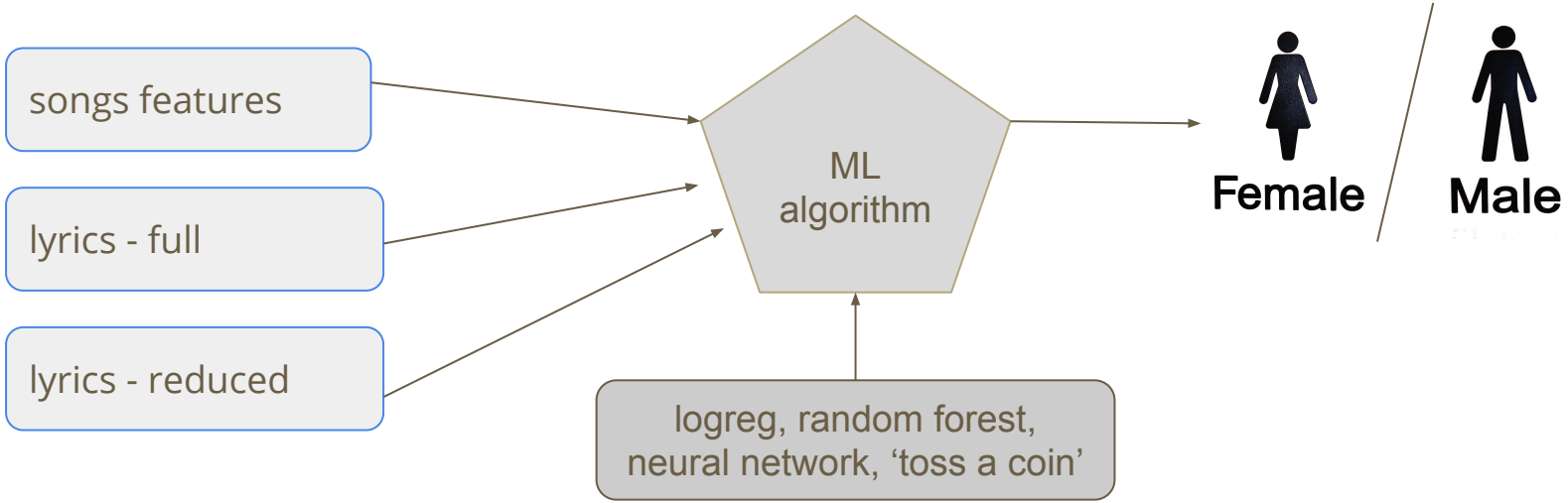


Data analysis - gender

Gender	Number
unknown	525709
Male	382694
Female	104242
Other	541



Gender prediction



Gender prediction - results

1. Baseline - ‘toss a coin’
   accuracy - 65%

2. Random forest - only songs features:
accuracy - 79%

	precision	recall	f1-score	support
0	0.57	0.13	0.21	8570
1	0.80	0.97	0.88	30335
avg / total	0.75	0.79	0.73	38905

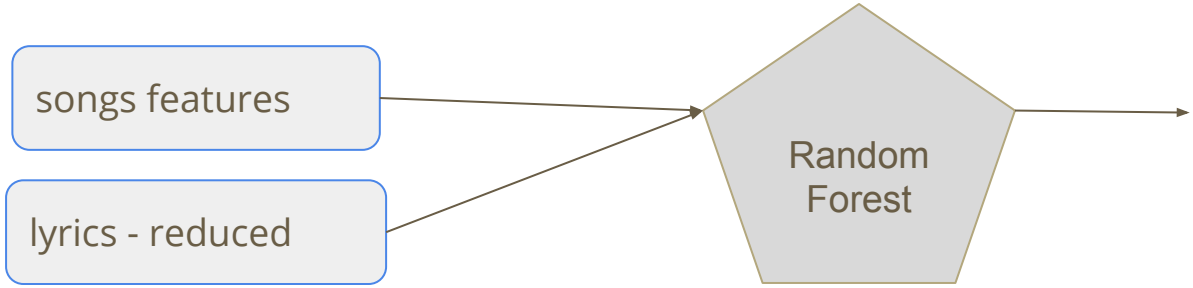
3. Random forest - songs and full lyrics:
accuracy - 71%

	precision	recall	f1-score	support
0	0.34	0.70	0.46	8570
1	0.88	0.62	0.73	30335
avg / total	0.76	0.64	0.67	38905

4. Random forest - songs and reduced lyrics:
accuracy -78%

	precision	recall	f1-score	support
0	0.61	0.12	0.20	8570
1	0.80	0.98	0.88	30335
avg / total	0.76	0.79	0.73	38905

Age prediction



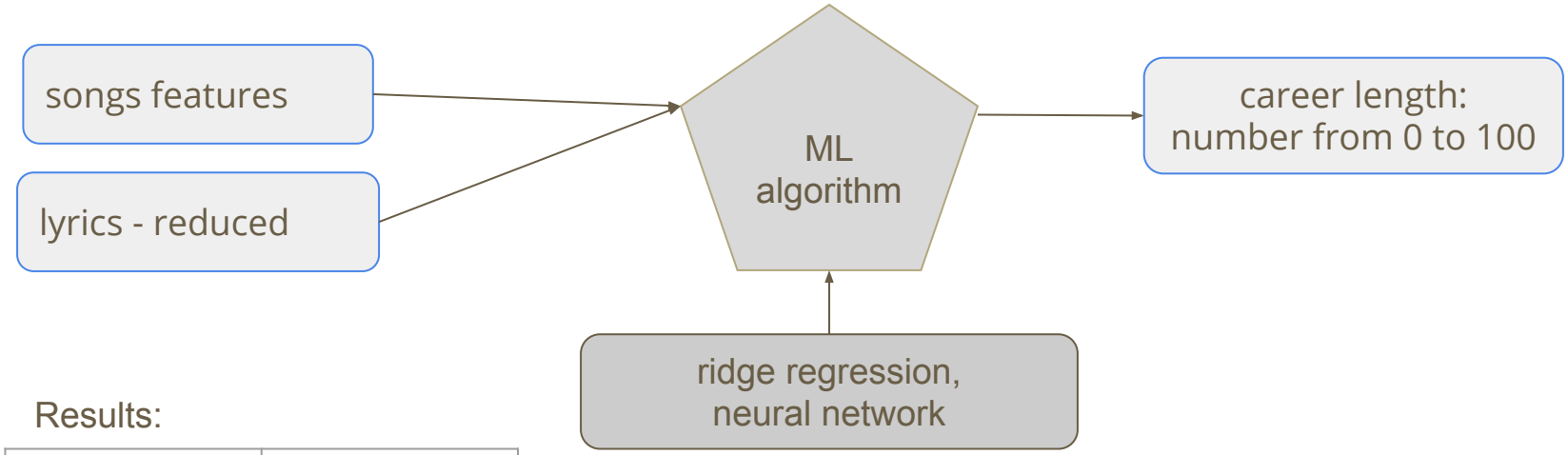
class	age
0	0-40
1	41-57
2	58-78
3	79-...

Confusion matrix:

```
[3821,   93,   17,   75]
[3932,  296,   67,  197]
[3593,   42,  274,  415]
[2468,    7,   67, 2019]
```

	precision	recall	f1-score	support
0	0.69	0.40	0.51	4006
1	0.68	0.07	0.12	4492
2	0.64	0.06	0.12	4324
3	0.75	0.44	0.56	4561
avg / total	0.69	0.24	0.32	17383

Career length predicition



Results:

algorithm	MAE
ridge reg.	6,04
NN	5,68

Summary

- Predicting age and gender - not an easy task
- Career length prediction - promising results, possible improvements
- More information and details of the project: https://github.com/mkuziora/PredictingAgeAndGenderOfArtists