Emma Franchino
Sid Laventhol

Exercise 2A: Character-based language models

Task 2B-II: One model; multiple periods of time

Introduction

In this exercise we aim to generate Italian names based on two periods of time. We used a character N-gram language model trained on 4 different datasets of Italian names (1999/2014 * M/F). We chose the years 1999 and 2014 because we were able to find datasets quite easily, and at the same time, we hoped to be distant enough in time to see a difference in the results from our language model.

Material and methods

Our original idea was not to test proper Italian names, but rather to conduct the experiment on popular Italian songs' titles from each decade from 1960-2010. Firstly, we collected the data from RateYourMusic, where we could create a customized chart, in which we selected top singles in the pop genre. To extract the names from the HTML file, a Python document was created (html_clearner.py), where the songs were saved in a dictionary and then exported into a JSON file. But we ran into an issue with the dataset. The number of songs for each decade was limited to 200-400 songs, and when finding the list of characters used they ranged between 14-31. So in order to use this data, we would need to either remove a number of characters to make them the same length or modify the model per dataset which would not be appropriate for the task. As a result we decided to find another dataset which is much larger.

On Figshare we found a dataset for every name assigned at birth and the frequency in Italy for each year from 1999-2014. Using a simple python script (preprocessing.py) we generated 4 txt files (1999/2014 * M/F) that contain each name printed by their frequency in order to represent the demographic. This produced between 200-250 thousand names for each file.

Results

After we trained our model on the original datasets of female and male names of the years 1999 and 2014, we were able to obtain outputs including 10000 names for each group (we have 4 groups in total generated from gender and year). After a first glance to the outputs we could see, both from a native and non-native speaker perspective that there are many correct names and some incorrect ones, which appear as a mix of common syllabus taken from different names: for instance, "Angelisa" could be a mix of "Angela" and "Lisa" or "Elentina" could be created from "Elena" and "Martina". Also, we didn't spot any particular difference between the output of the different years, nor between genders. To have better terms to evaluate our language model performance and to be able to compare it across different time periods, we calculated: the average length of names in each group, the frequency of names in order to analyze if the 5 most common ones generated

Emma Franchino
Sid Laventhol

by the model were the same as the original ones; and how many names were the same between the original dataset and the output. With these values we were able to see how well the model performed based on the training data and if there were any substantial diachronic differences.

In Figure 1 we can observe how the average names length for each group is almost the same going from $\sim 7,40$ to $\sim 8,23$ The biggest difference in avg. length can be observed in the first and third group, representing female names in the year 1999 and 2014 (they both have a difference of $\sim 0,38$). On the opposite hand, the male names seem to differ by $\sim 0,05$ in both years, being overall longer than the female names. From these results we can conclude that the names in the two different periods of time are pretty much of the same length, maintaining also the difference between gender, and that our language model had the same performance, which we could define as good, in both periods in keeping the length of the original names for output ones.

Another way we used to evaluate our model was observing if the 5 most used names according to the original database would be the same of the ones in the output. We can see our results in Figure 2 and Figure 3: for what concerns the year 1999 we can see that our model predicted the same most common female names (expect one), even thought in different orders; whilst all different names for the male ones. Instead, if we consider the 2014 names, we can observe a slight difference in the female names, with 3 names being the same between the original and output ones, and an improve in the male names, where 2 over 5 are the equal. From these results we can say that our model seems to work better with the female names, especially in the first time period, whilst with the second time period it looks like it increased its performance only with the male names.

Lastly, we calculated how many names were constant between the original and the output data, we collected the results in Figure 4. Here we can detect that the numbers comparing the time periods are very similar to each other, we can spot just one bigger difference regarding the female names: the predictions of 1999 names have a loss of 10 names compared to the 2014, which is a lot related to the 1 in the male names: this suggest a better prediction of the model for what concerns this group. This result is in conflict with the previous observation, where the predictions of the model regarding the female names returned most of the same most common names in the original dataset; this could be influenced by the fact that we considered just the first 5 most common, probably, seeing this last result, we would see more difference determining a lower performance if we compared more of them.

From all of these considerations we can state that our model had an overall good conduct in both time periods, with some differences between the gender groups; which we initially didn't expect, thinking the time period and dataset would have been more influencing. The performance of the model can also be asserted calculating the its training, validation and test loss: in our case all three of them gave us a result of $\sim 1,02\%$, which is pretty good, also considering the evaluation of the outputs. Afterwards, we also trained our model on the 1999 outputs to predict 2014 names, which, on the other hand returned us a 10% loss.

Finally, even though we can affirm that our model worked considerably well, we have to say that our results could represent some limitations, in fact we could have used more distant periods of time, that could have shown more variability in teh results and maybe in the performance; as well as larger datasets or more detailed results analysis.

Emma Franchino
Sid Laventhol

Code

The code used to produce the discussed results can be found in the file "final_model.ipynb", on Github.
The dependencies that need to be fulfilled to run the code:
- random
- torch
- torch.nn.functional as F
- matplotlib.pyplot as plt
- pandas as pd
- collections and Counter

List of contributions

Both of us worked together to conceptualize the task: to think about the dataset we were going to use and to overcome the initial problems we encountered with the song titles. Then, Sid focused more on the language model training, while Emma on the results analysis.

Figures

Avg. original names length	Avg. output names length
7.784137276676495	8.160467906418717
8.121726555124832	8.061487702459509
7.40974375638288	7.778477847784779
8.18798503295476	8.2301

Figure 1: Average length of names

Most common original names 1999	Most common output names
ANDREA, 3.91	'Martina, 3.79'
FRANCESCO, 3.60	'Sara, 3.58'
ALESSANDRO, 3.41	'Maria, 3.52'
MATTEO, 3.30	'Giulia, 3.26'
LUCA, 2.95	'Chiara, 3.19'
MARTINA, 3.50	'Marco, 4.0'
ALESSIA, 3.39	'Francesco, 3.71'
GIULIA, 3.36	'Luca, 3.63'
CHIARA, 3.13	'Matteo, 2.57'
SARA, 2.87	'Andrea, 2.34'

Figure 2: Common names 1999

Emma Franchino
Sid Laventhol

Most common original names 2014	Most common output names
FRANCESCO, 4.05	'Sofia, 3.11'
ALESSANDRO, 2.73	'Giulia, 2.22'
LORENZO, 2.62	'Gaia, 2.11'
ANDREA, 2.46	'Emma, 2.05'
LEONARDO, 2.39	'Maria, 1.71'
SOFIA, 3.06	'Francesco, 3.45'
GIULIA, 2.80	'Luca, 2.14'
AURORA, 2.38	'Matteo, 2.13'
GIORGIA, 1.88	'Gabriele, 1.94'
MARTINA, 1.78	'Lorenzo, 1.79'

Figure 3: Common names 2014

	Common names between original and output 1999	Common names between original and output 2014
Female	299	309
Male	193	194

Figure 4: Number of common names