

Assignment 3: LSTM

1 Dataset

For the given text-prediction problem, The Count of Monte Cristo has been chosen for creating the training dataset. Since the model is developed as a character-by-character predictor, we are interested in analyzing the source content at the character-level.

The book contains a total of about 2.65 million characters. The number of unique characters is 107: most of them are English letters and punctuation, but also numbers and Greek letters are present. Displayable characters are shown below.

```
\n ! " # $ % & ' ( ) * + , - . /
0 1 2 3 4 5 6 7 8 9 : ; ? @ [ ] _
a b c d e f g h i j k l m n o p q r s t u v w x y z
à â ã ä å ç è é ê ë ì í î ï ð ñ ò ó
% σ τ υ | ' , " " †
```

These are distributed according to the figure 1, from which it can be easily seen that the middle characters, the English letters, are the most prevalent ones together with the space character (the first one in the chart). On the other side, the second figure focus on showing how alphabetic-only characters (a-z) are distributed. For plotting both distributions and for all next computations, the entire book has been previously lowercased.

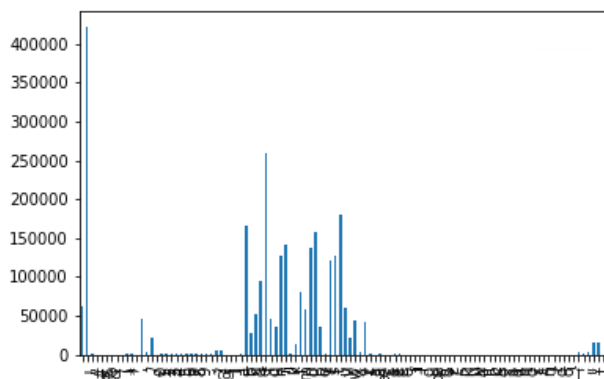


Figure 1: All characters

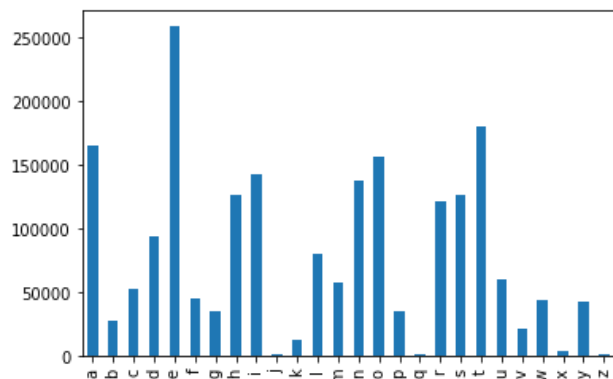


Figure 2: Letters (a-z) only

Figure 3: Characters distribution in The Count of Montecristo

For completeness, also the top 20 most common characters are shown in table 1.

Table 1: Top 20 most common characters in The Count of Montecristo

Position	Character	Frequency
#1	<i>space</i>	420940
#2	e	259094
#3	t	180542
#4	a	165546
#5	o	157076
#6	i	142302
#7	n	137584
#8	s	126590
#9	h	126368
#10	r	121407
#11	d	94099
#12	l	80730
#13	<i>newline</i>	61739
#14	u	60318
#15	m	57157
#16	c	52526
#17	f	45383
#18	<i>comma</i>	45246
#19	w	43892
#20	y	42642
#...

The machine learning dataset has been built by dividing the original text into batches composed by 16 sub-sequences, each 256 characters long. Doing so, a total of 647 batches has been obtained. The training set is composed of all the characters which appear ordered inside the book; the target of each character is the immediately following one, used to actually learning probabilistic sequences generation during the training phase.

2 Training

Training has been done using two LSTM cells of 256 units each, through 5 epochs with a learning rate of 10^{-2} . Figure 4 documents the evolution of the training loss (value on the y-axis) with respect to the epochs (x-axis). Starting from a loss of around 4.67, the fifth epoch reaches a loss that is around 1.30.

In the chart, it can be easily seen that the main improvement is done during the first two epochs, while the next ones are mainly useful for refining the predictions. Figure 5 further visualizes this phenomenon: first epoch reaches a loss of about 1.80, improved at 1.50 by the second epoch and decreased to 1.3 during the next three epochs.

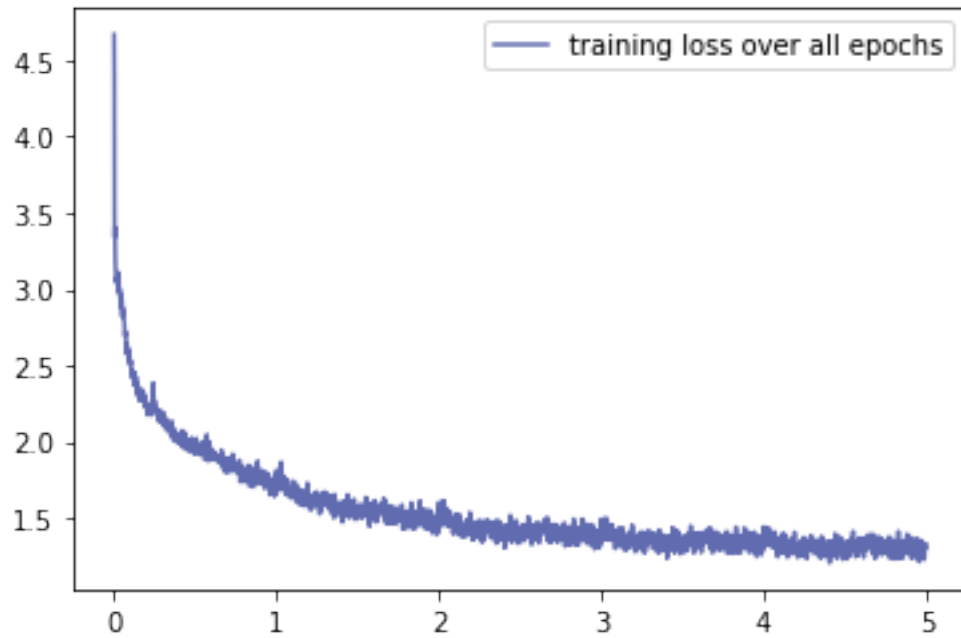


Figure 4: Training loss over 5 epochs

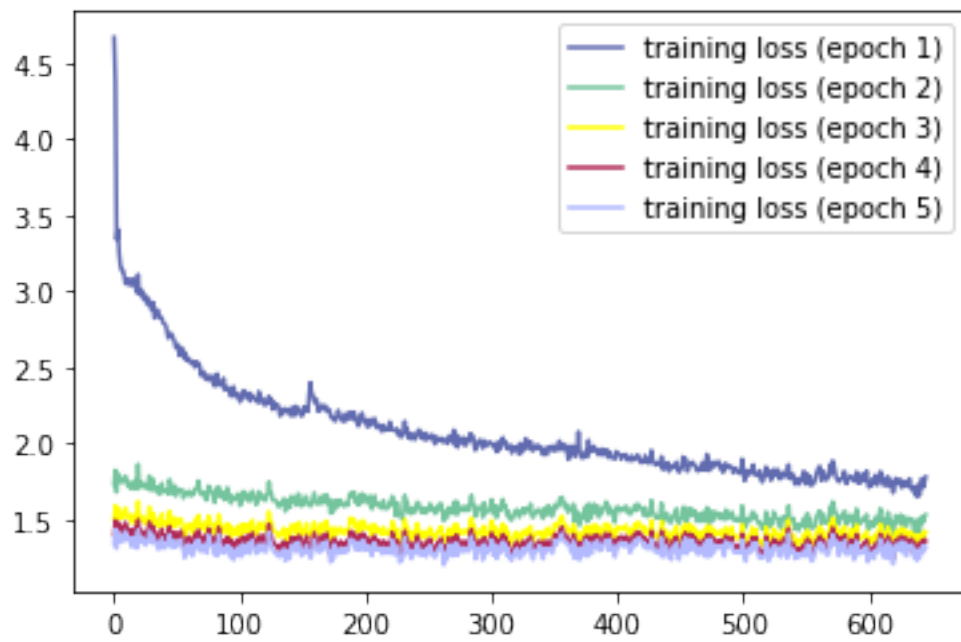


Figure 5: Training losses for each epoch

3 Text generation

After the training has finished, several sequences (each 256 characters long) have been generated to evaluate the model. The first character has been chosen accordingly to a probability distribution based on the frequency of the characters in the original text.

The result is quite satisfying since, even though the examples have no actual meaning, at least most of the generated tokens are English words. Some examples are reported in the following pages; it can be seen that although without any sense, many of them still have some reasonable structure: take a look at the usage of quotation marks to enclose a direct speech, for example. However, some of the presented generation are really bad.

Please note that few newline characters have been removed for brevity's sake.

Example 1

great that i shall propose.
these overed is not astonished your persons with a sive cime,
silence of all paid with astimuts this sleep as this blood, arrived to our
poor fehe. the shade, who tistering to provont
around."
"what speens with the oldament l

Example 2

phopped vonie.
"and whom vagered the loably prayers."
"but do you pleastleds."
"i understand."
"alandy was someone," this croe
as the count and darkness under, whice befallled him, which.
"you received him your mark in his life?"
"that knowsere, 1819,

Example 3

chuppers of the
as monte cristo at his letter wills which
last, as it may asked not disposed on molets, wanted this knowledge,
behind in it, he was cleaned of our fathe

300,000 counten of last surprison in
it. he added any ground. dantès, he opened to wi

Example 4

ttementor.

"yes," replied contrardationable.

did you little, sir," said leanning; "do you promised yourself, then, the
fauther world. there are the wrigue-own|dies,|sine?"

"yes, for you see."

"how corate, did i do we to receive
them; then, see so herse

Example 5

"proprain in de
diese

no one restesing.

speality, but what he knew meeting if she
learning.

he could not kend, marmit and fixed jeine.

"how?"

"yes; distrable understove." an insufferengness was we peragience.
andrea observed with a small black.

"pry

Example 6

ud her.

"harm!"

"i must will give you the contrary of the sentreage, your wife, as i speak,
before remained at the little dangle's stepless you have been acced
for by the same name my hundred tenea, you have here the scene,
from the garbigul."

"we shu

Example 7

ody new deep

assurtd by his

face albert,

and enchiverets decondern,

which after walked the carrial

since monte cristo.

"my lord." cried joined forgottenjes. there were i only

for dantès dangle's: and takingly been over, then, in all directions with a con

Example 8

alland clanted in a fining femous other ala, not knows anyone. do not
do mellois in possibing arms, and to
know whom you know when when my confidingly name he
pidled me, i come; above the sailed benedetto, i something and
beautifully bright,
and involuting

Example 9

sf.
"but did not exchana," said monte cristo. "and where a voice."
"and then anything carried, nay, monsieur," said he, in an once
sacession informations. and rosped him.

"that kepted
it all thes
promises! is
up with yo
why kind you stared like a
fat

Example 10

cère who always
complipal. professed to
writes filled into.

monte cristo's gentlemen was away, the door were alone fine door
addid monte cristo?"

18471 as there. "the light that."

"well, it is very horse|no was at sight of the lateorice you dies."
"s

Example 11

t7; three to the bescuite
carriagezing princess, nearly ventuace
came his brow.

"you are alone|come."

"well."

"yes," said he to the valunive
steward to use and madness either and father conversing the fixe them,
but to took that.

beings revengered, seeme

Example 12

ex, he a side, and whi
had gived him fernand.

"i do not."

"gentlemen, for his three door would not like any werr equal young."
they knews and smokes.

40168m

"yes," said the worthence which awger, the light, and poterewself him to the
new order. never

Example 13

refully, "no one she flone, which, door alone like a so
good accuranger prifalry at laboage."

"at the despercuis
for
procopy?|i magrandely, to ever experis; his isservants,
they must be well efforts."

30019m

"and twering three the teneht, and it, i ha

Example 14

nanded with villefort's light to this wedtings, and sparkless serv
this head of seeing.

"but i not?"

"now," said monte's streeek), and of the indering well-beated declentions
into t enfort of monte ceip, and whom m. 29 tempted evereted from blispold ex

Example 15

|"

"yes," said bentine increasing, stood innocribited naysuilh;
expressed being right, and dargeely its
riches
toscees
in his enemy. at, but th t
expleaus
gutful
and inform the
steps where
for an hour of his father that everything was
not alarmed
them a p

Example 16

asturmed, ah, dressen. shat sunled whether the
particulars. the acquaintess with their girl. then, on as demands thrught of the
exclay of things which

every
door, learns, and that comrades ece|willing
mad,
never overcapted morcerf on his sor so eyes tha

Example 17

im of the frwit one as bursost bluppomor, so with

information shall then spertic and
curious ritading that a privaxe a count's descarcely ginding mith these occhinated
swen
for severates to still
alid who
then returned to began to
liberables with extrah,

Example 18

sthosed on the particant confine, it is see into the very gardenum
forgeaveling chaborts of the

sensations which nervess
had even strawseres on the same approod from times.

morrel with asseding a whore to
unarkilated to become that he is that she looked an

Example 19

ngial
to a sign.

under nor. in a shadow
and public, which descended to the opera much; and he project gut
opens wherespealed
crounced by the
moment,
for as a hundred by a chibino
documents of occupied upains. not
the house, becomes t
came; through an
air,

Example 20

l
has deficied the storm, stronged to ask it with a side-podery l
off it, it is franz.
706112
031(m
sutmos caat," said
ever from,"
replied dantès, and
trauspiences, little veloneess get one to his hearth;
"but by the usual of the islemen, old expect