Deep Learning Lab
Università della Svizzera Italiana
Faculty of Informatics
November 28, 2019

Assignment 3: Long Short-Term Memory Network

Giorgia Adorni (giorgia.adorni@usi.ch)

1 Introduction

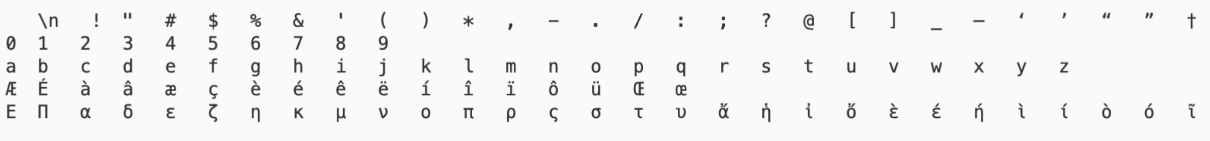
The goal of this project is to implement a text generator based on Long Short-Term Memory (LSTM).

2 Preprocessing

First of all, the book *The Count of Monte Cristo* has been downloaded in plain English text from Project Gutenberg.

The first operation performed consists in the conversion to lower case of all the text characters. Afterwards, a simple analysis of the data was performed.

The book contains about 2.65 million characters and in total there are 111 unique characters. The character list is displayed in Figure 1, and it can be seen that the book contains English letters, number, punctuation, symbols and Greek letters.



```
    \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
Æ   É   à   â   æ   ç   è   é   ê   ë   í   î   ï   ô   ü   Œ   œ
Ε   Π   α   δ   ε   ζ   η   κ   μ   ν   ο   π   ρ   ς   σ   τ   υ   ἅ   ἡ   ἰ   ὅ   ὲ   έ   ή   ὶ   ί   ὸ   ό   ῖ
```

Figure 1: Unique characters in *The Count of Monte Cristo*

In order to train the network, it is necessary to map the text strings to a numerical representation. First of all, the number of unique characters were counted as well as the absolute and relative frequencies. Hence, two dictionaries that act as lookup tables are created: one mapping characters to integers, and the other integers to characters. This will be crucial for the creation of a one-hot encoding.

In Table 1 are presented the 20 most frequent characters, with the corresponding encodings and frequencies.

Soon after, training examples and targets were created. The input examples correspond to a sequence of characters, while the target of each character corresponds to the following character in the original sequence. In this way, there is no target for the last element of the sequence.

Encoding	Characters	Absolute Frequencies	Relative Frequencies
8		420940	15.901%
5	e	259094	9.787%
7	t	180542	6.82%
21	a	165546	6.253%
3	o	157076	5.933%
18	i	142302	5.375%
11	n	137584	5.197%
14	s	126590	4.782%
15	h	126368	4.773%
2	r	121407	4.586%
24	d	94099	3.554%
22	l	80730	3.049%
0	\n	61739	2.332%
10	u	60318	2.278%
17	m	57157	2.159%
6	c	52526	1.984%
16	f	45383	1.714%
19	,	45246	1.709%
27	w	43892	1.658%
20	y	42642	1.611%
...	...	...	...

Table 1: Encoding and frequencies of the most frequent characters

3 Truncated backpropagation through time

Given a very long sequence, it is impossible to feed it entirely and then compute the backpropagation. Instead, it is possible to divide the sequence into n blocks and each block in m subsequences which contain a fixed number of characters of the text.

The $i-$th batch is created taking the $i-$th subsequence from each block and computing the backpropagation on it. In this way, each batch has multiple subsequences and the computations are more efficient.

This approach enables us to pass the output state of a batch as input state of the next batch. The state is reset to zero at the beginning of each epoch.

In the experiment that will be presented have been used 16 blocks with subsequences of size 256. To ensure that all batches have sequences of the same length, before creating the batches, the text is padded in such a way that sequences that are shorter than 256 are filled with 0 at the end, thus avoiding to truncate them.

A mask containing the index of the valid characters is created in order to compute the loss.

4 Network

The model implemented is composed of a MultiRNNCell with two LSTMCells each containing 256 units, following by a softmax output layer with k units, which correspond to the number of unique characters (one-hot encoding), in this case 111.

Table 2 summarises the architecture of the network used in the first experiment.

LSTMCell1	**LSTMCell2**	**softmax**
256	256	111

Table 2: Network architecture

5 Evolution of the training loss function

The training would take 5 epochs and Adam is used as optimiser with a learning rate of 10^2. As loss function, the Softmax Cross Entropy with Logits is used since the problem can be treated as a classification one.

All the models were implemented using TensorFlow and trained on an NVIDIA Tesla V100-PCIE-16GB GPU.

The training loss of this experiment is shown in Figure 2.

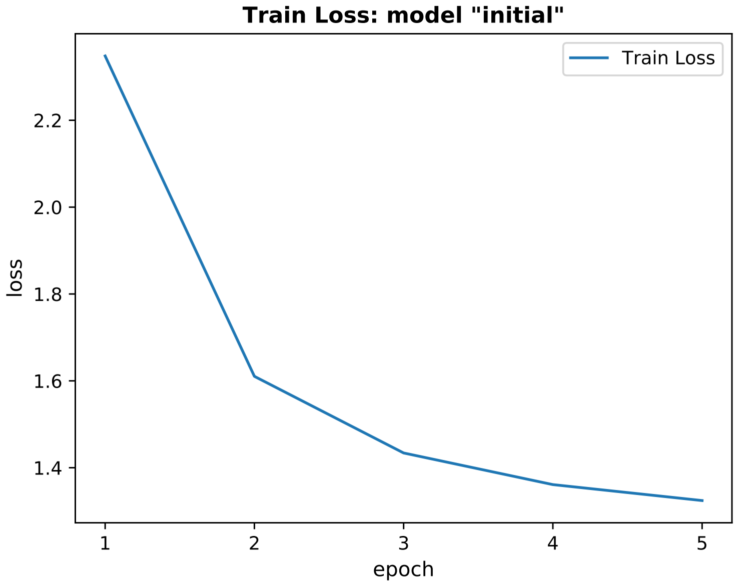


Figure 2: Training loss

The training loss at the end of the last epoch is 1.32. It is clearly visible that the training loss is still decreasing, so in a further modification of the model, presented in Section 7, will be analysed the possibility of increasing the number of training epochs in order to improve the results.

6 Generate and document 20 sequences

After training the model, in order to evaluate the network, 20 sequences composed of 256 characters are generated.

The procedure of generation of a sequence starts by choosing an initial character randomly, based on the relative frequencies presented in Section 2.

After that, the prediction distribution of the next character is given using the initial character and the state of the network. In order to calculate the index of the predicted character, a categorical distribution is used.

The predicted character is used as the following input of the model along with the previous hidden state.

Below, the 20 sequences generated are shown:

> 5sat\n
> promisence, clatood rumbs if he had been belong and the approached my general promises remained for me, not me, really seem you tractly second shoulders.\n
> \n
> “means to your other with them.”\n
> \n
> “do not spoke when the orssused assisting his color\n
> upon\n
> the s

> 6strel—shall\n
> remate (to?”\n
> \n
> “well,” said the don his deformined to spaftly had mig.”\n
> \n
> “wnaties; and he a name; but natuon mine of an enellem masters, making frequently smilings of many time?” she said. this great convinced with mucious accutth the dround th

> t\n
> dantès was\n
> in yous hears litenance. “he keyess withsser?”\n
> \n
> “well,” returned madame he has valentine was a man, to request from valentine.”\n
> \n
> “ah, do you will understand it not makes.\n
> \n
> 50257m\n
> \n
> \n
> \n
> “go,” edmond words agried. which five about of the scarce boo

> called\n
> byoustion\n
> since the old man, “whom best alone than assopeated the corner of the\n

> continued dansèled that instant themselves milling only constances i will debray has\n
> been\n
> albert unlived its at the same worthy table of which\n
> another through me to ong

> slock\n
> in their it by having appearer whom do not made that had impossible plance the _whole nighted box. “that he exciteed all wildieu,” replied\n
> debral presence combriction of the maxitelos in this parison\n
> and projective was took this deadly storp by at te

> stillow was creek of our entire?”\n
> \n
> “alas? dantès (and crime frécapans he, under their appressions of themselves found and you too goved you about conccarce a gendarmes.\n
> \n
> “of their cloud that offering emotion?”\n
> \n
> “oh, yes; brown of them, like today created

> respect of any composed me of cleaked on golden\n
> out of a word to after his mothed them with\n
> stidned with and murder of the count shill long memble and my days of announceplies celeased his man, and spazz marsen me.”\n
> \n
> “alaply feeling really\n
> below.”\n
> \n
> “and is

> lemmores a gazed to the\n
> shadon of brains to m. de villet, that serias,\n
> spoken to looking one of his from the six fately\n
> strange to\n
> seemed there of in-lacquent from\n
> nothing\n
> called to atpen what deserved that is his accous, for any placed my partes.”\n
> \n
> “undes

> led and still proof venigality left?”\n
> \n
> “have he has from medound, and when an occond detailed.”\n
> \n
> “complexcessed two voice corled that just have\n

> his pater fuxent and odden of gentlemen watched to ut leave heambsimally told me a yestrioled with anotanted att ”

> “ ded i have found understand,\n
> that to has genessed woodenate stalus which promise to gotn, breaunden one with an exclamatual\n
> possessed themselve than which not accustomed withspaper as followed me sister, to mensions to place in them converse was followed a ”

> “ exts his words riving a—wailing out, which beligualles had sufficed her conquponor had to consime offering with the count’s diamonds of franz; “my officent\n
> ship to villefort regend\n
> franz had been sailed a young\n
> in\n
> subject which were on boused by a fext to ”

> “ yselve to believe\n
> me by an\n
> innocend, “where he befoced over\n
> only here greeks, which\n
> always like me make,\n
> cherusbated 1.1..................6... ladded villefort; “what know on them for assolity to themselves them; you alough the leat or blood and lamber, st
>
> ”

> “ íbandrally orden-clouds.\n
> \n
> the\n
> travelling a crramts than the country with leaped them, come of this bird,” said monte cristo of the horsesforte white escaped by the came of a pair of all mesting his\n
> moning to introvide that me; if you reserved until make th ”

> “ hass was, fixed\n
> his\n
> goon.”\n
> \n
> “and opposite plovand her offers and shruck.”\n
> \n
> “to be place offul has just beautiful. but his know mystermatide when i explanations to raised all heaven had to\n
> offer, splond.\n
> \n
> “you blowent of them, well, had affairs of his ninqu ”

> asheddens, had elexpected that the asked with him.”`\n`
> `\n`
> “i will probable of accepted without cranced to frer partaken, duen and`\n`
> the tradiled him to`\n`
> more with waited to not possession`\n`
> of`\n`
> gyombore docdens pathed his`\n`
> second lasts is acquern liced to look little

> ys whise motements-glausse, my dear arize all that from the duty smwault, smuggling and cid.”`\n`
> `\n`
> “sly and my di?—xepreasure of his petulon`\n`
> of then that one on the idease conclusious heavonishment—what is sux telling them more promise of you have circles, whi

> 8 all clamp the put, bright with all from the artial wavasts he name deforting to that laster and different according; be`\n`
> mankets remained`\n`
> by her, i will be a`\n`
> moment, cap these storbed office dows inknorm; “do not believes today and razing most`\n`
> and seat wh

> walvess your`\n`
> corners, assammables. never all was himsist became visitor weelse was the latter in château-renaud or capes`\n`
> of his son which immedia had become for thencholate`\n`
> box. danglars had any gatess instead-clank informlinated,” said madnamon was,`\n`
> cut t

> asily`\n`
> constinct of hi, from the over that these women on prazes of sopeated charmzest strobd ieplied to seeming`\n`
> mademalstables men into`\n`
> cace`\n`
> then fellor.`\n`
> `\n`
> “you have carnocient exame of`\n`
> the fored to keep out killed outstents and owns of his persantly part o

> odelly.`\n`
> still any hall,” said monte cristo wrote archive ofrenex do not well asked his prop men large which her futully`\n`
> officing as i have project. another had foreyoness of whom for your older me,” replied maximilian water or saved him, and if these retai

Even if the sentences obtained have no real meaning, most of the generated characters compose existing English words. In fact, only 20% of the words generated don't exist, that is a result quite promising.

The sentences often present a common structure that is to put the direct speeches in quotation marks.

Unfortunately, many of the phrases reported present numerous new lines because of the lack of preprocessing of the text.

7 Improvement of the model

In the first model trained, also blanks, multiple new lines and other infrequent characters have been considered. It would be interesting, in order to improve the performance, to apply a more thorough preprocessing procedure to the texts and document the evolution of the training loss function of the new model.

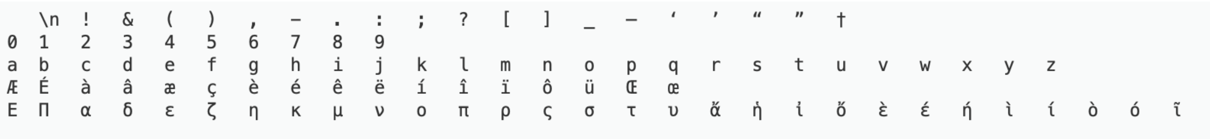
First of all, analysing the structure of the book, some parts of the text were removed at the beginning and at the end that was not actually part of the book (many of these were information concerning Project Gutenberg), and the table of contents.

After that, page breaks were removed from the text since they contain the page number, in the format `^\"[0-9]+m\"$` that was going to increase the frequency of numbers and of the letter *m*.

Finally, multiple new lines have been replaced with at most two, corresponding to a new line followed by an empty line.

Instead, were not taken into consideration the removal of punctuation, since the purpose is to generate pretty looking sequences, therefore also containing punctuation, as well as the removal of Greek letters, that does not make difference since they appeared with rarely frequency (one sentence in the whole text).

After the preprocessing, the book contains about 2.61 million characters and 103 unique characters. The character list is displayed in Figure 3.



```
  \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
Æ  É  à  â  æ  ç  è  é  ê  ë  í  î  ï  ô  ü  Œ  œ
Ε  Π  α  δ  ε  ζ  η  κ  μ  ν  ο  π  ρ  ς  σ  τ  υ  ἄ  ἡ  ἰ  ὅ  ὲ  έ  ή  ὶ  ί  ὸ  ό  ῖ
```

Figure 3: Unique characters in *The Count of Monte Cristo* after the preprocessing

Afterwards, analysing more carefully the frequencies of the characters, it was decided to replace rare characters with the token '`UNK`'. In particular, all those characters appearing with a frequency less than 100 have been considered "rare". In fact, in a text containing 2615988 occurrences, 100 are less than 0.004%.

Following, the two dictionaries containing the integer and integer mappings have been recreated.

In Table 3 are presented the new characters, which this time are only 53, with the corresponding encoding and frequency.

Encoding	Characters	Absolute Frequencies	Relative Frequencies
6		417112	15.945%
5	e	256526	9.806%
13	t	178459	6.822%
11	a	164051	6.271%
1	o	155404	5.941%
18	i	140911	5.387%
...	...	...	...
52	UNK	667	0.025%
...	...	...	...

Table 3: Encoding and frequencies of the characters after the preprocessing

The training loss of this experiment is shown in Figure 4. At the end of the training, the loss is 1.24, so is decreased respect to the experiment without preprocessing.

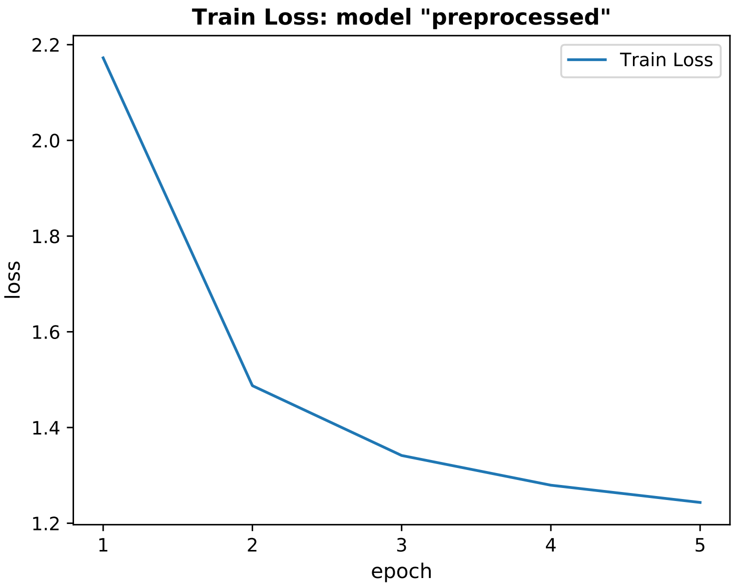


Figure 4: Training loss of the model with preprocessing

Below are shown some examples of sentences generated after the training of this model:

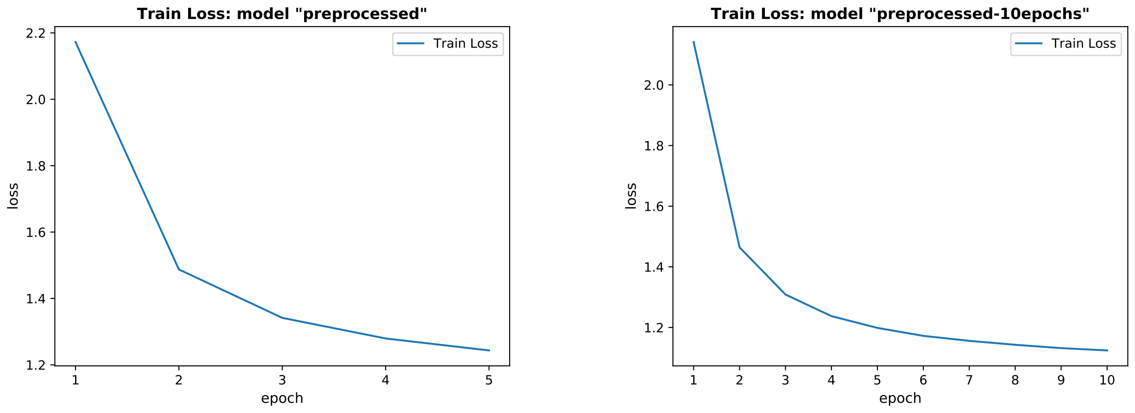
> “ madame de villefort\n
> passed before it glict, the ception of similar smile to the man here, those whisten to\n
> habituated in their if for him for me, by such about to take pity on chance from the\n
> young step, to have because entering the assampo, he has the co ”

> “ ’ took so wifterness. he is venecont respectly a labing.”\n
> \n
> “and do you are\n
> already known the first obstatrark seemed her with so partiery were death, only or his bed\n
> of the grotto_ gentlemen, ali risk to the days–”\n
> \n
> “the\n
> count,” said monte cristo, as eres ”

Also in this experiment, the sentences obtained have poor meaning, but the percentage of existing English words is increase from 80% to 86%. In this case, much less phrases generated contain multiple new lines. Moreover, most of the sentences include proper names such as ‘*mademoiselle de villefort*’ and ‘*monte cristo*’ that comes directly from the book.

One of the easiest thing that is possible to do to improve the results is increasing the number of epochs of the training.

A new experiment was then carried out: the previous model has been trained for 10 epochs instead of 5. The two trends of the losses are shown the Figures 5.



(a) (b)

Figure 5: Training loss of the preprocessed models with 5 and 10 epoch training

It is clearly visible that the training loss has continued to decrease in the following epochs, reaching 1.12. Table 4 documents the training losses of the first experiment and of the ones with the preprocessing.

Model	Epochs	training loss	Train Time
`initial`	5	1.32	768 sec
`preprocessing_5e`	5	1.24	1124 sec
`preprocessing_10e`	10	1.12	2267 sec

Table 4: Initial and preprocessed models performances

Below are shown some examples of sentences generated after the training of this model:

> knose eugénie so appreasciously come. my\n
> mother has entrobited her habyighten light for albert had heard of obligor\n
> to itiquer some shipowners, or if not that unhappy. besides, let me not get.
> in the ground old man are also, and now it is theref."\n
> \n
> "yet—sa

> mademoiselle decribidities, has their fortunes at verture\n
> monsieur, and that he has done by his impassic gegary\n
> dest on the yacht appeared."\n
> \n
> "you know no other, that there is dead that the plan of myself and the pleasure\n
> of the different charms, countess

Also in this experiment, the sentences obtained are very similar to the previous. Sometimes it can be seen the attempt to generate sentences with a more elaborate grammatical structure. Once again, there are some proper names derived from the book.

The following experiment proposes the use of a model regularisation technique, that is the addition of a dropout layer after each LSTM cell. In particular, during the training phase, the probability to keep each neuron is set to 0.5, while during the generation of the sentence is set to 1.

The experiment with the dropout has been executed on the initial model and on the one with preprocessing, in both cases for 5 training epochs. Another experiment has been carried out on the experiment with preprocessing for 10 training epochs.

In Table 5 are summarised the training losses of the models presented above.

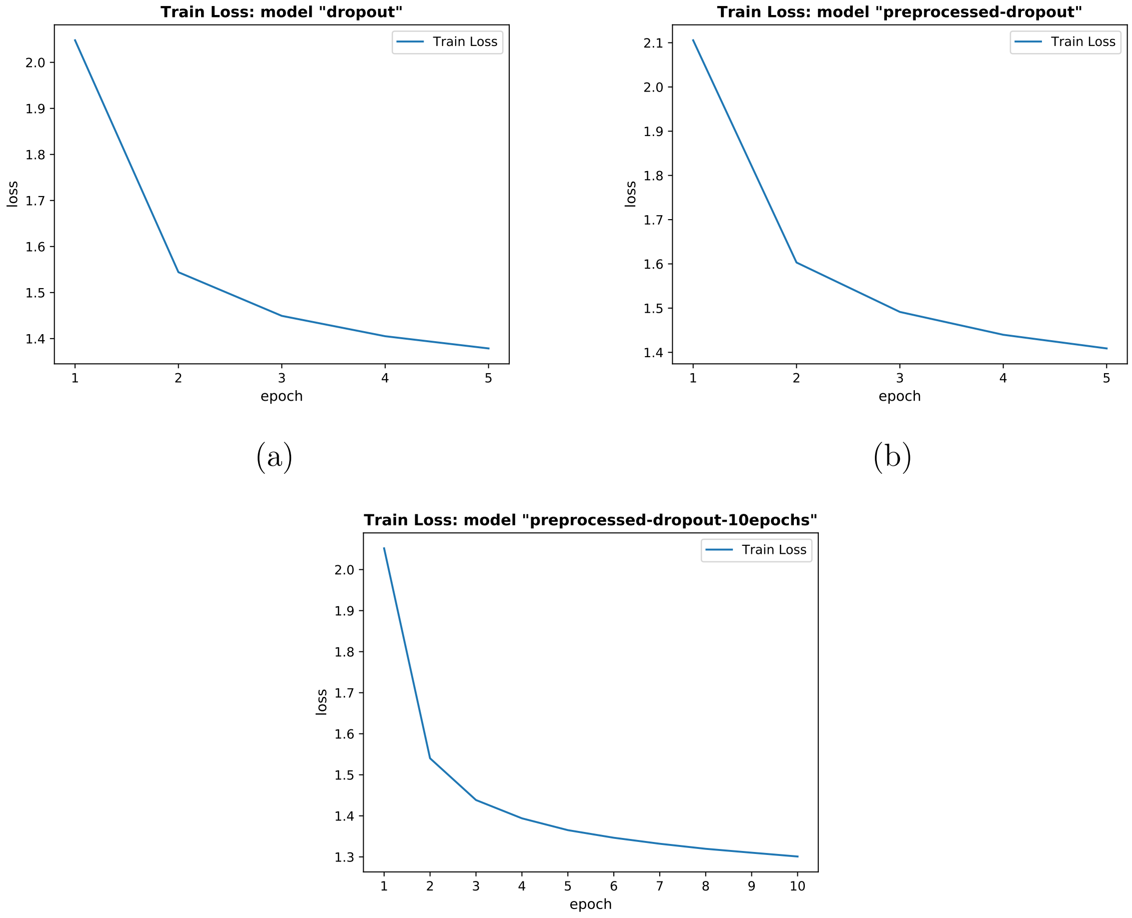
Model	**Epochs**	**training loss**	**Train Time**
`dropout_5e`	5	1.38	1203 sec
`dropout+preprocessing_5e`	5	1.41	1146 sec
`dropout+preprocessing_10e`	10	1.30	2268 sec

Table 5: Dropout models performances

The new performances are shown in Figure 6. Looking at the loss values, the performances did not improve as expected after the application of the dropout.

Below are shown some examples of sentences generated after the training of these models. First of all, the `dropout_5e` model sentences:

> with a fear of that thing; he counded, and\n
> that how but ere shall following the hour after himself with him, her\n
> about like a father in which i have made up wable where they, having not to the\n
> luminy of little officers, of real, who will a took lice that s



(a) (b)

(c)

Figure 6: Training loss of the models with dropout

> “ ble alone, and kill him and ceunal,\n
> he silently surprised through a shuspications in\n
> persons about him demins, they have been signor in a may but his times in ones examined the room elgering. a sordious pessage. one assured escape about done back—they beli ”

Following the `dropout+preprocessing_5e` generated sentences:

> “ ate a king and persons and occupidioned saments to the six months, ecchapting by the question. my horsing on the\n
> content\n
> to thrust in if the minister of\n
> done in the peapr was\n
> fortuno would enter all in all the gion, not, and they have dispased to her certa ”

> “ at this sifficience at\n
> them similary! a present paris, and there he\n

> exactly i have been breathtoms,\n
> as then it am flefts to the brother as the\n
> merening had been about towards the time of two honor, now,\n
> peeping, himself, who had piimed albert, shuddered s ”

Finally, the `dropout+preprocessing_10e` generated sentences:

> “ apped himself announcing laugh. like her more pray. he could not now visit with there the did the judge of series,\n
> promised into the mother. her sailors shuddered to the idea that he had seemed\n
> as\n
> given at murderers. having were ministerings because the ho ”

> “ has dispermined me, you were men he deet of you to filled, you cause that rush to her accommands,—now prayers i can find my\n
> sound which is\n
> not since\n
> all them to mention by dusty police.”\n
> \n
> “rean of the correst that was delighted as yourself take preparate, ”

Since the models seem to be underfitting, the following experiment aims is to improve the model's performances by increasing the complexity of the network, in particular, the number of parameters and in general the dimension of the network. To accomplish this, a new LSTM cell is added to the network after the previous, with the same number of hidden-units.

The experiment with the additional layer has been executed on the preprocessed model both with and without dropout. Since the results on the model with dropouts are the worst, only those on the model without dropouts will be shown.

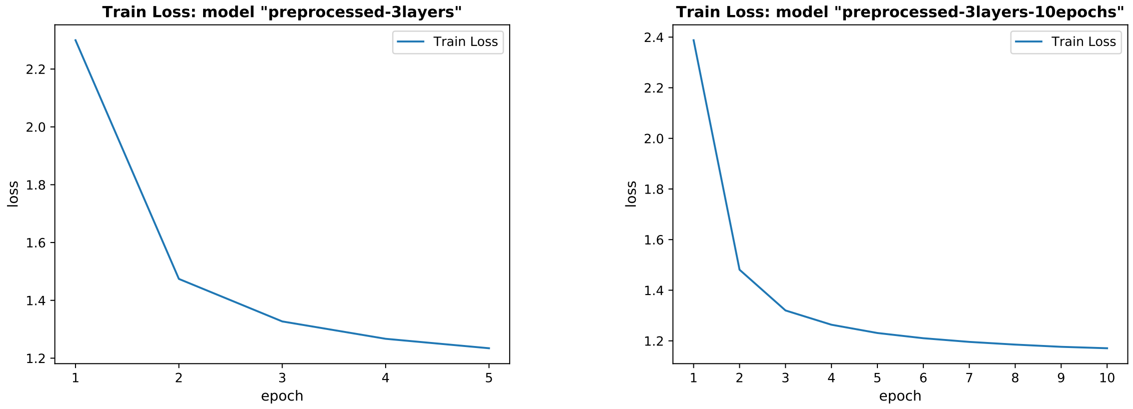
This experiment has been carried out for both 5 and 10 training epochs.

In Table 6 are summarised the training losses of the models presented above.

Model	**Epochs**	**training loss**	**Train Time**
`preprocessing_5e`	5	1.24	1124 sec
`preprocessing_10e`	10	1.12	2267 sec
`preprocessing+3layers_5e`	5	1.23	1574 sec
`preprocessing+3layers_10e`	10	1.17	3113 sec

Table 6: Comparison of the performance of the experiments with preprocessing as the number of layers varies

The performances are shown in Figure 7. Looking at the loss values, the performances seem to be very similar to the experiment without the additional layer.



(a)

(b)

Figure 7: Training loss of the models with an additional layer

Below are shown some examples of sentences generated after the training of these two models. First of all, the `preprocessing+3layers_5e` model sentences:

> this man mistrust, i know the valet, is a popularity, and the gilt on it. many, was well he remind him the fair, you are penehed to mention any more, sir, and crossed now nearly."\n
> \n
> "i meanh. the statue, who has a life, come; i\n
> may, then your recollection."

> but what will deceive the dream, and villefort did not concernment sitting a man, who in the larger together usuals! what dantès sail at the count it\n
> was but they princess each straw the understand by the third\n
> louis where i think of the taking notice? wh

Finally, the `preprocessing+3layers_10e` generated sentences:

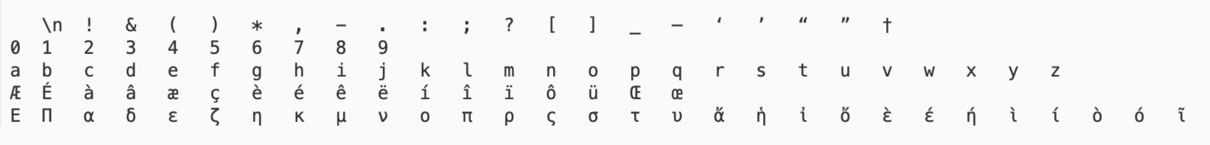
> don the preuisjest," exclaimed the count gave you, that\n
> he heard some years, englest he has impossively for messness of it?" chote me entirelyable successy to auteuil from the island whose theatantly\n
> pressed much\n
> anything at such at all the dying places,

> her son, asred you, you amed france, my mother, enerved with five?" asked d'Épinay, aud valentine," observed miner."\n
> \n
> dantès would be enough than your father's affair; it is quarones, and wild you have as bad the phys. and pass of\n
> the man\n
> in black careness

The sentences generated with this experiment seems to have many more existing English word. The structure of the sentences is improved too.

The last experiment performed, used multiple books to train the network. In particular, the books *The Count of Monte Cristo*, *The Three Musketeers* and *The Man In The Iron Mask*, which are all by the same author, namely Alexandre Dumas, have been downloaded in plain English text from Project Gutenberg.

These books contain a total of 4.88 million characters and 104 unique characters. The character list is displayed in Figure 8.



```
   \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
Æ  É   à   â   æ   ç   è   é   ê   ë   í   î   ï   ô   ü   Œ   œ
Ε  Π   α   δ   ε   ζ   η   κ   μ   ν   ο   π   ρ   ς   σ   τ   υ   ἄ   ἡ   ἰ   ὄ   ἐ   ἔ   ἤ   ὶ   ί   ὸ   ό   ῖ
```

Figure 8: Unique characters in the three books

After having replaced the rare characters with the token ‘`UNK`’, the unique characters are 55. Table 7 presents the characters with the corresponding encoding and frequency.

Encoding	Characters	Absolute Frequencies	Relative Frequencies
6		777271	15.928%
5	e	475075	9.736%
13	t	334958	6.864%
11	a	307634	6.304%
1	o	288840	5.919%
18	i	263412	5.398%
...	...	...	...
54	UNK	556	0.011%
...	...	...	...

Table 7: Encoding and frequencies of the characters of the three books preprocessed

The experiment consists in training the network after applying the previous presented preprocessing procedure. The training loss of this experiment is shown in Figure 9, and the end of the training it is 1.09, the best result reported so far.

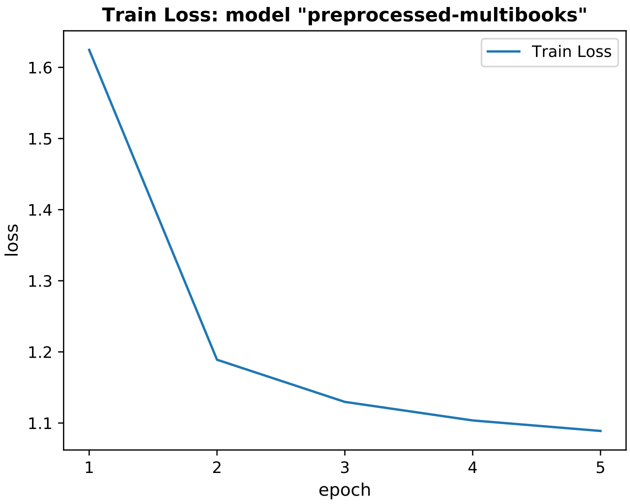


Figure 9: Training loss of the experiment with multiple books

Below are shown some examples of sentences generated by this model:

```
one saw so that one morning, really to say, ‘at\n
the first heep; it is my mother, dig theolord.”\n
\n
“harning to me? then, then, who could not appear not to send the\n
ill vignitady for him this.”\n
\n
“and i will not place against your hands and the present is\n
goin
```

```
, suddenly, i will be fated, and have we\n
cast and leave barries, but an assistance, and reconsuited myself\n
and dressed to hear my arms, and for women. instead of afflinity, whose\n
apartments of lines, madame morrel asked any—to wish to obtain the not\n
drague
```