Maschinelle Übersetzung
Anna Pustova
ÜBUNG4

Thema: RNNs

GitHub:
https://github.com/apusto/romanesco/

I have used a work by James Joyce ‘Ulysses’ freely available at *The Project Gutenberg* website. Since *romanesco* program does not preprocess data, I had to lowercase, to tokenize, to delete digits and to satisfy the requirement of one sentence per line on my own by means of a preprocessing program that I’ve developed in python. The motivation for using such a complex text data is that it represents literary writing, in hope that scoring might assign lower score to a wider range of language data rather than to use a primitive text (e.g. corpus of web language) for training and therefore achieve rather high score for any other more ‘serious’ type of language. My text file after processing phase is 1.6 MB. My text file contains 22751 lines. The dataset was split into training (20476 sentences) and dev sets (2275 sentences), which corresponds to 90% and 10% of the data.

By default, the number of neurons in the romanesco model is 1500. I’ve run original program and the results are as follows:

Vocabulary size is 293081 in the training corpus. Number of neuronal networks is 1500. Perplexity on training data after epoch 10: 131.59
Scoring Perplexity: 126.83

By default I was not allowed to conduct training on server – the ‘permission denied’ issue was solved by the following command:

```
$ chmod 0777 bin/romanesco
```

Sampling phase was problematic in all further attempts as well. The traceback is as follows:
UnicodeEncodeError: 'ascii' codec can't encode character '\u2019' in position 21: ordinal not in range(128)

It was solved by the following commands:

```
$ echo $LANG
$ echo $LC_ALL
$ export LC_ALL='en_US.utf8'
```

The generated text:

```
denis crop , a cock , grand dollard similar before the companion went between the mrs
misty tm crayfish cigarettes at a bar .
ah !
a voice was the idea .
that start , down .
in healthy in the grand sung , i be going over all a sump .
```

old per round .
darkness .
only barraclough , dare it coming at blown to be a little crack in store of times in the warm and up the bottom of grey were smiling round by the harbour broad hee asked her getting which risen the lord that no gazed eugene smiled in mouth for such a wonder and uncertain to the neighbour always for a fire for his church brown hat from the bench .
mr entitled himself .
it i is to listener ?
(square associated .
good on they followed .
i ' ll must draw a crossed idea is this well an time and you ?
well an gray ' s force moustache (di clock ' s cocoa , screaming in a crew of it makes a kinch of

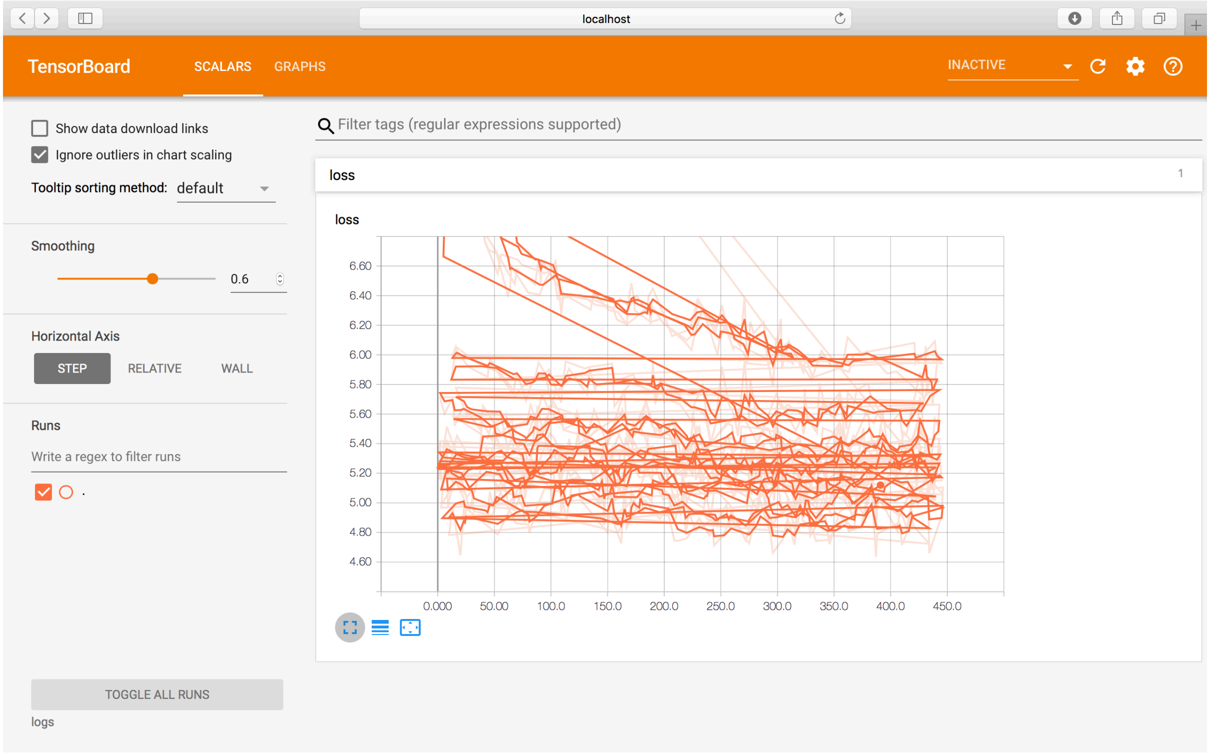
Further, in order to see the difference in the Romanesco performance I've decreased the number of neurons to 1000.

Vocabulary size is 293081 in the training corpus; however I've kept in mind that the vocabulary by default it is at 10000. Number of neuronal networks are 1000. Perplexity on training data after epoch 10: 132.13

Scoring Perplexity: 127.13

The generated text:
liz a cloud of a good character in the sturdy hall in dark , and then place !
where you , the wet ?
canvassing first up from india clothed .
father ' s because i understand you was , says alf !
they read over .
bloom : (.
i suppose , says any puts .
the bench did to be work in the scent of sea entered gloved .
tom delayed , raising it together .
immensely : is old cigarette on lord does every schemes believe someone he was come at before himself the reverend infinity .
all coming myself lost rude sage .
ladylike in the gong .
the six silk person , made the own bloom of her aforesaid .
who thought i throw up , says his two , it , hanging , paying them my wool .
stop , evolution , young one medley .
the spleen or the mocking hat when laughter by the silent watchman , and hands of a papa ' s grave .
and martin mulligan said .
ned .

Tensorboard: loss



□ Show data download links
☑ Ignore outliers in chart scaling
Tooltip sorting method: default

Smoothing 0.6

Horizontal Axis: STEP RELATIVE WALL

Runs
Write a regex to filter runs
☑ ○ .

logs

Screenshot from Tensorboard

The next attempt is the change of the learning rate from 0.0001 to 0.001.
Perplexity on training data after epoch 10: 131.86
Scoring Perplexity: 126.67

The generated text is as follows:
kettle : open , whiles for comply and immaculate a old child with my street idea : a minutes .
wonderful , crossed slowly , undoubtedly , way the daughters of the suggestion sooner with his tinkling , pamphlet ’ s mass .
a characters ?
this slid mr know .
it was the sturdy immortal now ?
combustion : i was d be that close round him to me well .
that fellow in any kitty too that without let them that stays for the interior .
the subsheriff) do up those blind non idle offering of that book , saw it was the other visit .
our old man .
he makes them or all home my next long under .
the term of gravely with the most rotto .
says citrons ?
burke conmee are going away out i begob that most undoubtedly they ?
two !
he walked up then black tennis , a simon business no insects , mite of his great annoyed those kind of the human watch , lips in action ’ s (s landing . <eos>

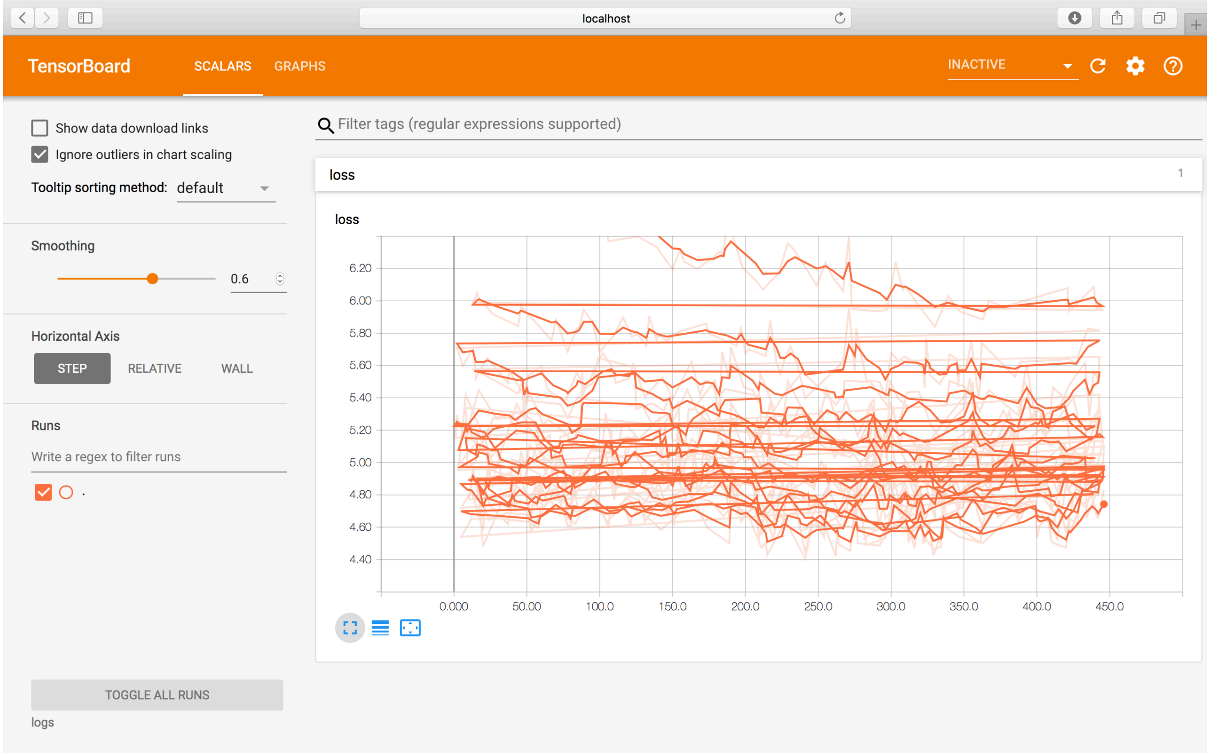
Submitted version of Romanesco program in the GitHub repositorium includes the combination of changes that helped to achieve the lowest scoring of the perplexity. To achieve 118.70 the following aspects were required: the preprocessing step of the dataset, then the number of neurons is 2500, learning rate is at 0.001, and the number of *epoch* was changed from the default 10 to 15 *epoch*.

As a result, the following perplexities were achieved:
Perplexity on training data after epoch 15: 103.90
Scoring Perplexity: 118.70

The generated text is as follows:
were or infinitesimal maria all the social master drawn in the public glen of sitting parts of full lane a old hint .
bloom : (by her eyes when that sure we provide on him .
but boylan five features ’ s band receiver movement .
we can rest away and you are ?
the hobgoblin with the gob .
distributing having half after go .) night , martin lambert ’ s shelter , and the sailor , to himself with his heart , though a cart from fez , were shift on evidence) at his repose) lynch .) why want to eat me , says joe .
that ’ s rooms , ben kate said such by plenty . you suppose at cork winks !
as you all not or his tweedy showed one tired to our lower with one then p.m. to master small views over his from attraction .
thank whack off the place with those diffusion of the favourite and included for his .
just do ?
lily with bliss .
four surgeon do , it ?
i ask

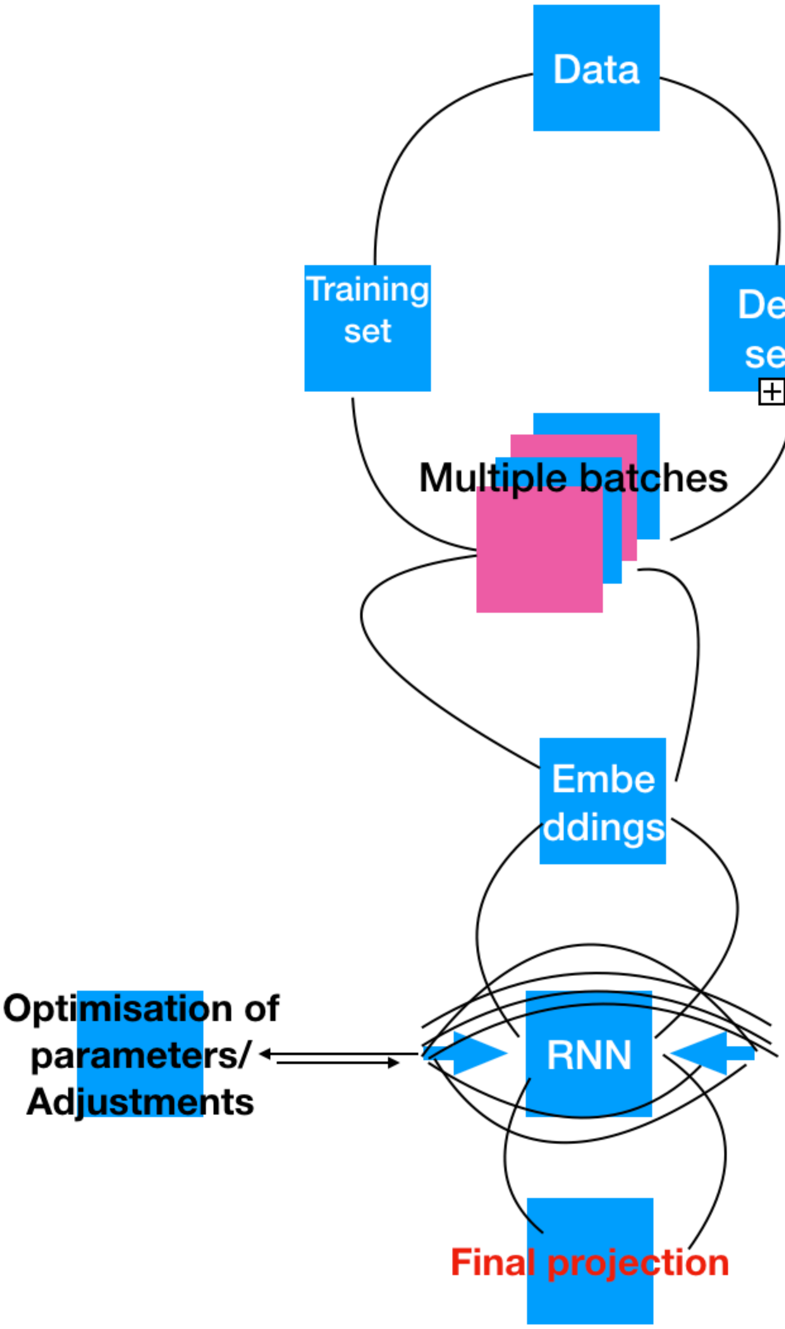
Also, Tensorboard was a challenge as well. Access to it was achieved by the following command:

```
$ tensorboard --logdir logs
```



☐ Show data download links
☑ Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing 0.6
Horizontal Axis: STEP
Filter tags (regular expressions supported)
Runs: Write a regex to filter runs
☑ ○ .

#8 Model Scheme


Data
Training set
Dev set
Multiple batches
Embe ddings
Optimisation of parameters/ Adjustments
RNN
Final projection