Chapter 3

**Project:**

**Building a ChatBot with Deep NLP**

Project Introduction

**3.1 ChatBot project intro**

We'll build the Chatbot from scratch following a step by step approach. We will build a super powerful ChatBot by implementing *Deep-NLP Seq2Seq* model with one of the best API the ***TensorFlow***. (We could use ***Pytorch***, but ***TensorFlow*** will give better result).

* Implementation steps: This implementation will be done in five parts:

1. Part 0:
2. *Installing Anaconda IDE and getting the data set:* We will actually use more precisely spider inside Anaconda.
3. *Data set:* Which is the ***Cornell movie corpus Data-set***. Basically it's a data set of more than *600 movies* containing thousands of *conversations* between lots of characters.

* We train our *ChatBot* on this dataset because it can have *general conversation* with us like a *friend*, instead of *specified* *ChatBot*.
* However, the model can be trained on other datasets for some other purposes, for example a more specific dataset like a *calendar assistant* or a *navigation assistant*. These are some more specific applications.
* But we will try a ***general ChatBot*** to talk about ***everyday conversations*** and that's why movies are perfect, because in movies you have a lot of ***random*** every day ***conversations*** between friends.

1. Part 1: Data processing.

* Data processing is inevitable whenever you build an AI or an ML model you have to make the data set compatible with the model you're going to build.
* We're going to build a *NN* based model and therefore the data will have to have a *special format* especially for the inputs.
* Besides we'll have to *clean the text* because the *less we clean* it and *simplify* it the *more difficult* it will be for a *model* to *train itself* to talk like a human.
* So we will have to do a *lot of data processing* this will not be the *funniest part*.

1. Part 2: Seq2Seq Architecture.

* We will actually build a brain composed of an encoder and then a decoder and we will assemble all of them to build the final brain which will be trained later.

1. Part 3: Training Seq2Seq

* It will be about training the Seq2Seq model built in part 2.
* We will set up a ***Loss*** function, the ***optimizer*** and then apply ***SGD*** to ***update*** the ***weight*** of the neurons of the brain so that it improves its ability to talk with us.

1. Part 4: Test the Seq2SEq model

* We will make some kind of a *code* to have an *interface* where we can ask *some questions* and then the *ChatBot* will answer that.
* By *observing* its *answers* we can see how its capable of conversing with us.

**3.2 Install Anaconda**

* Install anaconda from its site with spyder.
* Install TensorFlow using anaconda CLI. (Anaconda Prompt).

"conda create –n chatbot python3.5 anaconda"

Above command will install ***python3.5*** in the Anaconda Virtual Environment. It will be used to install TensorFlow in this VE.

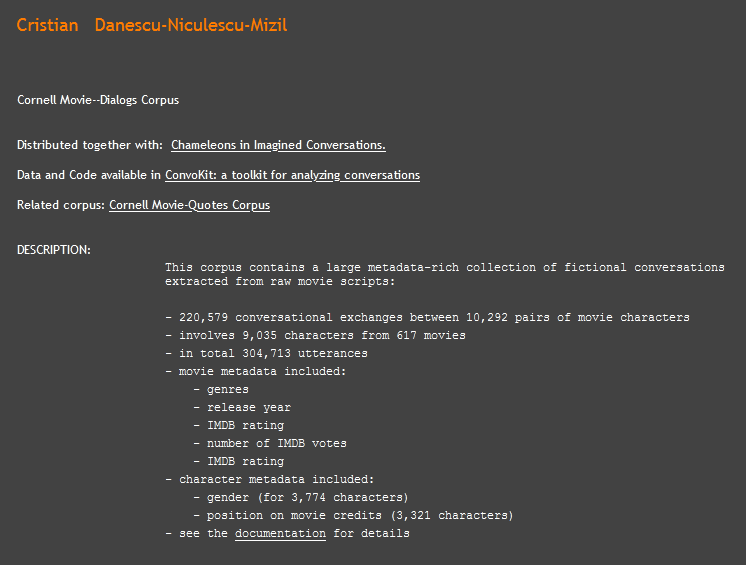
* To activate the VE: use command "**activate chatbot** "
* Then use this command: "**pip install tensorflow 1.0.0**"
* Connect to the VE: In the left corner Applications on>chatbot

**3.3 Collect the DATA**

The dataset is called the CORNELL MOVIE CORPUS DATA SET and it's basically a data-set containing thousands of conversations from 600 plus movies between many characters.

* Search in Google: "cornell movie dialog corpus"

Link: <https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html>



* Go to the link:

<https://www.cs.cornell.edu/~cristian/Chameleons_in_imagined_conversations.html>

And download "Legacy Dataset".

* Direct Link: <http://www.mpi-sws.org/~cristian/data/cornell_movie_dialogs_corpus.zip>

You can also download "Quote Corpus":

<https://www.cs.cornell.edu/~cristian/memorability_files/cornell_movie_quotes_corpus.zip>

* Real Data vs MetaData: After download and unzip the files, we have to make the difference between *metadata* and the *real data*. Metadata is the data we will not use for the training.
* Like the *ratings* or the *movie* that comes from (like some description or data). We just need the *conversations* between the characters.
* We just need two files: "**movie\_conversations.txt**" and "**movie\_lines.txt**".
* **movie\_lines.txt** contains some extracts of scripts taken from movies. It contains several columns separated by "**+++$+++**" strings. So let's have a look, use "Sublime text" to open it:

L1045 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ They do not!

L1044 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ They do to!

L985 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ I hope so.

L984 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ She okay?

L925 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Let's go.

L924 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ Wow

L872 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Okay -- you're gonna need to learn how to lie.

L871 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ No

L870 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ I'm kidding. You know how sometimes you just become this "persona"? And you don't know how to quit?

L869 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Like my fear of wearing pastels?

L868 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ The "real you".

L867 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ What good stuff?

L866 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ I figured you'd get to the good stuff eventually.

L865 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ Thank God! If I had to hear one more story about your coiffure...

L864 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Me. This endless ...blonde babble. I'm like, boring myself.

L863 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ What crap?

L862 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ do you listen to this crap?

* First column is just an ID. ID of the line said by some character in some specific movie.
* The second column is actually the character. U means "user" here. There are u0, u1, u2 etc.
* If we take these first two lines here that a **conversation** between **u0** and **u2**.

L1045 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ They do not!

L1044 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ They do to!

* The third column is the movie. Represented by ***m0***, ***m1***, ***m2*** etc.
* The fourth column is the *name of the character* and this is actually some metadata which we *won't use* to *train* our ***ChatBot***.
* Finally the last column is the line. These lines are the actual conversation.
* *Notice the IDs, there are some patterns:* For example ***L1044*** is ***L1045*** before, it means **u2** asks **u0**, and answers **"They do not!**".
* Also notice there is a long conversations between **u0** and **u2** from ***L862*** to ***L872***.
* It's important for you to understand well the structure of the data set. This is movie lines that basically contains several conversations between some characters taken from different movies.
* Now let's have a look at the other interesting dataset, which is **movie\_conversations.txt** open it with Sublime Text as well.

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L194', 'L195', 'L196', 'L197']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L198', 'L199']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L200', 'L201', 'L202', 'L203']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L204', 'L205', 'L206']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L207', 'L208']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L271', 'L272', 'L273', 'L274', 'L275']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L276', 'L277']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L280', 'L281']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L363', 'L364']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L365', 'L366']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L367', 'L368']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L401', 'L402', 'L403']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L404', 'L405', 'L406', 'L407']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L575', 'L576']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L577', 'L578']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L662', 'L663']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L693', 'L694', 'L695']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L696', 'L697', 'L698', 'L699']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L860', 'L861']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L862', 'L863', 'L864', 'L865']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L866', 'L867', 'L868', 'L869']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L870', 'L871', 'L872']

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L924', 'L925']

* This kind of lists **['L194', 'L195', 'L196', 'L197']** are the conversations between the characters, here it is represented as the *list of movie-lines* from the previous dataset **movie\_lines.txt**.
* *Each row* of this dataset *corresponds* to *one single conversation*. So as you can see in each *conversation* we have *several* *lines* and these lines in *one same conversation* are all the lines composing the conversation.
* And this is important because the *previous dataset didn't somehow separate the conversation* so we need this data set to make sure we *distinguish* two *different conversations* on which ***ChatBot*** will be trained.