# TV Script Generation (The Office)

NLP Sessional 3 Project

Submitted By:

Parikshit Saikia
CSB17035

CONTENTS

# 1.INTRODUCTION:

Over the past few years, NLP and language generation has experienced a renaissance moment. Researchers are finding new efficient techniques and applications of NLP.

In this project, we are going to explore one of the applications of NLP which is Script Generation. In script or text generation machines are trained on a large amount of text from a particular domain which will later generate similar but unique text of that particular domain.
Thanks to major advancements in the field of Natural Language Processing (NLP), machines are able to understand the context from text and spin up tales all by themselves.

So that's a general introduction to the problem statement.
In this project, the goal is to generate a unique  TV script for the Sitcom "The Office" which feels like a real handwritten script having a plot, storyline, the essence of the characters, and humor.

# 2.DATASET:

The first challenge I faced is getting an authentic dataset. Finally, a found a dataset with an organized format having a character-wise script of each episode from the starting to final season.
The_Office_Dataset

However, it had a minor issue the symbol  " ‘ " was replaced by some random garbage numbers.
So, I had written a separate script to process those.
Below there is a preview of the overall format.

| id | season | episode | scene | line_text | speaker | deleted |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | All right Jim. Your quarterlies look very good. How are things at the library? | Michael | FALSE |
| 2 | 1 | 1 | 1 | Oh, I told you. I couldn't close it. So... | Jim | FALSE |
| 3 | 1 | 1 | 1 | So you've come to the master for guidance? Is this what you're saying, grasshopp | Michael | FALSE |
| 4 | 1 | 1 | 1 | Actually, you called me in here, but yeah. | Jim | FALSE |
| 5 | 1 | 1 | 1 | All right. Well, let me show you how it's done. | Michael | FALSE |
| 6 | 1 | 1 | 2 | [on the phone] Yes, I'd like to speak to your office manager, please. Yes, hello. Th | Michael | FALSE |
| 7 | 1 | 1 | 3 | I've, uh, I've been at Dunder Mifflin for 12 years, the last four as Regional Manage | Michael | FALSE |
| 8 | 1 | 1 | 3 | Well. I don't know. | Pam | FALSE |
| 9 | 1 | 1 | 3 | If you think she's cute now, you should have seen her a couple of years ago. [grov | Michael | FALSE |
| 10 | 1 | 1 | 3 | What? | Pam | FALSE |
| 11 | 1 | 1 | 3 | Any messages? | Michael | FALSE |
| 12 | 1 | 1 | 3 | Uh, yeah. Just a fax. | Pam | FALSE |
| 13 | 1 | 1 | 3 | Oh! Pam, this is from Corporate. How many times have I told you? There's a spec | Michael | FALSE |
| 14 | 1 | 1 | 3 | You haven't told me. | Pam | FALSE |
| 15 | 1 | 1 | 3 | It's called the wastepaper basket! Look at that! Look at that face. | Michael | FALSE |
| 16 | 1 | 1 | 4 | People say I am the best boss. They go, 'God we've never worked in a place like th | Michael | FALSE |
| 17 | 1 | 1 | 5 | [singing] Shall I play for you? Pa rum pump um pum [Imitates heavy drumming] I h | Dwight | FALSE |
| 18 | 1 | 1 | 6 | My job is to speak to clients on the phone about... uh, quantities and type of copi | Jim | FALSE |
| 19 | 1 | 1 | 7 | Whassup! | Michael | FALSE |
| 20 | 1 | 1 | 7 | Whassup! I still love that after seven years. | Jim | FALSE |
| 21 | 1 | 1 | 7 | Whassup! | Michael | FALSE |
| 22 | 1 | 1 | 7 | Whassup! | Dwight | FALSE |
| 23 | 1 | 1 | 7 | Whass...up! | Michael | FALSE |
| 24 | 1 | 1 | 7 | Whassup. | Dwight | FALSE |
| 25 | 1 | 1 | 7 | [Strains, grunts] What? | Michael | FALSE |
| 26 | 1 | 1 | 7 | Nothing. | Jim | FALSE |
| 27 | 1 | 1 | 7 | OK. All right. See you later. | Michael | FALSE |
| 28 | 1 | 1 | 7 | All right. Take care. | Jim | FALSE |
| 29 | 1 | 1 | 7 | Back to work. | Michael | FALSE |

However, this is not the final version of the dataset that we will be using. We just need the line_text and the speaker column from the dataset and convert it into text format like this.

```
Michael: All right Jim. Your quarterlies look very good. How are things at the library?

Jim: Oh, I told you. I couldn't close it. So...

Michael: So you've come to the master for guidance? Is this what you're saying, grasshopper?

Jim: Actually, you called me in here, but yeah.

Michael: All right. Well, let me show you how it's done.

Michael: [on the phone] Yes, I'd like to speak to your office manager, please. Yes, hello. This is Michael Scott. I am the Regional Manager of

Michael: I've, uh, I've been at Dunder Mifflin for 12 years, the last four as Regional Manager. If you want to come through here... See we hav

Pam: Well. I don't know.

Michael: If you think she's cute now, you should have seen her a couple of years ago. [growls]

Pam: What?

Michael: Any messages?

Pam: Uh, yeah. Just a fax.

Michael: Oh! Pam, this is from Corporate. How many times have I told you? There's a special filing cabinet for things from corporate.

Pam: You haven't told me.

Michael: It's called the wastepaper basket! Look at that! Look at that face.
```

# 3. APPROACH:

Our core program is a neural network that will take input as text and will give output as text. But we know that a neural network will only take numbers as input.
So, we have to make some data structure that will map our text to some unique number and vice versa.
NOTE: The mapping must be one to one.
Hence, we need to do some more data preprocessing to make it feasible to give input for the neural network.

We need to  Implement the following pre-processing functions below:

- Lookup Table
- Tokenize Punctuation

Lookup Table
To create a word embedding, you first need to transform the words into int ids. In this function, create two dictionaries:

- Dictionary to go from the words to an id, we'll call vocab_to_int
- Dictionary to go from the id to word, we'll call int_to_vocab

Return these dictionaries in the following tuple (vocab_to_int, int_to_vocab)

Token Punctuation
This function will look for punctuation in word and return the root word

Now we can split the dataset into words using some delimiter and apply the functions above to get the bag of words and their mapping to int.

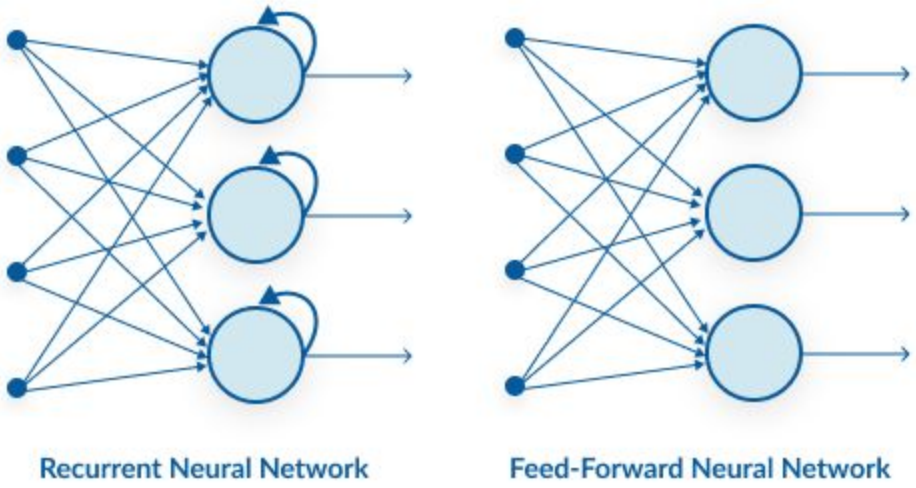
Int_text = integer representation of the script
Vocab_to_int = mapping from words to int ids

Since our script is converted to int, now we are ready to input it into the neural network.

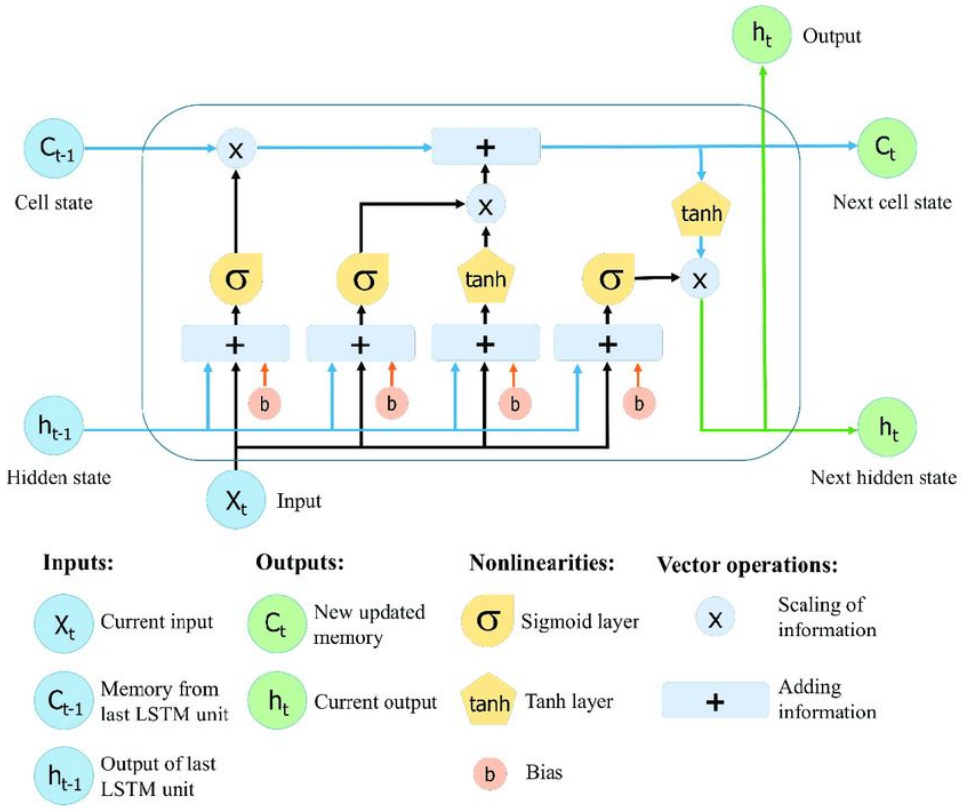
# 4.MODELING AND TRAINING

Now our data is ready, it's time to make the model. In generating a script the context of the words is important, So we must consider what inputs we gave to the model before. The output should not only depend on the current input but also the previous input it has seen.

Normal Neural Neural network does not provide this feature because it doesn't take time under consideration but Recurrent Neural Network does.

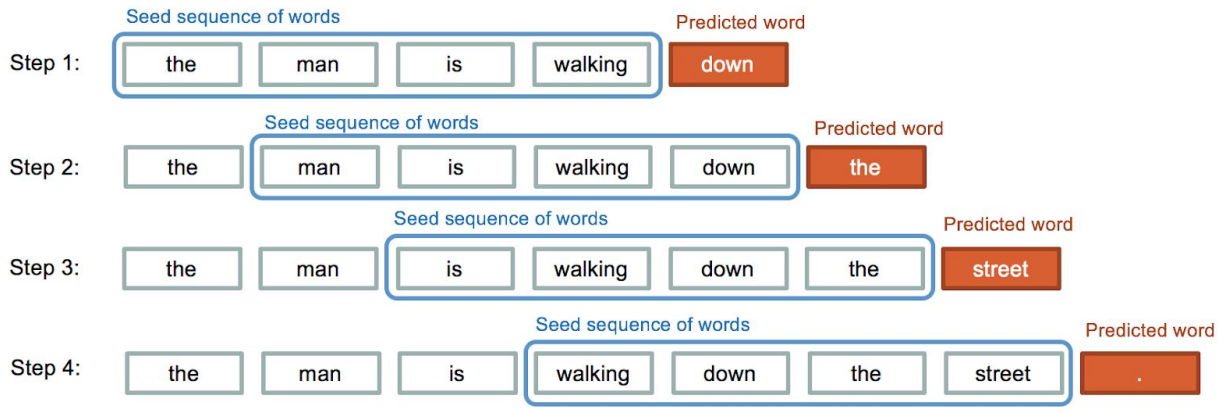



In RNN we will use LSTM (Long Short Term Memory) because it has the ability to extract the contextual meaning better because of its architecture it doesn't suffer backpropagation decay unlike RNN, Thus provide a better result.




So our model is ready and its time for training, here how it works

The model will take a batch of input sequences from the training data and try to predict the next word, and add to the input for the prediction of the next word and thus the script is generated.



# 5.RESULT AND FURTHER SCOPE

The generated script is not perfect but the result is quite satisfactory. The generated script has a contextual meaning in the dialogues and it had the essence of the character in their respective dialogues. Finally, the overall script has a plot in it.

Further, We may design for complex and effective architecture so that it can extract more important contextual important info, like Transformer, Attention model, even GPT

We can also do some effective data preprocessing so that it makes the job model easy to extract features.

Result:

**andy: i think you're a great man, and you don't have to be able to be able to be a hero. you can have a job, and you are.**

**jim: no, i don't want you.**

**jim: i don't think you should do that.**

**dwight: i don't know what i do.**

**jim: oh, no, i just don't think i should go to work.**

**michael: no.**

**dwight: i am.**

**michael: i am going to go to the bathroom.**

**dwight: i don't care.**

**michael: oh, i don't know.**

**andy: i am so proud of that. i mean i think i'm a little nervous. but i have been in the bathroom for a few months.**

**andy: yeah, well, it's just a big idea, so you should know that i was a lot better than that... i don't know, i'm sorry, i was thinking about it and i don't care about the rules. i think you should go back.**

**andy: i am so sorry, i don't know, i know.**

**michael: i don't know what to do, i will do that.**

**michael: i don't. i think we should get a couple minutes. we should get a chance to get the best christmas party for you. and i am going to be in a few minutes...**

**dwight: you know what? i am so sorry, i just... i don't even know what it means.**

**dwight: you know what? i don't have to be the first person in the world.**

**michael: i am a little concerned.**

**michael: oh my god, oh my god! oh my god!**

**michael: oh, oh, hey. i have a good day, i will see you in a bit.**

**michael: i think you should just...**

**jim: you know what? i am so happy that we are not a little bit of the same ones.**

**dwight: i don't think i could do this.**

**jim: no, it's not the same way, it's just, i just don't think so.**

**andy: no, i'm not going to be here.**

**andy: i have an idea that you are going to have a lot of money, and i think i have to get it.**

**jim: no, i just don't.**

**michael: okay. well, i am going to be a little confusing.**

**dwight: i know i can.**

**michael: you know what? i think i should go home with a little bit.**

**dwight: i think it's a great idea,**

# 6. Conclusion

The goal of the project is successfully fulfilled, Generated a TV Script for “The Office” having some plot and some contextual meaning although there is much room for improvement.
If anyone is interested, they can contribute to GitHub:
https://github.com/parikshitsaikia1619/TV-Script-Generation-The-Office