# NATURAL LANGUAGE STATISTICAL FEATURES OF NEURAL LANGUAGE GENERATED PSEUDO TEXT

Submitted by: Sumedha

M.Tech - S.P.D.D (Third Semester)

Roll No.: 2K19/SPD/17

Department of Electronics & Communication,
Delhi Technological University

Under the supervision of: Prof. Rajesh Rohilla

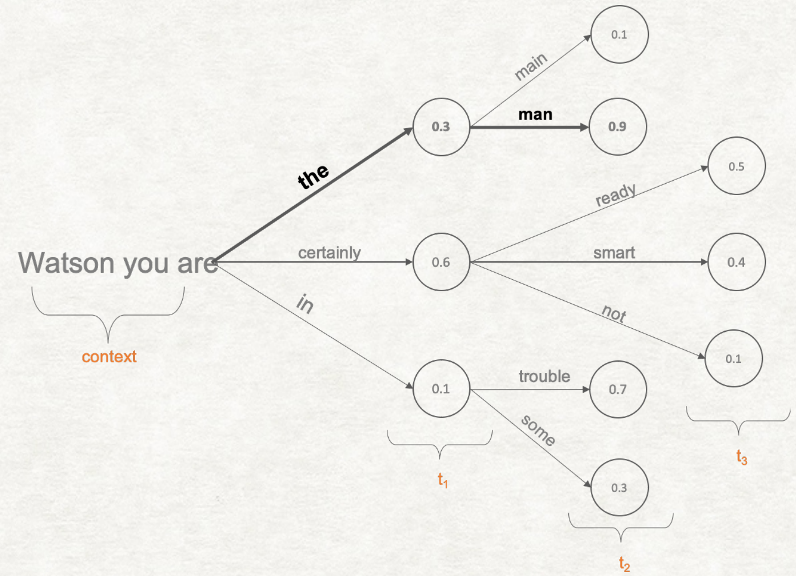
Department of Electronics & Communication,
Delhi Technological University

# CONTENTS

Github Link to Project https://github.com/sumedhachugh/Natural-Language-Statistical-Features-of-neural-language-Generated-pseudo-Text

# AIM

- To generate pseudo text using LSTMs (Long Short Term Memory networks)
- Evaluate how close this machine generated text is to human generated text by checking if they follow statistical features followed by human generated text such as Zipf's and Heap's Laws for Words

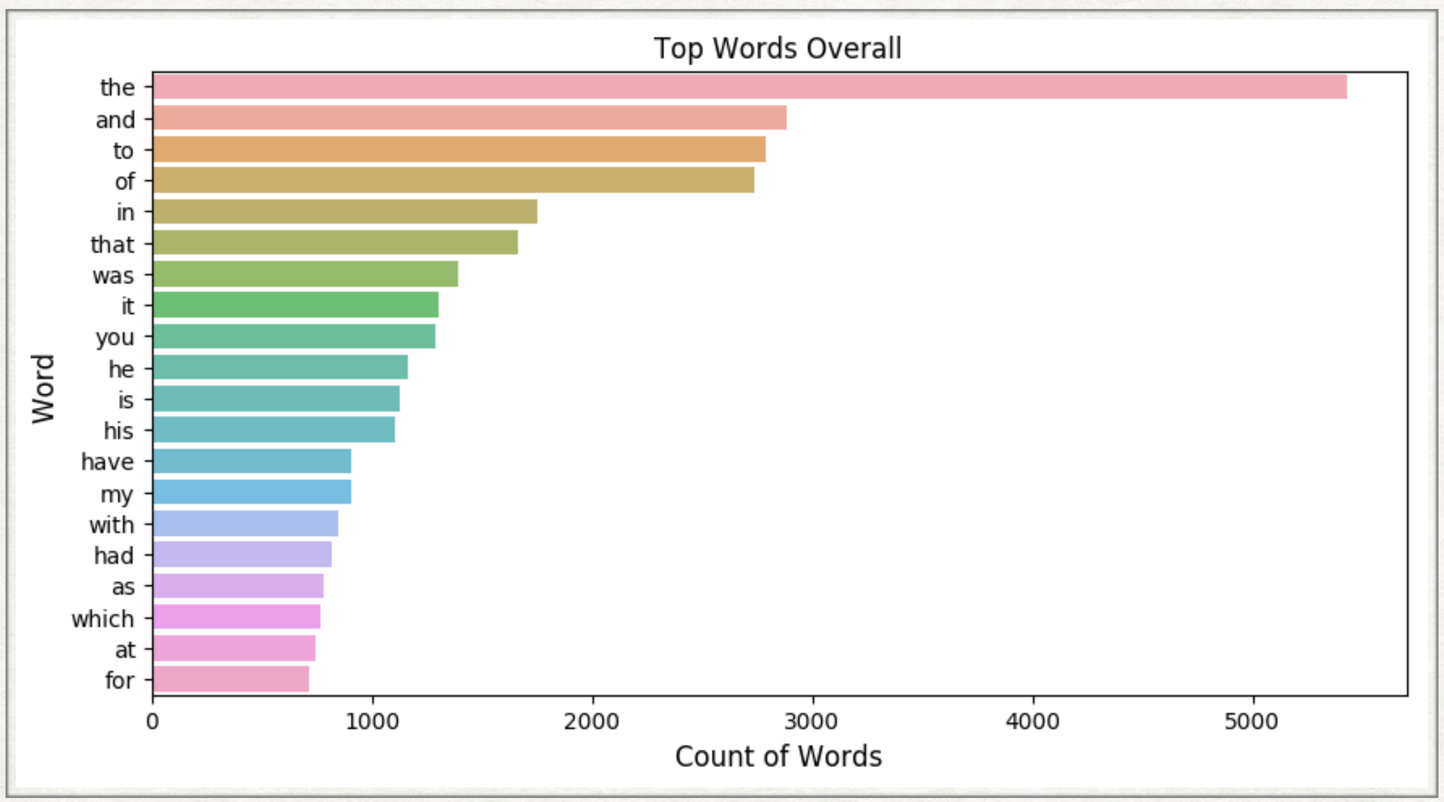



# RESEARCH PAPER IMPLEMENTED

- Natural Language Statistical Features of LSTM-Generated Texts by Marco Lippi , Marcelo A. Montemurro , Mirko Degli Esposti, and Giampaolo Cristadoro
- Published in IEEE Transaction on Neural Networks and Learning Systems, VOL. 30, NO. 11, dated NOVEMBER 2019
- Link: https://ieeexplore.ieee.org/document/8681285

# DATASET USED

- As dataset I utilized the famous book "The Adventures of Sherlock Holmes" by Sir Arthur Conan Doyle.
- The book is made available through Project Gutenberg
- Dataset Link: https://www.gutenberg.org/files/1661/1661-0.txt
- Dataset Stats: Book contains a total of 594197 characters




```
the       5428
and       2888
to        2790
of        2737
in        1750
that      1664
was       1394
it        1303
you       1286
he        1168
is        1129
his       1103
have       909
my         907
with       850
had        822
as         780
which      770
at         742
for        716
Name: 0, dtype: int64
```

Fig: Most common words and their occurring frequency

First two lines of the dataset: I had seen little of Holmes lately. My marriage had drifted us away from each other.

# INTRODUCTION TO NATURAL LANGUAGE STATISTICAL FEATURES

## ZIPF'S LAWS FOR WORDS

*All human generated texts in all the languages follow zipf's law and heap's law. If our machine generated text also follows these laws we say that this neural language generated text is close to natural language generated text*

According to Zipf's law for all human generated texts the rank-frequency distribution is an inverse relation (frequency is proportional to inverse of rank).

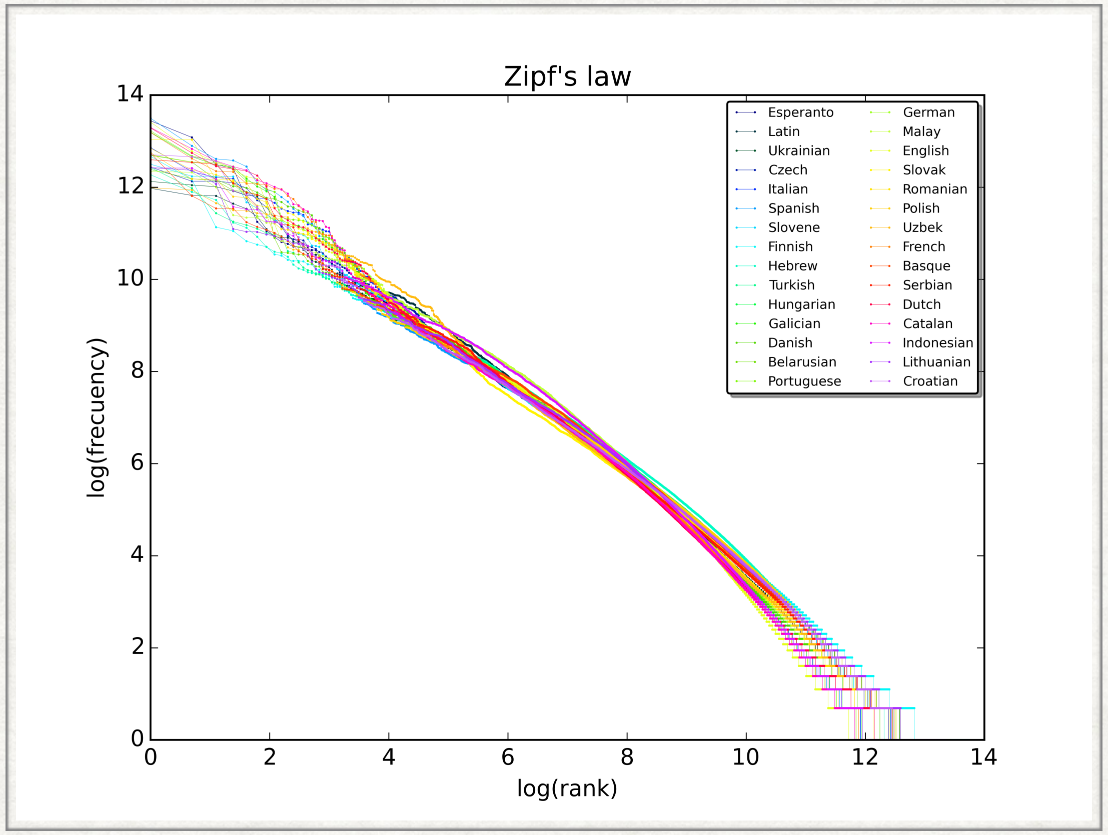



*Fig: A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias in a log-log scale*

# INTRODUCTION TO NATURAL LANGUAGE STATISTICAL FEATURES

## HEAP'S LAWS FOR WORDS

The law can be described like as the number of words in a document increases, the rate of the count of distinct words available in the document slows down.

e.g: Suppose in a document with 1000 words no. of unique words are 100, then for a document with 2000 words no. of unique words will be less than 200, for a document with 3000 words no. of unique words will be much less than 300 etc.

- The documented definition of Heaps' law says that the number of unique words in a text of n words is approximated by: V(n) = K n^**β**

where K is a positive constant and **β** is between 0 and 1, K is often upto 100 and **β** is often between 0.4 & 0.6

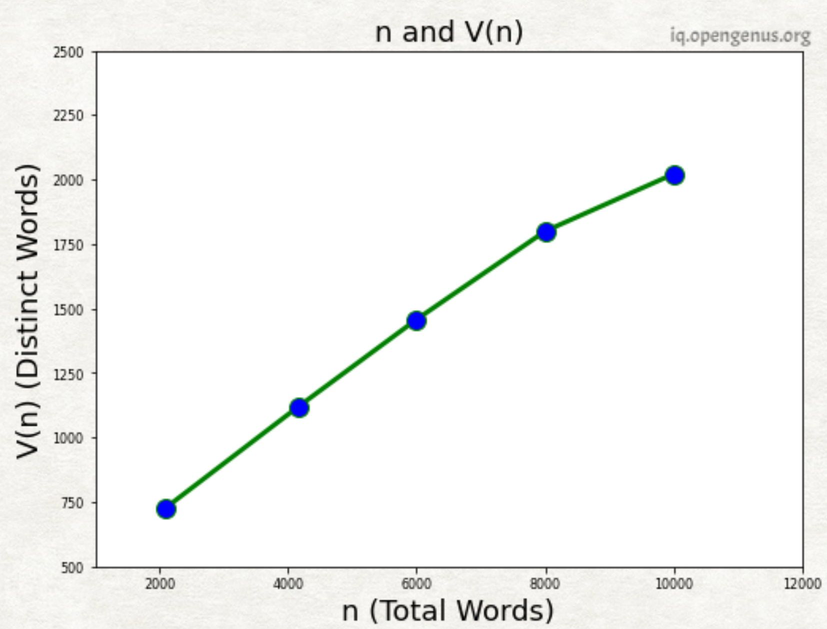



*Fig: A typical Heaps-law plot. The x-axis represents the text size, and the y-axis represents the number of distinct vocabulary elements present in the text*

# INTRODUCTION TO RNN (RECURRENT NEURAL NETWORKS)

- Recurrent Neural Networks (RNNs) are a type of Neural Network where the output from previous step is fed as input to the current step
- As a result RNNs remember the past and it's decisions are influenced by what it has learnt from the past
- These are designed to work with sequential data like sentences audio, video etc.

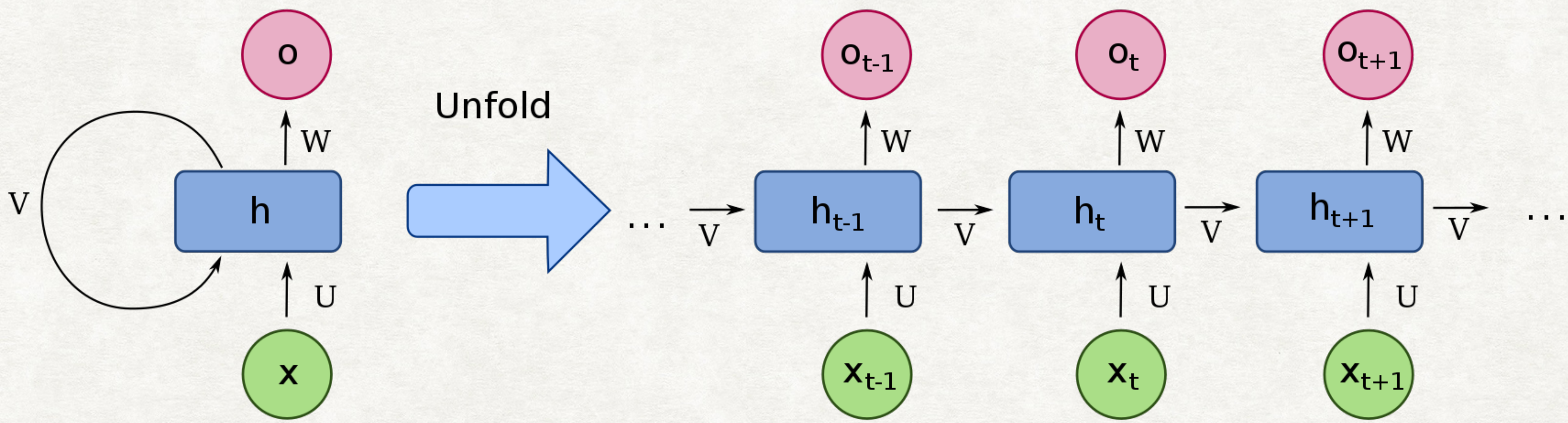



*Fig: Unfolded RNN Cell*

In the fig. x(t) is current state input

V is output from previous state

# INTRODUCTION TO LSTM (LONG SHORT TERM MEMORY NETWORKS)

*While training really large sequence with the RNNs weights of sequences at end start overwriting the weights of word in the beginning, hence they start forgetting the starting part (a.k.a. Vanishing Gradient problem), so LSTMs come into play*

- Long Short Term Memory networks  are a special kind of RNN, capable of learning long-term dependencies

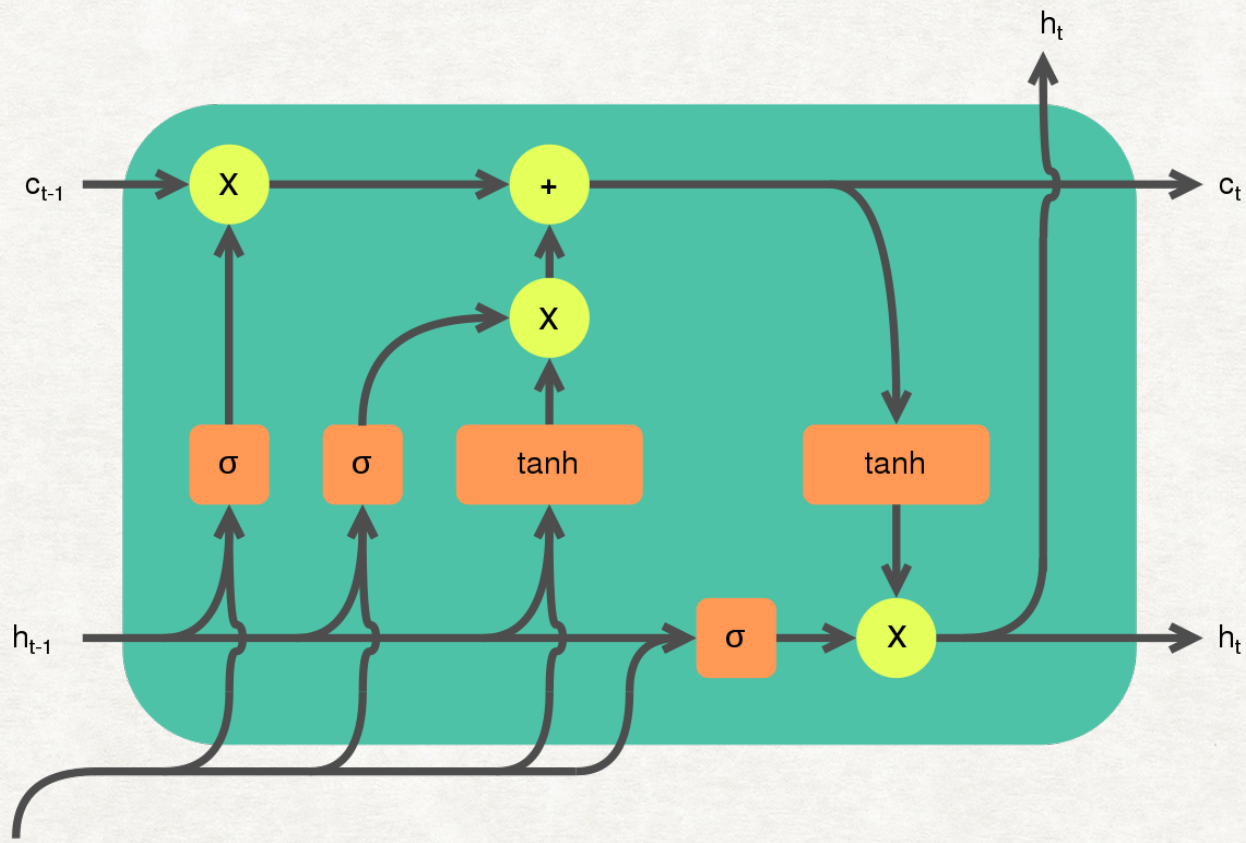



*Fig: LSTM Cell*

In the fig. x(t) is current state input

h(t-1) is output from previous state

c(t-1) is cell state

# INTRODUCTION TO LSTM (LONG SHORT TERM MEMORY NETWORKS)

Its working can be explained in 4 steps

1. Decide what information to throw away from the cell state.This decision is made by a sigmoid layer. “1” as output of this layer is represents “keep” while a “0” represents “throw

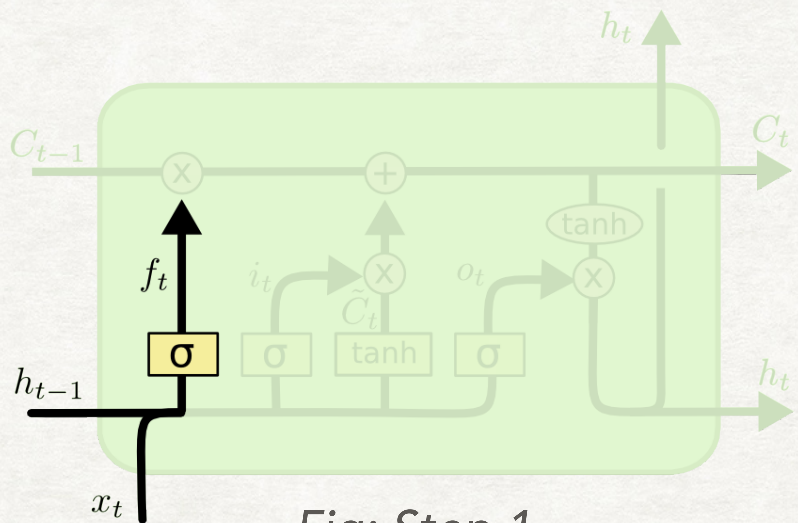



Fig: Step 1

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

2. Decide what new information to store in the cell state. This has two parts. First, a sigmoid layer called the “input gate layer” decides which values we’ll update. Next, a tanh layer creates a vector of new candidate values, C̃ t, that could be added to the state. In the next step

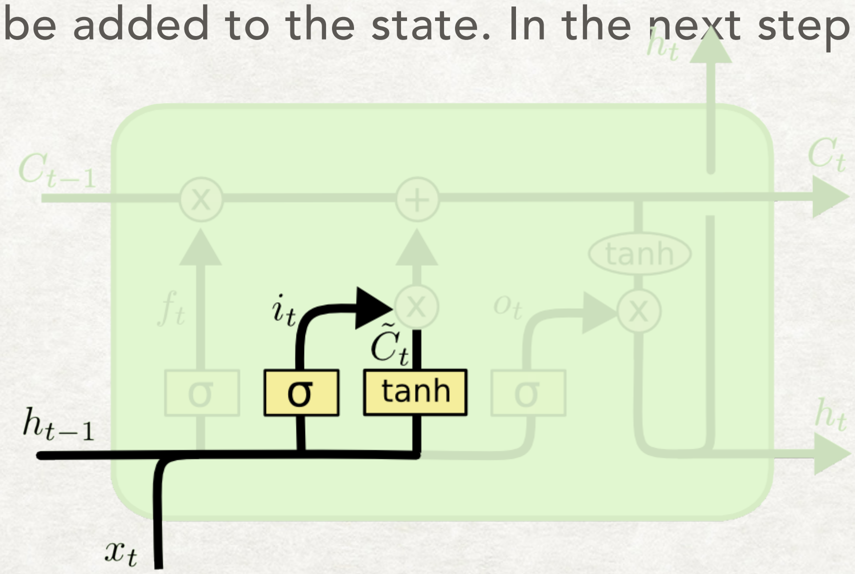



Fig: Step 2

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

# INTRODUCTION TO LSTM (LONG SHORT TERM MEMORY NETWORKS)

3. Update the old cell state, C(t−1), into the new cell state C(t).

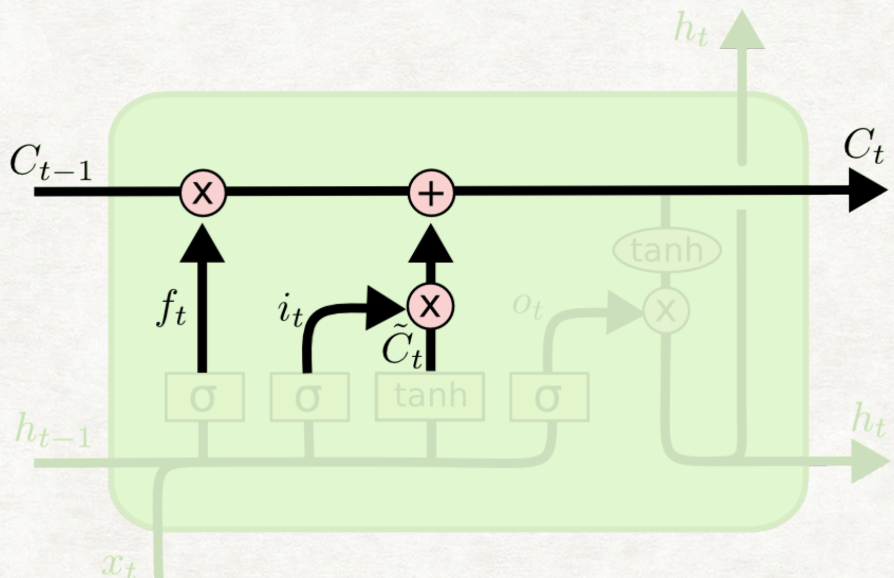



*Fig: Step 3*

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

4. Updating the output H(t). First a sigmoid layer decides what parts of the cell state go to output. Then, cell state is put through tanh and is multiplied by the output of the sigmoid gate

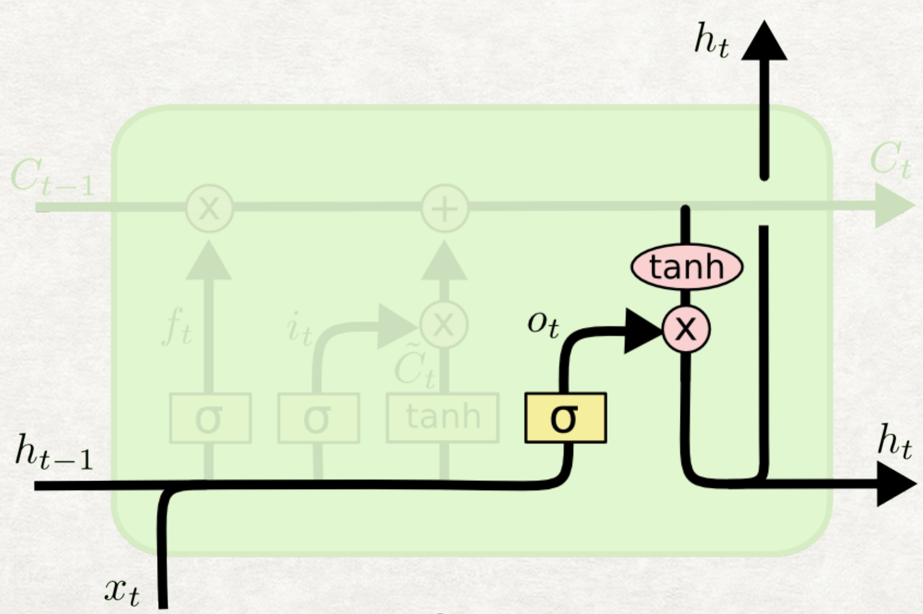



*Fig: Step 4*

$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

# FLOWCHART

IMPORTING LIBRARIES AND DATASET

↓

EXPLORING THE DATASET (FINDING TOTAL NO. OF CHARACTERS, MOST COMMONLY USED WORDS)

↓

PRE-PROCESSING THE DATASET (CLEANING, TOKENIZING, REMOVING PUNCTUATIONS AND STOP WORDS)

↓

PREPARING DATASET FOR TRAINING (APPLYING SLIDING WINDOW TECHNIQUE ,CHARACTER TO INTEGER MAPPING)

↓

DECIDING HYPER-PARAMETERS (TRIED DIFFERENT COMBINATIONS OF BATCH SIZE, NUMBER OF EPOCHS, LAYERS AND CHOSE THE ONES GIVING MAX. ACCURACY & MIN. LOSS)

↓

CREATING LSTM BASED MODEL, FITTING THE MODEL AND CALCULATING ACCURACY

↓

GENERATING NEW TEXT FOR DIFFERENT VALUES OF TEMPERATURE (T=0.1 TO T=3)

↓

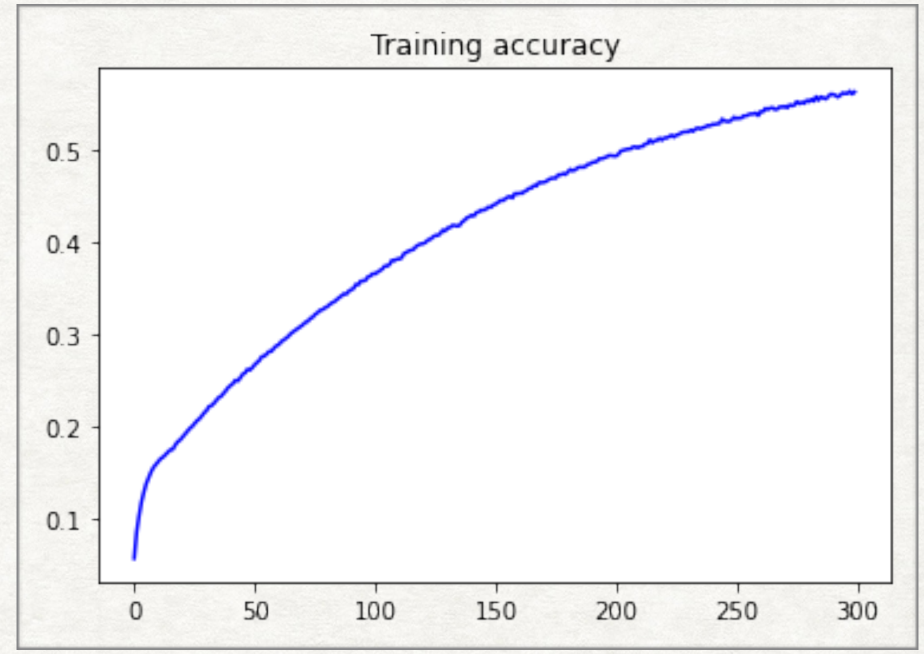
COMPARING HOW WELL LSTM GENERATED TEXT FOR DIFFERENT VALUES OF T FOLLOW ZIPF'S AND HEAP'S LAW

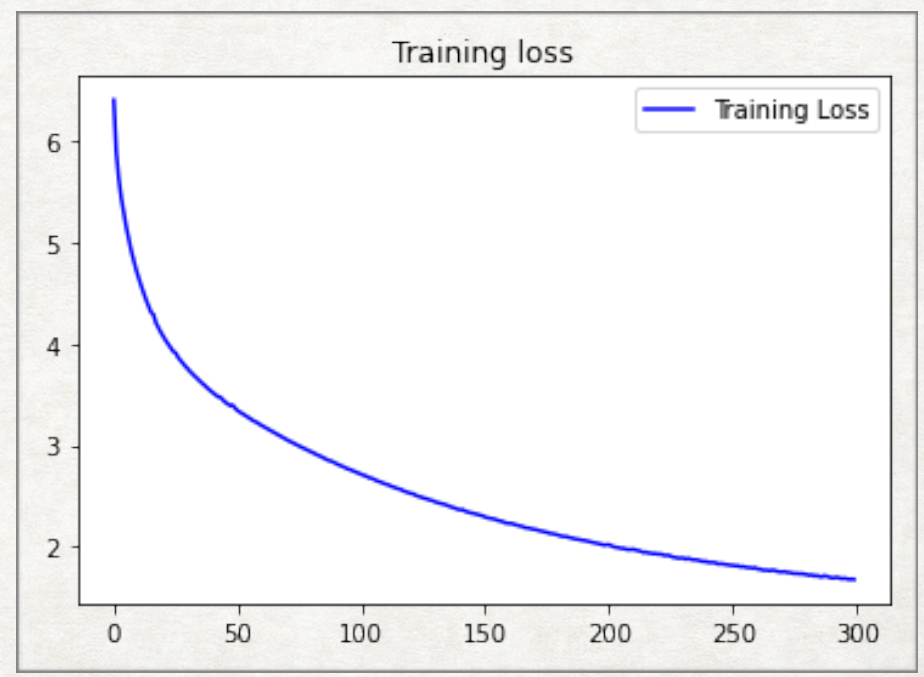
# HYPER-PARAMETERS USED

## TRIAL - 1

- No. of neurones in hidden layer = 150, No. Of epochs = 300, input_length = 25, activation='relu', loss='categorical_crossentropy', optimizer='adam'
- No. Of LSTM Layers = 2
- No. Of Dense Layers = 2
- No. Of Dropout layers = 1
- Training Accuracy Obtained = 55%
- Training Loss Obtained < 1%

```
# define model
model = create_model(vocabulary_size+1, seq_len)

Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 25, 25)            206425
_________________________________________________________________
lstm_5 (LSTM)                (None, 25, 150)           105600
_________________________________________________________________
dropout_3 (Dropout)          (None, 25, 150)           0
_________________________________________________________________
lstm_6 (LSTM)                (None, 150)               180600
_________________________________________________________________
dense_4 (Dense)              (None, 150)               22650
_________________________________________________________________
dense_5 (Dense)              (None, 8257)              1246807
=================================================================
Total params: 1,762,082
Trainable params: 1,762,082
Non-trainable params: 0
_________________________________________________________________
```



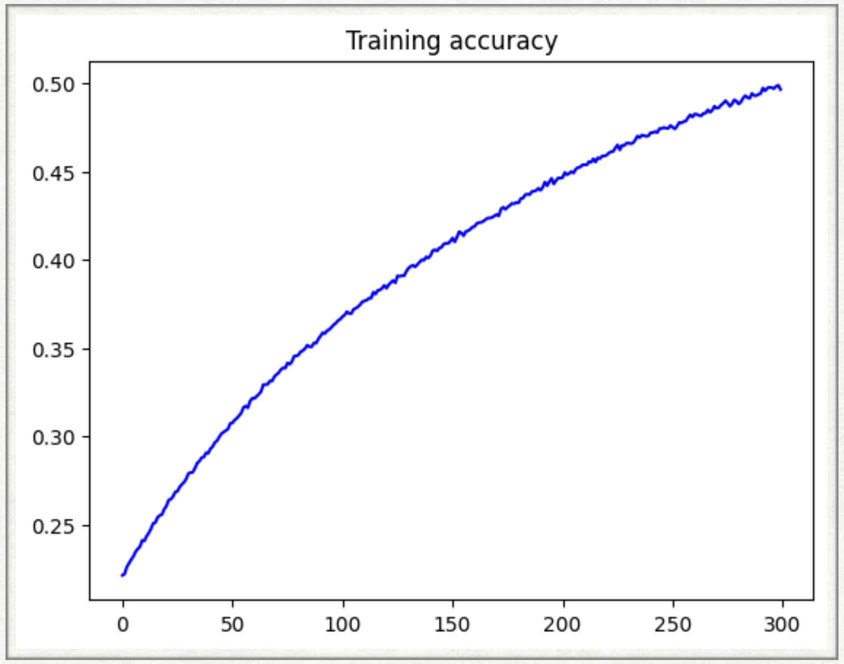
# HYPER-PARAMETERS USED

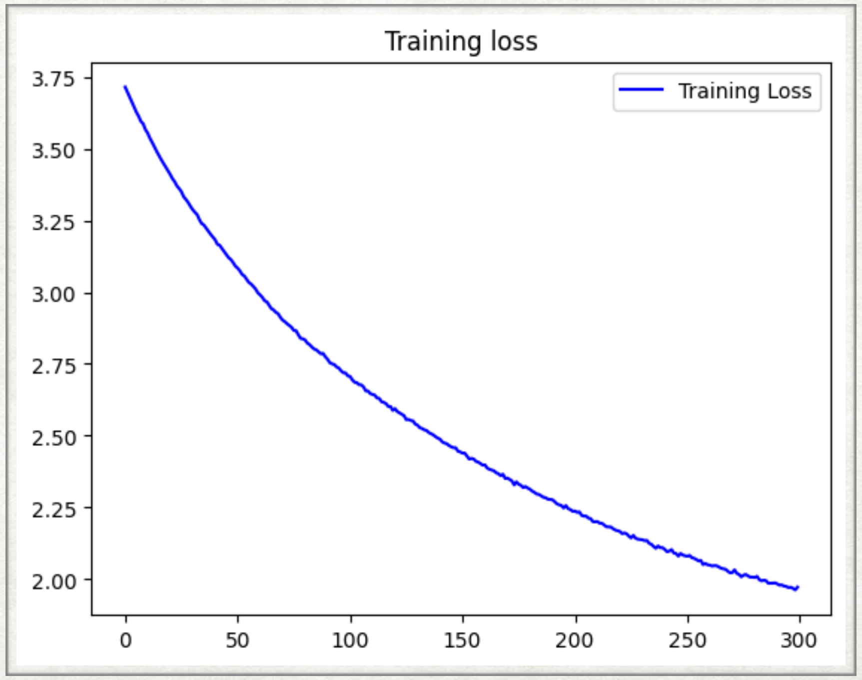
## TRIAL - 2

### No of LSTM & Dropout layers increased, Training Accuracy Decreases, Training Loss Increases

- No. of neurones in hidden layer = 150, No. Of epochs = 300, input_length = 25, activation='relu', loss='categorical_crossentropy', optimizer='adam'
- No. Of LSTM Layers = 3
- No. Of Dense Layers = 2
- No. Of Dropout layers = 2
- Training Accuracy Obtained = 50%
- Training Loss Obtained  =2 %

```
# define model
model = create_model(vocabulary_size+1, seq_len)

Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 25, 25)            206425
_________________________________________________________________
lstm (LSTM)                  (None, 25, 150)           105600
_________________________________________________________________
dropout (Dropout)            (None, 25, 150)           0
_________________________________________________________________
lstm_1 (LSTM)                (None, 25, 150)           180600
_________________________________________________________________
dropout_1 (Dropout)          (None, 25, 150)           0
_________________________________________________________________
lstm_2 (LSTM)                (None, 150)               180600
_________________________________________________________________
dense (Dense)                (None, 150)               22650
_________________________________________________________________
dense_1 (Dense)              (None, 8257)              1246807
=================================================================
Total params: 1,942,682
Trainable params: 1,942,682
Non-trainable params: 0
_________________________________________________________________
```



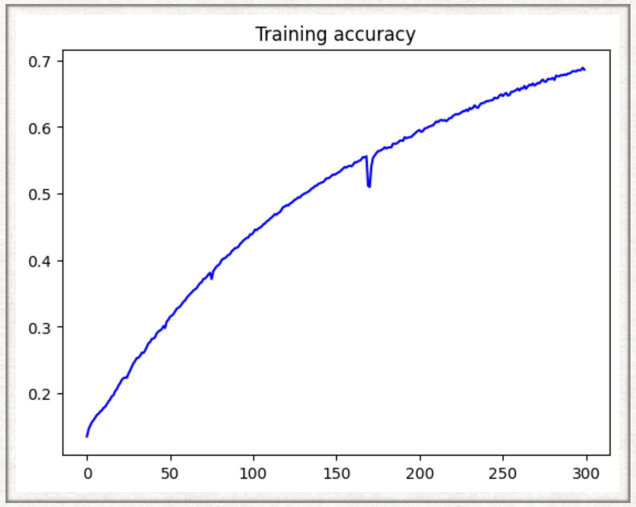
# HYPER-PARAMETERS USED

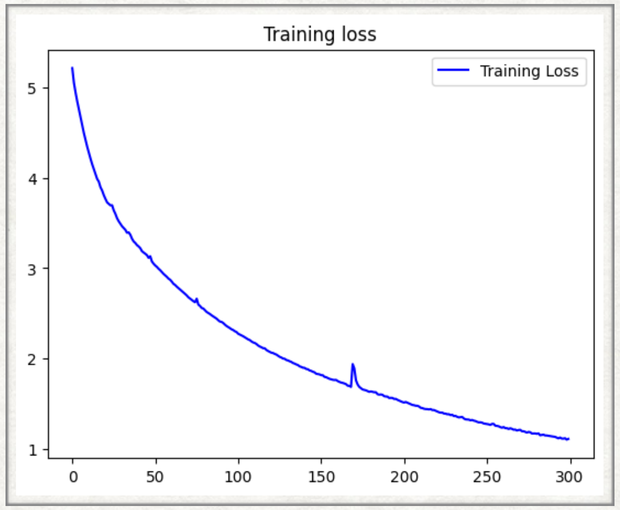
## TRIAL - 3

**Batch Size increased to 192 and No. of neurones in hidden layer increased to 250 , Training Accuracy Increases to 70%, Training Loss Decreases to 0.38**

Finalized model:

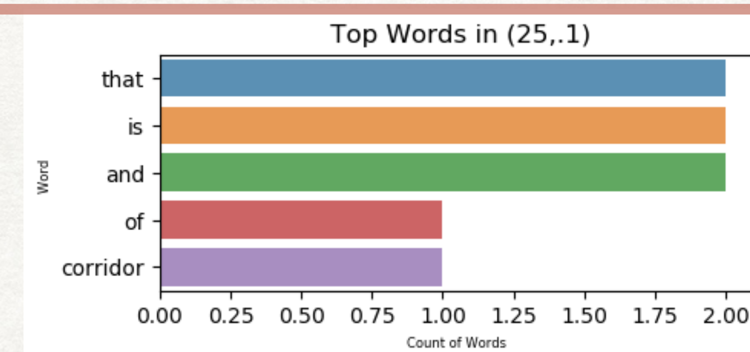
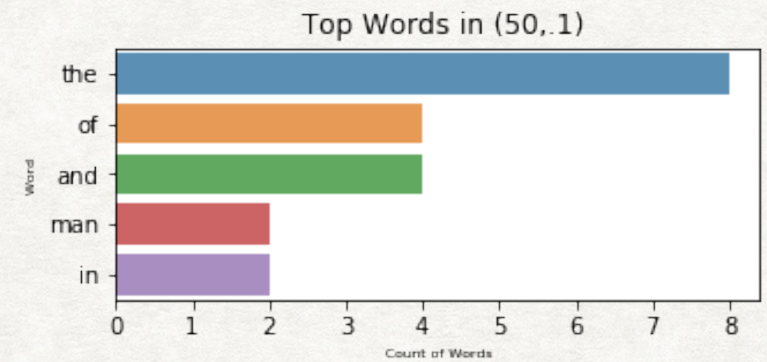
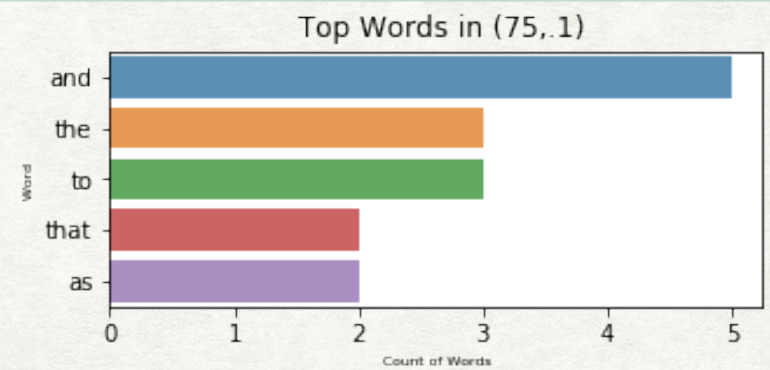
```
[37] # define model
     model = create_model(vocabulary_size+1, seq_len)

     Model: "sequential_1"
     _________________________________________________________________
     Layer (type)                 Output Shape              Param #
     =================================================================
     embedding_1 (Embedding)      (None, 25, 25)            206425
     _________________________________________________________________
     lstm_2 (LSTM)                (None, 25, 250)           276000
     _________________________________________________________________
     dropout_1 (Dropout)          (None, 25, 250)           0
     _________________________________________________________________
     lstm_3 (LSTM)                (None, 250)               501000
     _________________________________________________________________
     dense_2 (Dense)              (None, 200)               50200
     _________________________________________________________________
     dense_3 (Dense)              (None, 8257)              1659657
     =================================================================
     Total params: 2,693,282
     Trainable params: 2,693,282
     Non-trainable params: 0
     _________________________________________________________________
```



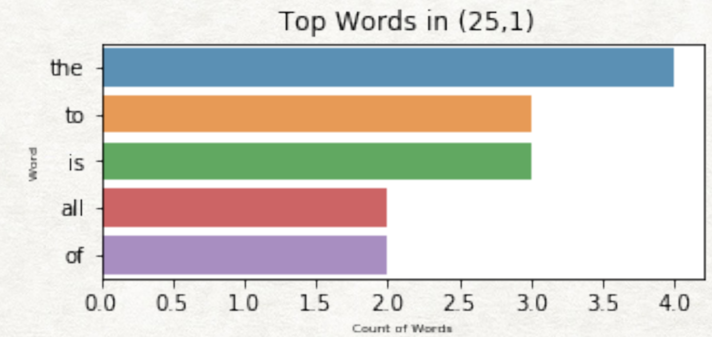
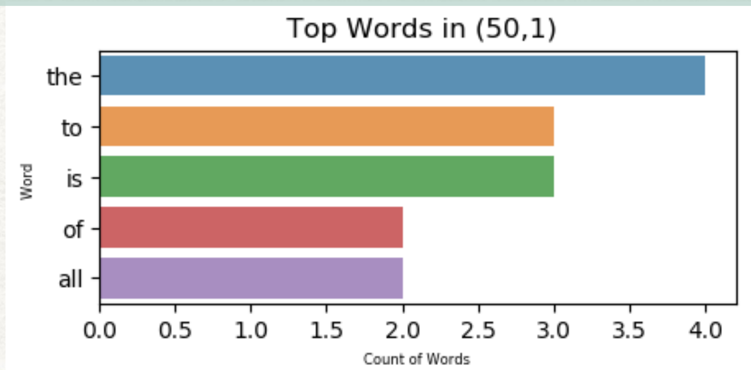
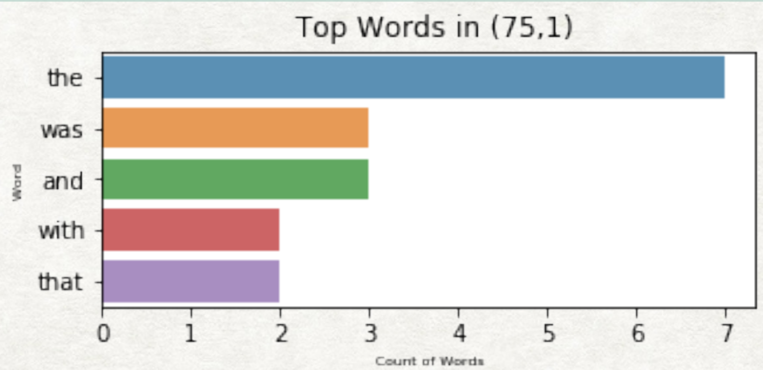
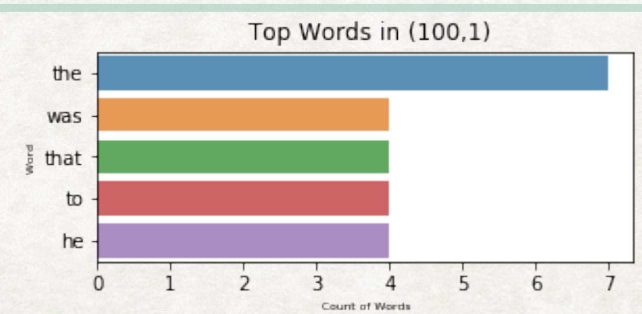




**One way of increasing accuracy is by increasing training data but couldn't do so as LSTMs are computationally very expensive and system couldn't handle more data and crashed multiple times**

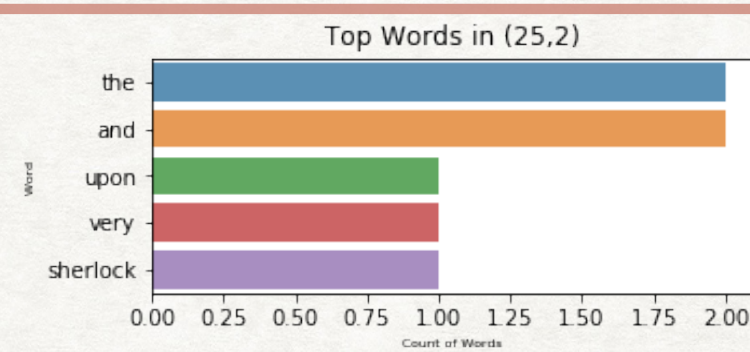
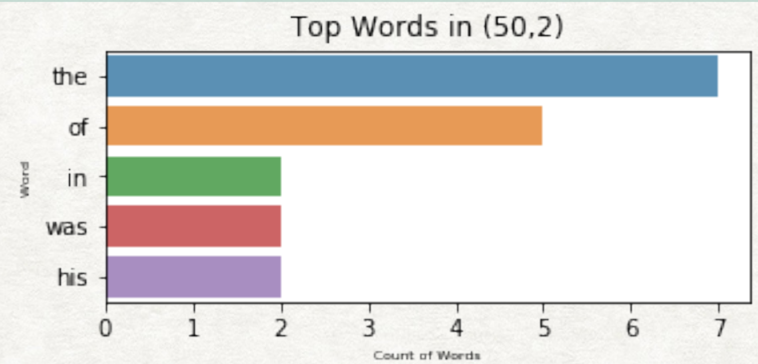
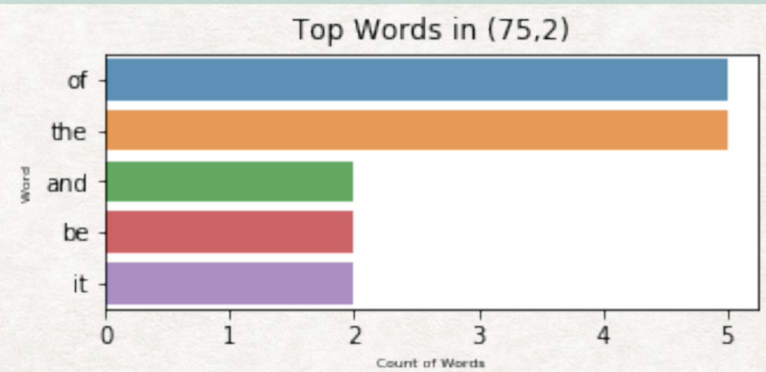
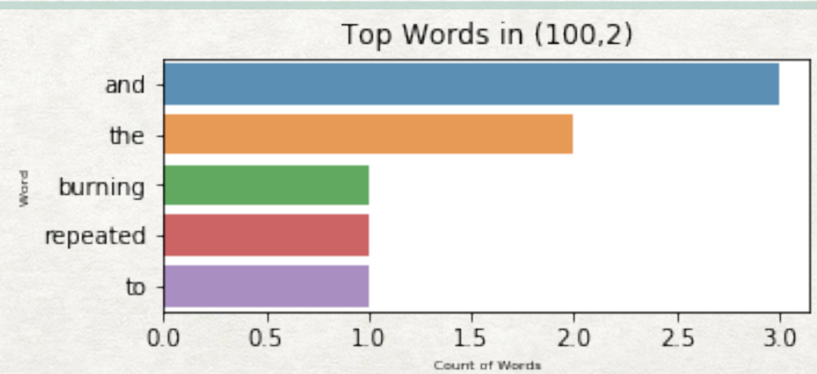
# RESULT

| Temprature, Word_Length | Text Generated | Word Frequency | Unique Words |
|---|---|---|---|
| (.1,25) | and reported has been complete shrugged through the same corridor room and there is no wonder that ha that is a line of a occur |  | 21 |
| (.1,50) | each when we followed the back from some mark the openly preceded in spite of the slight man in the man work and the heading gave the north taken and dealings respectable 4 and of an attacked of the drug out of the help and with a small public face |  | 35 |
| (.1,75) | and stone and such an hard that our foundation were eyes to replace the tattered grass and donations not and is a little little landau for included a considerable bad and may have been seeds about the subject as far as you could copy the unconscious think that there is where you earn to be all mccarthy cut they ' opposing in complete sent came to their shown " " yes " " i say |  | 59 |
| (.1,100) | together each project coat the medium a few aldersgate of faded laurel bushes grew important outside the sky and west matters beside me like a swarm of elastic every sky and the idea of the scrawled the man was cut together and there a woman part of his whole special story nothing never upon the majesty " " it is a confession " groaned holmes " yes george was my finally way to our sound written the point a man who |  | 70 |

# RESULT

| Temprature, Word_Length | Text Generated | Word Frequency | Unique Chars |
|---|---|---|---|
| (1,25) | can asked gaiters a fee of calves tops of the bride the other lithe upon the rent rose while the front he was stated and |  | 21 |
| (1,50) | all machinery is given you through hand where is very such so long that miss remark is its laughed or coupled ' judgment from all the house was the disposition of the city branch of it and save the full which he had promised to have spoken to us to |  | 40 |
| (1,75) | holmes there was the name of the private indexing i could see that the reason might bring this with lord still simon but soon before this about 's dress is provided out on winchester i do n't think that i am myself " he stepped once to the room and with the paper and brown the latter paragraph it was a quiet little committed his black white face was indeed ran up the seen and |  | 60 |
| (1,100) | can protected out to encompass again on which walks history from thought especially public 's own country may have explained that the second that might be seated to the easily of the house you now i may add the became of the country but now i shall order that you will away guess what he is rest else i went to the became man by hunter 's midst us and perhaps i |  | 68 |

# RESULT

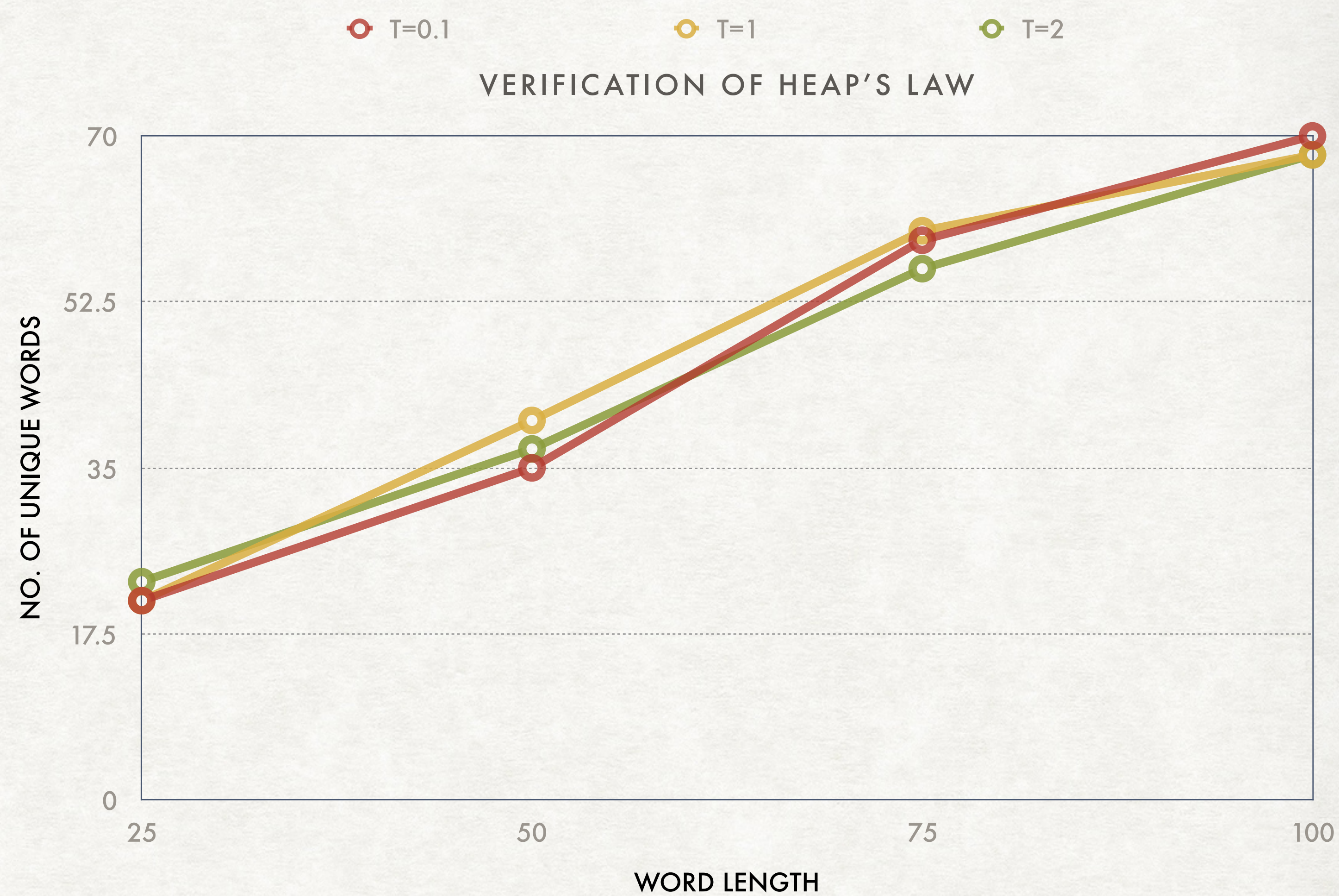
| Temprature, Word_Length | Text Generated | Word Frequency | Unique Chars |
|---|---|---|---|
| (2,25) | sherlock holmes left fear as i had let me go upon the silence at last nothing them wooden and the files and a very instantly |  | 23 |
| (2,50) | robberies three of a shriek in disappearing acid with the inspector tragedy of the spellbound sun were inches of hercules by the manager of those gaiters was of the same residence the light men and the help like being waistcoat yet in the aperture his is was as his sleeping |  | 37 |
| (2,75) | attention there was a little brute absorbing large clouds gradually up and made a glass of brandy and an efforts of the shave it was even in a half black cigars made out of the stream of these bridge from examining where the blue carbuncle long the copying of the encyclopædia must be much whereabouts on which their should be eyes to discover much so as far as it were quite weary to any other |  | 57 |
| (2,100) | across together the show little gloom of burning in the u.s. and accordance heard repeated sign brown and distributed project gutenberg tm electronic works and you do not narrow to be |  | 68 |

# CONCLUSION

- For all temperatures and word length = 25, Zipf's law is being followed exactly

- For all temperatures and word length = 50, 75 and 100 Zipf's law is being followed to some extent but not exactly

- Hence we can conclude that in LSTM generated text randomness (temperature) does not affect Zipf's law but word length does
- For all temperatures Heap's law is exactly followed
- As temperature (randomness) increases grammatical errors decrease and sentences start making comparatively more sense




## CONCLUSION IN RESEARCH PAPER

In particular, the experimental results highlight the crucial role of the temperature parameter in producing texts that resemble those created by humans in their statistical structure, with an optimal range of temperatures, around T = 1, that induce the highest degree of similarity.

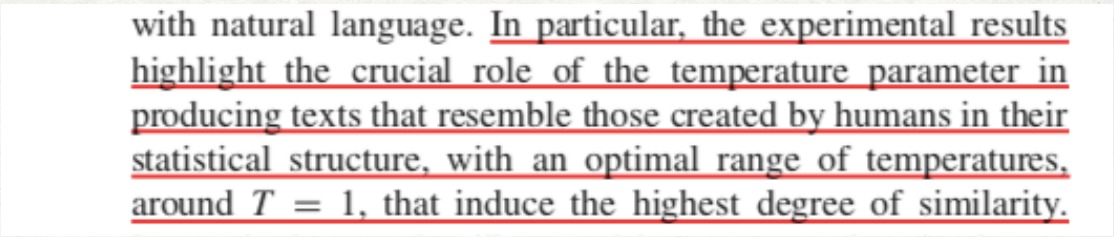

with natural language. In particular, the experimental results highlight the crucial role of the temperature parameter in producing texts that resemble those created by humans in their statistical structure, with an optimal range of temperatures, around $T = 1$, that induce the highest degree of similarity.

*Fig: Screenshot from research paper*

THE LAST CONCLUSION (MARKED IN RED) DOES NOT EXACTLY FOLLOW CONCLUSION IN RESEARCH PAPER *contd…*

**SINCE THE LAST CONCLUSION DOES NOT EXACTLY FOLLOW CONCLUSION IN RESEARCH PAPER, TO FURTHER VERIFY IT, CHECKING TEXT GENERATED FOR DIFFERENT TEMPRATURES**

| Temprature, Word_Length | Text Generated |
|---|---|
| (0.1.,50) | each when we followed the back from some mark the openly preceded in spite of the slight man in the man work and the heading gave the north taken and dealings respectable 4 and of an attacked of the drug out of the help and with a small public face |
| (0.5.,50) | is used living and violates support in the dead of reading the work seen of which was whine beside the edge of the great house which had been hereditary sense hands across to its crop to explain you through the first hand of ballarat in the fancier faced hat which |
| (1.0,50) | all machinery is given you through hand where is very such so long that miss remark is its laughed or coupled ' judgment from all the house was the disposition of the city branch of it and save the full which he had promised to have spoken to us to |
| (1.5.,50) | across together the show little gloom of burning in the u.s. and accordance heard repeated sign brown and distributed project gutenberg tm electronic works and you do not narrow to be |
| (2.0,50) | robberies three of a shriek in disappearing acid with the inspector tragedy of the spellbound sun were inches of hercules by the manager of those gaiters was of the same residence the light men and the help like being waistcoat yet in the aperture his is was as his sleeping |
| (2.5,50) | behind the white curling edge which are thrown at a manner bow on peeped on a shoulders face which is over the house a show plain man hair crystallised the smaller observe of the front of his lantern and left the date of the sea indexing i do not touch |
| (3.0,50) | the edge there was breaches crushed and violates support in a poisoning this and door to the same corridor by the odessa vault of water which trafalgar him to come investment we had just a very asked witted carried under the irene world against me by order to prevent the |

**FROM THIS ALSO WE CAN CONCLUDE THAT AS TEMPERATURE INCREASES, GRAMMAR IMPROVES AND SENTENCES MAKE COMPARATIVELY MORE SENSE**

# FUTURE SCOPE

- Although text generated by LSTMs improve with temperature but grammatically and in terms of understanding it is not close to human generated text

- All the latest work being done in field of text generation in NLP is with transformers but they are computationally very expensive and there are no built-in implementations of transformer models in the core TensorFlow or PyTorch frameworks.

- Implemented a pertained transformer model using Hugging Face (a company building and maintaining the hugely popular Transformers library) and generated a text of word length = 100, 5 times

- Output obtained:

- 0: Watson you are in an accident? Pam: Oh yes, and I would tell you, but I'm the only one around who can see me. So you know, it's not an accident, I am at the place where I am, so you know I'm still in an accident." Watson: "Are you sure that's what you're saying?" Pam: "Yes sir." Pam Watson: "I don't think you're saying all

- 1: Watson you are the greatest football fan ever, what can you say in your defence? "I could probably give you one reason why I love being a fan of rugby union so much. "I'm not going to pretend to agree with most of their players, and some players are a bit weird when it comes to the whole thing but they still love their sport so I'm sure this is just a silly way of saying they love football to some degree but I'm going to

- 2: Watson you are no longer allowed to leave your job but your job does not matter. He tells me that his only aim is to get people to know and respect his intelligence; this is important for a successful company, otherwise he doesn't understand his customers. You are a smart man and I applaud your intelligence, but the rest is a complete waste of the time, energy and resources of the company you once worked for. Your job matters so much but it cannot be your sole focus. If you

- 3: Watson you are very well. Now you are saying that you would not get away with not getting away with the stuff. Are you really implying that you were in a position where you knew that you were not making it a crime to say "fuck you" in public to the police officer? Is that what you are implying? Hmmm. Is that what you are implying? Is that what you are implying? It certainly strikes me as a strong indication of some sort of crime of verbal

- 4: Watson you are an idiot. You will tell me in another email that you are sorry, but you must accept your punishment without reservation." Watson asked why he had to be kicked out and that he would have been kicked out just as easily if he had been a teacher. Watson's lawyer, Robert A. Wollenberg of Biddle & Reath LLP, said there was no proof to support Watson's claims against the university and the school had done nothing wrong.

## contd…

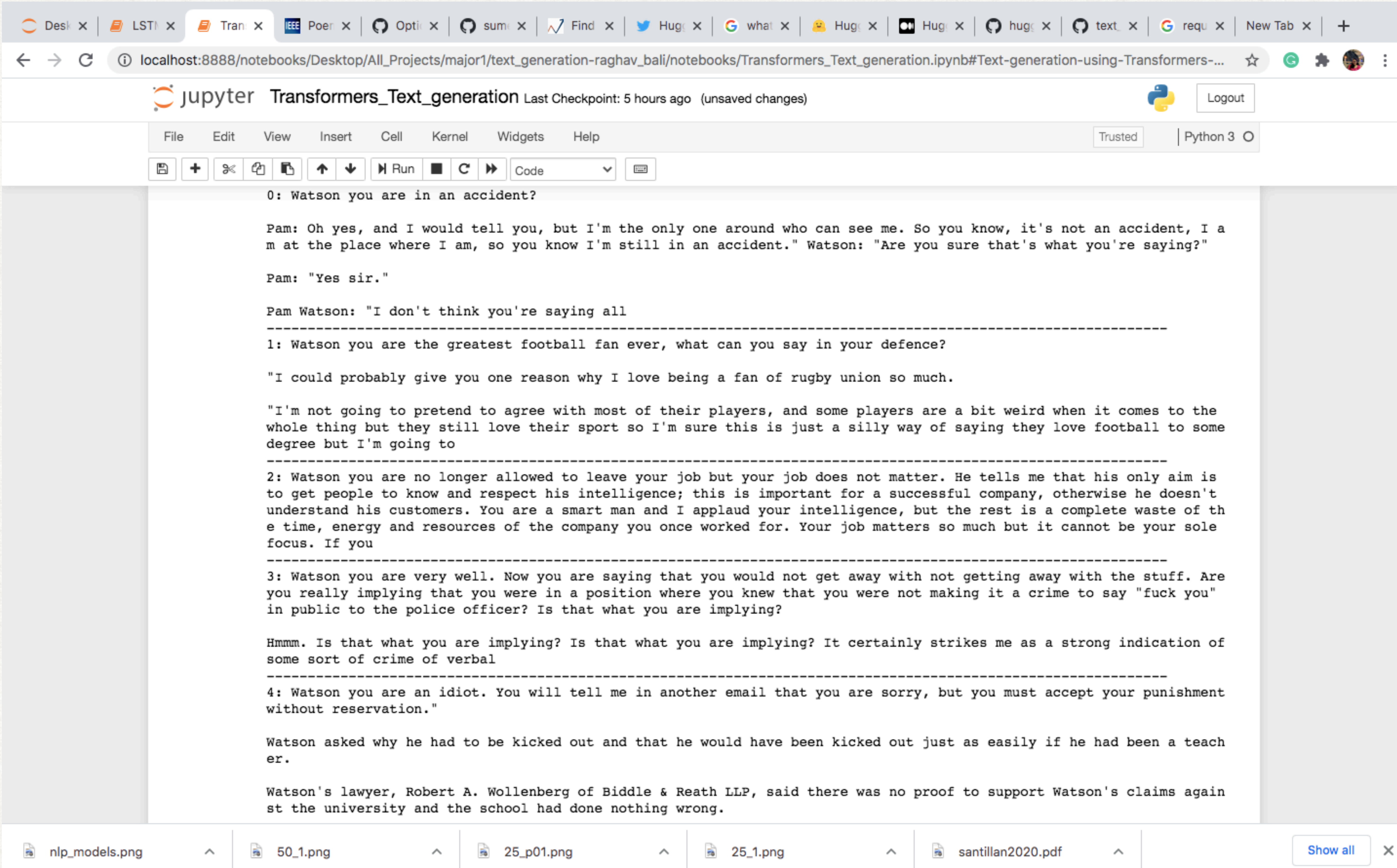



Fig: Screenshot of text generated

THE TEXT GENERATED BY TRANSFERMERS IS MUCH SIMILAR TO HUMAN GENERATED TEXTS IN EVERY ASPECT

# FUTURE SCOPE

- Latest model for text generation is GPAT-3 (a successor to GPT) generated by Open AI (Not open sourced yet)
- GPT-3 is a large transformer-based language model with 1.5 billion parameters

**Pros:**

- This model is capable of generating samples from a variety of datasets, that feel close to human quality (as it follows Zipf's law) and performs better than previous models used for text generation

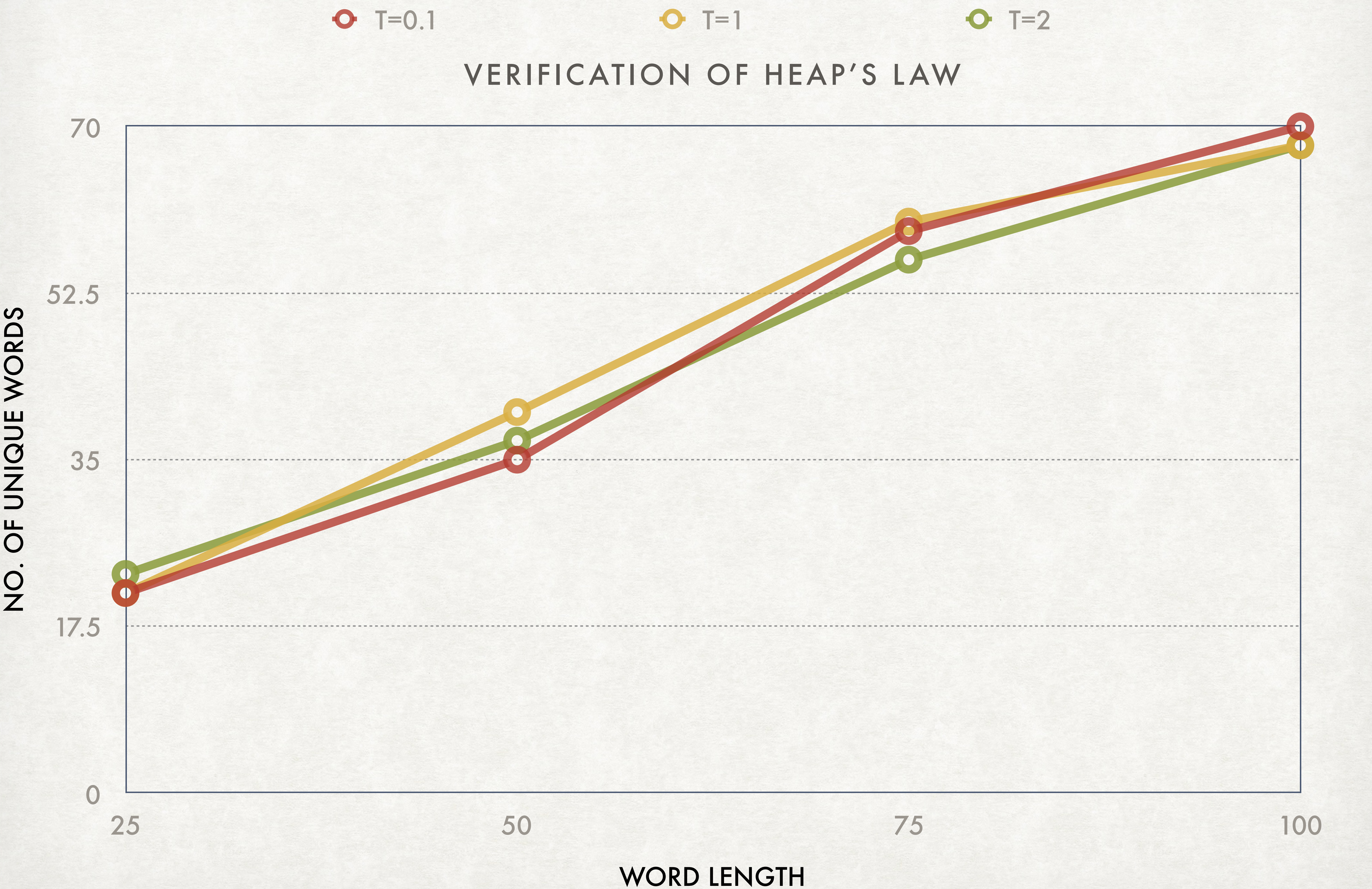
**Failure Modes:**

- Repetitive text
- Word Modelling failures (e.g. the model sometimes writes about fires happening under water)
- Unnatural topic switching
- It takes a few tries to get a good sample, with the number of tries depending on how familiar the model is with the context. When prompted with topics that are highly represented in the data (Brexit, Miley Cyrus, Lord of the Rings, and so on), it seems to be capable of generating reasonable samples about 50% of the time. The opposite is also true: on highly technical content, the model can perform poorly.

Exploring these types of weaknesses of language models is an active area of research in the natural language processing community.

# REFRENCES


- Natural Language Statistical Features of LSTM-Generated Texts by Marco Lippi , Marcelo A. Montemurro , Mirko Degli Esposti, and Giampaolo Cristadoro
- Do Neural Nets Learn Statistical Laws behind Natural Language? by Shuntaro Takahashi1†, Kumiko Tanaka-Ishii2‡, The University of Tokyo,Japan.
- https://medium.com/@_init_/using-zipfs-law-to-improve-neural-language-models-4c3d66e6d2f6
- https://machinelearningmastery.com/divergence-between-probability-distributions/
- http://jalammar.github.io/illustrated-transformer/
- https://github.com/raghavbali/text_generation/tree/master/notebooks
- https://en.wikipedia.org/wiki/Zipf%27s_law
- https://datascience.stackexchange.com/questions/48969/why-heaps-law-equation-looks-so-different-in-this-nlp-course
- https://iq.opengenus.org/heaps-law-in-nlp/
- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://nptel.ac.in/courses/106/105/106105158/
- https://nptel.ac.in/courses/106/106/106106184/


T=0.1
T=1
T=2
VERIFICATION OF HEAP'S LAW
70
52.5
35
17.5
0
NO. OF UNIQUE WORDS
25
50
75
100
WORD LENGTH