Text Generation for The Big Bang Theory

Raghav Raj Mittal, Austin Jiang, Satwik Mekala

Motivation / Introduction

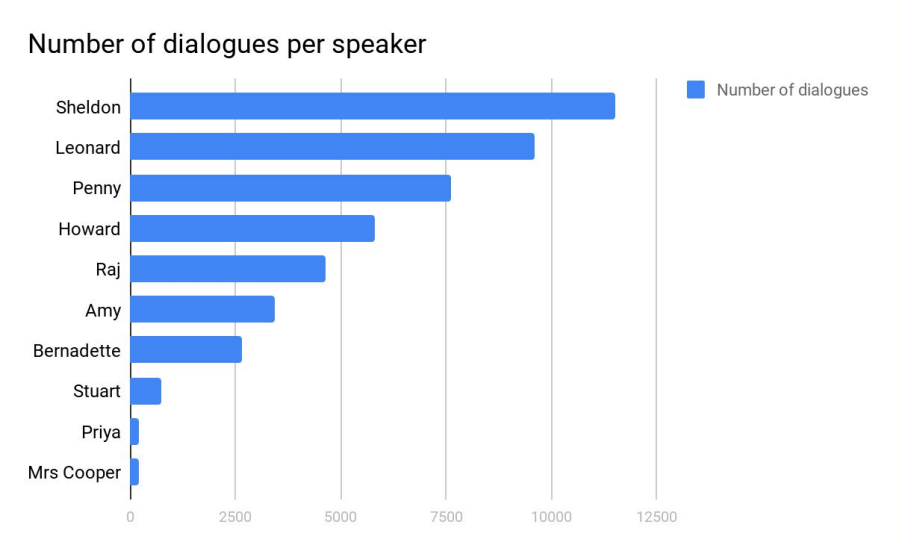
Our objective is to create a model that will be able to generate scripts for the TV sitcom ‘The Big Bang Theory’.

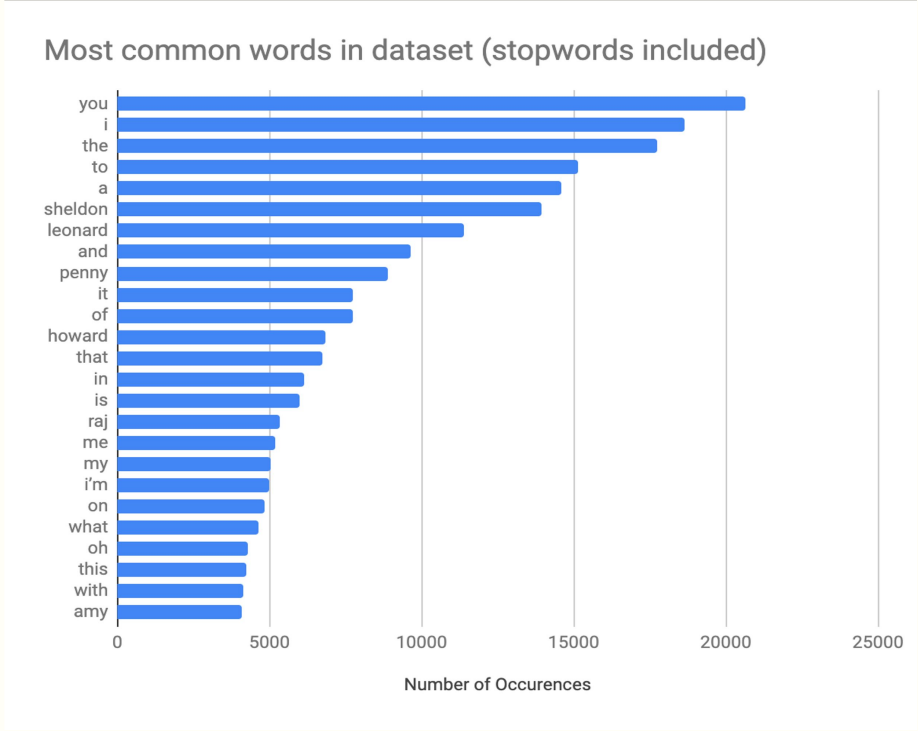
In the entertainment industry, directors and writers are being pushed to produce original ideas and new content. This is the most time consuming portion of a production is the writing process. Script generators help provide a starting point, thereby increasing efficiency and the quality of writing.

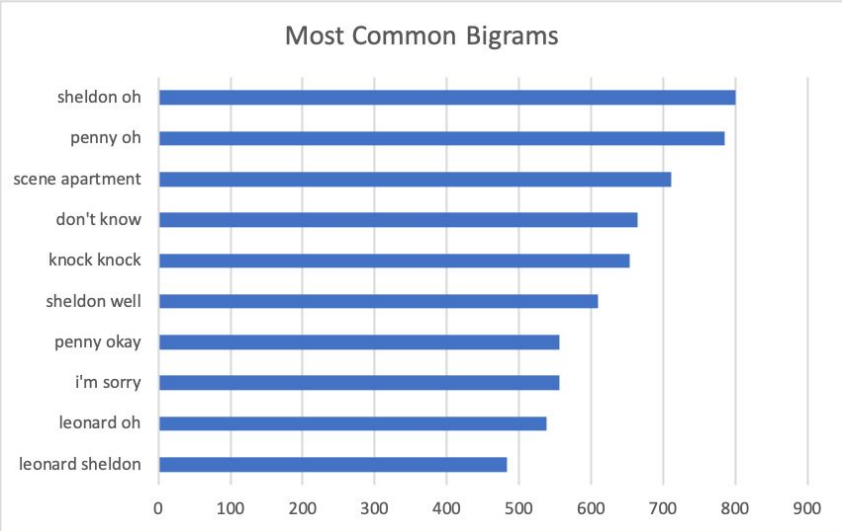
Data

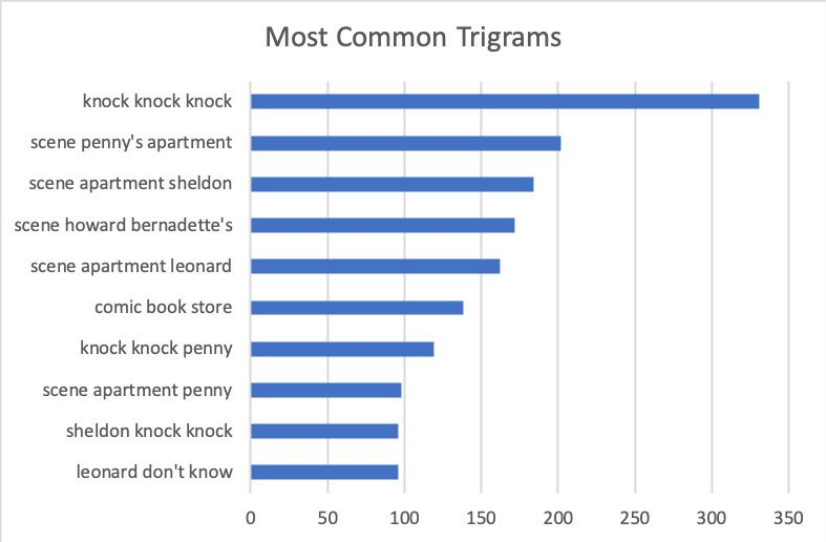
Scripts were scraped from a public website using BeautifulSoup!

- 10 seasons
- 229 episodes
- 53997 lines
- 2847 different scenes
- 22709 unique words


Number of dialogues per speaker
Number of dialogues
Sheldon
Leonard
Penny
Howard
Raj
Amy
Bernadette
Stuart
Priya
Mrs Cooper
0
2500
5000
7500
10000
12500

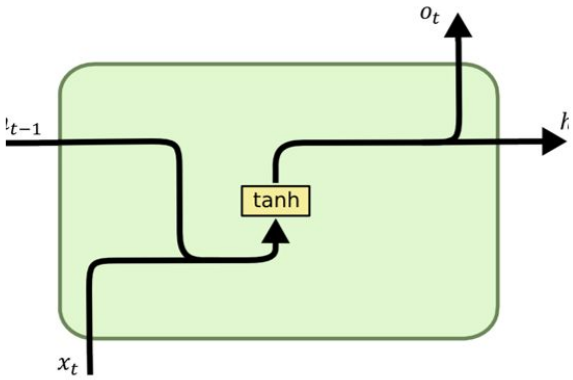

Most common words in dataset (stopwords included)
you
i
the
to
a
sheldon
leonard
and
penny
it
of
howard
that
in
is
raj
me
my
i'm
on
what
oh
this
with
amy
0
5000
10000
15000
20000
25000
Number of Occurences


Most Common Bigrams
sheldon oh
penny oh
scene apartment
don't know
knock knock
sheldon well
penny okay
i'm sorry
leonard oh
leonard sheldon
0
100
200
300
400
500
600
700
800
900

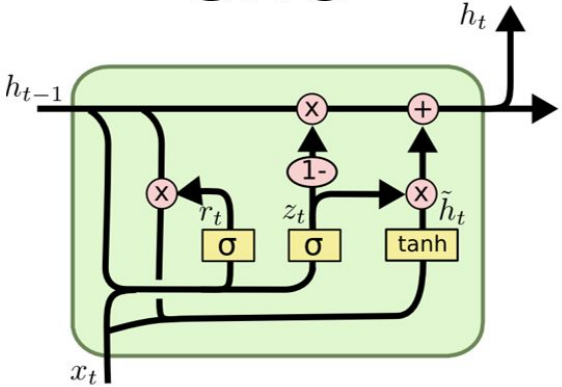

Most Common Trigrams
knock knock knock
scene penny's apartment
scene apartment sheldon
scene howard bernadette's
scene apartment leonard
comic book store
knock knock penny
scene apartment penny
sheldon knock knock
leonard don't know
0
50
100
150
200
250
300
350

Approach

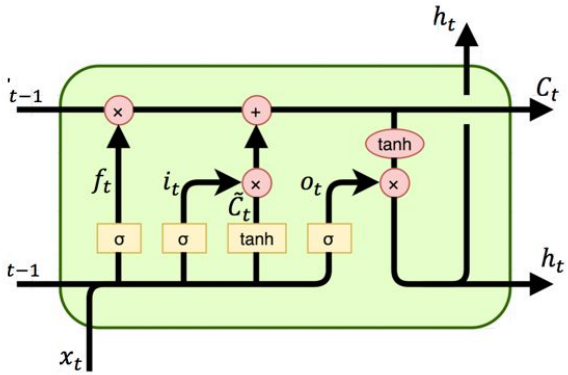
RNN



GRU



LSTM



Experimental Setup / Evaluation

Model training and text generation done using Google Colab and PyTorch 1.4.

Model training:

- GPU setting - NVIDIA Tesla K80 GPU.
- GPU usage is 1GB and RAM usage is 2.46GB
- 20 minutes

Text Generation:

- CPU setting
- 4-5 seconds

Evaluation: BLEU and Self-BLEU

Results (Baselines + Approach)

Model	BLEU	Self-BLEU
Vanilla RNN (Seq length 3)	0.746	0.645
Vanilla RNN (Seq length 5)	0.756	0.660
Vanilla RNN (Seq length 6)	0.703	0.748
Vanilla RNN (Seq length 7)	**0.800**	0.700
GRU (Seq length 3)	0.758	0.715
GRU (Seq length 5)	0.678	0.611
GRU (Seq length 6)	0.579	0.632
GRU (Seq length 7)	0.700	0.634
LSTM (Seq length 3)	0.698	0.648
LSTM (Seq length 5)	0.724	0.589
LSTM (Seq length 6)	0.659	**0.467**
LSTM (Seq length 7)	0.702	0.572
Word2Vec + Vanilla RNN	0.644	0.739
Word2Vec + GRU	0.633	0.633
Word2Vec + LSTM	0.672	0.670

Example of Generated Text

scene: the apartment.
sheldon: i’m sorry. i just wanted to talk to him.
leonard: oh, i can’t tell you what?
amy: i don’t know what you meant, i have no idea about it.
penny: okay. i guess i was thinking of your own culture and the pope who has crossed his own desk?
howard: i don’t have coffee.
sheldon(on video): hey, i don’t know why i’m on a little.

Future Work

- Vary the initializations of the word-to-embedding matrix with different word embeddings, including Glove and Elmo

- Train on limited data

- Generate text given a sentiment

Thank you!