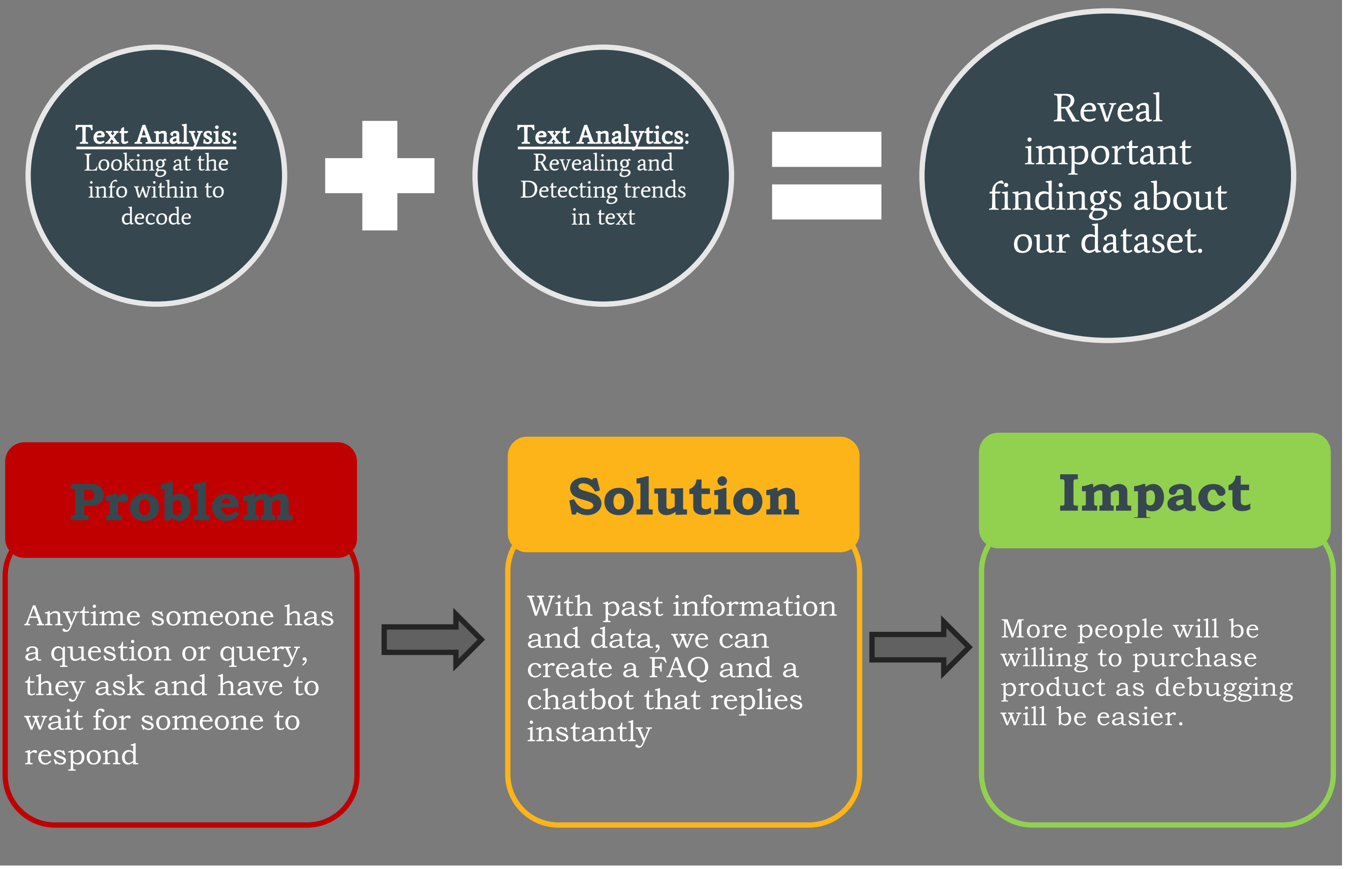
UBUNTU - A TEXT ANALYSIS EXPLORATION

BY: ABBAS PARDAWALLA

OUR GOAL

Text Analysis that will create and embed chatbot to assist users debug their queries .

In order for us to achieve such a task, first we have to:



Text Analysis: Looking at the info within to decode

+

Text Analytics: Revealing and Detecting trends in text

=

Reveal important findings about our dataset.

Problem

Anytime someone has a question or query, they ask and have to wait for someone to respond

Solution

With past information and data, we can create a FAQ and a chatbot that replies instantly

Impact

More people will be willing to purchase product as debugging will be easier.

Overview of Dataset & Pre-processing

Our Dataset is made up of 6 Main columns

1.Folder:
The folder the query was retrieved from

2.DialogueID:
ID that a set of text corpus is part of

3. Date:
Timestamp when query was submitted

4. From:
The user who sent the query

5. To:
The user whom the answer is for

6. Text - Our heart
The line of text that is going to help our analysis. The question and response

Initial Dataset

1,038,325 Lines

11,035,331 Words

116,070,597 Characters

- Removed1630 Duplicate rows
- Filled in 87 empty cells in text column
- Removed Stop words , Punctuation, Spaces
- Removed Digits
- Added Parts of Speech
- Lemmatized the dataset and
- Made all our letters into lower cases

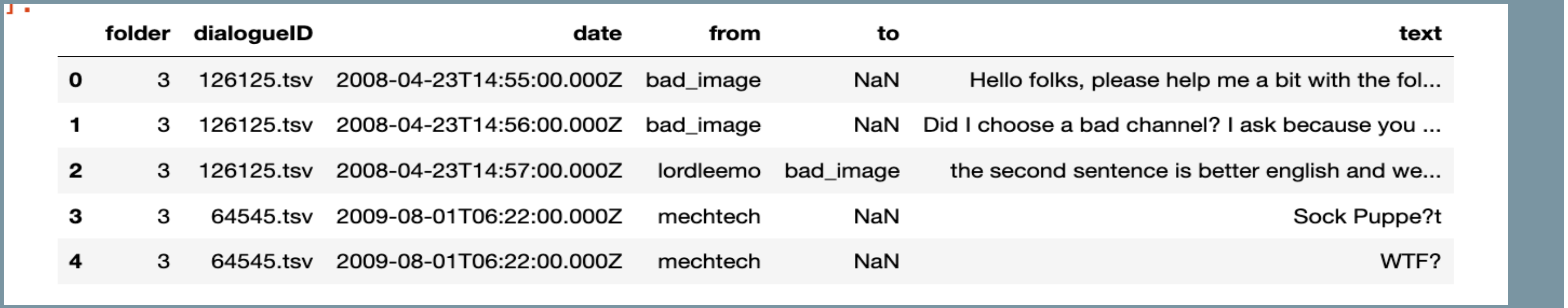
Final Dataset

1,038,237 Lines

6,101,882 Words

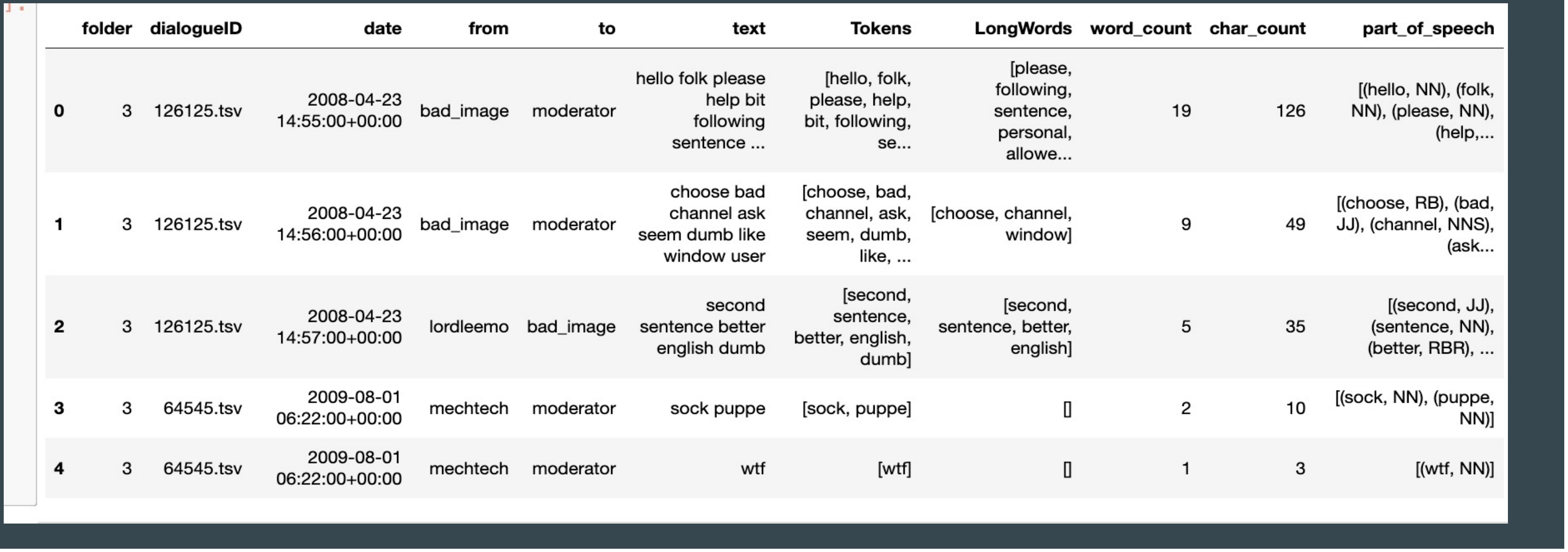
37,451,381 Characters

BEFORE



	folder	dialogueID	date	from	to	text
0	3	126125.tsv	2008-04-23T14:55:00.000Z	bad_image	NaN	Hello folks, please help me a bit with the fol...
1	3	126125.tsv	2008-04-23T14:56:00.000Z	bad_image	NaN	Did I choose a bad channel? I ask because you ...
2	3	126125.tsv	2008-04-23T14:57:00.000Z	lordleemo	bad_image	the second sentence is better english and we...
3	3	64545.tsv	2009-08-01T06:22:00.000Z	mechtech	NaN	Sock Puppe?t
4	3	64545.tsv	2009-08-01T06:22:00.000Z	mechtech	NaN	WTF?

AFTER



	folder	dialogueID	date	from	to	text	Tokens	LongWords	word_count	char_count	part_of_speech
0	3	126125.tsv	2008-04-23 14:55:00+00:00	bad_image	moderator	hello folk please help bit following sentence ...	[hello, folk, please, help, bit, following, se...	[please, following, sentence, personal, allowe...	19	126	[(hello, NN), (folk, NN), (please, NN), (help,...
1	3	126125.tsv	2008-04-23 14:56:00+00:00	bad_image	moderator	choose bad channel ask seem dumb like window user	[choose, bad, channel, ask, seem, dumb, like, ...	[choose, channel, window]	9	49	[(choose, RB), (bad, JJ), (channel, NNS), (ask...
2	3	126125.tsv	2008-04-23 14:57:00+00:00	lordleemo	bad_image	second sentence better english dumb	[second, sentence, better, english, dumb]	[second, sentence, better, english]	5	35	[(second, JJ), (sentence, NN), (better, RBR), ...
3	3	64545.tsv	2009-08-01 06:22:00+00:00	mechtech	moderator	sock puppe	[sock, puppe]	[]	2	10	[(sock, NN), (puppe, NN)]
4	3	64545.tsv	2009-08-01 06:22:00+00:00	mechtech	moderator	wtf	[wtf]	[]	1	3	[(wtf, NN)]

Key Insights

Crucial Findings Important for Modeling and Implementation

Computational Power
Compulsory asset for the model to predict correctly

More Data > Accuracy
The more data we are able to feed into our model - the more accurately we will be able to run it.

Emphasis on ReadMe
The most popular topics consists of people asking how to install software

Expertise On-site
Currently there are a lot more questions being asked and not enough answers being provided

Model Comparison and Interpretation

Our models consists of **Pre-Trained** models built with **Unsupervised** Learning.

BERT

Bidirectional Encoder Representations from Transformer (BERT) - Transformer learning, Epoch loss and Entropy

Latent Dirichlet Allocation (LDA)

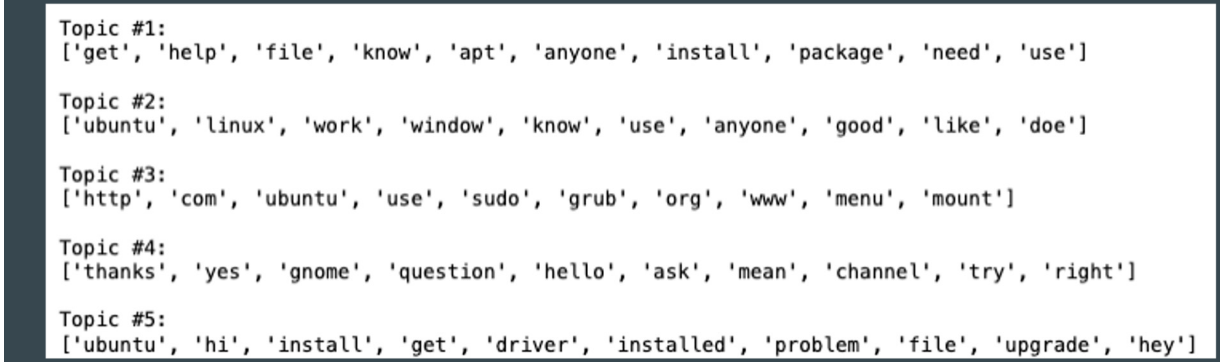
Topic Modelling via Statistics. Document is topics and topics is words

Spacy

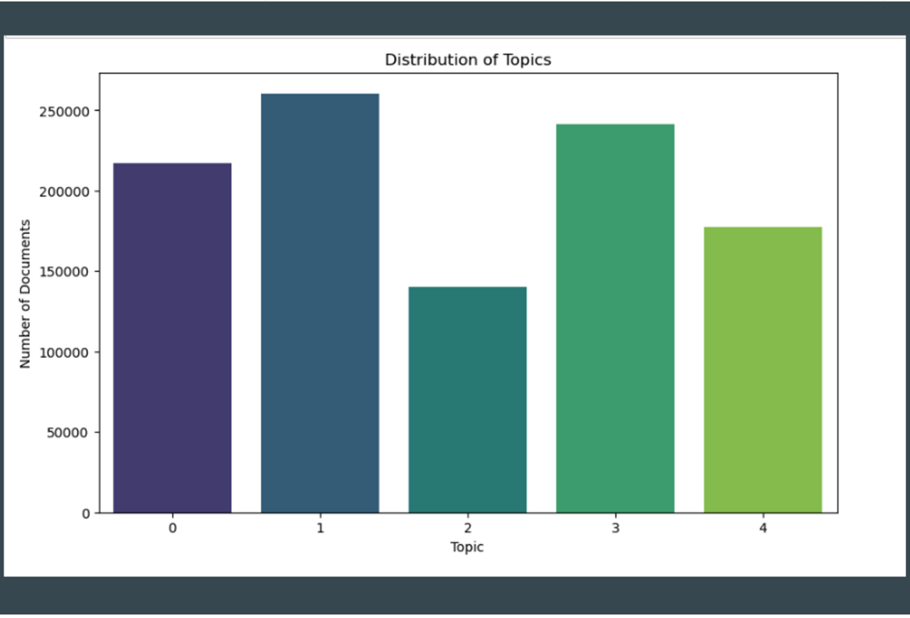
Open Source library to help and understand large volumes of text

LDA

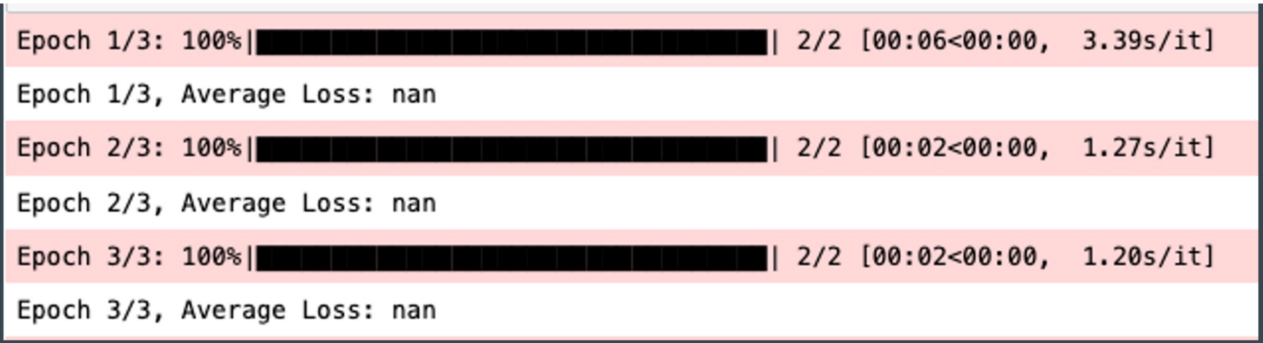
Break down our text into Topics and then visualize the usage of those topics over the corpus



```
Topic #1:
['get', 'help', 'file', 'know', 'apt', 'anyone', 'install', 'package', 'need', 'use']

Topic #2:
['ubuntu', 'linux', 'work', 'window', 'know', 'use', 'anyone', 'good', 'like', 'doe']

Topic #3:
['http', 'com', 'ubuntu', 'use', 'sudo', 'grub', 'org', 'www', 'menu', 'mount']

Topic #4:
['thanks', 'yes', 'gnome', 'question', 'hello', 'ask', 'mean', 'channel', 'try', 'right']

Topic #5:
['ubuntu', 'hi', 'install', 'get', 'driver', 'installed', 'problem', 'file', 'upgrade', 'hey']
```



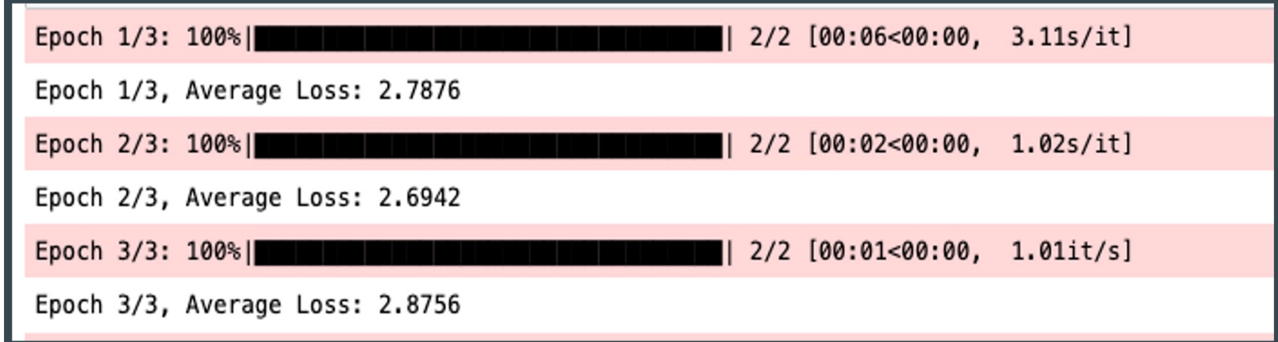
BERT



```
Epoch 1/3: 100%|██████████████████████████████| 2/2 [00:06<00:00,  3.39s/it]
Epoch 1/3, Average Loss: nan
Epoch 2/3: 100%|██████████████████████████████| 2/2 [00:02<00:00,  1.27s/it]
Epoch 2/3, Average Loss: nan
Epoch 3/3: 100%|██████████████████████████████| 2/2 [00:02<00:00,  1.20s/it]
Epoch 3/3, Average Loss: nan
```

Sample Size 10

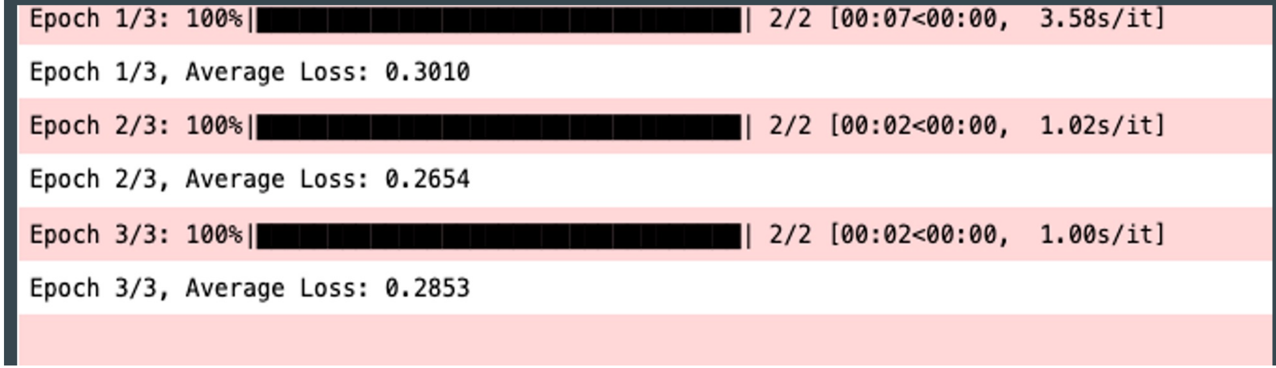
No Loss



```
Epoch 1/3: 100%|██████████████████████████████| 2/2 [00:06<00:00,  3.11s/it]
Epoch 1/3, Average Loss: 2.7876
Epoch 2/3: 100%|██████████████████████████████| 2/2 [00:02<00:00,  1.02s/it]
Epoch 2/3, Average Loss: 2.6942
Epoch 3/3: 100%|██████████████████████████████| 2/2 [00:01<00:00,  1.01it/s]
Epoch 3/3, Average Loss: 2.8756
```

Sample Size 100

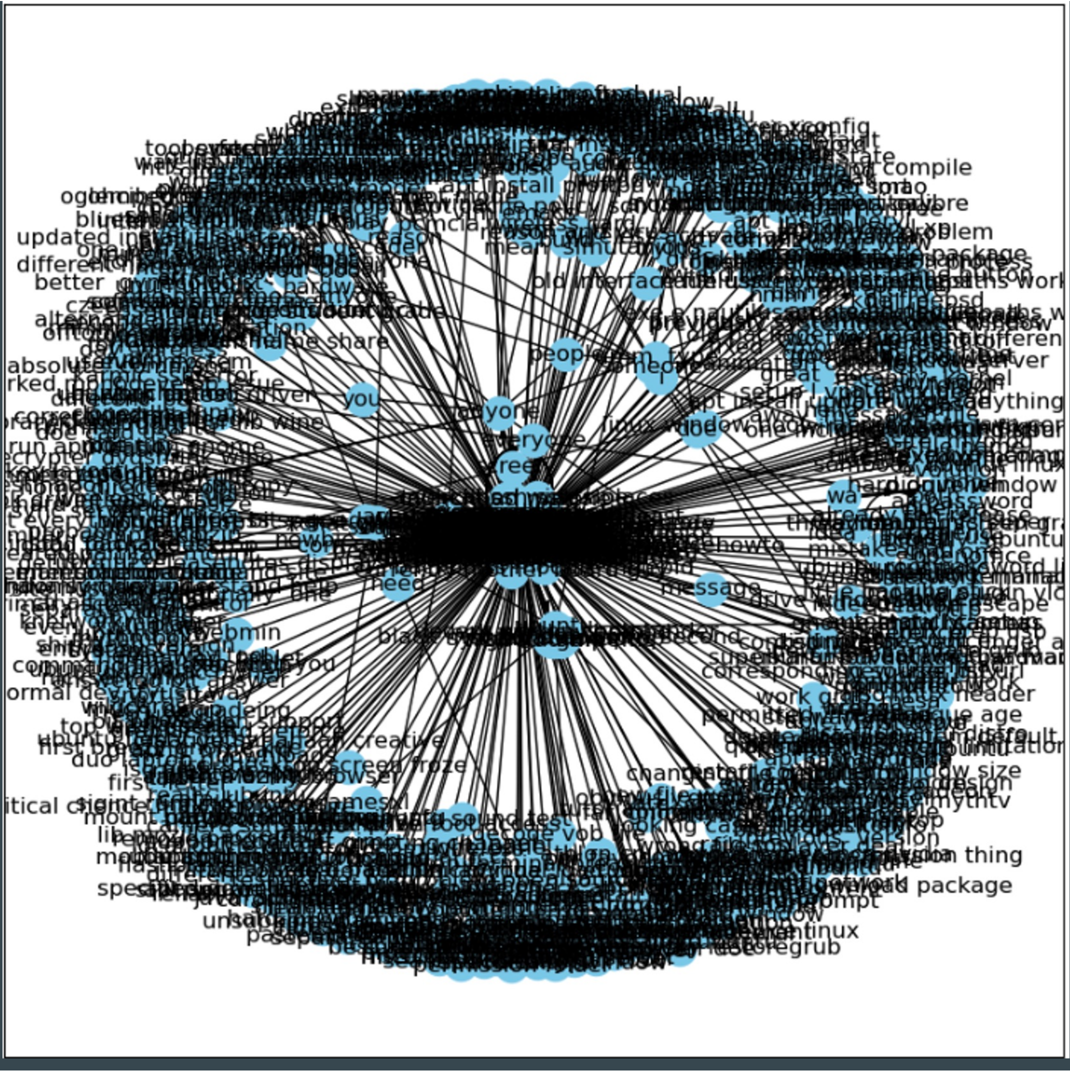
An increasing loss rate, but still minimal



```
Epoch 1/3: 100%|██████████████████████████████| 2/2 [00:07<00:00,  3.58s/it]
Epoch 1/3, Average Loss: 0.3010
Epoch 2/3: 100%|██████████████████████████████| 2/2 [00:02<00:00,  1.02s/it]
Epoch 2/3, Average Loss: 0.2654
Epoch 3/3: 100%|██████████████████████████████| 2/2 [00:02<00:00,  1.00s/it]
Epoch 3/3, Average Loss: 0.2853
```

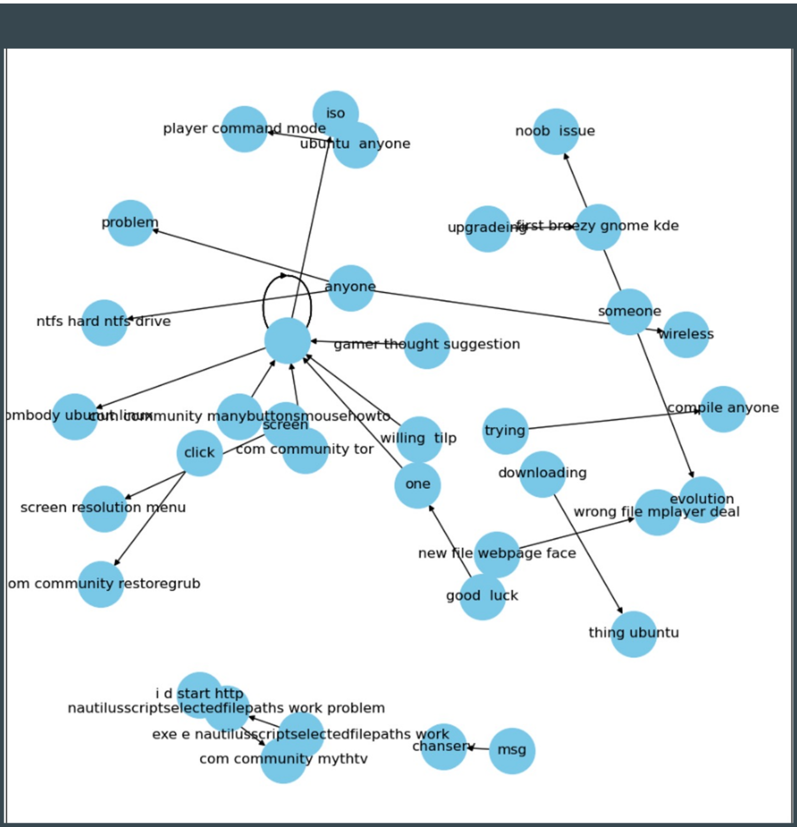
Sample Size 1000

Close to zero loss, but increasing over each epoch

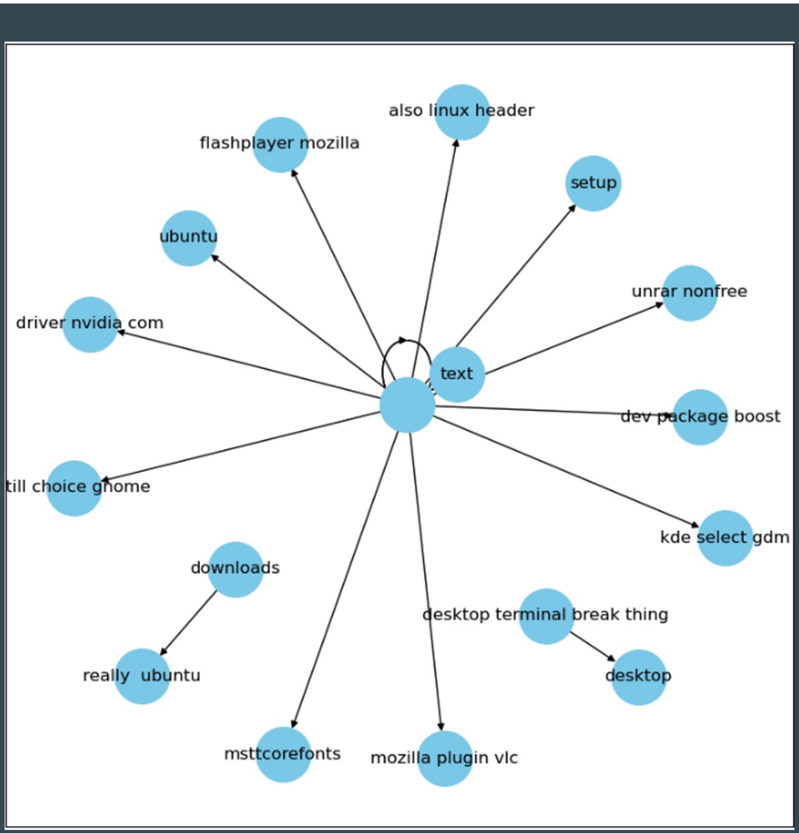


SPACY

All our topics combined and connected to each other



Connections for Help



Connections for Install

Product Design & Next Steps

- Now that we have trained our dataset with the different models (LDA, BERT & Spacy)
- Next step will be to use more of the dataset that we have to its full capacity to be able to extract more information.
- Currently our model used around 1000 rows out of a corpus of millions rows. Will need to run our model on the full set to proceed.

Based on the linkage and implementation of the corpus, we will then be able to do the following steps:

FAQs

Create a FAQ of the top 20 questions asked by users

GENERATIVE TEXT

Use generative text to predict what a user will type - and if predicted correct, link them to the desired question

CHATBOT

Create a rule based chatbot to answer the most common questions.

SPAM

Remove spam and trolling comments