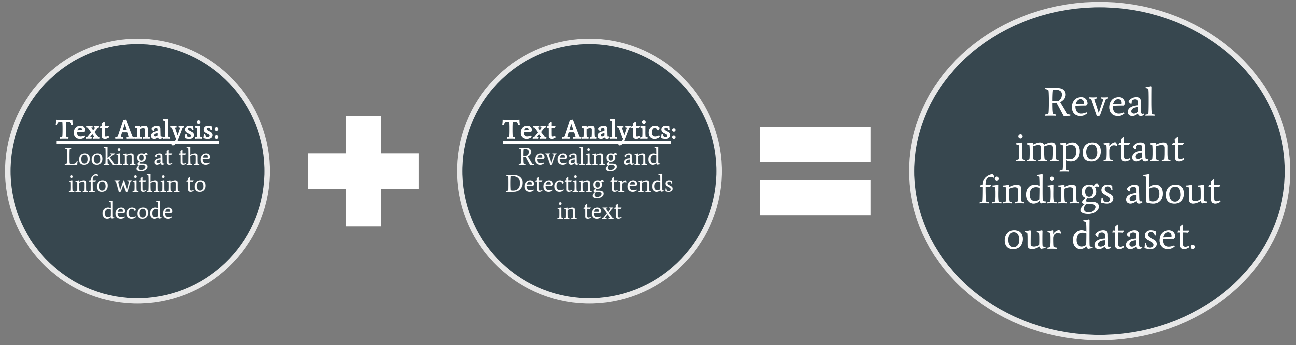
UBUNTU - A TEXT ANALYSIS EXPLORATION

BY: ABBAS PARDAWALLA

OUR GOAL

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

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



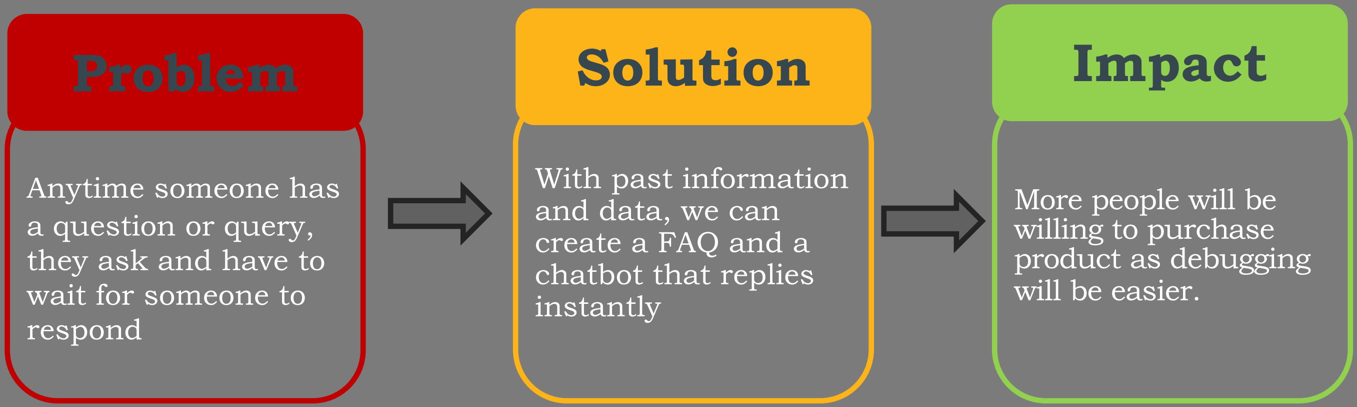
Text Analysis: Looking at the info within to decode

+

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

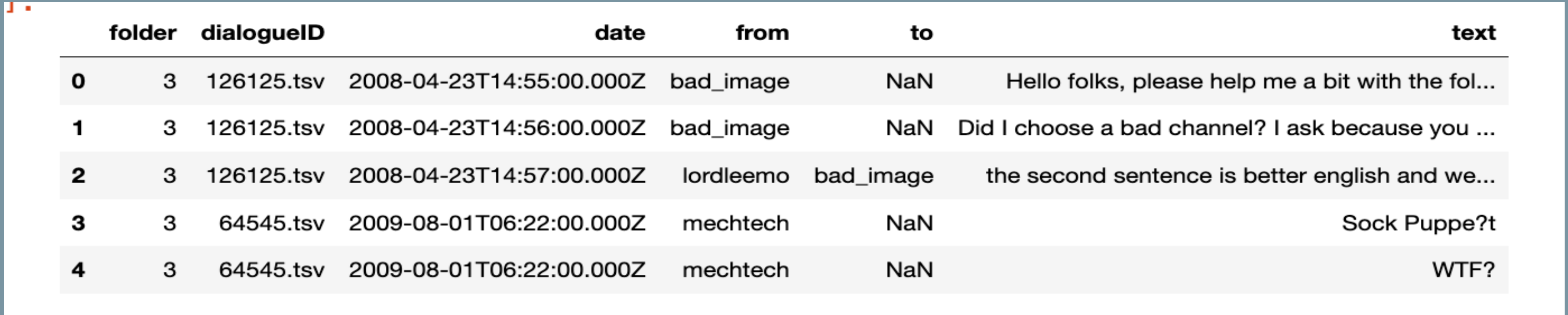
Final Dataset

1,038,237 Lines

6,101,882 Words

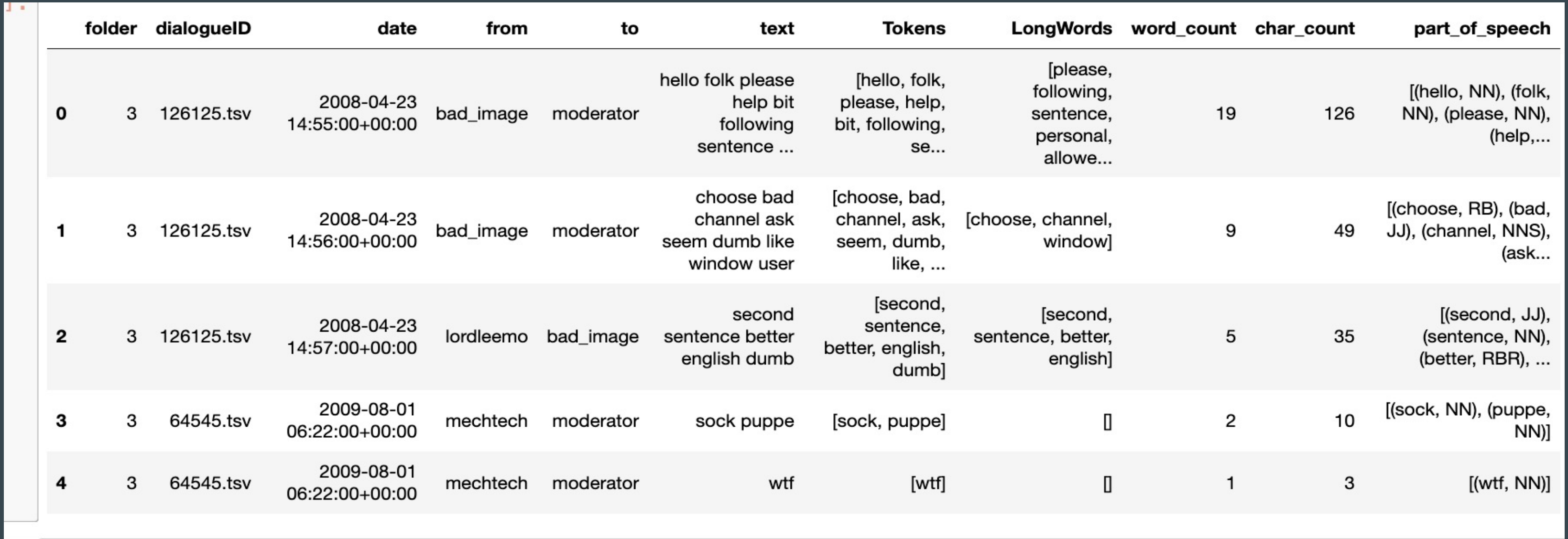
37,451,381 Characters

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)]

EDA Findings

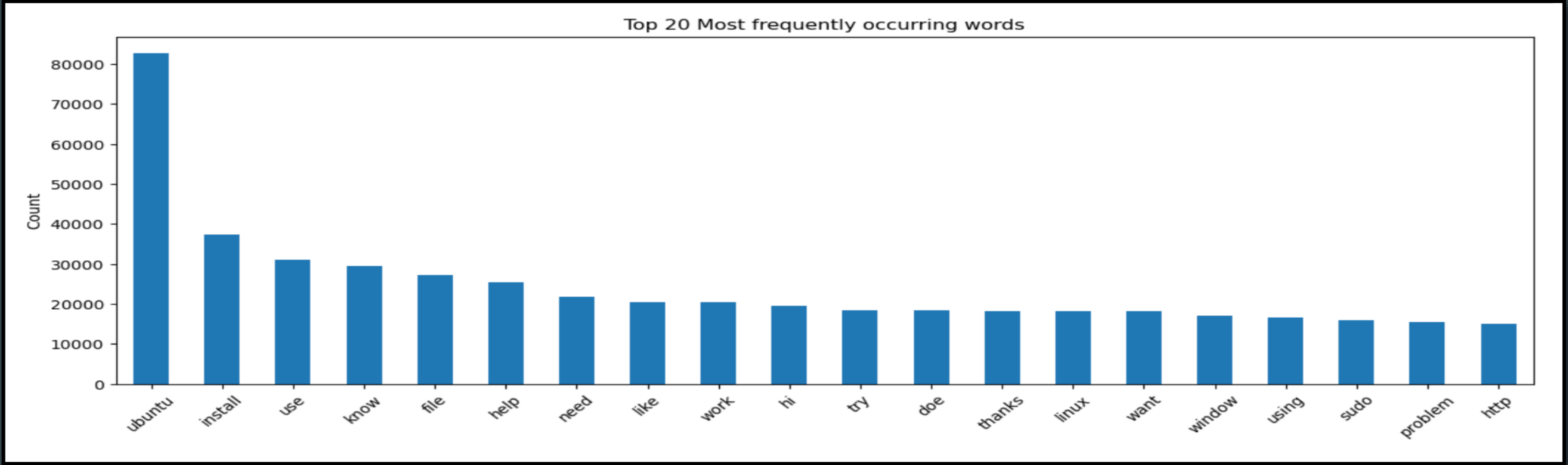
Crucial Findings Important for Modeling

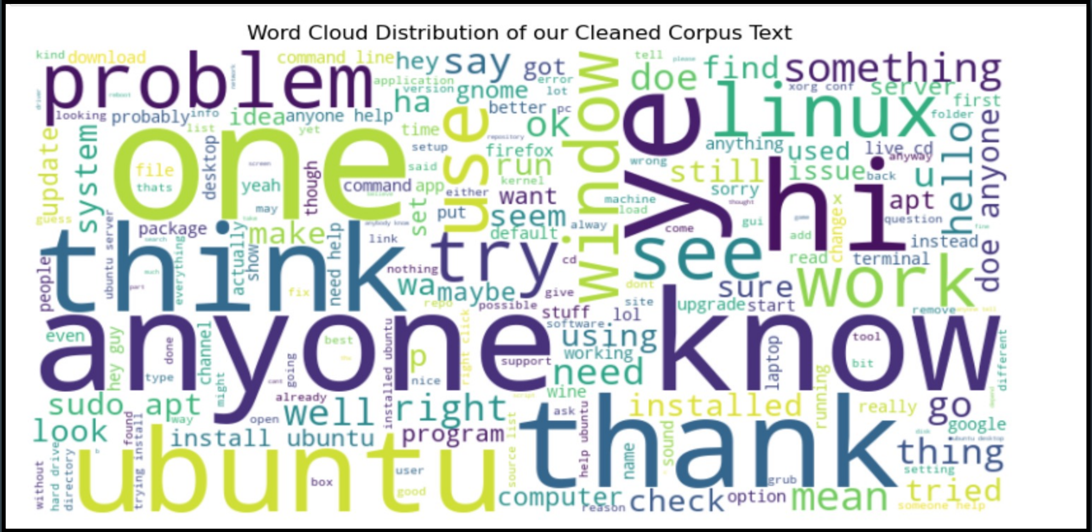
Most Used Words –
Help us determine most popular topics

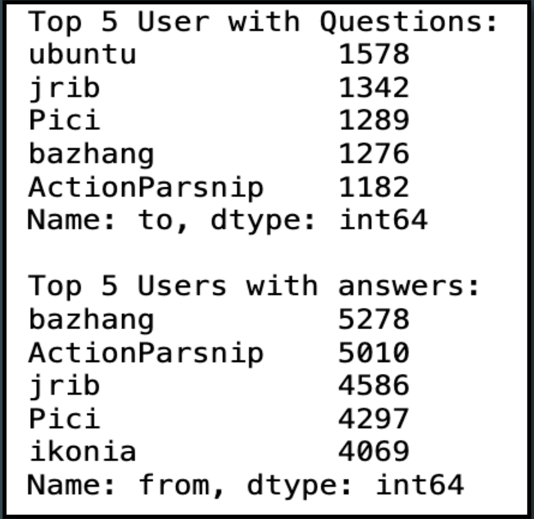
Active Users –
Knowing who is helping us answer user enquiries

Years –
2008-2012.
Work with Old Data to Better understand New Data

DialogueId –
Total of 1,036,607 conversations had, out which 346,108 are unique.


Top 20 Most frequently occurring words


Word Cloud Distribution of our Cleaned Corpus Text


```
Top 5 User with Questions:
ubuntu              1578
jrib                1342
Pici                1289
bazhang             1276
ActionParsnip       1182
Name: to, dtype: int64

Top 5 Users with answers:
bazhang             5278
ActionParsnip       5010
jrib                4586
Pici                4297
ikonia              4069
Name: from, dtype: int64
```

Baseline Model & Evaluations

We have a unique dataset from other Machine Learning models, based on how we don't have a target. Just working with the Text Column

Due to this we will do Unsupervised Learning.

Word2Vector

Generate Word Embeddings. Vector Representation of words capturing similarities between words in context

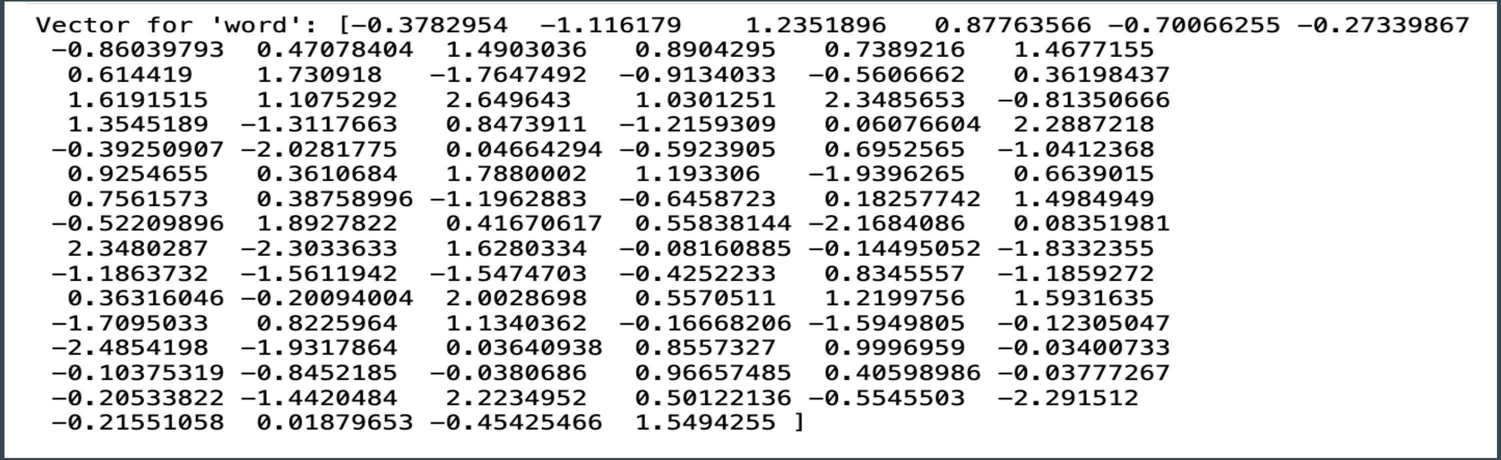
Latent Dirichlet Allocation (LDA)

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

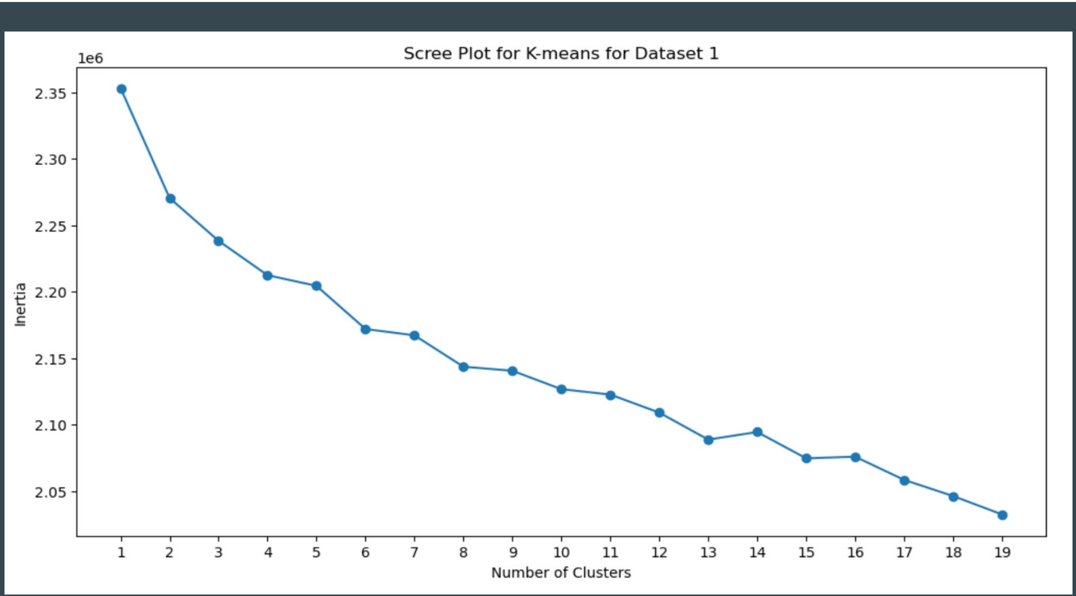
K-Means

Principal Component Factor Analysis to retain Dimensionality reduction

Word Embeddings for `Ubuntu`



```
Vector for 'word': [-0.3782954  -1.116179    1.2351896   0.87763566 -0.70066255 -0.27339867
 -0.86039793  0.47078404  1.4903036   0.8904295   0.7389216   1.4677155
  0.614419    1.730918   -1.7647492  -0.9134033  -0.5606662   0.36198437
  1.6191515   1.1075292   2.649643    1.0301251   2.3485653  -0.81350666
  1.3545189  -1.3117663   0.8473911  -1.2159309   0.06076604  2.2887218
 -0.39250907 -2.0281775   0.04664294 -0.5923905   0.6952565  -1.0412368
  0.9254655   0.3610684   1.7880002   1.193306   -1.9396265   0.6639015
  0.7561573   0.38758996 -1.1962883  -0.6458723   0.18257742  1.4984949
 -0.52209896  1.8927822   0.41670617  0.55838144 -2.1684086   0.08351981
  2.3480287  -2.3033633   1.6280334  -0.08160885 -0.14495052 -1.8332355
 -1.1863732  -1.5611942  -1.5474703  -0.4252233   0.8345557  -1.1859272
  0.36316046 -0.20094004  2.0028698   0.5570511   1.2199756   1.5931635
 -1.7095033   0.8225964   1.1340362  -0.16668206 -1.5949805  -0.12305047
 -2.4854198  -1.9317864   0.03640938  0.8557327   0.9996959  -0.03400733
 -0.10375319 -0.8452185  -0.0380686   0.96657485  0.40598986 -0.03777267
 -0.20533822 -1.4420484   2.2234952   0.50122136 -0.5545503  -2.291512
 -0.21551058  0.01879653 -0.45425466  1.5494255 ]
```



K-Means Inertia Plot

Next Steps for Advanced Modeling

We have gotten to know our dataset even more now and have understood what it is made up of and our limitations.

PYTORCH

Natural Language Processing

TENSORFLOW

Training Neural Networks

LDA

Further Test for Perplexity

BERT

Transformer Learning

- Adding new features in our code and Constantly training the model, we should have enough analysis to create our bot and push it through production

- The models we wish to create, there exists multiple other version in different classes. We can even use what exists there to better our model and learn from them.

- In addition to all of this, Deep Learning and Language Learning Models will help us make our model stronger.