Contents

[The Real Problem 2](#_Toc90050428)

[Business Domain Value 2](#_Toc90050429)

[Summary of problem statement, data, and findings 3](#_Toc90050430)

[Other business use case of text classification 3](#_Toc90050431)

[Summary of the Approach to EDA and Pre-processing: Pre-processing 4](#_Toc90050432)

[Data Shape 4](#_Toc90050433)

[Some more info on the dataset 4](#_Toc90050434)

[Quick peek into data 4](#_Toc90050435)

[Check for nulls 4](#_Toc90050436)

[Handling Nulls 5](#_Toc90050437)

[Mojibake 5](#_Toc90050438)

[Presence of non-English language 6](#_Toc90050439)

[Checking for Patterns: "received from: *eylqgodm.ybqkwiam@gmail.com*" 6](#_Toc90050440)

[Checking for Patterns: "email ids" 7](#_Toc90050441)

[Checking for Patterns: "Mail Format" 7](#_Toc90050442)

[Checking for Patterns: <mailto:> 8](#_Toc90050443)

[Checking for Patterns: template with name, language, browser, etc. 8](#_Toc90050444)

[Checking for Patterns: Checking for embedded images text 9](#_Toc90050445)

[Checking for Patterns: Phrases 10](#_Toc90050446)

[Checking for Patterns: Expanding acronyms such as “pls” and replacing phrase such as “please help to” 11](#_Toc90050447)

[Expanding contractions 11](#_Toc90050448)

[Summary of the Approach to EDA and Pre-processing: EDA 12](#_Toc90050449)

[Distribution of tickets as per Assignment Group 12](#_Toc90050450)

[Word Cloud of Description (whole data set) for max 5000 words 12](#_Toc90050451)

[Word Cloud of Short Description (whole data set) for max 5000 words 13](#_Toc90050452)

[Deciding Models and Model Building 13](#_Toc90050453)

[Rule based model 13](#_Toc90050454)

[ML based model 14](#_Toc90050455)

[Hybrid model 14](#_Toc90050456)

[Traditional ML Model 15](#_Toc90050457)

[Neural Network based Model 16](#_Toc90050458)

[Summary 17](#_Toc90050459)

[How to improve your model performance? 17](#_Toc90050460)

[Use of embeddings 18](#_Toc90050461)

[Reduction of noise 18](#_Toc90050462)

# The Real Problem

One of the key activities of any IT function is to “Keep the lights on” to ensure there is no impact to the Business operations. IT leverages Incident Management process to achieve the above Objective. An incident is something that is unplanned interruption to an IT service or reduction in the quality of an IT service that affects the Users and the Business. The main goal of Incident Management process is to provide a quick fix / workarounds or solutions that resolves the interruption and restores the service to its full capacity to ensure no business impact. In most of the organizations, incidents are created by various Business and IT Users, End Users/ Vendors if they have access to ticketing systems, and from the integrated monitoring systems and tools. Assigning the incidents to the appropriate person or unit in the support team has critical importance to provide improved user satisfaction while ensuring better allocation of support resources. The assignment of incidents to appropriate IT groups is still a manual process in many of the IT organizations. Manual assignment of incidents is time consuming and requires human efforts. There may be mistakes due to human errors and resource consumption is carried out ineffectively because of the misaddressing. On the other hand, manual assignment increases the response and resolution times which result in user satisfaction deterioration / poor customer service.

# Business Domain Value

In the support process, incoming incidents are analyzed and assessed by organization’s support teams to fulfill the request. In many organizations, better allocation and effective usage of the valuable support resources will directly result in substantial cost savings. Currently the incidents are created by various stakeholders (Business Users, IT Users and Monitoring Tools) within IT Service Management Tool and are assigned to Service Desk teams (L1 / L2 teams). This team will review the incidents for right ticket categorization, priorities and then carry out initial diagnosis to see if they can resolve. Around ~54% of the incidents are resolved by L1 / L2 teams. Incase L1 / L2 is unable to resolve, they will then escalate / assign the tickets to Functional teams from Applications and Infrastructure (L3 teams). Some portions of incidents are directly assigned to L3 teams by either Monitoring tools or Callers / Requestors. L3 teams will carry out detailed diagnosis and resolve the incidents. Around ~56% of incidents are resolved by Functional / L3 teams. Incase if vendor support is needed, they will reach out for their support towards incident closure. L1 / L2 needs to spend time reviewing Standard Operating Procedures (SOPs) before assigning to Functional teams (Minimum ~25-30% of incidents needs to be reviewed for SOPs before ticket assignment). 15 min is being spent for SOP review for each incident. Minimum of ~1 FTE effort needed only for incident assignment to L3 teams. Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited During the process of incident assignments by L1 / L2 teams to functional groups, there were multiple instances of incidents getting assigned to wrong functional groups. Around ~25% of Incidents are wrongly assigned to functional teams. Additional effort needed for Functional teams to re-assign to right functional groups. During this process, some of the incidents are in queue and not addressed timely resulting in poor customer service. Guided by powerful AI techniques that can classify incidents to right functional groups can help organizations to reduce the resolving time of the issue and can focus on more productive tasks.

# Summary of problem statement, data, and findings

Automatic Ticket Assignment (ATA) is a classification problem which comes under the Supervised Machine Learning category & plays a key role for successfully running any Incident Management System, especially in very large system that provides numerous services, and each service has multiple categories and sub-categories. Manually tagging of task to specific category and sub-category requires user training, manpower and also prone to human error that can impact over all service delivery. ATA uses machine learning technique to assign task to appropriate group automatically that can improve overall turnaround time of service delivery.

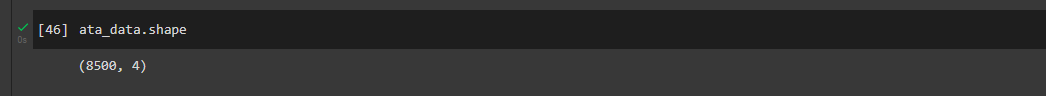
## Other business use case of text classification

1. categorize Code review comments so that patterns of review comments can be identified and automated
2. Post incident resolution in incident management system like SNOW, a user has to tag resolution comments to certain category. For e.g. in software incidents these categories may be (code issue, environment issue, Auto resolved, user training issue etc). Most often user misses to tag the comments to appropriate category. We can automate this process by creating model that can predict appropriate category for any resolution comment.
3. After sales support in product based companies, assignment of correct service personnel so that cost can be optimized and customer satisfaction can be enhanced

# Summary of the Approach to EDA and Pre-processing: Pre-processing

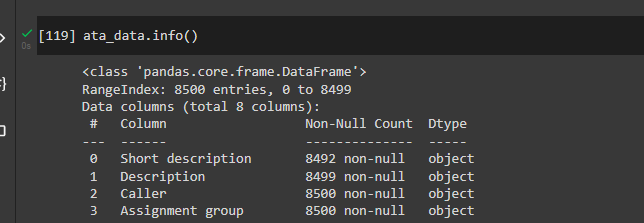
## Data Shape

There are 8,500 rows and 4 columns in the base data set

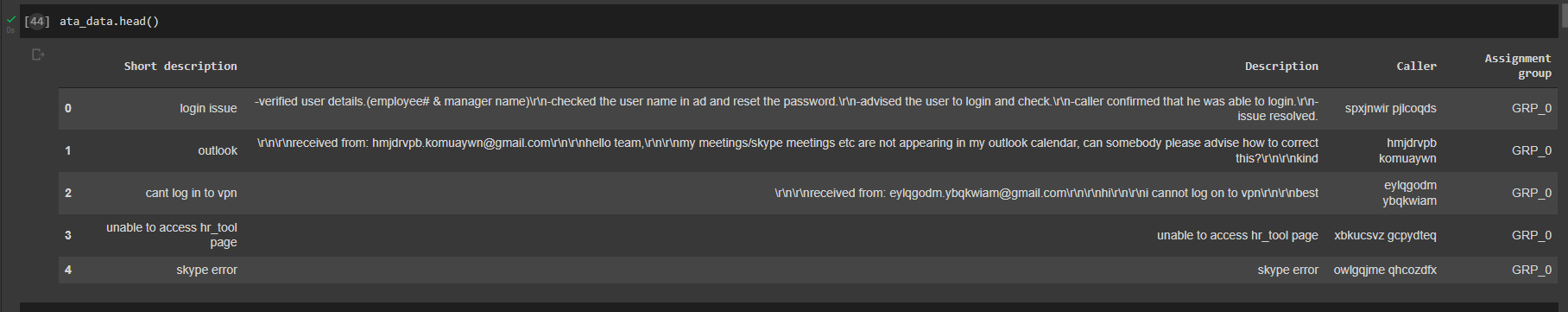


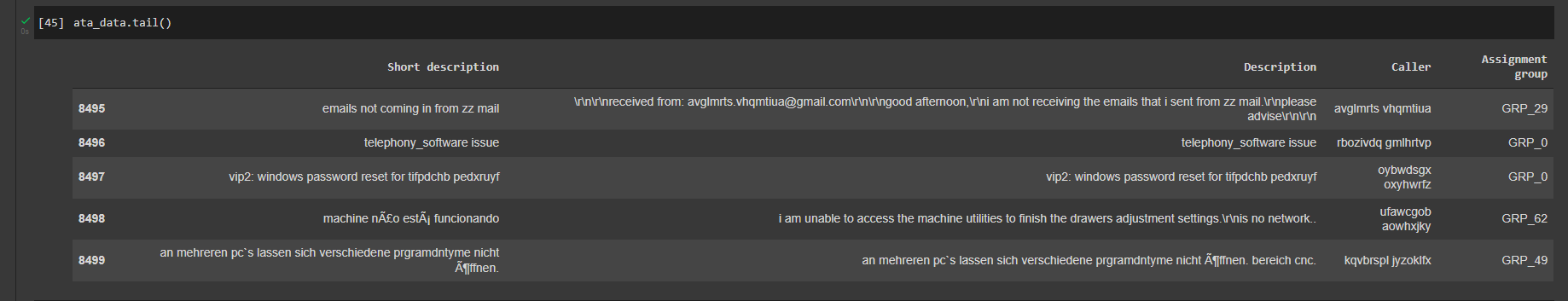
## Some more info on the dataset

The four columns are:-



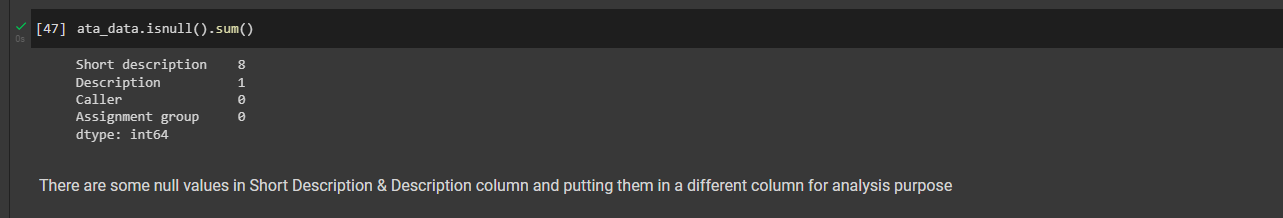
## Quick peek into data

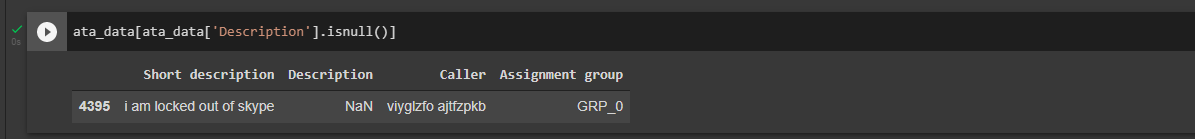


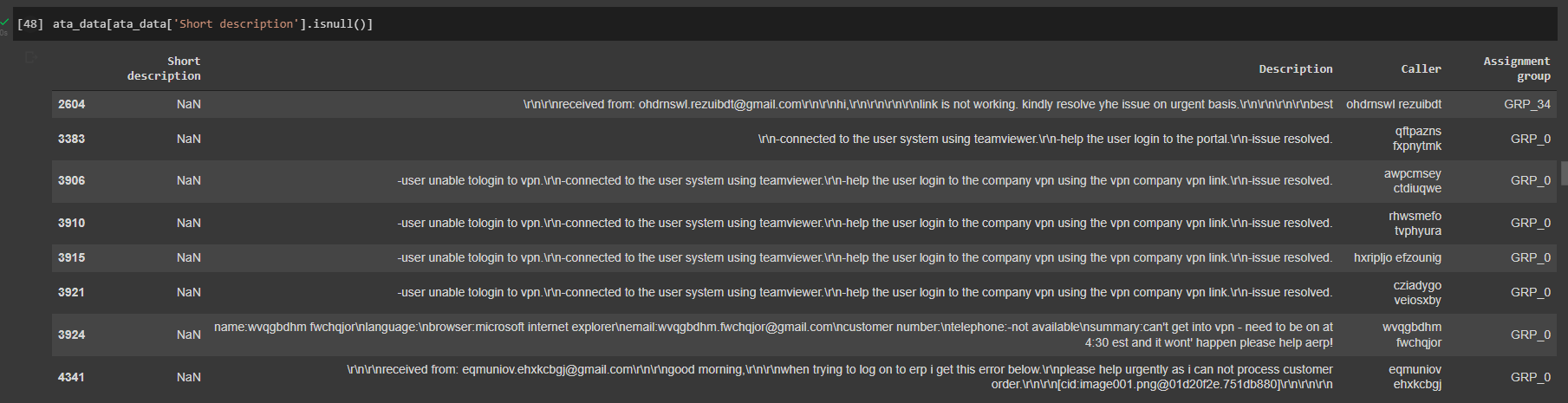


## Check for nulls

8 values in Short Description and 1 in Description are null values

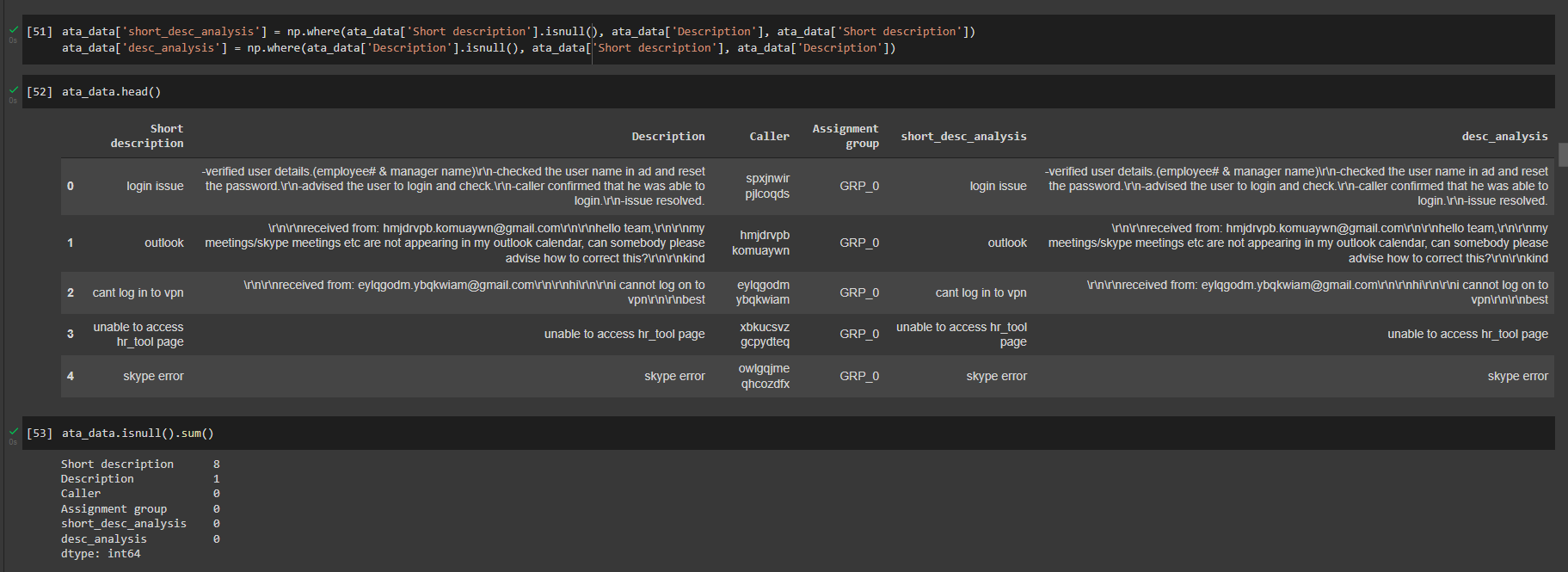






## Handling Nulls

Wherever short description is null, replacing it with description and vice versa. The data is put in another column short\_desc\_analysis and desc\_analysis which will be used for further analysis. The base columns are kept intact



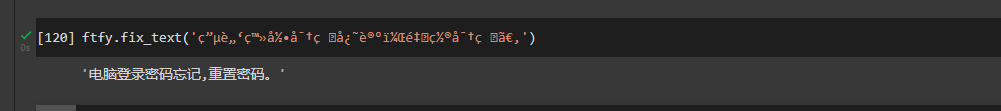
## Mojibake

The base data has presence of scrambled text called Mojibake. Example given below. It occurs when we try to read text in some other encodings



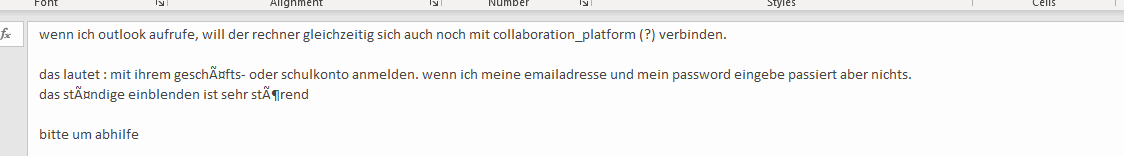
Reference: <https://www.kaggle.com/rtatman/data-cleaning-challenge-character-encodings>

**Package FTFY** is used to clean the Mojibake text. The below code snippet shows that Mojibake texts are indeed non-english text

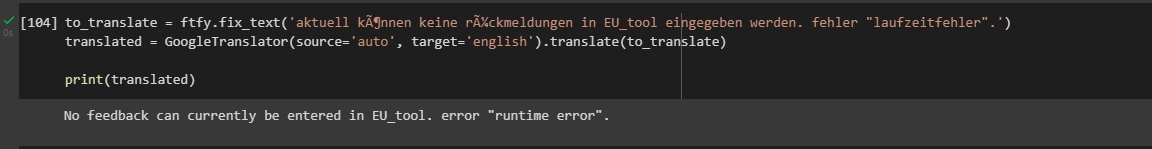


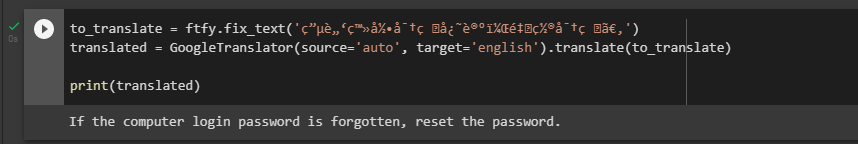
## Presence of non-English language

Cleaning Mojibake text helps us understand that there are non-English texts. But apart from cleaned Mojibake as well, we can find non-English text in corpus



**GoogleTranslator** is used to translate non-English text to English

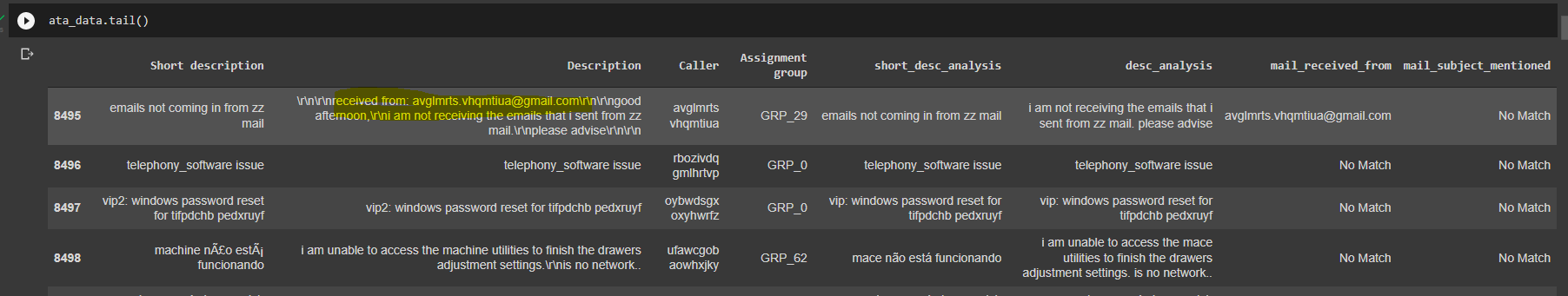




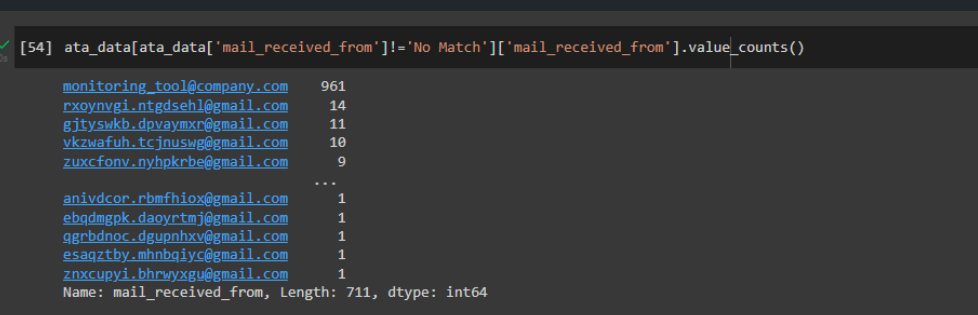
## Checking for Patterns: "received from: [*eylqgodm.ybqkwiam@gmail.com*](mailto:eylqgodm.ybqkwiam@gmail.com)"

The portion in italics can be any mail id. Further investigation showed there are 2251 such records out of 8500.

The pattern is removed from the text and the email id is added to another mail\_received\_from

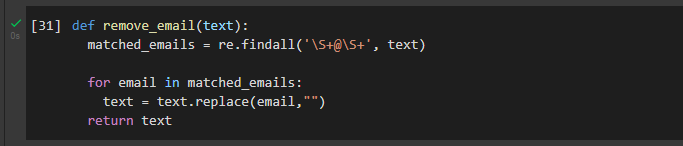


Analyzing the mails received from indicates that there are 961 records where initially issue was triggered due to system generated mail from monitoring\_tool@company.com



## Checking for Patterns: "email ids"

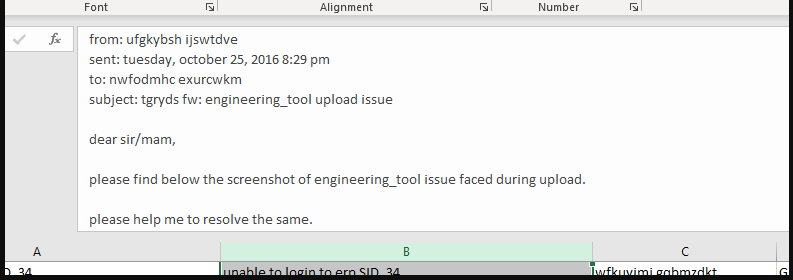
A description could contain multiple email ids. Removing that pattern and replacing with blank



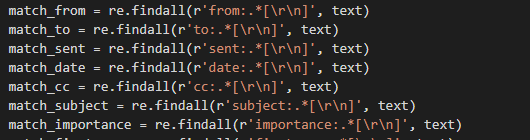
## Checking for Patterns: "Mail Format"

Some of the Description have pattern like that of mail

1. From
2. Sent
3. To
4. Subject
5. Cc
6. importance



Removing such patterns using regex and replacing with blanks. Also the subject is copied into another column mail\_subject\_mentioned



Clean text



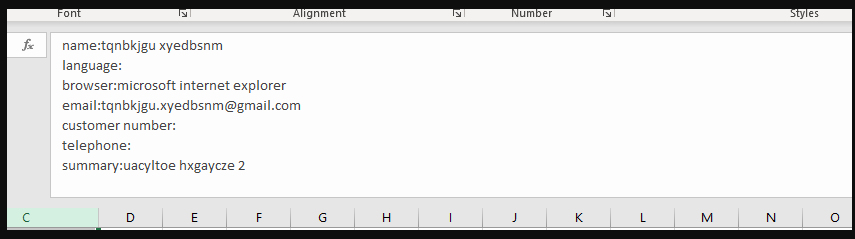
## Checking for Patterns: <mailto:>

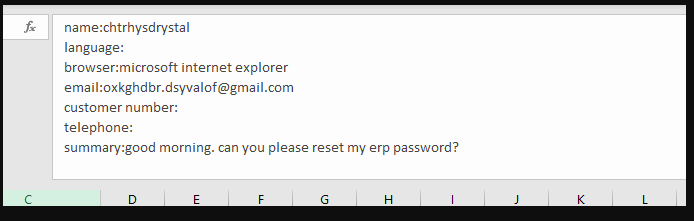
Removing occurrences of mailto: from the text



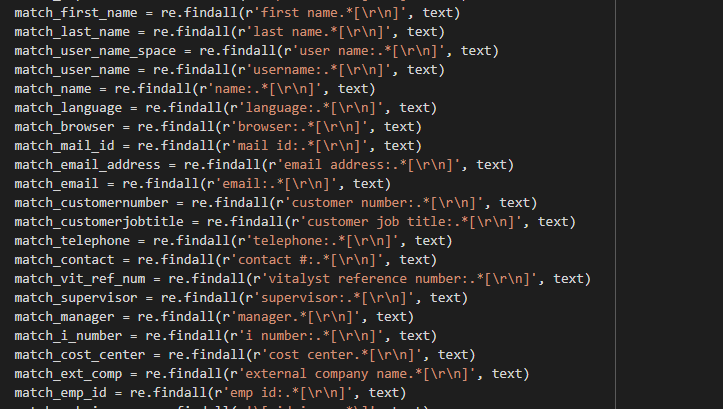
## Checking for Patterns: template with name, language, browser, etc.

The description column has a template for e.g.





Removing such pattern using regex



## Checking for Patterns: Checking for embedded images text

The data contains reference of embedded images as shown in the below image



Regex is used to replace such patterns with blanks

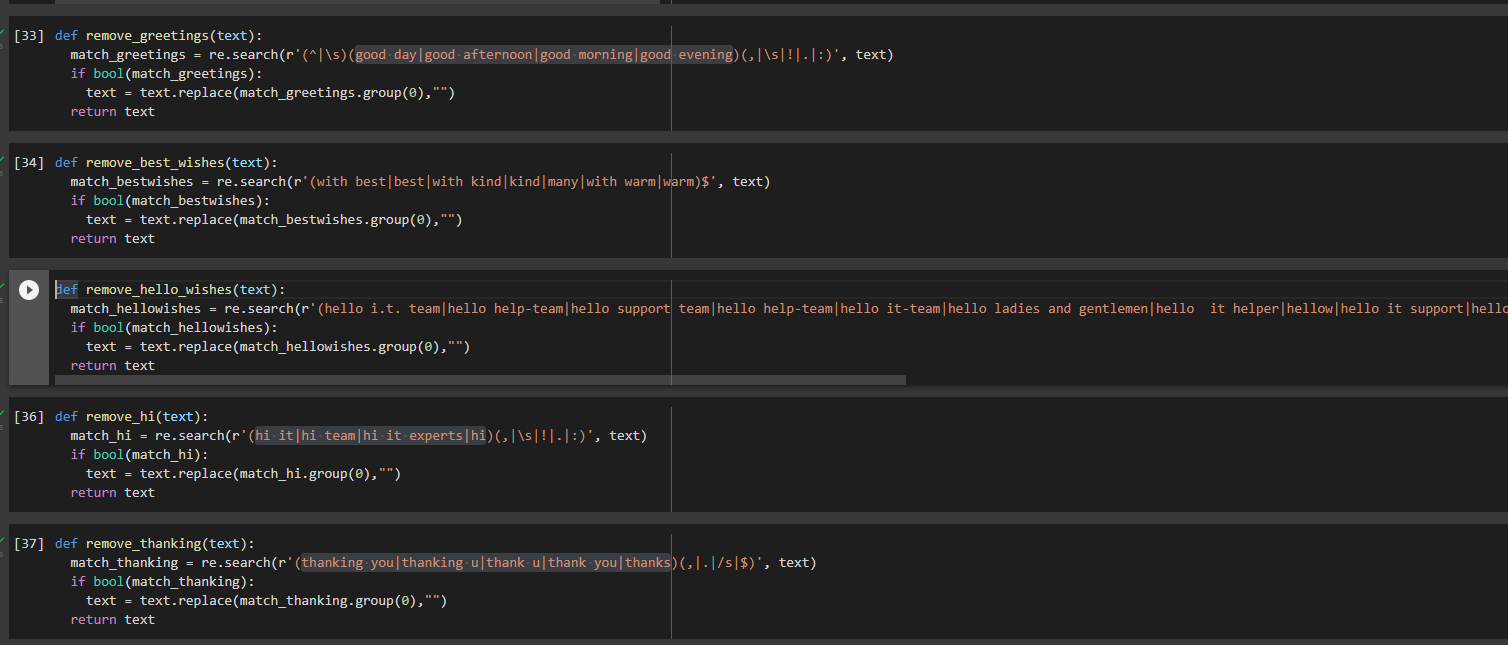


## Checking for Patterns: Phrases

Certain pattern of text found in corpus

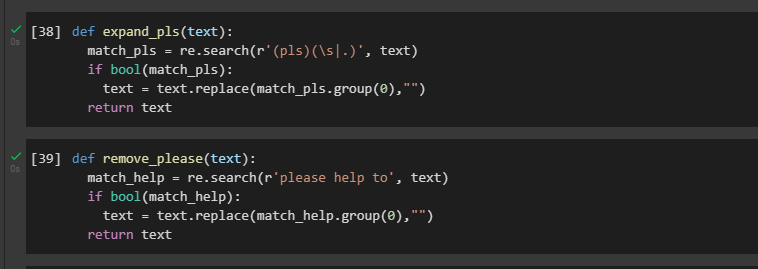
* begin forwarded message:
* sent from my iphone
* sent from my ipad
* “sir or madam,” or “sir/mam,” or “sir,”
* yes/no/na
* good day or good afternoon or good morning or good evening
* hello i.t. team or hello help-team or hello support team or hello help-team or hello it-team or hello ladies and gentlemen or hello it helper or hellow or hello it support or hello all or hello colleagues or hi there or hello it team or hello sir or hello it service or hello it or hello helpdesk or hello team or hello all or hello it desk or hello it helper or hello dac or hello or gentles or it team or dear all or dear it or dear or hallo or all groups or it help or team ith best or best or with kind or kind or many or with warm or warm
* hi it or hi team or hi it experts or hi
* thanking you or thanking u or thank u or thank you or thanks

Such patterns are removed using regex



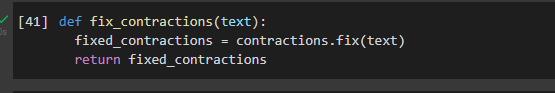
## Checking for Patterns: Expanding acronyms such as “pls” and replacing phrase such as “please help to”

Short form such as pls is replaced with please and then “please help to” is replaced with blanks

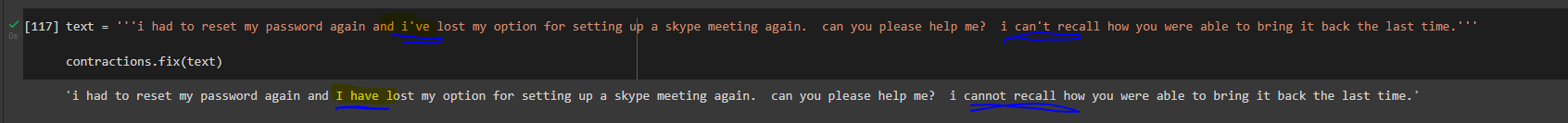


## Expanding contractions

Contractions such as isn’t, can’t, doesn’t can be expanded to is not, cannot & does not resp. This is done using the package **contractions**

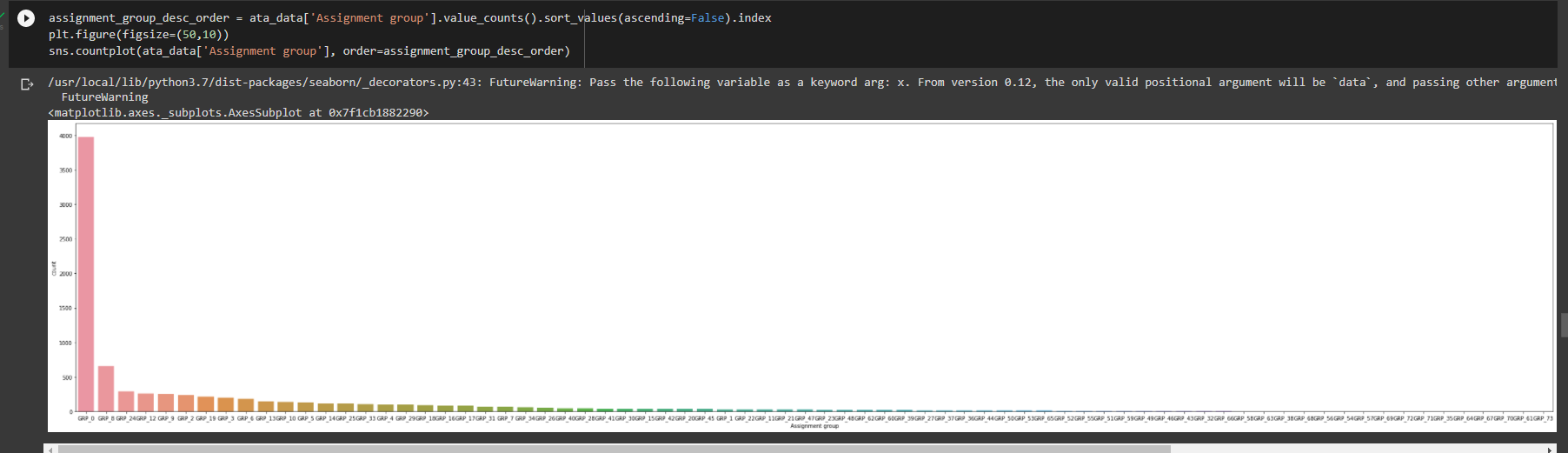


Applying contractions.fix



# Summary of the Approach to EDA and Pre-processing: EDA

## Distribution of tickets as per Assignment Group

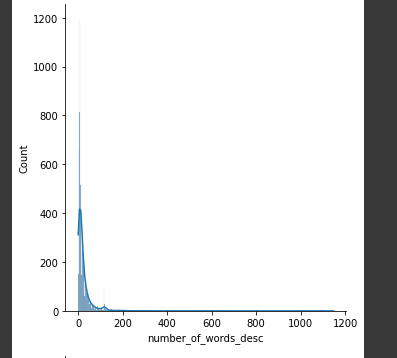


Group\_0 has 3976 tickets (roughly 47% of tickets). Dataset is imbalanced

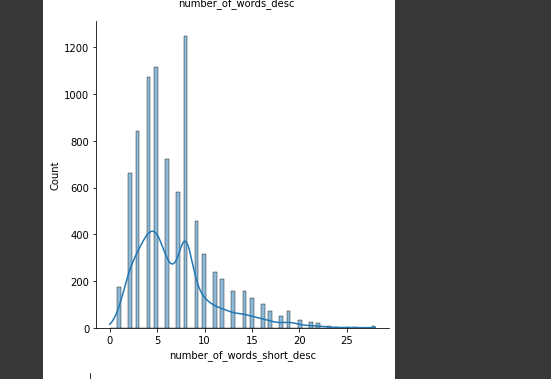
* no. of group less than 10 tickets- 25
* no. of groups in which records between 10 and 100 - 37
* no. of group greater than 100 tickets – 11

## Words/Characters analysis

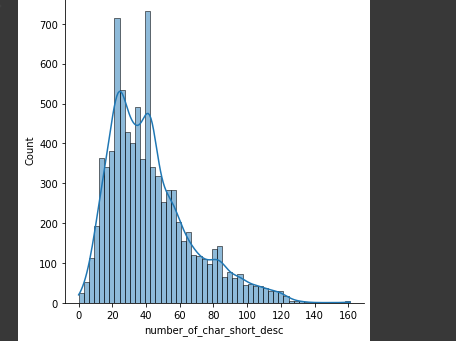
Majority of tickets have upto 100 words in Description. However, there are some tickets ranging from 200 – 1200 words



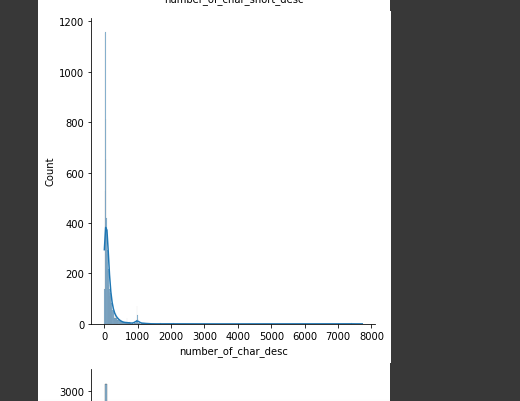
Short Description has a much compact word distribution



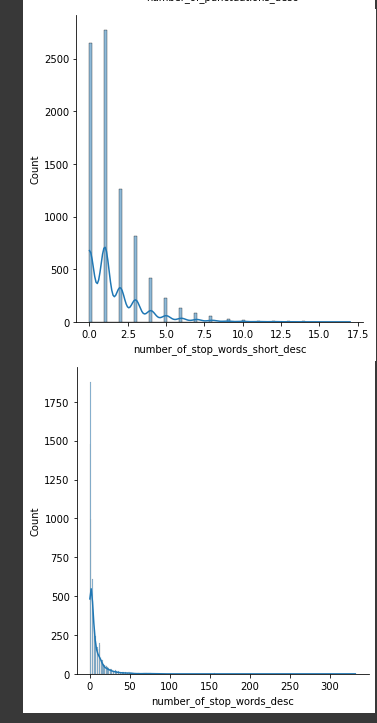
Character distribution for Short Description is also very compact



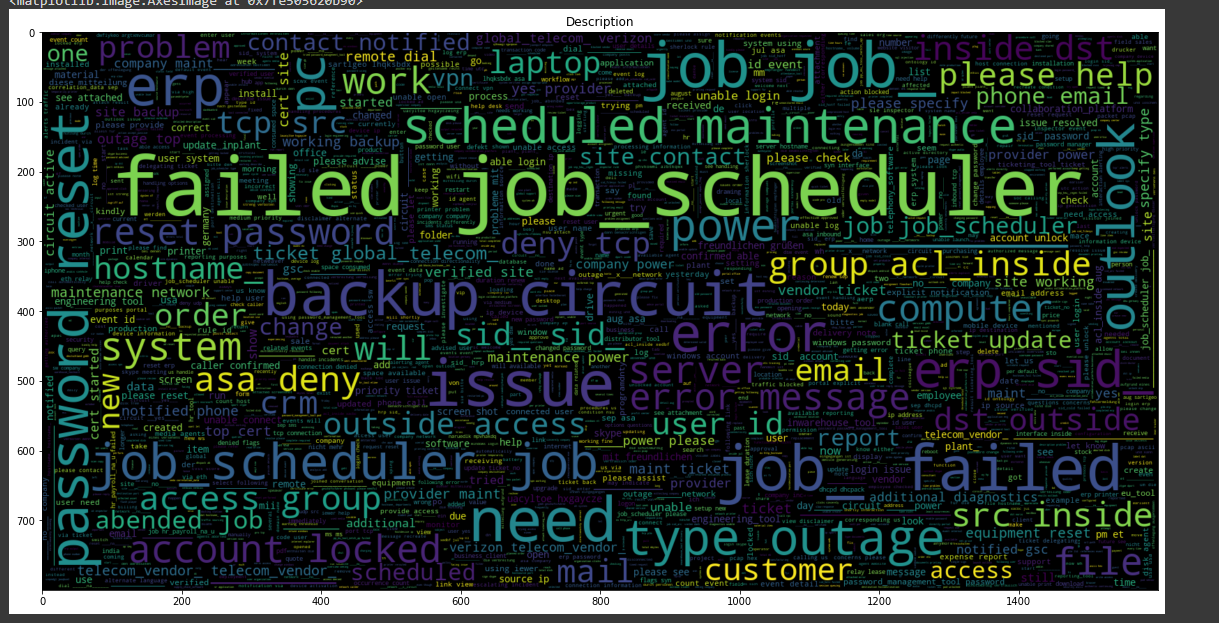
For description, in case of some tickets the character length ranges from 1000-8000 characters



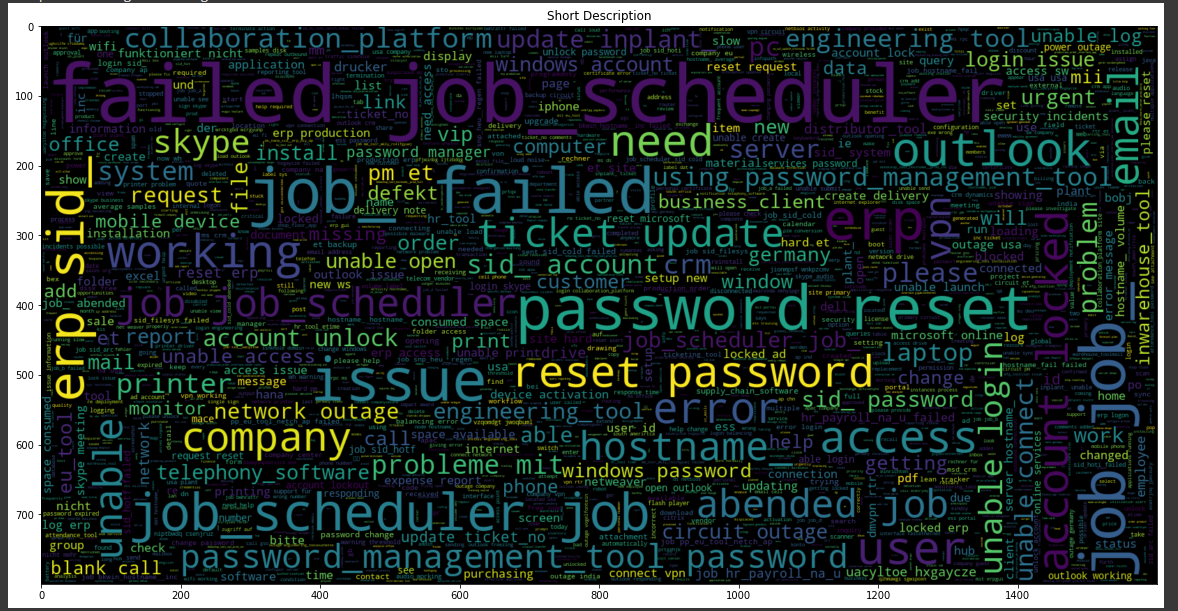
The presence of stop words is also significant in Description



## Word Cloud of Description (whole data set) for max 5000 words



## Word Cloud of Short Description (whole data set) for max 5000 words

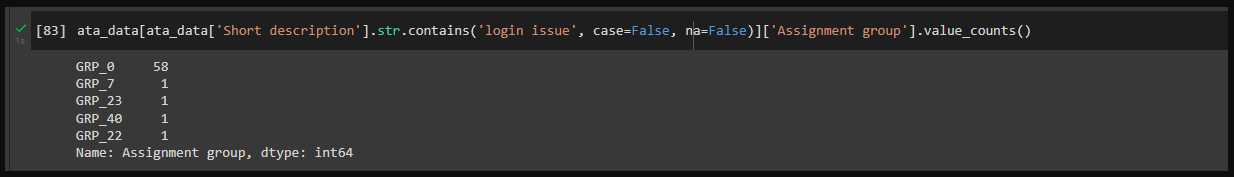


# Deciding Models and Model Building

Approach towards Ticket Classification

## Rule based model

This method classifies tickets into groups using certain rules. For example, if we consider a Short Description containing "login issue" and see the base data - we find most of the cases belong to Group 0. Hence, we may come up with a rule that if text contains "login issue" then assign it to Group 0



Though this may look simpler for a small dataset, but there are disadvantages to this approach:-

* It is very time consuming to come up with such rules. Someone has to go through all the tickets and should have good domain knowledge to come up with such rules
* this may lead to biasness, as the person creating the rules will give suggestions based on the way they perceive a ticket
* this indicates that maintaining such a list will be difficult and to scale up such a system will be difficult (consider a scenario where million tickets are received)
* In case we get a text which has never come up in past, Rule Based system will not be able to suggest a new grouping

## ML based model

ML based models eliminates the need of manually creating rules. Instead, they classify text based on past observations (labeled training data set). The important aspect of ML based model is proper cleaning of text and then converting those text into a format which machine can understand i.e., in form of vector.

There are various ways to convert text into vector - from as basic as One Hot Encoding, TFIDF to complex embeddings using Glove/Bert.

The vectors are then fed into a classification model. Again, this could be a traditional ML Classifier such as SVC, Naive Bayes, or Neural network-based models such as GRU, LSTM.

There are certain limitations though:

Biasness in various stage of model building

1. At data collection stage

- Misclassification bias - tickets tagged to incorrect group

- Sampling bias - scenario where certain group of population may have a lower/higher probability to be picked up during sampling. Our dataset is imbalanced dataset and there are scenarios where certain groups have one 1 entry.

2. At data processing stage

- Labeling bias - the dataset is not fully representative of all the labels/groups available in the universe

3. Algorithm building

- underfitting

- overfitting

Though there are so many biases which can creep up during model development, ML based model is still superior to the rule-based model

## Hybrid model

Combination of both Rule Based and ML based model.

Studies have show that Hybrid models perform good. We would be going for a hybrid model. Explaining the ML/AI based model

Classification of text can be done either by: -

## Traditional ML Model

1. Naive Bayes

* A multinomial naive bayes classifier can be used to classify tickets into different groups. The model uses frequency of words to calculate probability of the group to which it will belong to. It will not consider the context of the statements.
* Here, the ‘naive’ assumption is that every word in a sentence is independent of the other ones and hence the Bayes Theorem could be applied. This assumption will not be true in real world scenario. In addition, the multinomial model makes an assumption of positional independence. (The position of a term in a document by itself does not carry information about the class. Although there is a difference between China sues France and France sues China, the occurrence of China in position 1 versus position 3 of the document is not useful in NB classification because it looks at each term separately. The conditional independence assumption commits to this way of processing the evidence.) However, NB models perform well despite the conditional independence assumption
* Even if it is not the method with the highest accuracy for text, NB has many virtues that make it a strong contender for text classification. It excels if there are many equally important features that jointly contribute to the classification decision
* it can have decent performance when using fewer than a dozen terms. The most important indicators for a class are less likely to change. Thus, a model that only relies on these features is more likely to maintain a certain level of accuracy
* NB's main strength is its efficiency: Training and classification can be accomplished with one pass over the data. Because it combines efficiency with good accuracy it is often used as a baseline in text classification research. It is often the method of choice if (i) squeezing out a few extra percentage points of accuracy is not worth the trouble in a text classification application, (ii) a very large amount of training data is available and there is more to be gained from training on a lot of data than using a better classifier on a smaller training set, or (iii) if its robustness to concept drift can be exploited.

The performance of Naive Bayes depends on the accuracy of the estimated conditional probability terms. It is hard to accurately estimate these terms when the training data is scarce.

(ref: https://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html)

1. SVM

An SVM is a kind of large-margin classifier: it is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data (possibly discounting some points as outliers or noise).

SVMs are inherently two-class classifiers. A “one-versus-rest” classifier can be used for multi-class classification. However, adjusting the weights will be difficult.

Traditional ML Model would prefer the embeddings such as

1. TF-IDF
2. Word2Vec
3. Glove

## Neural Network based Model

1. RNN

Recurrent Neural Networks (RNN) are designed to work with sequential data. RNN uses the previous information in the sequence to produce the current output. At the **last step**, the RNN has information about all the previous words. RNN’s face short-term memory problem. It is caused due to vanishing gradient problem. As RNN processes more steps it suffers from vanishing gradient more than other neural network architectures.

1. GRU

The workflow of GRU is same as RNN but the difference is in the operations inside the GRU unit. Inside GRU it has two gates 1)reset gate 2)update gate. Gates are nothing but neural networks, each gate has its own weights and biases.

**Update gate**

Update gate decides if the cell state should be updated with the candidate state(current activation value)or not.

**Reset gate**

The reset gate is used to decide whether the previous cell state is important or not. Sometimes the reset gate is not used in simple GRU.

**Candidate cell**

It is just simply the same as the hidden state(activation) of RNN.

**Final cell state**

The final cell state is dependent on the update gate. It may or may not be updated with candidate state. Remove some content from last cell state, and write some new cell content.

In GRU the final cell state is directly passing as the activation to the next cell.

In GRU,

* If reset close to 0, ignore previous hidden state (allows the model to drop information that is irrelevant in the future).
* If gamma(update gate) close to 1, then we can copy information in that unit through many steps
* Gamma Controls how much of past state should matter now.

1. LSTM

LSTMs are pretty much similar to GRU’s, they are also intended to solve the vanishing gradient problem. Additional to GRU, here there are 2 more gates 1) forget gate 2) output gate.

All 3 gates (input gate, output gate, forget gate) use sigmoid as activation function so all gate values are between 0 and 1.

**Forget gate**

It controls what is kept vs forgotten, from previous cell state. In laymen terms, it will decide how much information from the previous state should be kept and forget remaining.

**Output gate**

It controls which parts of the cell are output to the hidden state. It will determine what the next hidden state will be.

(ref: https://medium.com/analytics-vidhya/rnn-vs-gru-vs-lstm-863b0b7b1573)

Neural Network Model would prefer the embeddings such as

1. Glove
2. BERT

## Summary

Choosing the type of classifier will depend on the amount of data. If volume of data is less, then a classifier with high bias (Naive Bayes) should do well.

If the volume of data is good and context of data has to be considered then Neural Network based model can be applied.

In the input data we have Short Description & Description.

* Traditional ML based model can be used if only Short Description is used.
* Neural Network based model can be used if Description or combination of Short Description & Description is used

# How to improve your model performance?

## Use of embeddings

Different kinds of embeddings can be used to see the performance of model

1. TFIDF
2. Word2Vec
3. Glove
4. BERT – Can be used on Description
5. ELMO – Can be used on Description

## Reduction of noise

Identifying and eliminating word patterns which are not helpful