# AUTOMATIC TICKET ASSIGNMENT

Assigning ticket to appropriate group

# Contents

**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

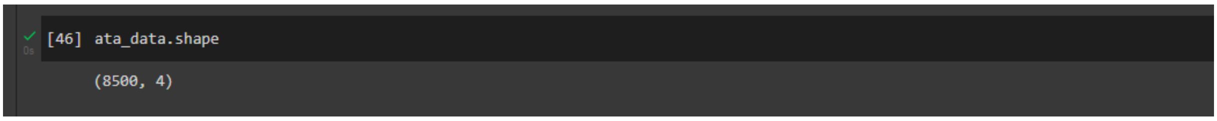
(2) Post incident resolution in incident management system like SNOW, a user has to tag resolution comments to certain category. For e.g. in software incidents these categories may be (code issue, environment issue, Auto resolved, user training issue etc). Most often user misses to tag the comments to 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

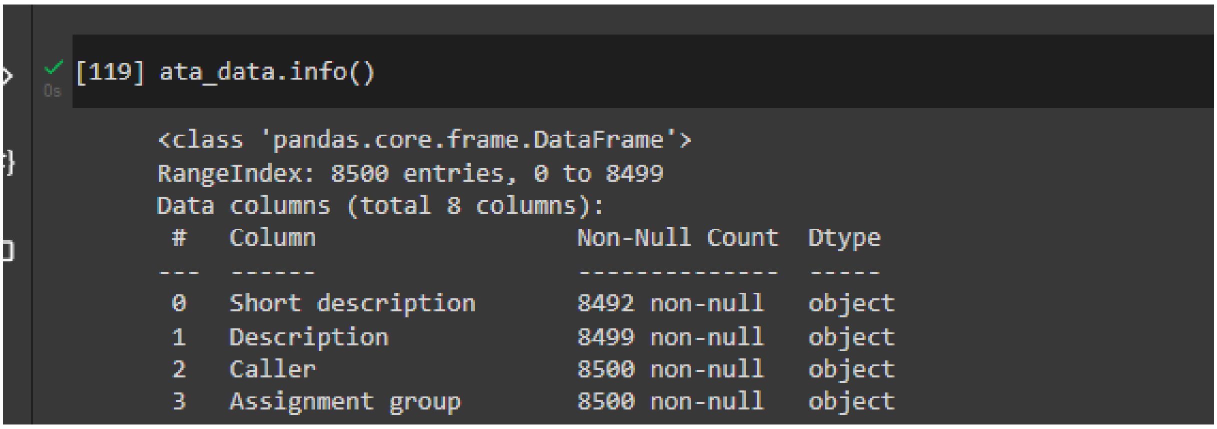
There are 8,500 rows and 4 columns in the base data set



```
[46] ata_data.shape

     (8500, 4)
```

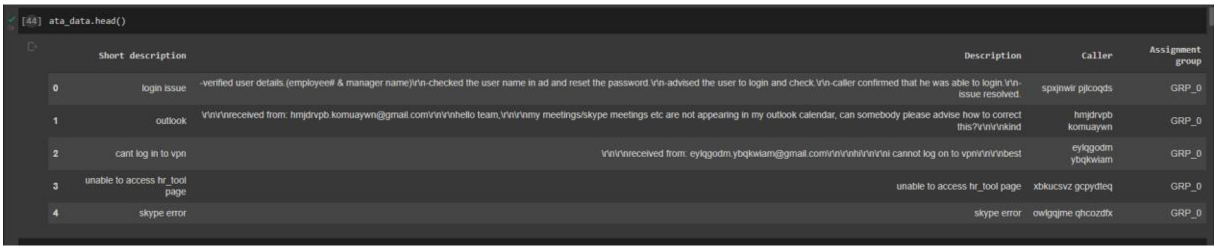
## Some more info on the dataset

The four columns are:-



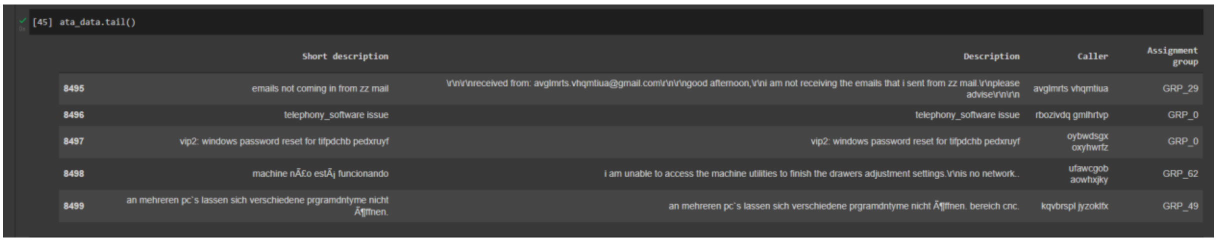
```
[119] ata_data.info()

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 8500 entries, 0 to 8499
      Data columns (total 8 columns):
       #   Column              Non-Null Count  Dtype
      ---  ------              --------------  -----
       0   Short description   8492 non-null   object
       1   Description         8499 non-null   object
       2   Caller              8500 non-null   object
       3   Assignment group    8500 non-null   object
```

## Quick peek into data



```
[44] ata_data.head()
```

| | Short description | Description | Caller | Assignment group |
|---|---|---|---|---|
| 0 | login issue | -verified user details.(employee# & manager name)\r\n-checked the user name in ad and reset the password.\r\n-advised the user to login and check.\r\n-caller confirmed that he was able to login.\r\n-issue resolved. | spxjnwir pjlcoqds | GRP_0 |
| 1 | outlook | \r\n\r\nreceived from: hmjdrvpb.komuaywn@gmail.com\r\n\r\nhello team,\r\n\r\nmy meetings/skype meetings etc are not appearing in my outlook calendar, can somebody please advise how to correct this?\r\n\r\nkind | hmjdrvpb komuaywn | GRP_0 |
| 2 | cant log in to vpn | \r\n\r\nreceived from: eylqgodm.ybqkwiam@gmail.com\r\n\r\nhi\r\n\r\ni cannot log on to vpn\r\n\r\nbest | eylqgodm ybqkwiam | GRP_0 |
| 3 | unable to access hr_tool page | unable to access hr_tool page | xbkucsvz gcpydteq | GRP_0 |
| 4 | skype error | skype error | owlgqjme qhcozdfx | GRP_0 |

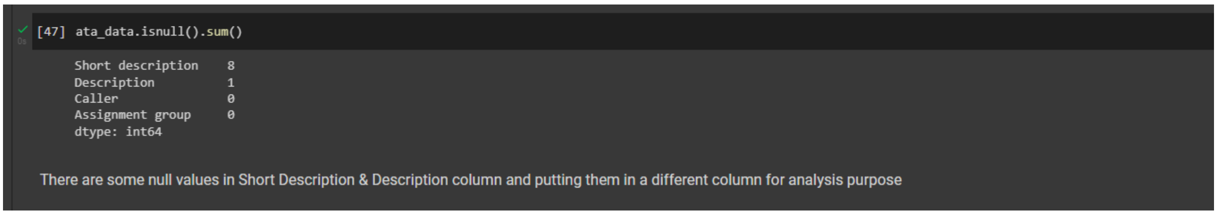


```
[45] ata_data.tail()
```

| | Short description | Description | Caller | Assignment group |
|---|---|---|---|---|
| 8495 | emails not coming in from zz mail | \r\n\r\nreceived from: avglmrts.vhqmtiua@gmail.com\r\n\r\ngood afternoon,\r\ni am not receiving the emails that i sent from zz mail.\r\nplease advise\r\n\r\n | avglmrts vhqmtiua | GRP_29 |
| 8496 | telephony_software issue | telephony_software issue | rbozivdq gmlhrtvp | GRP_0 |
| 8497 | vip2: windows password reset for tifpdchb pedxruyf | vip2: windows password reset for tifpdchb pedxruyf | oybwdsgx oxyhwrfz | GRP_0 |
| 8498 | machine nÃ£o estÃ¡ funcionando | i am unable to access the machine utilities to finish the drawers adjustment settings.\r\nis no network.. | ufawcgob aowhxjky | GRP_62 |
| 8499 | an mehreren pc`s lassen sich verschiedene prgramdntyme nicht Ã¶ffnen. | an mehreren pc`s lassen sich verschiedene prgramdntyme nicht Ã¶ffnen. bereich cnc. | kqvbrspl jyzoklfx | GRP_49 |

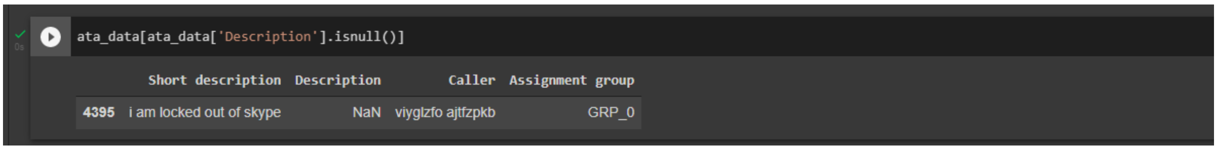
## Check for nulls

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



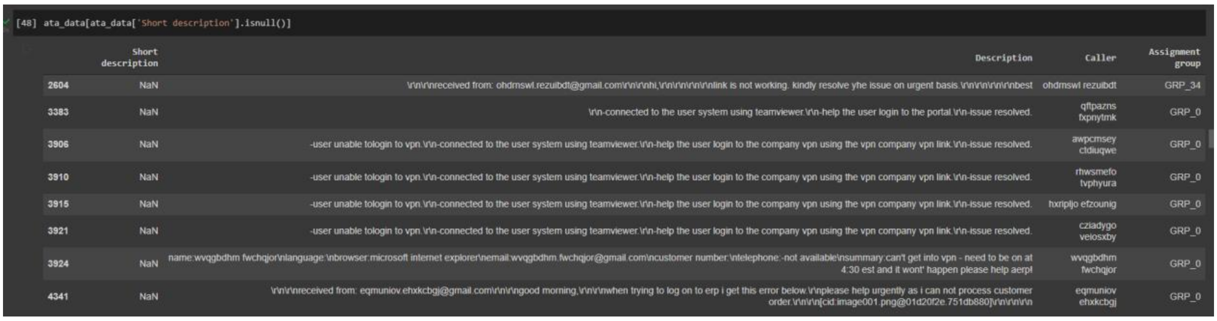
```
[47] ata_data.isnull().sum()

     Short description    8
     Description          1
     Caller               0
     Assignment group     0
     dtype: int64
```

There are some null values in Short Description & Description column and putting them in a different column for analysis purpose



```
ata_data[ata_data['Description'].isnull()]
```

| | Short description | Description | Caller | Assignment group |
|---|---|---|---|---|
| 4395 | i am locked out of skype | NaN | viyglzfo ajtfzpkb | GRP_0 |

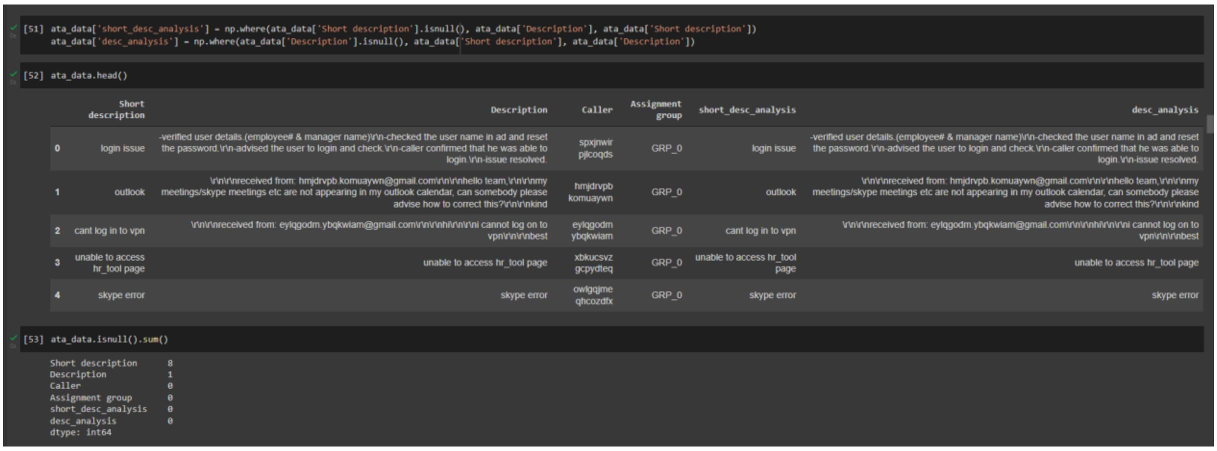


```
[48] ata_data[ata_data['Short description'].isnull()]
```

| | Short description | Description | Caller | Assignment group |
|---|---|---|---|---|
| 2604 | NaN | \r\n\r\nreceived from: ohdmswl.rezuibdt@gmail.com\r\n\r\nhi,\r\n\r\n\r\n\r\nlink is not working. kindly resolve yhe issue on urgent basis.\r\n\r\n\r\n\r\nbest | ohdmswl rezuibdt | GRP_34 |
| 3383 | NaN | \r\n-connected to the user system using teamviewer.\r\n-help the user login to the portal.\r\n-issue resolved. | qftpazns fxpnytmk | GRP_0 |
| 3906 | NaN | -user unable tologin to vpn.\r\n-connected to the user system using teamviewer.\r\n-help the user login to the company vpn using the vpn company vpn link.\r\n-issue resolved. | awpcmsey ctdiuqwe | GRP_0 |
| 3910 | NaN | -user unable tologin to vpn.\r\n-connected to the user system using teamviewer.\r\n-help the user login to the company vpn using the vpn company vpn link.\r\n-issue resolved. | rhwsmefo tvphyura | GRP_0 |
| 3915 | NaN | -user unable tologin to vpn.\r\n-connected to the user system using teamviewer.\r\n-help the user login to the company vpn using the vpn company vpn link.\r\n-issue resolved. | hxripljo efzounig | GRP_0 |
| 3921 | NaN | -user unable tologin to vpn.\r\n-connected to the user system using teamviewer.\r\n-help the user login to the company vpn using the vpn company vpn link.\r\n-issue resolved. | czladygo velosxby | GRP_0 |
| 3924 | NaN | name:wvqgbdhm fwchqjor\nlanguage:\nbrowser:microsoft internet explorer\nemail:wvqgbdhm.fwchqjor@gmail.com\ncustomer number:\ntelephone:-not available\nsummary:can't get into vpn - need to be on at 4:30 est and it wont' happen please help aerp! | wvqgbdhm fwchqjor | GRP_0 |
| 4341 | NaN | \r\n\r\nreceived from: eqmuniov.ehxkcbgj@gmail.com\r\n\r\ngood morning,\r\n\r\nwhen trying to log on to erp i get this error below.\r\nplease help urgently as i can not process customer order.\r\n\r\n[cid:image001.png@01d20f2e.751db880]\r\n\r\n\r\n | eqmuniov ehxkcbgj | GRP_0 |

## 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



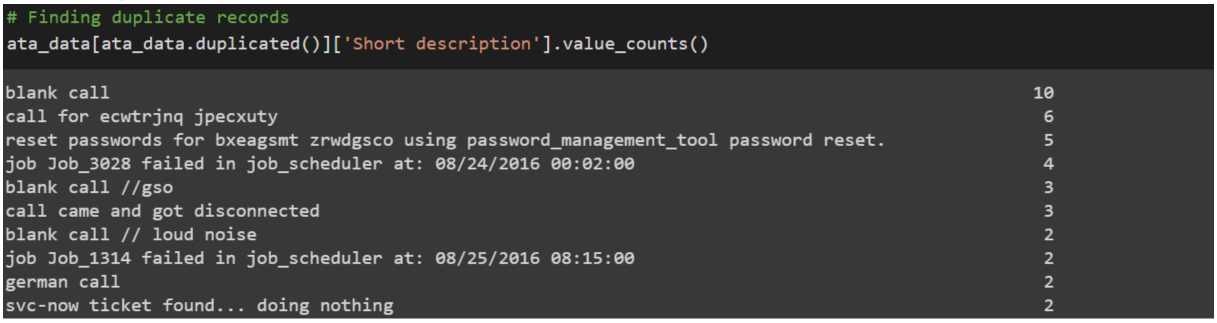
```
[51] ata_data['short_desc_analysis'] = np.where(ata_data['Short description'].isnull(), ata_data['Description'], ata_data['Short description'])
     ata_data['desc_analysis'] = np.where(ata_data['Description'].isnull(), ata_data['Short description'], ata_data['Description'])
```

```
[52] ata_data.head()
```

| | Short description | Description | Caller | Assignment group | short_desc_analysis | desc_analysis |
|---|---|---|---|---|---|---|
| 0 | login issue | -verified user details.(employee# & manager name)\r\n-checked the user name in ad and reset the password.\r\n-advised the user to login and check.\r\n-caller confirmed that he was able to login.\r\n-issue resolved. | spxjnwir pjlcoqds | GRP_0 | login issue | -verified user details.(employee# & manager name)\r\n-checked the user name in ad and reset the password.\r\n-advised the user to login and check.\r\n-caller confirmed that he was able to login.\r\n-issue resolved. |
| 1 | outlook | \r\n\r\nreceived from: hmjdrvpb.komuaywn@gmail.com\r\n\r\nhello team,\r\n\r\nmy meetings/skype meetings etc are not appearing in my outlook calendar, can somebody please advise how to correct this?\r\n\r\nkind | hmjdrvpb komuaywn | GRP_0 | outlook | \r\n\r\nreceived from: hmjdrvpb.komuaywn@gmail.com\r\n\r\nhello team,\r\n\r\nmy meetings/skype meetings etc are not appearing in my outlook calendar, can somebody please advise how to correct this?\r\n\r\nkind |
| 2 | cant log in to vpn | \r\n\r\nreceived from: eylqgodm.ybqkwiam@gmail.com\r\n\r\nhi\r\n\r\ni cannot log on to vpn\r\n\r\nbest | eylqgodm ybqkwiam | GRP_0 | cant log in to vpn | \r\n\r\nreceived from: eylqgodm.ybqkwiam@gmail.com\r\n\r\nhi\r\n\r\ni cannot log on to vpn\r\n\r\nbest |
| 3 | unable to access hr_tool page | unable to access hr_tool page | xbkucsvz gcpydteq | GRP_0 | unable to access hr_tool page | unable to access hr_tool page |
| 4 | skype error | skype error | owlgqjme qhcozdfx | GRP_0 | skype error | skype error |

```
[53] ata_data.isnull().sum()

Short description      8
Description            1
Caller                 0
Assignment group       0
short_desc_analysis    0
desc_analysis          0
dtype: int64
```

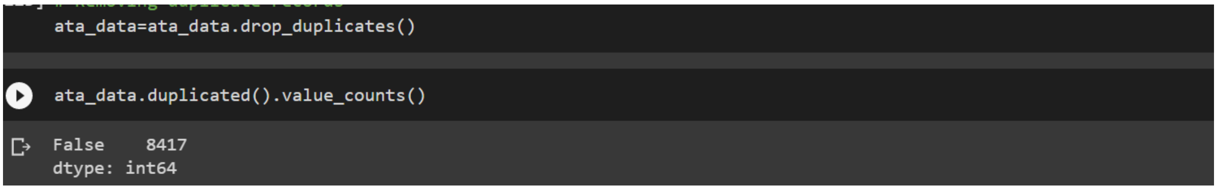
## Identifying Duplicates



```
# Finding duplicate records
ata_data[ata_data.duplicated()]['Short description'].value_counts()

blank call                                                                          10
call for ecwtrjnq jpecxuty                                                           6
reset passwords for bxeagsmt zrwdgsco using password_management_tool password reset.  5
job Job_3028 failed in job_scheduler at: 08/24/2016 00:02:00                         4
blank call //gso                                                                     3
call came and got disconnected                                                       3
blank call // loud noise                                                             2
job Job_1314 failed in job_scheduler at: 08/25/2016 08:15:00                         2
german call                                                                          2
svc-now ticket found... doing nothing                                                2
```

## Removing duplicates

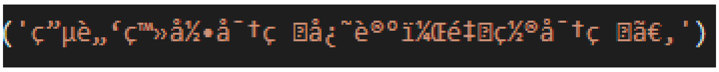
After removing duplicates remaining records: 8417



```
ata_data=ata_data.drop_duplicates()

ata_data.duplicated().value_counts()

False    8417
dtype: int64
```

## Mojibake

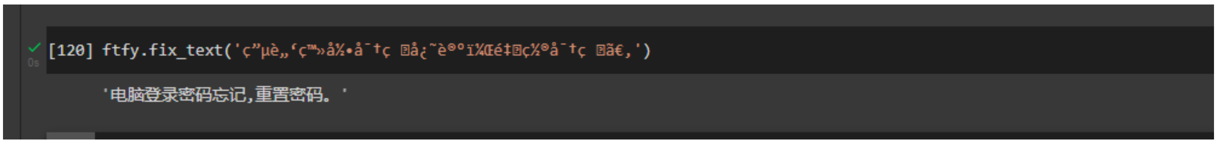
The base data has presence of scrambled text called Mojibake. Example given below. It occurs when we try to read text in some other encodings



```
('ç”µè„‘ç™»å½•å¯†ç □å¿˜è®°ï¼Œé‡□ç½®å¯†ç □ã€‚')
```

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

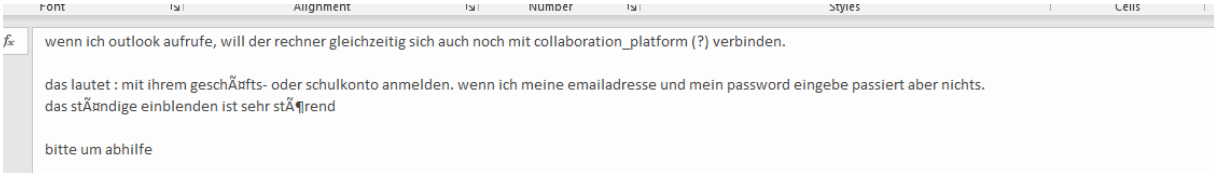
**Package FTFY** is used to clean the Mojibake text. The below code snippet shows that Mojibake texts are indeed non-english text



```
[120] ftfy.fix_text('ç”µè„‘ç™»å½•å¯†ç □å¿˜è®°ï¼Œé‡□ç½®å¯†ç □ã€‚')

      '电脑登录密码忘记,重置密码。'
```

## 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

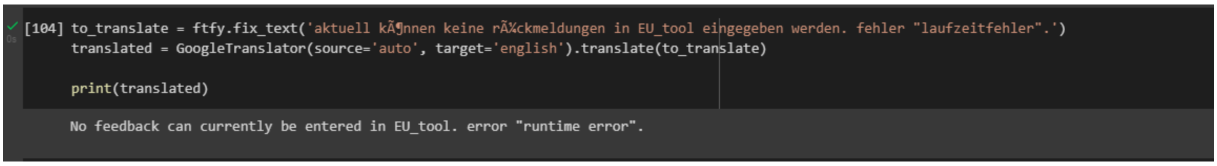


wenn ich outlook aufrufe, will der rechner gleichzeitig sich auch noch mit collaboration_platform (?) verbinden.

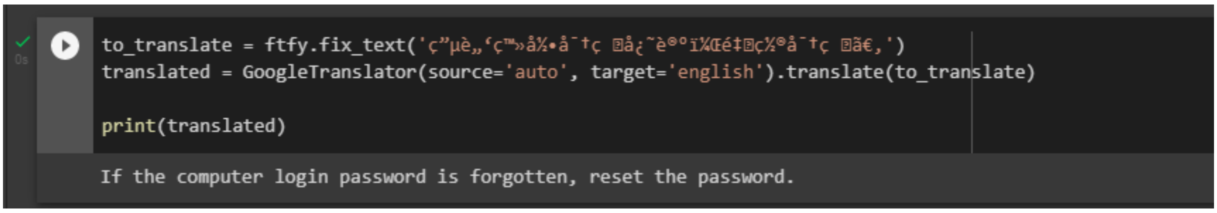
das lautet : mit ihrem geschÃ¤fts- oder schulkonto anmelden. wenn ich meine emailadresse und mein password eingebe passiert aber nichts.
das stÃ¤ndige einblenden ist sehr stÃ¶rend

bitte um abhilfe

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



```
[104] to_translate = ftfy.fix_text('aktuell kÃ¶nnen keine rÃ¼ckmeldungen in EU_tool eingegeben werden. fehler "laufzeitfehler".')
      translated = GoogleTranslator(source='auto', target='english').translate(to_translate)

      print(translated)

      No feedback can currently be entered in EU_tool. error "runtime error".
```

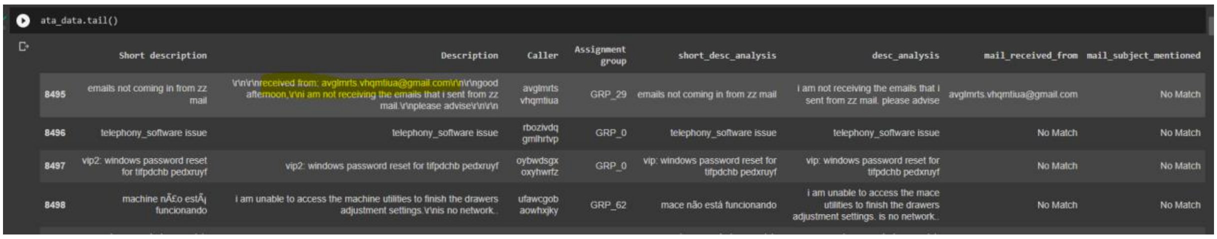


```
to_translate = ftfy.fix_text('ç”µè„‘ç™»å½•å¯†ç ▯å¿˜è®°ï¼Œé‡▯ç½®å¯†ç ▯ã€,')
translated = GoogleTranslator(source='auto', target='english').translate(to_translate)

print(translated)

If the computer login password is forgotten, reset the password.
```

## Checking for Patterns: "received from: *eylqgodm.ybqkwiam@gmail.com*"

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

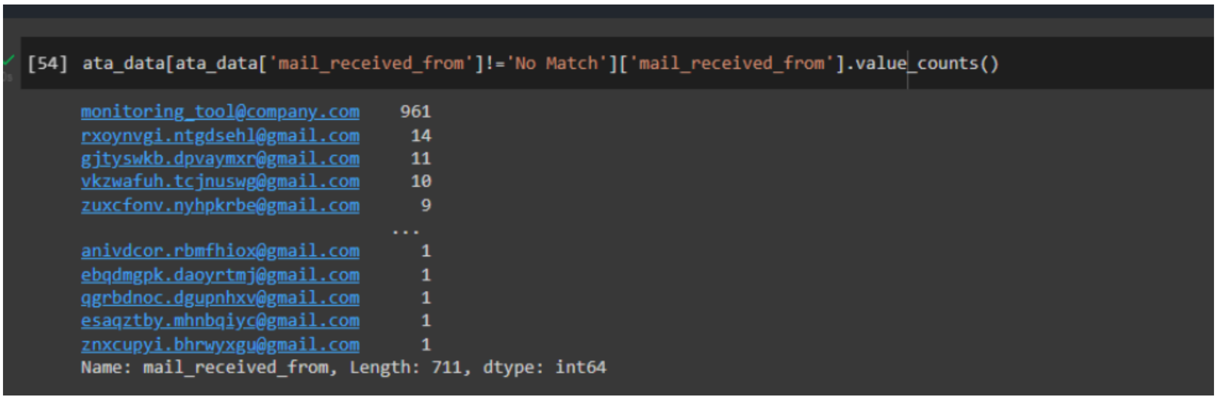
The pattern is removed from the text and the email id is added to another mail_received_from



```
ata_data.tail()
```

| | Short description | Description | Caller | Assignment group | short_desc_analysis | desc_analysis | mail_received_from | mail_subject_mentioned |
|---|---|---|---|---|---|---|---|---|
| 8495 | emails not coming in from zz mail | \r\n\r\nreceived from: avglmrts.vhqmtiua@gmail.com\r\n\r\ngood afternoon,\r\ni am not receiving the emails that i sent from zz mail.\r\nplease advise\r\n\r\n | avglmrts vhqmtiua | GRP_29 | emails not coming in from zz mail | i am not receiving the emails that i sent from zz mail. please advise | avglmrts.vhqmtiua@gmail.com | No Match |
| 8496 | telephony_software issue | telephony_software issue | rbozivdq gmlhrtvp | GRP_0 | telephony_software issue | telephony_software issue | No Match | No Match |
| 8497 | vip2: windows password reset for tifpdchb pedxruyf | vip2: windows password reset for tifpdchb pedxruyf | oybwdsgx oxyhwrfz | GRP_0 | vip: windows password reset for tifpdchb pedxruyf | vip: windows password reset for tifpdchb pedxruyf | No Match | No Match |
| 8498 | machine nÃ£o estÃ¡ funcionando | i am unable to access the machine utilities to finish the drawers adjustment settings.\r\nis no network. | ufawcgob aowhxjky | GRP_62 | mace não está funcionando | i am unable to access the mace utilities to finish the drawers adjustment settings. is no network. | No Match | No Match |

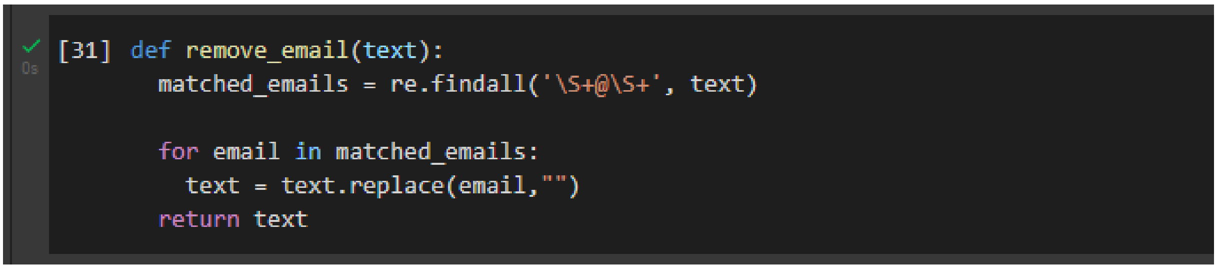
Analyzing the mails received from indicates that there are 961 records where initially issue was triggered due to system generated mail from monitoring_tool@company.com



```
[54] ata_data[ata_data['mail_received_from']!='No Match']['mail_received_from'].value_counts()

     monitoring_tool@company.com      961
     rxoynvgi.ntgdsehl@gmail.com       14
     gjtyswkb.dpvaymxr@gmail.com       11
     vkzwafuh.tcjnuswg@gmail.com       10
     zuxcfonv.nyhpkrbe@gmail.com        9
                                     ...
     anivdcor.rbmfhiox@gmail.com        1
     ebqdmgpk.daoyrtmj@gmail.com        1
     qgrbdnoc.dgupnhxv@gmail.com        1
     esaqztby.mhnbqiyc@gmail.com        1
     znxcupyi.bhrwyxgu@gmail.com        1
     Name: mail_received_from, Length: 711, dtype: int64
```

## Checking for Patterns: "email ids"

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



```
[31] def remove_email(text):
       matched_emails = re.findall('\S+@\S+', text)

       for email in matched_emails:
         text = text.replace(email,"")
       return text
```

## Checking for Patterns: "Mail Format"

Some of the Description have pattern like that of mail

(1) From

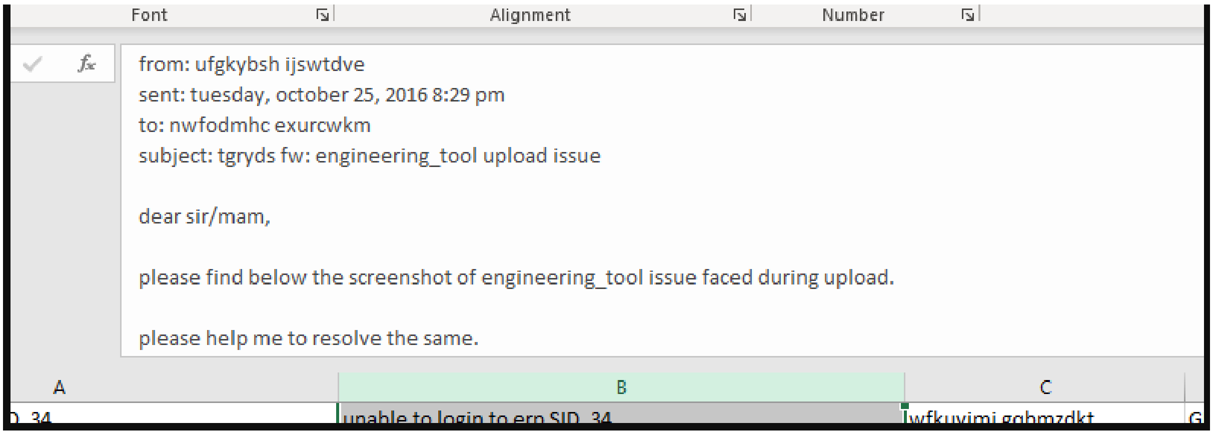
(2) Sent

(3) To

(4) Subject

(5) Cc

(6) importance



from: ufgkybsh ijswtdve
sent: tuesday, october 25, 2016 8:29 pm
to: nwfodmhc exurcwkm
subject: tgryds fw: engineering_tool upload issue

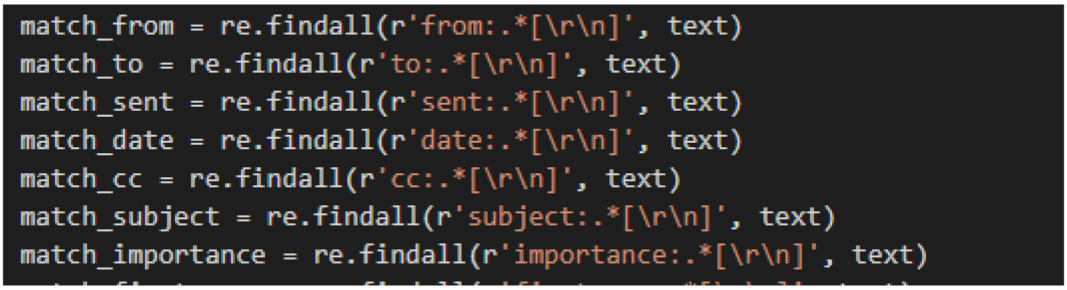
dear sir/mam,

please find below the screenshot of engineering_tool issue faced during upload.

please help me to resolve the same.

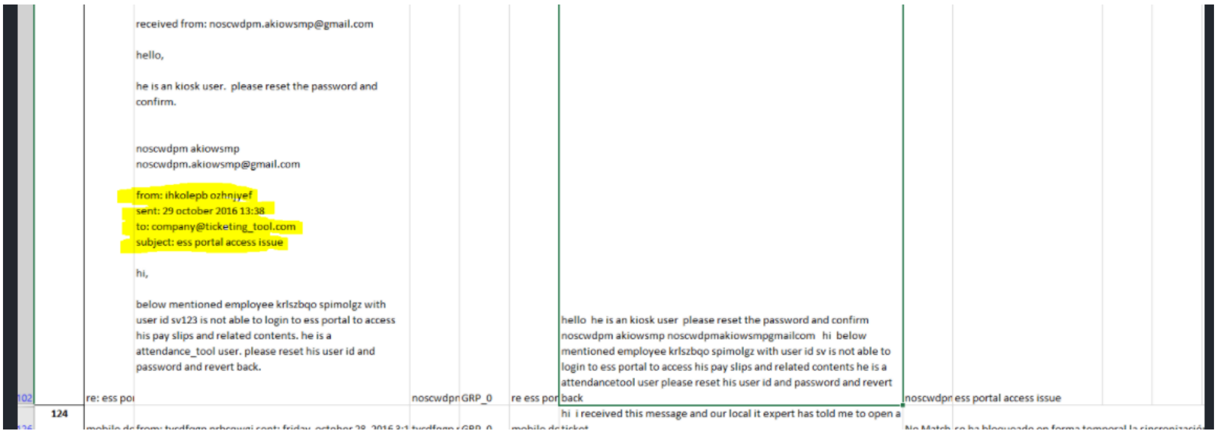
| A | B | C |
|---|---|---|
| D_34 | unable to login to erp SID_34 | wfkuvimi gqhmzdkt |

Removing such patterns using regex and replacing with blanks.  Also the subject is copied into another column mail_subject_mentioned



```
match_from = re.findall(r'from:.*[\r\n]', text)
match_to = re.findall(r'to:.*[\r\n]', text)
match_sent = re.findall(r'sent:.*[\r\n]', text)
match_date = re.findall(r'date:.*[\r\n]', text)
match_cc = re.findall(r'cc:.*[\r\n]', text)
match_subject = re.findall(r'subject:.*[\r\n]', text)
match_importance = re.findall(r'importance:.*[\r\n]', text)
```

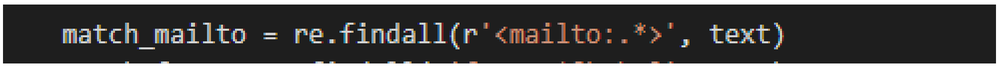
Clean text



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 102 | re: ess poi | received from: noscwdpm.akiowsmp@gmail.com<br>hello,<br>he is an kiosk user.  please reset the password and confirm.<br>noscwdpm akiowsmp<br>noscwdpm.akiowsmp@gmail.com<br>from: ihkolepb ozhnjyef<br>sent: 29 october 2016 13:38<br>to: company@ticketing_tool.com<br>subject: ess portal access issue<br>hi,<br>below mentioned employee krlszbqo spimolgz with user id sv123 is not able to login to ess portal to access his pay slips and related contents. he is a attendance_tool user. please reset his user id and password and revert back. | noscwdpn | GRP_0 | re ess por | hello  he is an kiosk user  please reset the password and confirm noscwdpm akiowsmp noscwdpmakiowsmpgmailcom   hi  below mentioned employee krlszbqo spimolgz with user id sv is not able to login to ess portal to access his pay slips and related contents he is a attendancetool user please reset his user id and password and revert back | noscwdpn | ess portal access issue |
| 124 | | | | | | hi  i received this message and our local it expert has told me to open a | | |

## Checking for Patterns: <mailto:>

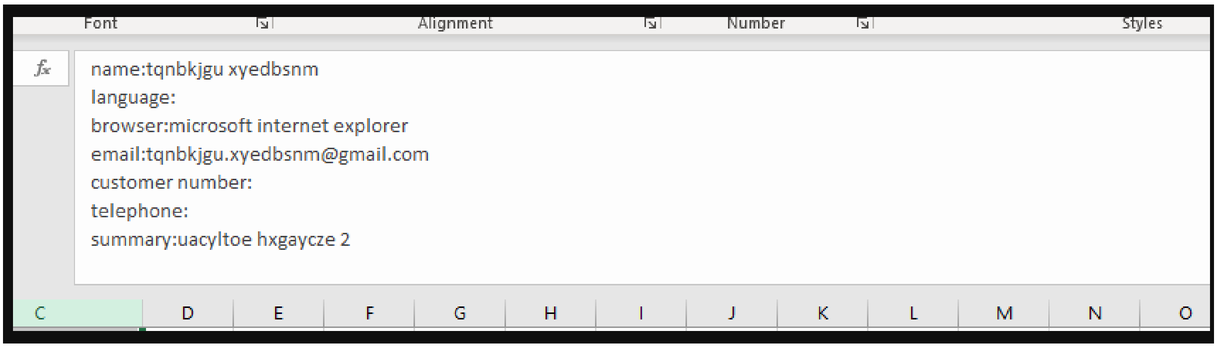
Removing occurrences of mailto: from the text



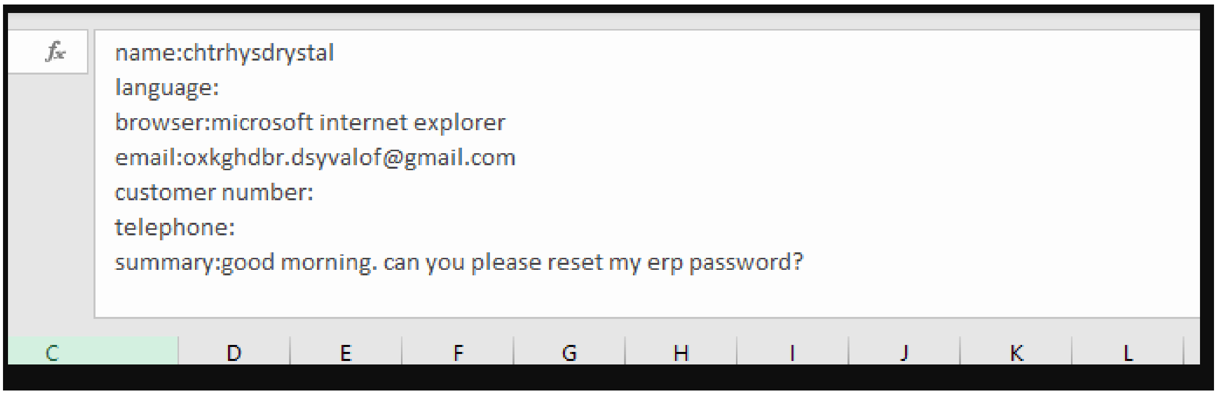
```
match_mailto = re.findall(r'<mailto:.*>', text)
```

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

The description column has a template for e.g.

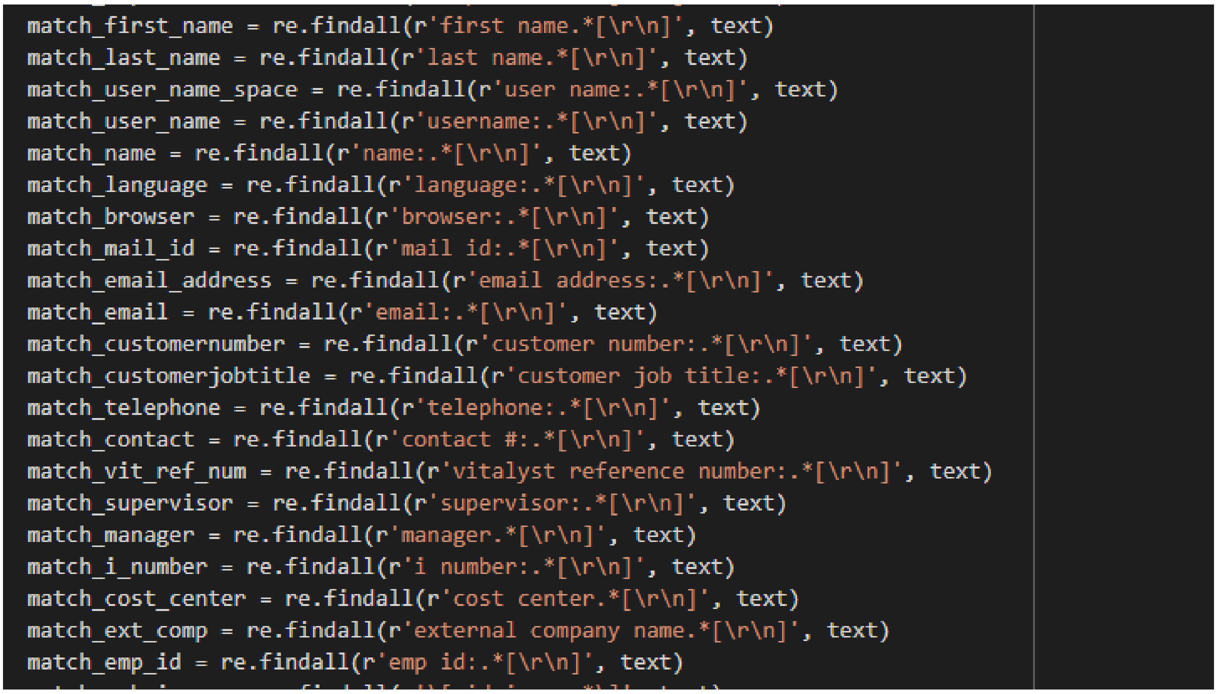


name:tqnbkjgu xyedbsnm
language:
browser:microsoft internet explorer
email:tqnbkjgu.xyedbsnm@gmail.com
customer number:
telephone:
summary:uacyltoe hxgaycze 2



name:chtrhysdrystal
language:
browser:microsoft internet explorer
email:oxkghdbr.dsyvalof@gmail.com
customer number:
telephone:
summary:good morning. can you please reset my erp password?

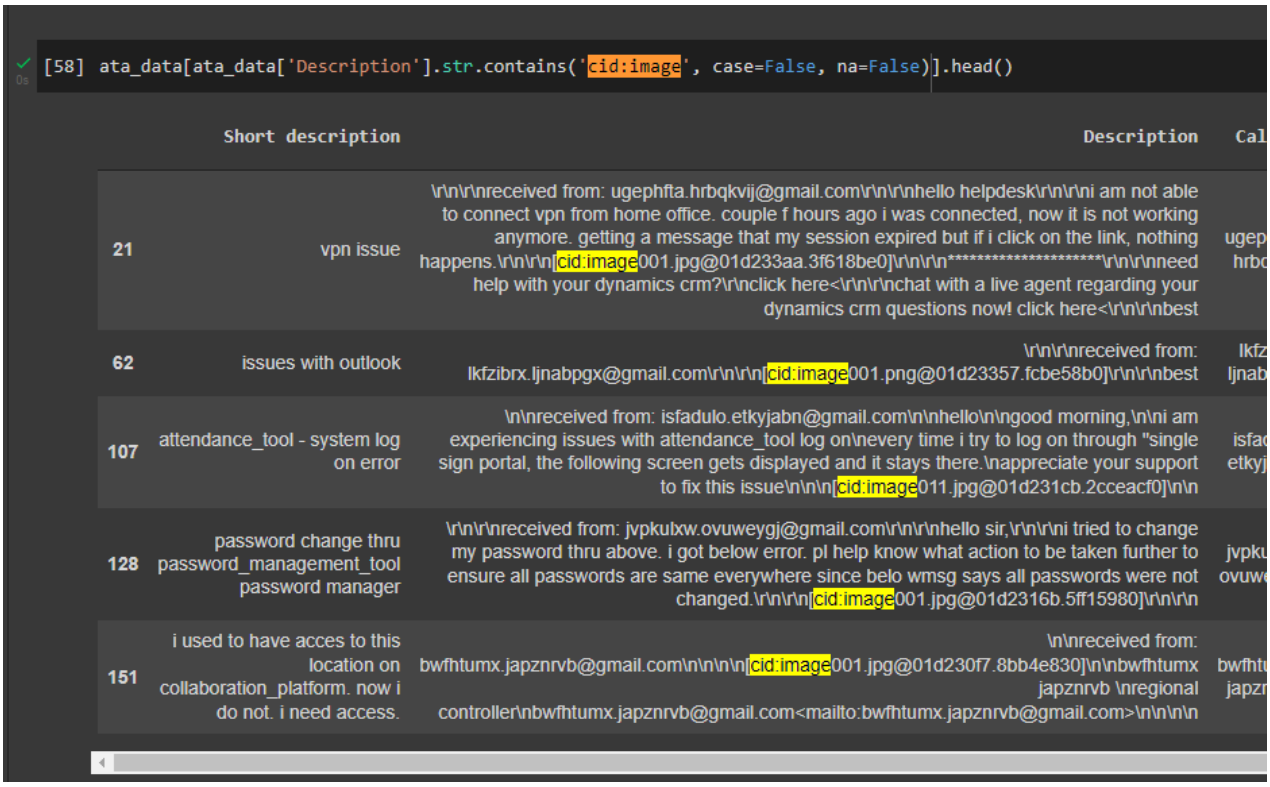
Removing such pattern using regex



```
match_first_name = re.findall(r'first name.*[\r\n]', text)
match_last_name = re.findall(r'last name.*[\r\n]', text)
match_user_name_space = re.findall(r'user name:.*[\r\n]', text)
match_user_name = re.findall(r'username:.*[\r\n]', text)
match_name = re.findall(r'name:.*[\r\n]', text)
match_language = re.findall(r'language:.*[\r\n]', text)
match_browser = re.findall(r'browser:.*[\r\n]', text)
match_mail_id = re.findall(r'mail id:.*[\r\n]', text)
match_email_address = re.findall(r'email address:.*[\r\n]', text)
match_email = re.findall(r'email:.*[\r\n]', text)
match_customernumber = re.findall(r'customer number:.*[\r\n]', text)
match_customerjobtitle = re.findall(r'customer job title:.*[\r\n]', text)
match_telephone = re.findall(r'telephone:.*[\r\n]', text)
match_contact = re.findall(r'contact #:.*[\r\n]', text)
match_vit_ref_num = re.findall(r'vitalyst reference number:.*[\r\n]', text)
match_supervisor = re.findall(r'supervisor:.*[\r\n]', text)
match_manager = re.findall(r'manager.*[\r\n]', text)
match_i_number = re.findall(r'i number:.*[\r\n]', text)
match_cost_center = re.findall(r'cost center.*[\r\n]', text)
match_ext_comp = re.findall(r'external company name.*[\r\n]', text)
match_emp_id = re.findall(r'emp id:.*[\r\n]', text)
```

## Checking for Patterns: Checking for embedded images text

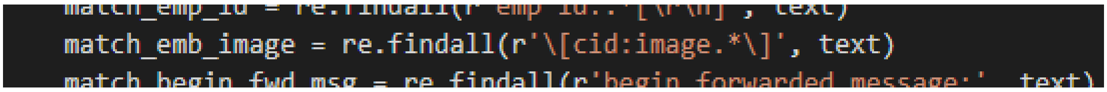
The data contains reference of embedded images as shown in the below image



```
[58] ata_data[ata_data['Description'].str.contains('cid:image', case=False, na=False)].head()
```

| | Short description | Description | Cal |
|---|---|---|---|
| 21 | vpn issue | \r\n\r\nreceived from: ugephfta.hrbqkvij@gmail.com\r\n\r\nhello helpdesk\r\n\r\ni am not able to connect vpn from home office. couple f hours ago i was connected, now it is not working anymore. getting a message that my session expired but if i click on the link, nothing happens.\r\n\r\n[cid:image001.jpg@01d233aa.3f618be0]\r\n\r\n********************\r\n\r\nneed help with your dynamics crm?\r\nclick here<\r\n\r\nchat with a live agent regarding your dynamics crm questions now! click here<\r\n\r\nbest | ugep<br>hrbq |
| 62 | issues with outlook | \r\n\r\nreceived from: lkfzibrx.ljnabpgx@gmail.com\r\n\r\n[cid:image001.png@01d23357.fcbe58b0]\r\n\r\nbest | lkfz<br>ljnab |
| 107 | attendance_tool - system log on error | \n\nreceived from: isfadulo.etkyjabn@gmail.com\n\nhello\n\ngood morning,\n\ni am experiencing issues with attendance_tool log on\nevery time i try to log on through "single sign portal, the following screen gets displayed and it stays there.\nappreciate your support to fix this issue\n\n\n[cid:image011.jpg@01d231cb.2cceacf0]\n\n | isfac<br>etkyj |
| 128 | password change thru password_management_tool password manager | \r\n\r\nreceived from: jvpkulxw.ovuweygj@gmail.com\r\n\r\nhello sir,\r\n\r\ni tried to change my password thru above. i got below error. pl help know what action to be taken further to ensure all passwords are same everywhere since belo wmsg says all passwords were not changed.\r\n\r\n[cid:image001.jpg@01d2316b.5ff15980]\r\n\r\n | jvpku<br>ovuwe |
| 151 | i used to have acces to this location on collaboration_platform. now i do not. i need access. | \n\nreceived from: bwfhtumx.japznrvb@gmail.com\n\n\n\n[cid:image001.jpg@01d230f7.8bb4e830]\n\nbwfhtumx japznrvb \nregional controller\nbwfhtumx.japznrvb@gmail.com<mailto:bwfhtumx.japznrvb@gmail.com>\n\n\n\n | bwfhtu<br>japzr |

Regex is used to replace such patterns with blanks



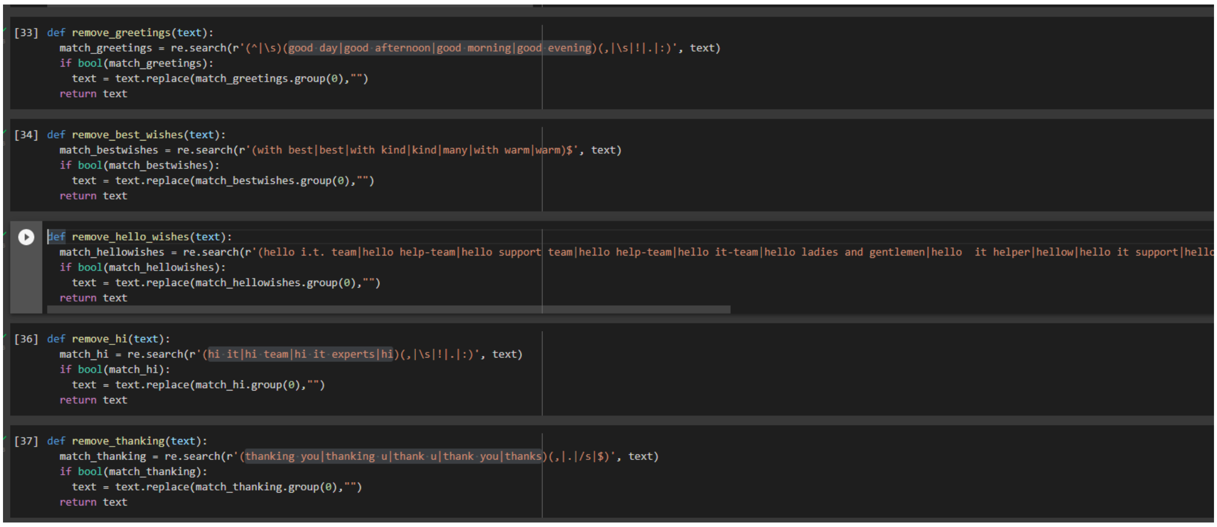
```
match_emp_id = re.findall(r'emp id:.*[\r\n]', text)
match_emb_image = re.findall(r'\[cid:image.*\]', text)
match_begin_fwd_msg = re.findall(r'begin forwarded message:', text)
```

## 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

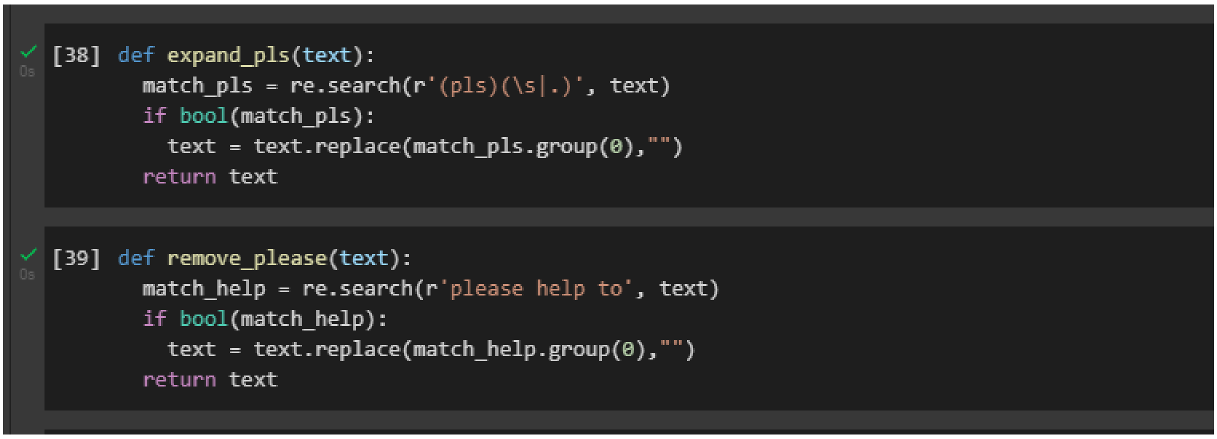
Such patterns are removed using regex



```
[33] def remove_greetings(text):
       match_greetings = re.search(r'(^|\s)(good day|good afternoon|good morning|good evening)(,|\s|!|.|:)', text)
       if bool(match_greetings):
         text = text.replace(match_greetings.group(0),"")
       return text

[34] def remove_best_wishes(text):
       match_bestwishes = re.search(r'(with best|best|with kind|kind|many|with warm|warm)$', text)
       if bool(match_bestwishes):
         text = text.replace(match_bestwishes.group(0),"")
       return text

     def remove_hello_wishes(text):
       match_hellowishes = re.search(r'(hello i.t. team|hello help-team|hello support team|hello help-team|hello it-team|hello ladies and gentlemen|hello  it helper|hellow|hello it support|hello
       if bool(match_hellowishes):
         text = text.replace(match_hellowishes.group(0),"")
       return text

[36] def remove_hi(text):
       match_hi = re.search(r'(hi it|hi team|hi it experts|hi)(,|\s|!|.|:)', text)
       if bool(match_hi):
         text = text.replace(match_hi.group(0),"")
       return text

[37] def remove_thanking(text):
       match_thanking = re.search(r'(thanking you|thanking u|thank u|thank you|thanks)(,|.|/s|$)', text)
       if bool(match_thanking):
         text = text.replace(match_thanking.group(0),"")
       return text
```

## 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

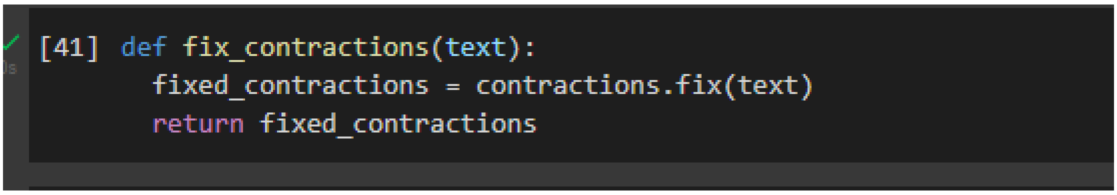


```
[38] def expand_pls(text):
       match_pls = re.search(r'(pls)(\s|.)', text)
       if bool(match_pls):
         text = text.replace(match_pls.group(0),"")
       return text
```

```
[39] def remove_please(text):
       match_help = re.search(r'please help to', text)
       if bool(match_help):
         text = text.replace(match_help.group(0),"")
       return text
```

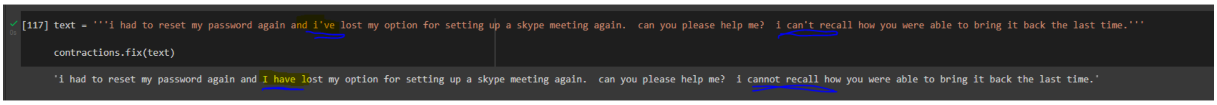
## Expanding contractions

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**



```
[41] def fix_contractions(text):
       fixed_contractions = contractions.fix(text)
       return fixed_contractions
```

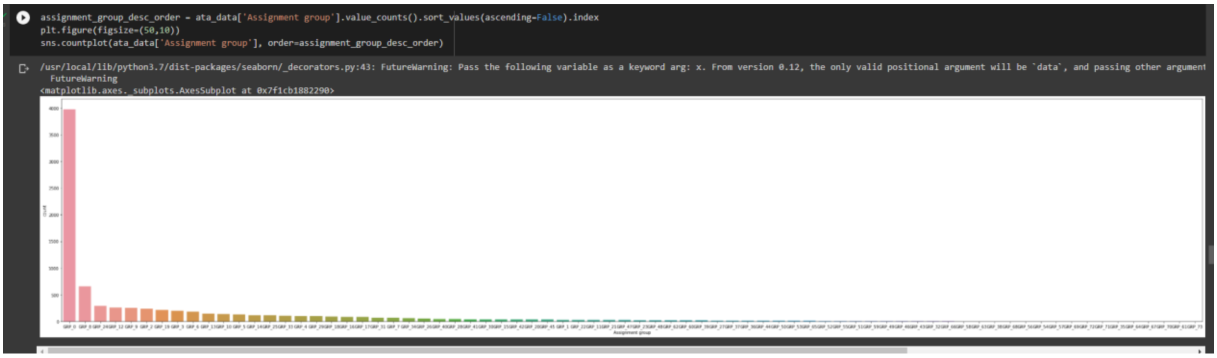
Applying contractions.fix



```
[117] text = '''i had to reset my password again and i've lost my option for setting up a skype meeting again.  can you please help me?  i can't recall how you were able to bring it back the last time.'''

      contractions.fix(text)

      'i had to reset my password again and I have lost my option for setting up a skype meeting again.  can you please help me?  i cannot recall how you were able to bring it back the last time.'
```

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

## Distribution of tickets as per Assignment Group



```
assignment_group_desc_order = ata_data['Assignment group'].value_counts().sort_values(ascending=False).index
plt.figure(figsize=(50,10))
sns.countplot(ata_data['Assignment group'], order=assignment_group_desc_order)
```

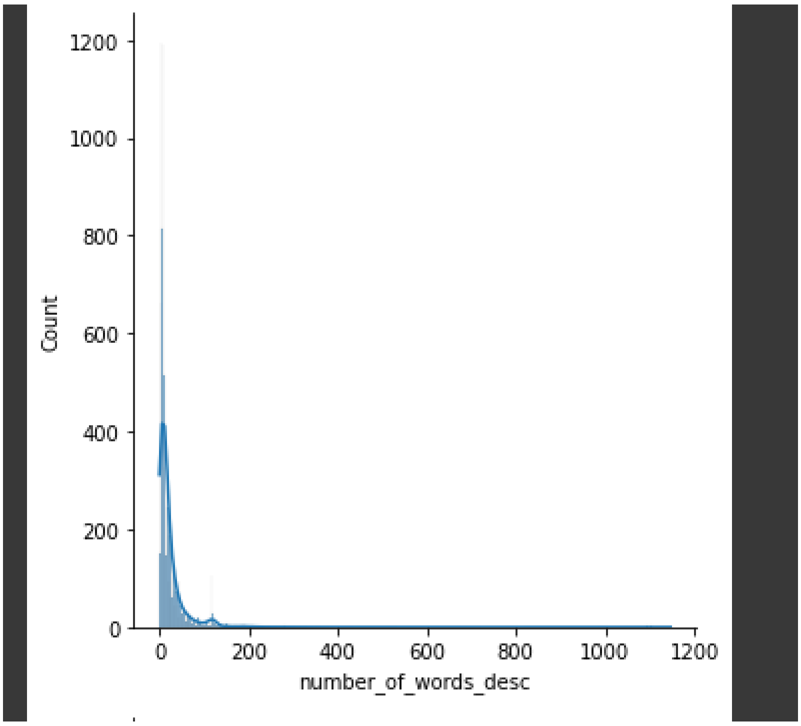
```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other argument
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7f1cb1882290>
```

Group_0 has 3976 tickets (roughly 47% of tickets). Dataset is imbalanced

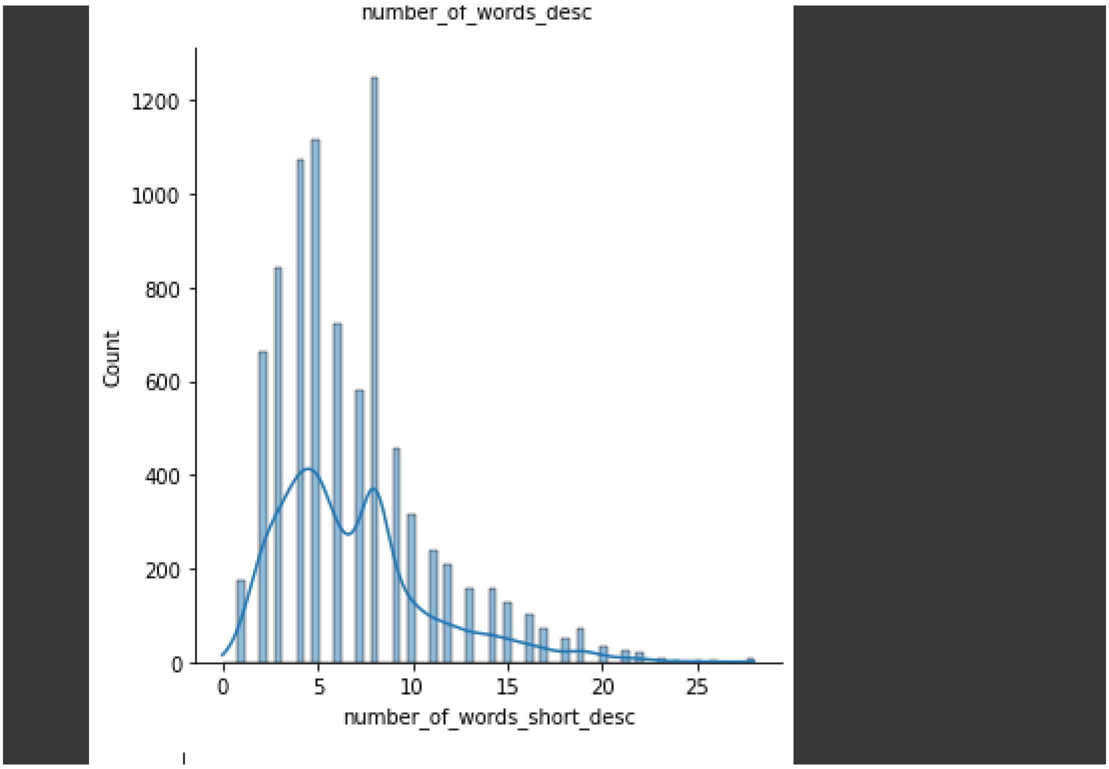
- no. of group less than 10 tickets- 25
- no. of groups in which records between 10 and 100 - 37
- no. of group greater than 100 tickets – 11

## Words/Characters analysis

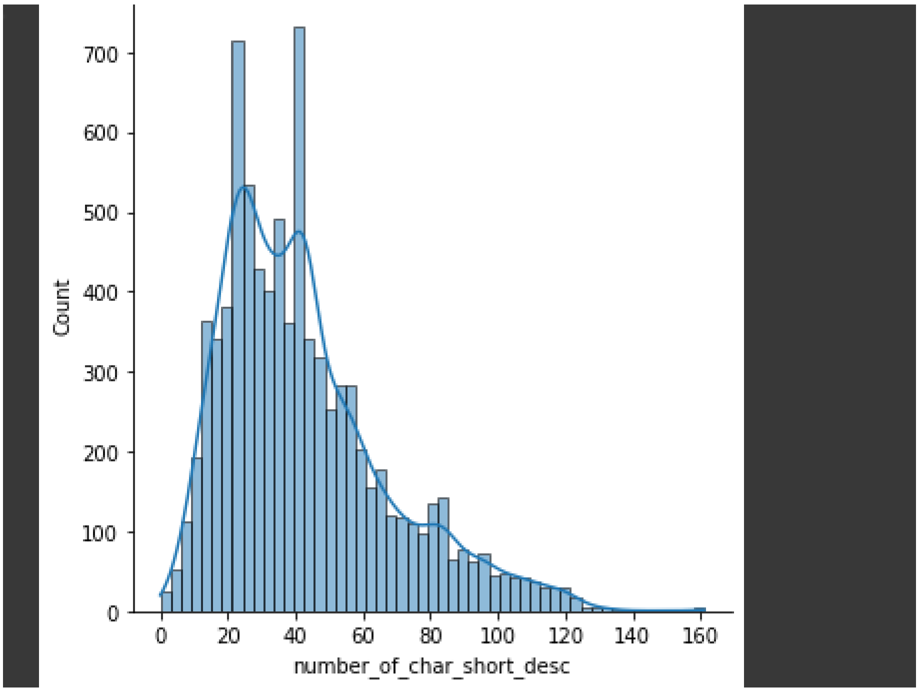
Majority of tickets have upto 100 words in Description. However, there are some tickets ranging from 200 – 1200 words



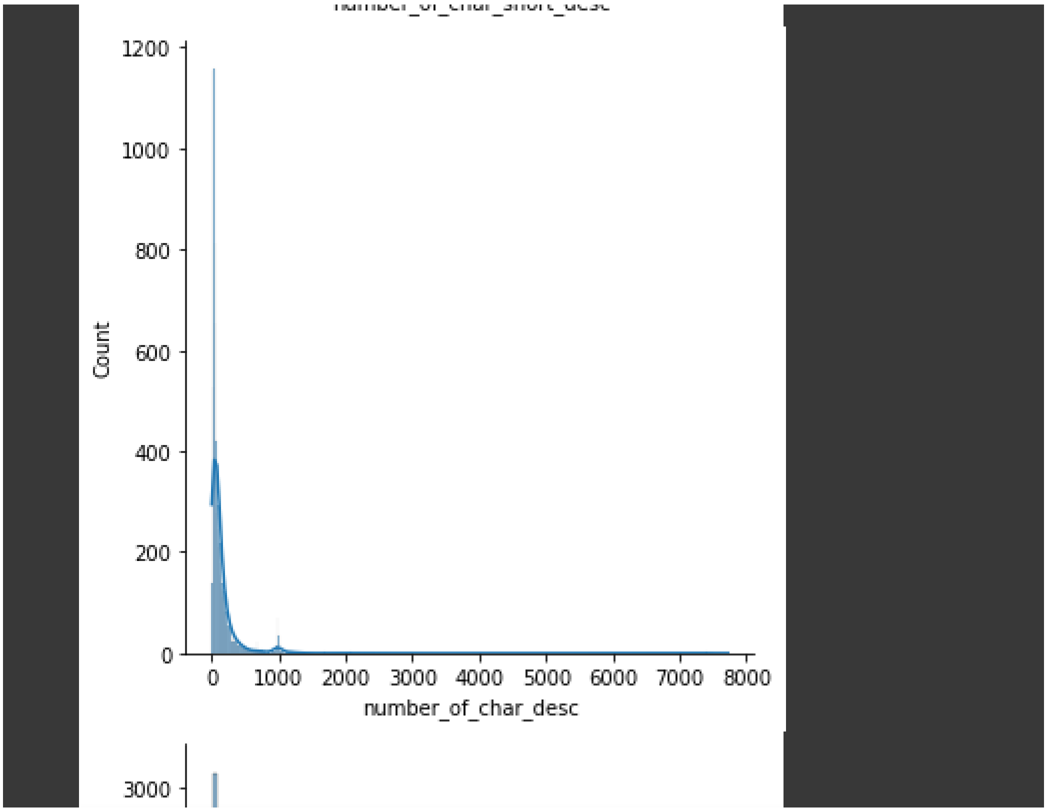
Short Description has a much compact word distribution



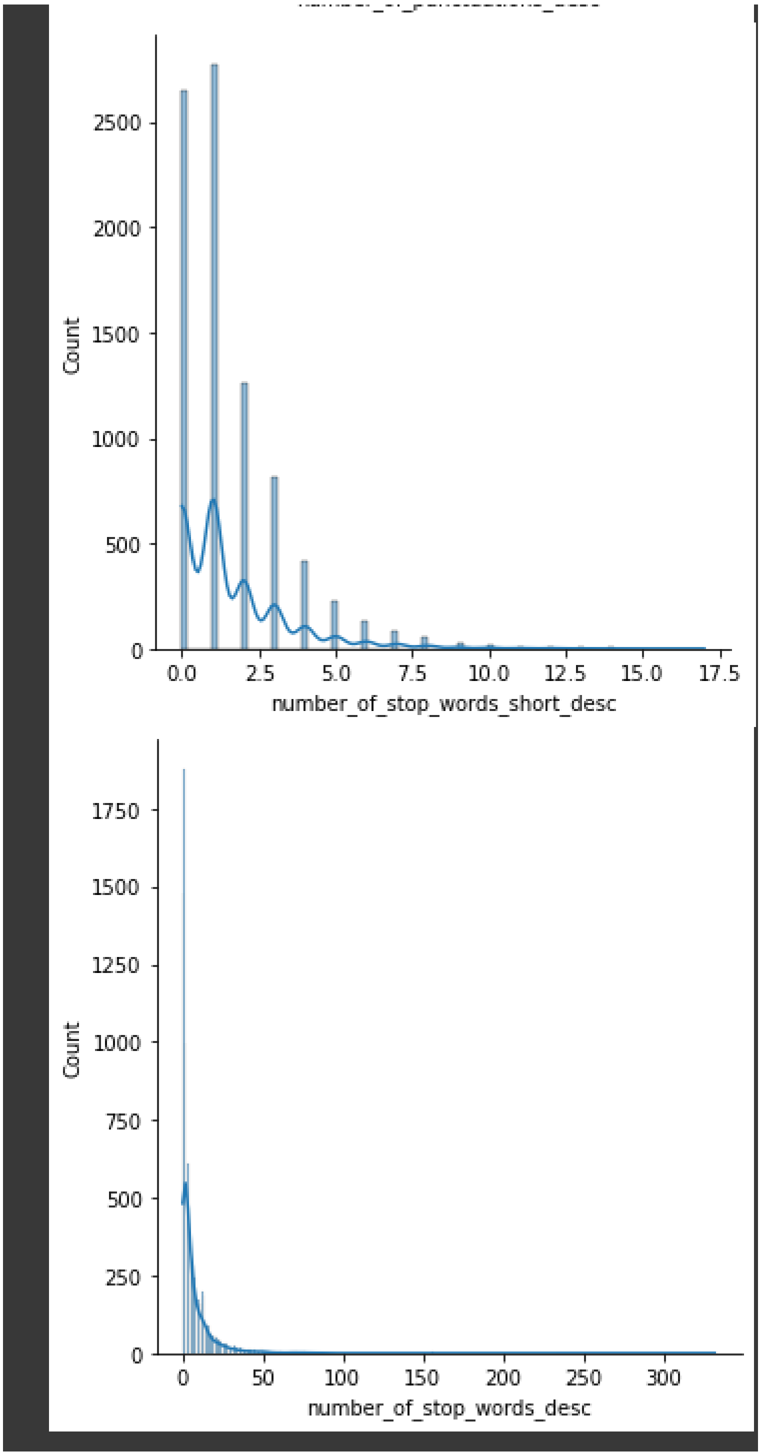
Character distribution for Short Description is also very compact



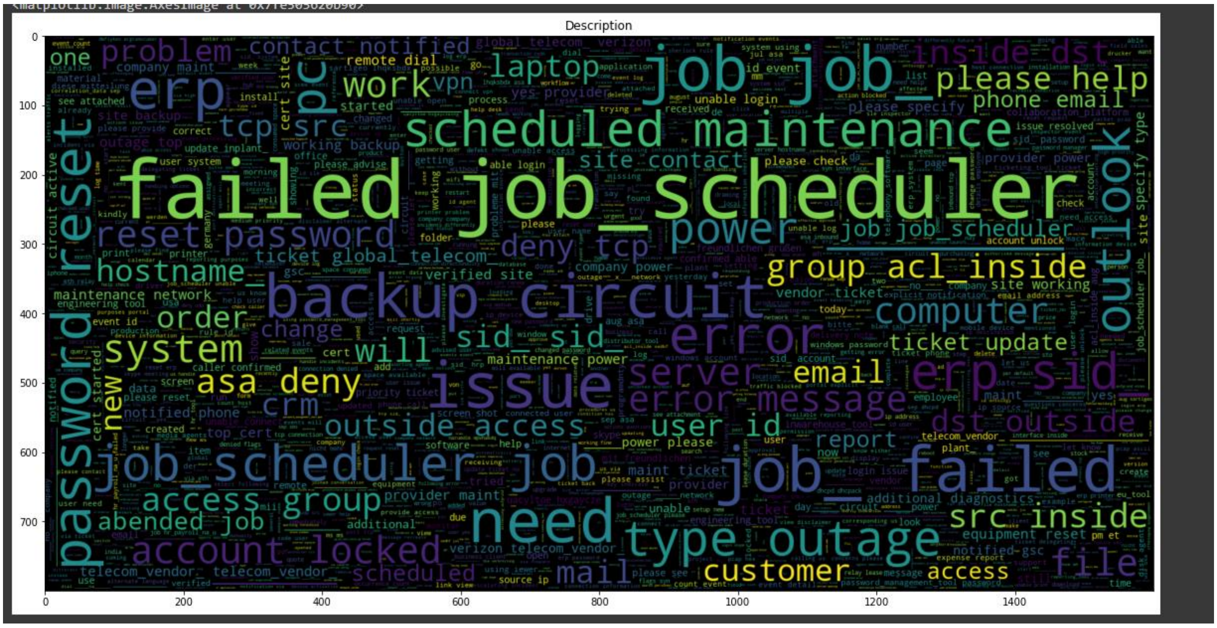
For description, in case of some tickets the character length ranges from 1000-8000 characters



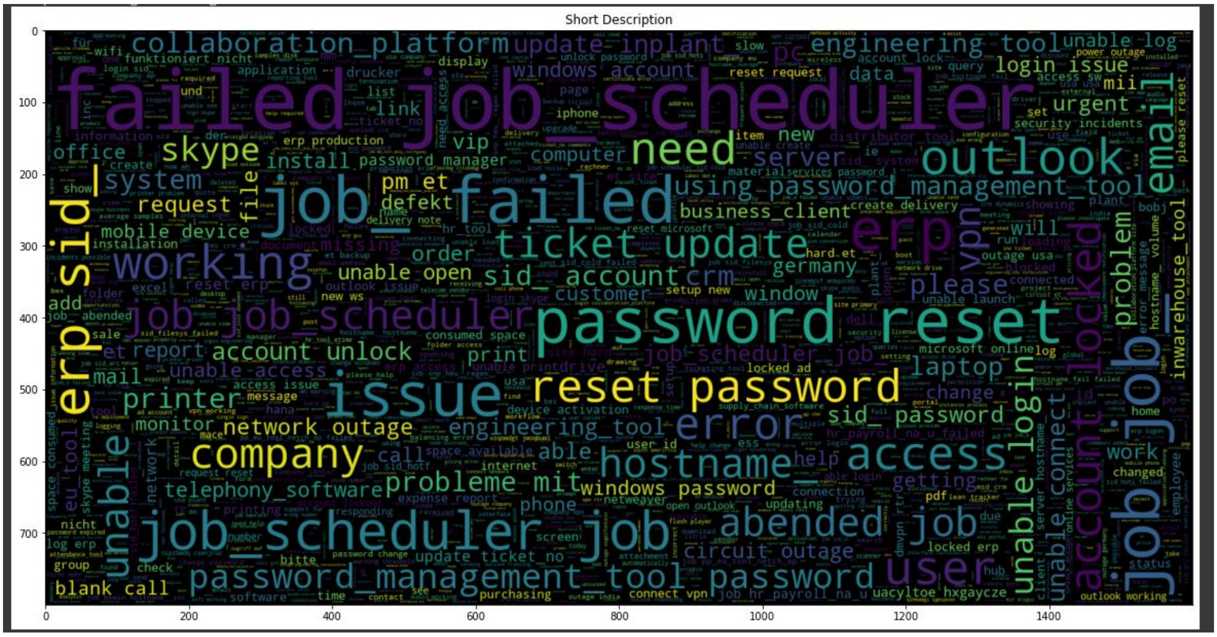
The presence of stop words is also significant in Description


Count
2500
2000
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
number_of_stop_words_short_desc
Count
1750
1500
1250
1000
750
500
250
0
0
50
100
150
200
250
300
number_of_stop_words_desc

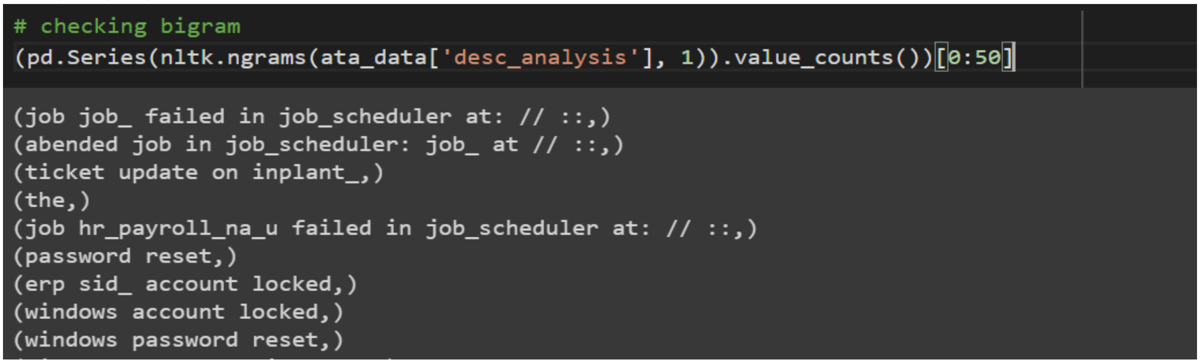
## Word Cloud of Description (whole data set) for max 5000 words



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

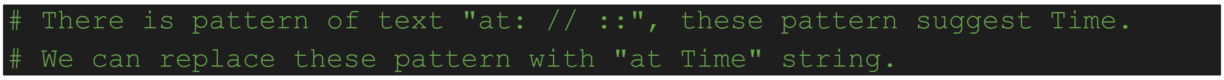


## Analyzing Ngram:



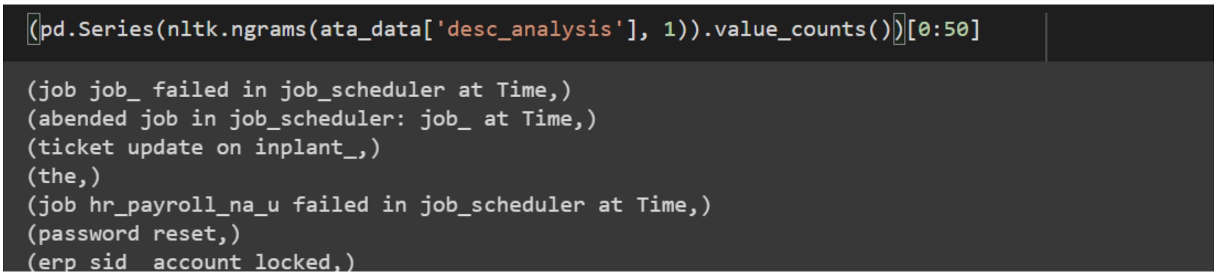
```
# checking bigram
(pd.Series(nltk.ngrams(ata_data['desc_analysis'], 1)).value_counts())[0:50]
```

```
(job job_ failed in job_scheduler at: // ::,)
(abended job in job_scheduler: job_ at // ::,)
(ticket update on inplant_,)
(the,)
(job hr_payroll_na_u failed in job_scheduler at: // ::,)
(password reset,)
(erp sid_ account locked,)
(windows account locked,)
(windows password reset,)
```



```
# There is pattern of text "at: // ::", these pattern suggest Time.
# We can replace these pattern with "at Time" string.
```

Replacing "at: // ::" pattern with “Time”



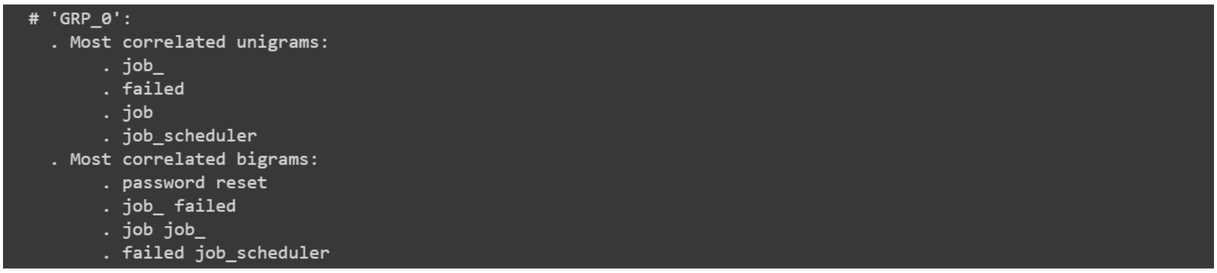
```
(pd.Series(nltk.ngrams(ata_data['desc_analysis'], 1)).value_counts())[0:50]

(job job_ failed in job_scheduler at Time,)
(abended job in job_scheduler: job_ at Time,)
(ticket update on inplant_,)
(the,)
(job hr_payroll_na_u failed in job_scheduler at Time,)
(password reset,)
(erp sid  account locked,)
```

Feature Engineering

Checking if any assignment group is related with any other assignment group

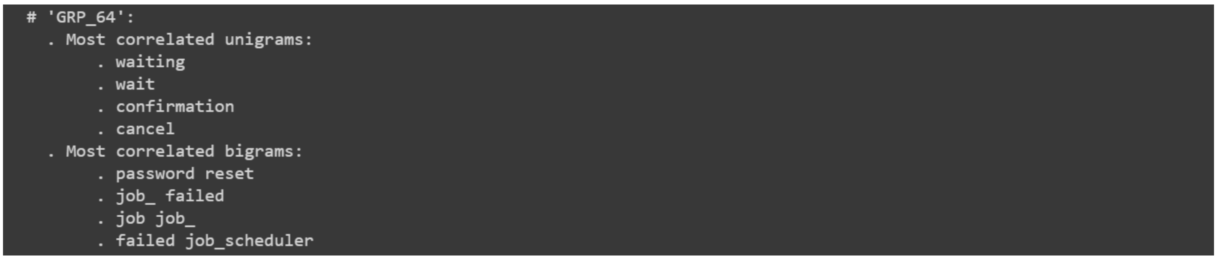
Finding correlated unigram and bigram between assignment group using Chi2 and TFIDF vectorization:

For GRP_0 below are most correlation unigram and bigram



```
# 'GRP_0':
  . Most correlated unigrams:
        . job_
        . failed
        . job
        . job_scheduler
  . Most correlated bigrams:
        . password reset
        . job_ failed
        . job job_
        . failed job_scheduler
```

For GRP_64, below are most correlated unigram and bigram, these are similar to GRP_0. These two group can be merged together.



```
# 'GRP_64':
  . Most correlated unigrams:
        . waiting
        . wait
        . confirmation
        . cancel
  . Most correlated bigrams:
        . password reset
        . job_ failed
        . job job_
        . failed job_scheduler
```

After applying above analysis (on both description and short description) on whole dataset we found these assignment groups can be merged together.

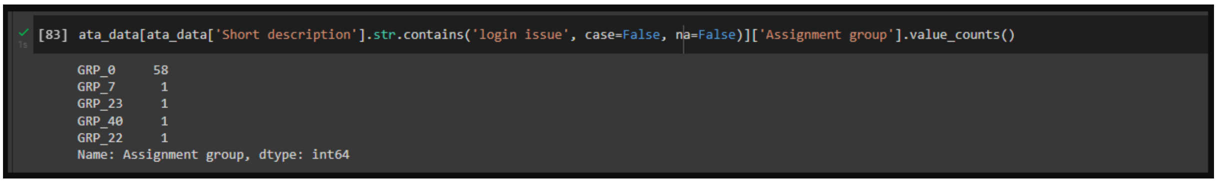
1. GRP_0,GRP_35,GRP_54,GRP_58,GRP_61,GRP_64,GRP_67,GRP_70,GRP_71,GRP_17,GRP_32, GRP_38,GRP_46,GRP_49,GRP_51,GRP_52,GRP_53,GRP_54,GRP_55,GRP_58,GRP_63,GRP_66
2. GRP_1,GRP_12,GRP_47,GRP_39
3. GRP_13,GRP_29
4. GRP_10,GRP_68

**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



```
[83] ata_data[ata_data['Short description'].str.contains('login issue', case=False, na=False)]['Assignment group'].value_counts()

     GRP_0     58
     GRP_7      1
     GRP_23     1
     GRP_40     1
     GRP_22     1
     Name: Assignment group, dtype: int64
```

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

Classification of text can be done either by: -

## Traditional ML Model

a. 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)

b. 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

a. Statistical approch
    a. Count Vectorizer
    b. TF-IDF
b. Predictive approch
    a. Word2Vec
    b. Glove

## Neural Network based Model

a. 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.

b. 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.

c. 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

a. Predictive approch
    a. Glove
b. Transformer based approch
    a. BERT

## Summary

We can do comparative study of both Traditional based model as well as Neural network based model to arrive at conclusion.

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 as NN based model can retain context over large number of sentences.

**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

## Use of Correlated unigram and bigram between assignment group

We intend to reduce the no of assignment group from existing 74 by using Chi2 and TFIDF feature extraction.

We will analyze model performance with Original dataset with 74 assignment groups as well as with reduced no of assignment group after applying correlation analysis.