Automatic IT Ticket Assignment

Project Notes 1

Nargess Ghahremani

Introduction

Nowadays the vast majority of companies and businesses have an IT department to support the technological infrastructure they need. Depending on the nature of the business, the IT department has to provide support for the employees or business end users and there are costs - often significant - associated with any issues which either of them may face. Therefore, responding to and solving the problems in a timely manner is critical. For an IT department to be able to resolve any issues efficiently, early detection of issues is paramount which is why most IT departments employ monitoring tools to identify any issues proactively, hopefully before they are noticed and reported by other users.

Once an issue is reported (by monitoring tools or users), the issue should be assigned to the right team/staff to be dealt with. Currently, in majority of cases, this is a manual process which if automated, could save significant time and money for the business as (a) it will assign the issue in the shortest time possible, (b) will save time from the support team, allowing them to help with resolution of issues and (c) if done well can be more accurate and as such saving resources which would have otherwise been wasted.

For this project, the current support process for the business is for the Level and Level 2 users to assign the issues they cannot resolve to groups in Level 3 support. They need to review the SOP process, spending at least 15 minutes per ticket and at least 1 full time employee need to be dedicated to this. Their accuracy in assigning the ticket correctly is around 75%. Therefore, if by automating this process we achieve a higher accuracy than 75%, the business will:

1. Ensure issues are not waiting in queues before they can be assigned
2. Save having to dedicate 1 employee to ticket assignment
3. Achieve better results in assigning tickets

Exploratory Data Analysis

Initial Observations

Initial observation of the data reveals the following.

- The data consists of 8500 entries and 4 columns: short description, description, caller and group assignment.
- It consists of tickets raised by users (people) and by monitoring tools. It is not clear how frequent the data was collected and over which period of time.
- **Input Variables**: The data is not clean or consistent. It contains:
    - a few blank fields
    - symbols (both readable and illegible)
    - email formats
    - filled form format
    - references to image files
    - hyperlinks
    - URL's
    - references to caller names
    - non-English languages

  all of which need to be cleaned or treated.
- **Target Variable**: Tickets are assigned to 74 groups in a highly imbalanced manner. More than 50% of the tickets are assigned to Group 0 which by an initial assessment seems to cover generic account/password issues as well as being the fall back for anything that is not otherwise assigned.

Data Pre-processing Steps

The following steps were taken to clean and pre-process the data for better understanding it and to prepare it for feed it to models.

1. Data Cleaning

- Convert to string (some fields were treated as float)
- Fill NA values
- Detect the language of the 'Description' field and anylse the outcome (consider only German as correctly detected non-English language)

- Clean:
    - Symbols
    - Digits
    - Email formats
    - HTML, hyperlinks, image file references
    - Caller names
    - Additional white spaces
- Merge 'Short description' and 'Description' fields and drop the former

2. Translation:

Translated German entries to English.

3. Remove Stop Words

Remove English Stopwords after adding the word 'please'. It was discovered by visualising the word distributions that 'please' was frequently appearing for some assignment groups while it clearly does not add any context.

4. Lemmatisation

Lemmatisation was chosen over stemming for better readability and compatibility with pre-trained models.

Data Analysis Steps

1. Topic Modelling

Forming dictionaries and corpus, then using both Bag of Words and TFIDF for LDA modelling yield the following 10 topics.

Topics using BOW

Topic	Terms
1	Tool, connect, Microsoft, screen, engineering, unable, internet, language, summary, name
2	Erp, account, sid, ticket, lock, update, team, create, inc, computer
3	Outside, software, usa, unlock, may, Germany, day, average, supply, sample

4	Access, yes, na, site, deny, circuit, power, company, vendor, network
5	Password, reset, id, outlook, tool, request, change, error, file, service
6	Job, scheduler, abended, fail, sid, drive, portal, reporting, hana, folder
7	Group, window, ip, call, order, phone, number, sale, problem, show
8	Issue, hostname, user, event, company, work, device, email, server, use
9	Unable, error, pc, plant, login, add, printer, print, per, warn
10	Inside, tcp, asa, dst, src, acl, aug, exe, jul, internal

Topics using TFIDF

Topic	**Terms**
1	Request, skype, log, screen, time, bex, report, set, let, meeting
2	Update, software, client, team, inc, computer, create, inplant, supply, chain
3	Access, error, open, use, try, hi, ethic, hr, load, reporting
4	Yes, na, site, telephony, circuit, power, backup, outage, warehouse, status
5	Outlook, collaboration, platform, launch, wifi, location, excel, lean, freeze, project
6	Job, abended, scheduler, fail, es, hana, etime, portal, net, dp
7	Password, reset, erp, tool, ticket, work, new, email, user, need
8	Account, sid, lock, unable, issue, login, window, hostname, setup, connect
9	Add, address, switch, freundlichen, sw, two, one, display, ip, ap
10	Mit, printer, usa, disk, problem, print, connection, audio, total, today

Although both models had similar coherence scores, BOW has segregated the topics better in terms of the inter-topic distances.

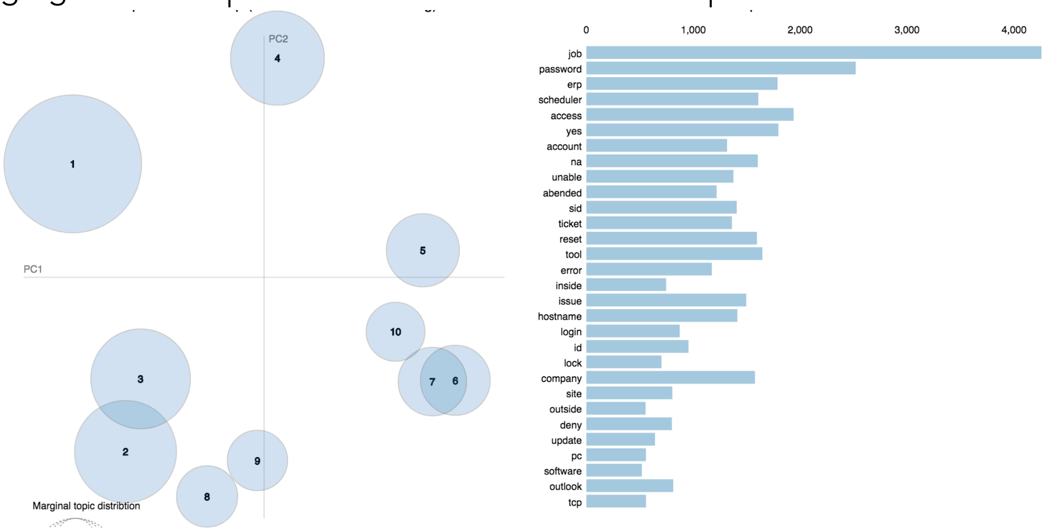


Figure 1- Topic Modelling using BOW

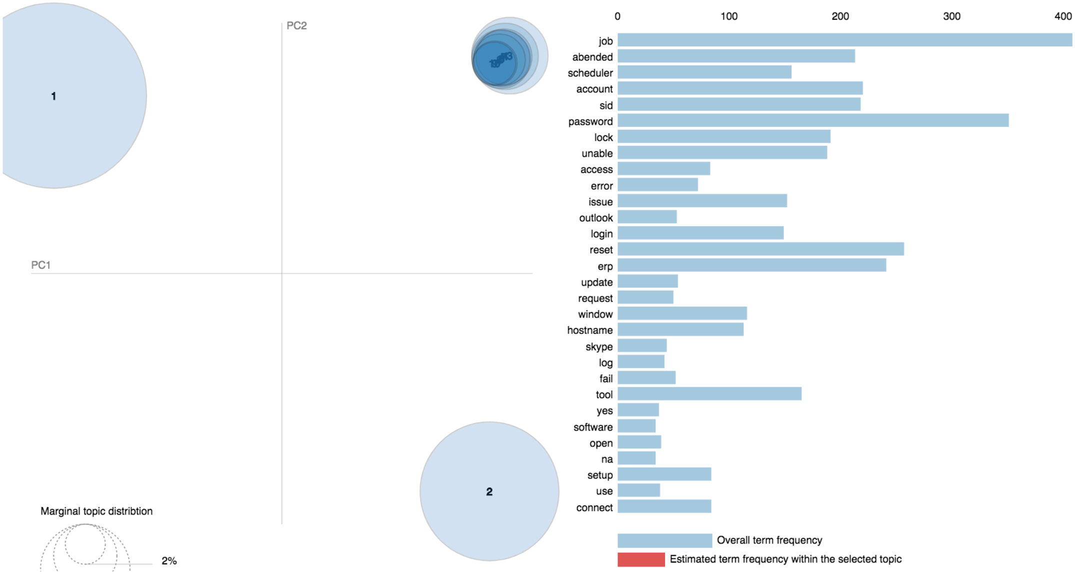


Figure 2 - Topic Modelling using TFIDF

2. Word distribution

Wordclouds of the whole corpus as well as those of the Assignment groups were created.

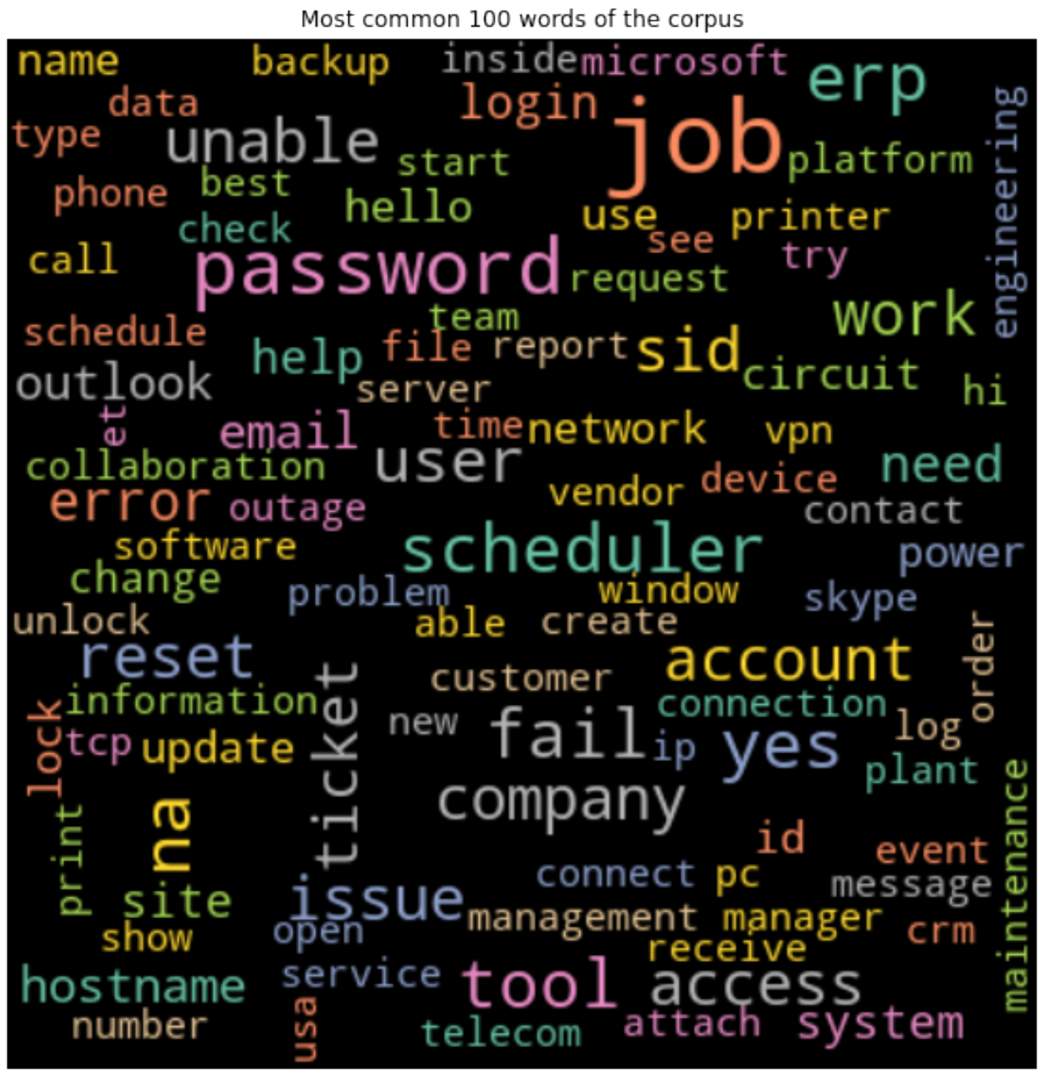
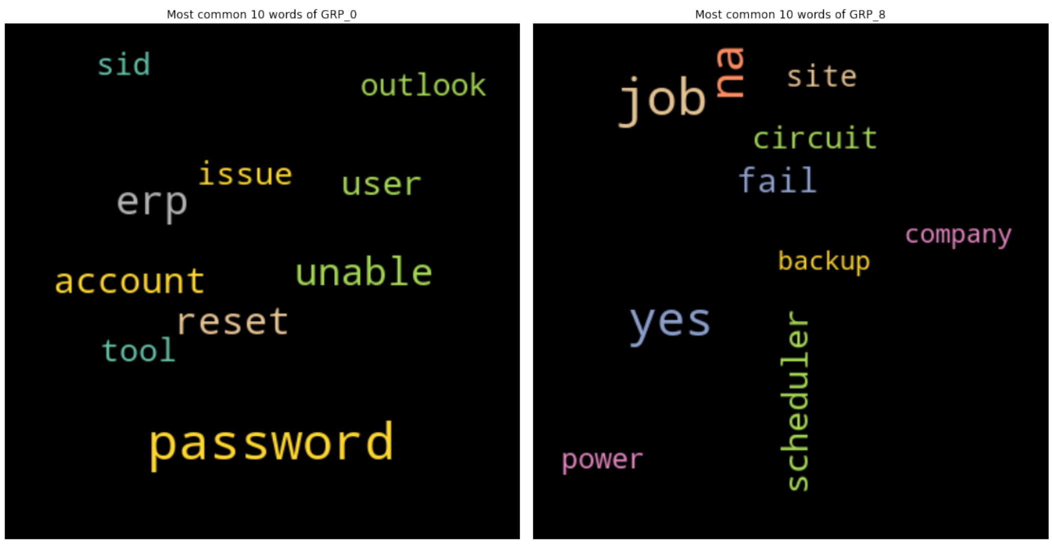


Figure 3 - Word Clouds



The 5 most assigned groups topic analysis is as follows.

Group	Some of the most frequent terms	Topic
Group 0	Password, unable, account, rese t, outlook, user, erp	Account, password, outlook
Group 8	Job, scheduler, circuit, fail, site, backup, power, company	Power, network, monitoring tools
Group 24	Problem, setup, calculator, printer, tool, new	Printer and tools issues
Group 12	Hostname, server, access, drive, deny, available, disk, space	Server and hardware issues
Group 9	Scheduler, job, failed	Monitoring tools

3. Treating imbalance in target variable

Groups with less than 10 tickets assigned were merged together and then the data was resampled to treat the imbalance between different groups.

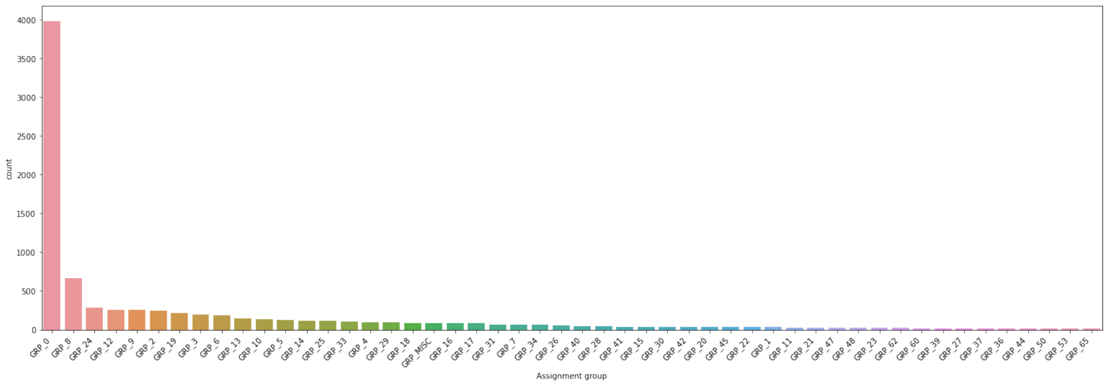


Figure 4 - Group Assignment Frequency

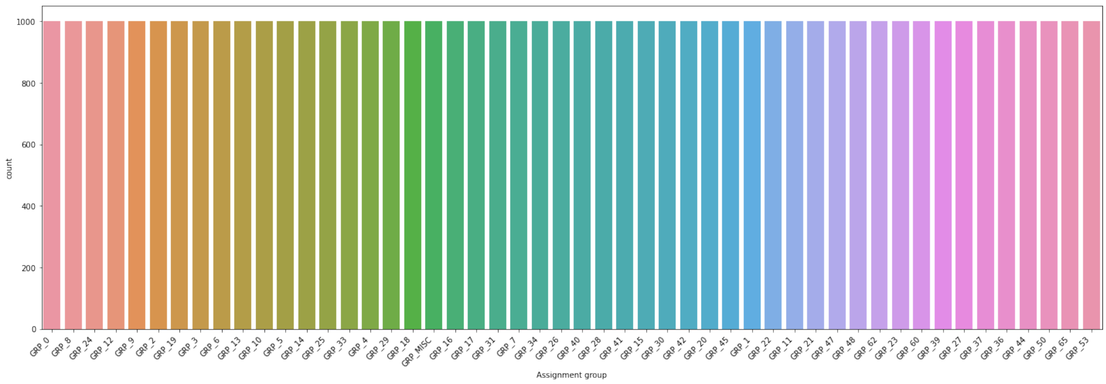


Figure 5 - Group Assignment Frequency after resampling

4. ‘Caller’ variable analysis

Analysing the ‘Caller’ input variable suggests that there are certain callers whose tickets are assigned to a smaller number of groups. We therefore will consider including the influential callers to see if they have an impact on our models’ performance.

Next steps would be to convert the input into word embeddings and prepare the data for feeding into various models and compare the results.