**Incident clusters**

Problem Statement: Cluster incidents, to identify common traits among different groups

Processing: Parsed symptoms through a novel parser to extract the body content from the message, and remove junk like signatures, header, repetitive footer text etc. Followed by tf-idf vectorization for textual data representation.

Clustering Techniques applied:

1. LDA (topic modelling)
2. K-means++
3. DBSCAN
4. Affinity Propagation

Results:

We are getting decent topics form LDA as mentioned below:

Topic 0:

heatsoftware com,following error,able access,access denied,support heatsoftware

Topic 1:

reset password,change password,self service,ip address,account locked

Topic 2:

heat software,getting error,distribution list,connect vpn,unable login

Topic 3:

make sure,service request,new hire,send email,thanks autoclosed

Topic 4:

ivanti com,lumension com,visual studio,unable connect,new laptop

Topic 5:

error message,need access,expense report,don know,sales order

Topic 6:

password reset,landesk com,received email,doesn work,new password

Topic 7:

let know,webex com,need able,know need,let know need

Topic 8:

email address,looks like,webex account,need help,ve tried

Topic 9:

skype business,soon possible,salesforce com,office 365,able log

However, when we check the cluster wise distribution of incidents, it is pretty skewed. Although the topics seems quite disjoint and inline with business expectations, one probable reason for skewness could be since most of the incidents contains words like “error message” hence, they are being grouped together. Also most of the incidents have Actual category as “Service Desk”, hence having a skewed distribution would not be much of a concern as it would be true representation of the data.

0 15219

5 1133

7 971

8 948

3 825

2 734

9 727

4 633

1 612

6 561

Kmeans++ :

Top terms per cluster:

Cluster 0: error message,heatsoftware com,reset password,heat software,ivanti com,

Cluster 1: heatsoftware com,support heatsoftware,support heatsoftware com,let know,email address,

Cluster 2: webex com,heatsoftware webex com,heatsoftware webex,com site,com account,

Cluster 3: let know,know need,let know need,know questions,let know questions,

Cluster 4: create new,ivanti com,office 365,looks like,following error,

Cluster 5: heat software,service management,self service,heatsoftware com,expense reports,

Cluster 6: maintenance support,heat software,service management,email address,support heatsoftware,

Cluster 7: need access,let know,heatsoftware com,corp domain,new hire,

Cluster 8: email address,ivanti com,com email,heatsoftware com,landesk com,

Cluster 9: lumension com,heatsoftware com,distribution list,service request,user password,

Here the results are similar to LDA.

DBSCAN:

Estimated number of clusters: 103

Silhouette Coefficient: 0.788

High silhouette co-efficient with high number of clusters, suggest there are a high number of outliers or low frequency clusters, explaining the skewness in previous methods.

Further Research:

Try out different methods for text representation, like Word2Vec, GLOVE, etc.