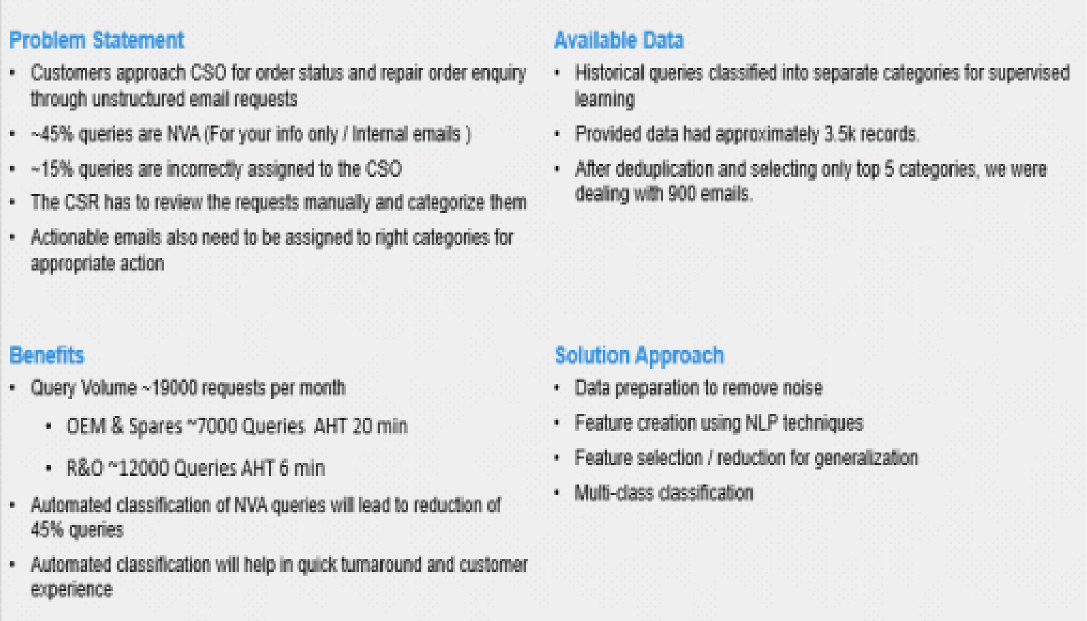
Intro

Thursday, October 18, 2018 9:14 AM

14 yrs + in industry
Did not start as Analytics professional
Mainframe -ETL-Big Data-DS
Due to onshore could not do full time professional program, completed it in India
Initially DS was part of my project scope but not the only scope and hence to get more exposure moved to individual contributor role confined to DS only in AI unit
Have got certified in Big Data, Microsoft certified DS, Automation Anwhere

############
Various classification and Regression problems .Offlate Text analytics
POCs on CRM 360 data in past. Also have done projects a part of academics

###########
Email Classification:-
7500
FYI/Internal/Quotation/shipment/expedited



Problem Statement

- Customers approach CSO for order status and repair order enquiry through unstructured email requests
- ~45% queries are NVA (For your info only / Internal emails)
- ~15% queries are incorrectly assigned to the CSO
- The CSR has to review the requests manually and categorize them
- Actionable emails also need to be assigned to right categories for appropriate action

Available Data

- Historical queries classified into separate categories for supervised learning
- Provided data had approximately 3.5k records.
- After deduplication and selecting only top 5 categories, we were dealing with 900 emails.

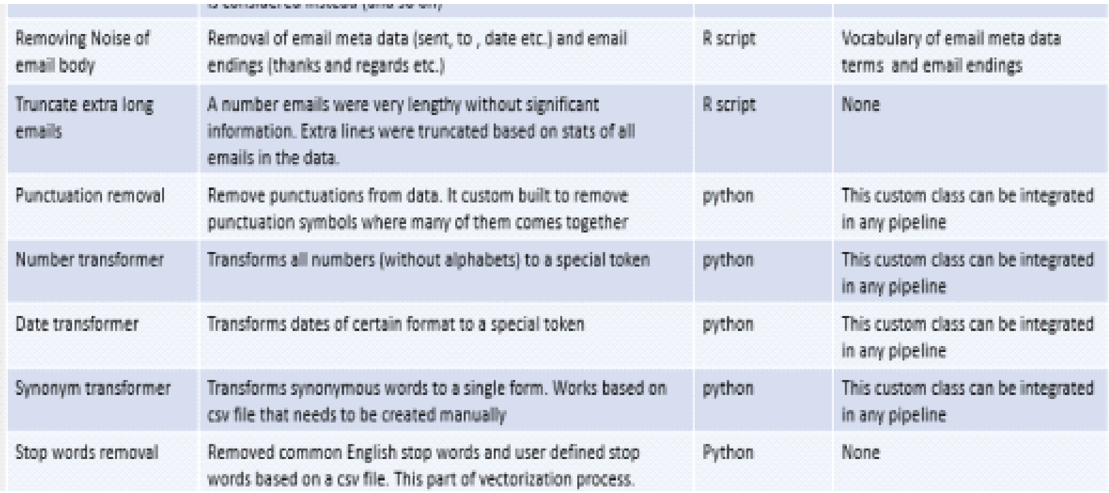
Benefits

- Query Volume ~19000 requests per month
  - OEM & Spares ~7000 Queries AHT 20 min
  - R&O ~12000 Queries AHT 6 min
- Automated classification of NVA queries will lead to reduction of 45% queries
- Automated classification will help in quick turnaround and customer experience

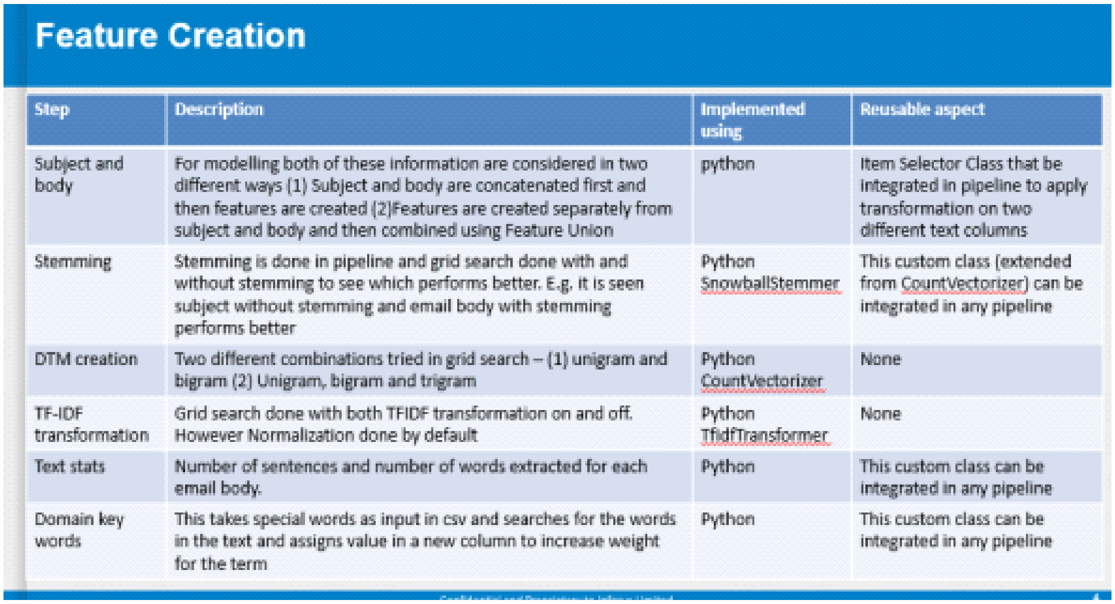
Solution Approach

- Data preparation to remove noise
- Feature creation using NLP techniques
- Feature selection / reduction for generalization
- Multi-class classification

Noise Removal



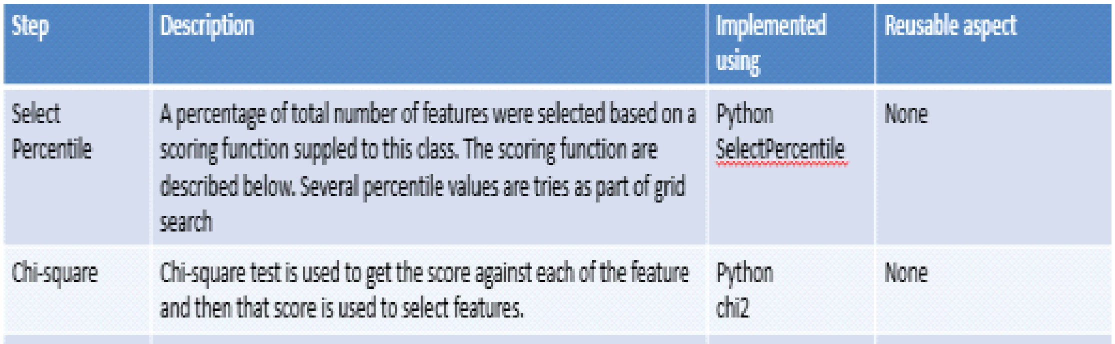
	is considered instead (and so on)		
Removing Noise of email body	Removal of email meta data (sent, to , date etc.) and email endings (thanks and regards etc.)	R script	Vocabulary of email meta data terms and email endings
Truncate extra long emails	A number emails were very lengthy without significant information. Extra lines were truncated based on stats of all emails in the data.	R script	None
Punctuation removal	Remove punctuations from data. It custom built to remove punctuation symbols where many of them comes together	python	This custom class can be integrated in any pipeline
Number transformer	Transforms all numbers (without alphabets) to a special token	python	This custom class can be integrated in any pipeline
Date transformer	Transforms dates of certain format to a special token	python	This custom class can be integrated in any pipeline
Synonym transformer	Transforms synonymous words to a single form. Works based on csv file that needs to be created manually	python	This custom class can be integrated in any pipeline
Stop words removal	Removed common English stop words and user defined stop words based on a csv file. This part of vectorization process.	Python	None



Feature Creation

Step	Description	Implemented using	Reusable aspect
Subject and body	For modelling both of these information are considered in two different ways (1) Subject and body are concatenated first and then features are created (2)Features are created separately from subject and body and then combined using Feature Union	python	Item Selector Class that be integrated in pipeline to apply transformation on two different text columns
Stemming	Stemming is done in pipeline and grid search done with and without stemming to see which performs better. E.g. it is seen subject without stemming and email body with stemming performs better	Python SnowballStemmer	This custom class (extended from CountVectorizer) can be integrated in any pipeline
DTM creation	Two different combinations tried in grid search – (1) unigram and bigram (2) Unigram, bigram and trigram	Python CountVectorizer	None
TF-IDF transformation	Grid search done with both TFIDF transformation on and off. However Normalization done by default	Python TfidfTransformer	None
Text stats	Number of sentences and number of words extracted for each email body.	Python	This custom class can be integrated in any pipeline
Domain key words	This takes special words as input in csv and searches for the words in the text and assigns value in a new column to increase weight for the term	Python	This custom class can be integrated in any pipeline

Feature selection



Step	Description	Implemented using	Reusable aspect
Select Percentile	A percentage of total number of features were selected based on a scoring function suppled to this class. The scoring function are described below. Several percentile values are tries as part of grid search	Python SelectPercentile	None
Chi-square	Chi-square test is used to get the score against each of the feature and then that score is used to select features.	Python chi2	None

https://www.analyticsvidhya.com/blog/2016/02/bigmart-sales-solution-top-20/