



Text Analytics and Automation

Manoj Kumar Dobbali – u1053996
Sai Deepthi Matam – u1053997

Table of Contents

Executive Summary:	3
Project Overview:	4
Languages and Tools used:	4
Brief description about data:	4
Process:	4
Automation:	7
Entire Chat screen shot:	9
Results:	10
Lessons Learned:	12
Next Steps:	13

Executive Summary:

Currently, a customer selects the issue category who will then be directed to a chat agent. Some customers chose phone call over chat. Either way, there is a certain wait time until the customer can be connected to an agent and the time to verify the details and explain the issue. Most of this requires a human resource to attend to, which is where the third part agency comes in to play. Given the volume of chat requests and phone calls the customer service team takes every single day, it is obvious that the operational expense is high. The Customer Service team at Overstock is planning to implement various strategies to change the user experience. Few being - integrating various channels of communication, saving the customer history thus making it easier to pick up from where he/she left off, providing better and accurate self-support information. From this, Overstock can reduce the operational expense and ensure customer satisfaction. The initial step towards this is a more informative and extensive self-support/FAQ s page related to the issue description that customer provides. This can be achieved by analyzing the historical chat transcripts and identifying the Key Words that relate to each issue.

Project Overview:

Languages and Tools used:

Teradata, Python (scikit-learn, Natural Language Toolkit, Pandas) and Microsoft Excel, AWS, Slack Bot

Brief description about data:

A sample of 501, 351 chat transcripts were extracted each for the following categories. Few categories of issues are combined together to. For example, shipping policy and incomplete tracking are combined into “Shipping”

1. Returns (Return status, Refunds)
2. Shipping (Shipping policy, Invalid tracking, Incomplete tracking, damaged in transit, Damaged item, Never delivered, Wrong item delivered)
3. International
4. Orders (Cancel, Taxes)
5. Using your account (Changing account info, Accessing account)
6. Missing parts
7. Coupon (Coupon didn't work, Coupon restrictions/complaints, Promo coupon didn't work)
8. Club O
9. Store card
10. Mastercard
11. Payment

Process:

The data we are dealing with is a human conversation which implies a highly noisy data that would yield low accuracy no matter which algorithm we choose for our need. As expected, there are greetings, wishes, spelling and grammatical errors, locations, email addresses, contact information, names, time stamps, jargon, and a lot shortcut words from texting language. These

do not reflect the customer's issue in any way and thus do not contribute at all to the process. These were taken out during the data cleaning.

Data cleaning and preparation:

As mentioned, given the nature of text, majority of time and effort has been spent on data cleaning. This also involved manually digging in to the data to figure out patterns in the conversations related to names, time stamps, email addresses, contact information, locations and other authentication details. Later, the regular expressions were modified accordingly to remove these phrases. A customized stopwords lexicon is created with non-significant domain specific words, city/state/country names(occurring independently but not as part of location), jargon, greetings, product names, item numbers etc.,.

TF_IDF:

TF-IDF (Term Frequency - Inverse Document Frequency) has been considered initially. This was applied to unigram, bigram and trigram datasets to factor in Key words as well as Key Phrases in weighing(link to: <http://www.tfidf.com/>). However, despite the extensive cleaning efforts, the final vocabulary obtained had tens of thousands (hundreds of thousands in case of more popular categories) words.

Rapid Automatic Keyword Extraction:

A new approach, Rapid Automatic Keyword Extraction(RAKE), has been implemented which gave in more accurate and concise results. The basic foundation of this unsupervised method(link to: <https://www.mathworks.com/discovery/unsupervised-learning.html?requestedDomain=www.mathworks.com>) is the observation that keywords frequently contain multiple words but rarely contain punctuation marks or stopwords. Therefore, stop words and punctuations are considered as phrase delimiters. All the phrases between the phrase delimiters called "Candidate Keywords" are generated which are then scored across the documents (in this case chats)

In this process keywords have been extracted from customer's chat. Here keywords mean, words or phrases that are appearing frequently in the chat. This is very common problem working when working in the text. Keywords can give us an idea on what the entire chat is about without going through the entire chat transcript.

By definition, keywords describe the main topics expressed in a document. The terminology can get a little confusing, so the image below compares related tasks in terms of the source of terminology and number of topics selected per document.

There are two specific tasks in this process:

- 1.Cleaning the text data.
- 2.Extracting the most significant words and phrases that appear in given text

A typical keyword extraction algorithm has three main components:

Candidate selection: Initial step would be to extract all phrase that could possibly be keywords.

Properties calculation: For each candidate, selecting properties which indicate which keyword might be indicating the importance. Observe the keyword - "prepaid labels shipping". This is an example of "Adjoining Keywords"(pairs of keywords that adjoin one another). If the phrase "*prepaid labels for shipping*" occurs at least twice in the documents, then the phrase delimiter "for" is considered as an "interior stopword" and thus is not split in to two candidate keywords.

Scoring and selecting keywords: A scoring method is selected based upon our need. Here we used degree/frequency. The keyword score is then calculated for every candidate keyword based frequency and degree of member words ("return" and "policy" would be the member words for the candidate keyword "return policy")

Degree of member word ($\text{deg}(w)$) = number of longer (multi-word) candidate keywords it is present in

Frequency of member word ($\text{freq}(w)$) = number of times a word occurs in the entire corpus regardless of the number of words with which they co-occur

Score of the member word = $\text{deg}(w) / \text{freq}(w)$

Finally, parameters such as the minimum frequency of a candidate, its minimum and maximum length in words, or the stemmer used to normalize the candidates help tweak the algorithm's performance to a specific dataset.

To summarize, RAKE is a simple keyword extraction library which focuses on finding multi-word phrases containing frequent words. Its strengths are its simplicity and the ease of use, whereas its weaknesses are its limited accuracy, the parameter configuration requirement, and the fact that it throws away many valid phrases and doesn't normalize candidates.

Automation:

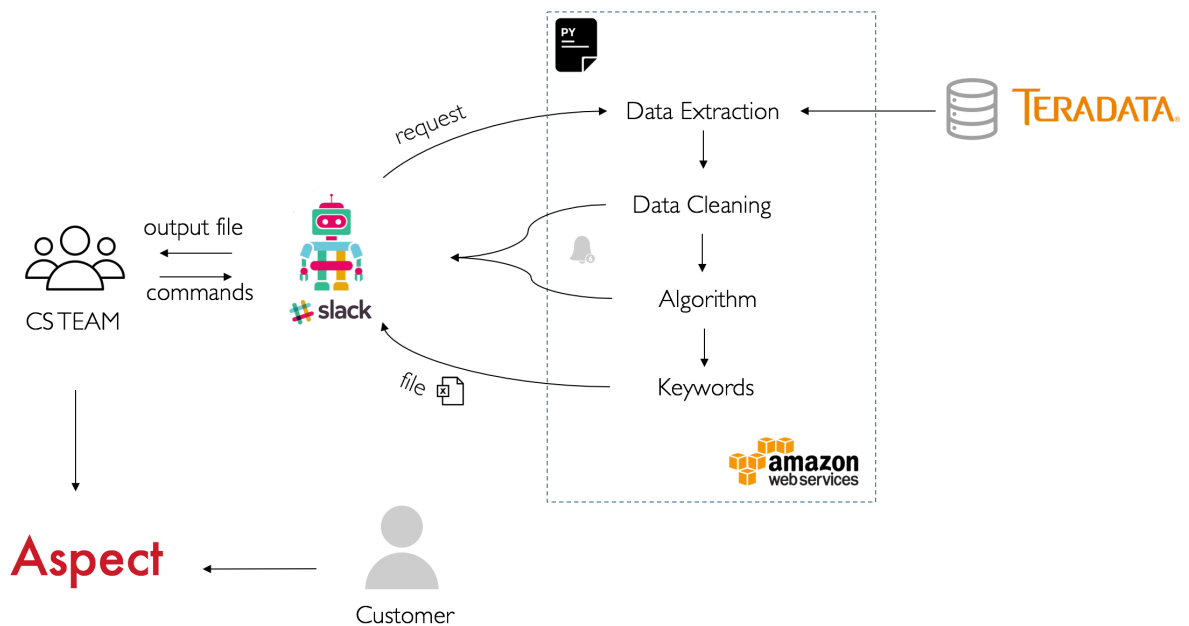
The entire process was created in Python. To deliver this to customer service, we had to create an application for a non technical person to understand and extract keyword upon his need. We thought of creating an web application initially but it will not be worth the effort for a process which is done occasionally. We wanted to create a solution which can be very useful, quick and reusable. As this idea was well appreciated many teams at overstock wanted to see how it works, so demo of the slack bot video has also been made.(Attached with submission)

This entire application is hosted on AWS so that slack bot can listen to user continuously.


When customer team member pings slackbot with respective commands. It sends request to slack API which inturn sends a request to application hosted on AWS. The application draws

data from Teradata warehouse. Entire Algorithm and Data cleaning parts are done on aws Ubuntu server. After the entire process has been executed, the extracted keywords are delivered in an excel file with each issue in different sheets. This entire chat extract will be fed into Aspect software which can form a link to the FAQs page depending upon the customer entering issue.



Architecture Diagram




Entire Chat screen shot:


**@textract**

☆ | ○ away | Text Extraction



**manojraj** 11:18 AM

@textract hi


**textract** BOT 11:18 AM

Hi, I am Textract. I can do Keyword extraction.
Please enter start date and end date of data to be extracted seperated by comma
Here are few details about algorithm


Manoj and Deepthi

Keyword Extraction
This will extract keywords from the data provided




Parameters requiried
Date Range(YYYY-MM-DD),Sample size, issues


**manojraj** 11:21 AM

@textract 2016-09-01,2016-10-01


**textract** BOT 11:21 AM ☆

will extract data for 2016-09-01, 2016-10-01 What is the sample size that you want me to consider?



**manojraj** 11:22 AM


@textract 78

**textract** BOT 11:22 AM


Will extract data for requested sample size
These are the issues available

List of Issues
Accessing_Account | ClubO | Coupon|
International | Keyword_Extracted | Master_Card |
Missing_Parts | Orders|
Payments | Refund |Store_Card | Tracking
Select one or more issues seperated by comma


Select one or more issues
Eg: ClubO,Coupon,Orders,Refund


**manojraj** 11:22 AM

@textract clubo,missing_parts




Message @textract






textract BOT 11:22 AM
 Basic cleaning on data is done
 Extracting customer chat in a list

11:22 ☆ Identified and removed address from Chat transcripts

Removed One word sentences
 Removed Numbers
 Customer Chat in a Paragraph
 Customer Chat in a Single Paragraph for each issue
 Rake Cleaning done
 Rake process started for missing_parts
 rake is done for missing_parts
 Rake process started for clubo
 rake is done for clubo


textract BOT 11:22 AM
 uploaded a file ▾


Keyword Extract
 20KB Excel Spreadsheet


textract BOT 11:22 AM
 Keyword Extraction is DONE

+

Message @textract

😊

Results:

Sample data of one issue of one chat transcript between customer and customer service agent:

There are 501, 351 chats for 11 issues(tracking, international orders, refunds, missing parts, coupon, store card, master card, club o, accessing account, payment , Damaged items) for a time period of 90 days.

*Full Transcript (includes private messages)

*

*[10:03:58 AM] Hi, my name is Briana. How may I help you?
 [10:04:08 AM] Jennie On: Hi*

Briana
 [10:04:25 AM] Jennie On: I have storage ottoman shipped to my apartment

[10:04:30 AM] Briana: Hello there.
 [10:04:36 AM] Jennie On: but I don't know when
it come
 [10:05:04 AM] Jennie On: Can you note with fedex that leave the stuff in front of
my door
 [10:05:22 AM] Jennie On: It will be ok not leave at the mail box
 [10:05:47
AM] Jennie On: It will be ok if they leave in front of my door
 [10:05:58 AM] Briana: I will be
happy to help you with the shipping status of your order.
 [10:07:10 AM] Jennie On: thanks
what information do you need?
 [10:07:48 AM] Jennie On: Jennie On
 13032 Monroe
Street
 apt 5
 Garden Grove, CA 92844
 [10:07:54 AM] Briana: For security
purposes, could you please verify your full name, email address and the complete billing
address?
 [10:08:05 AM] Briana: Thanks for the details, Jennie.
 [10:08:06 AM] Jennie
On: Jennie On
 [10:08:21 AM] Jennie On: onjennie@yahoo.com
 [10:08:36 AM] Jennie
On: the above is billing/shipping address
 [10:08:37 AM] Briana: I see that the item
shipped via FEDEX on 10/04/2015.
 [10:08:51 AM] Jennie On: but I don't see the
tracking number
 [10:08:54 AM] Briana: Yes, I got your message for the billing address,
Thanks.
 [10:09:09 AM] Briana: It will take 1-2 business days to update the tracking.

[10:09:31 AM] Briana: I will follow up the order and send you the email confirmation once the
tracking will update.
 [10:09:33 AM] Jennie On: can you update with the fedex that please
leave in front of my door
 [10:09:52 AM] Briana: Not to worry, the carrier will drop the
item in front of your door only.
 [10:10:05 AM] Jennie On: not at the mail box, it is outside
of the gate
 [10:10:15 AM] Briana: They do not have the option to leave the item in the
mail box.
 [10:10:25 AM] Briana: Not to worry, the carrier will drop the item in front of
your door only.
 [10:10:44 AM] Jennie On: do they require signature? no one at home at
that time
 [10:11:07 AM] Jennie On: in front of my door is secure, no signature need

[10:11:11 AM] Briana: They do not require any signature.
 [10:11:21 AM] Jennie On:
great
 [10:11:25 AM] Jennie On: thanks
 [10:11:32 AM] Briana: You are most
welcome.
 [10:11:59 AM] Jennie On: please send me the update tracking number once you
have it.
 [10:12:05 AM] Briana: In case of any need you can contact with the carrier FEDEX
at 800-463-3339.
 [10:12:15 AM] Jennie On: thanks again and have a good day

[10:12:23 AM] 'Jennie On' disconnected ('Concluded by End-user').

Sample of Keywords along with the Score that has been extracted from the data set

	Keyword	Score
0	clubo reward dollars expire	7.981369295
1	active duty military	6.450174216
2	overstock master card	6.355414546
3	clubo reward dollars	6.336167275
4	clubo silver member	6.319472947
5	clubo membership fee	6.3094953
6	clubo membership expires	6.271843307
7	clubo silver subscription	6.256593574
8	clubo gold member	6.243385238
9	clubo silver membership	6.231009189
10	credit card statement	6.220196973
11	clubo gold membership	6.15492148
12	gold clubo membership	6.15492148
13	clubo silver rewards	6.14337195
14	clubo rewards dollars	6.03136057
15	clubo expiration date	6.01419138
16	clubo rewards program	5.942470531
17	overstock gift card	5.935425179
18	free clubo membership	5.880181579
19	join clubo gold	5.830553349
20	reward dollars expire	5.813789029
21	clubo rewards membership	5.759543489
22	clubo rewards money	5.747104822
23	clubo rewards points	5.707484461
24	clubo rewards balance	5.692571766

Lessons Learned:

1. It has been a great learning experience. While starting the project, we never had python experience but by the end of the project we became very comfortable with it.
2. All the team members are from non technical back ground, creating a Chatbot helped us understand many networking concepts, devOps concepts etc.,

3. By using Pycharm tool for development, we learned not just to write code, but also to could deliver in production ready format
4. Used AWS Ubuntu server to host the slack api application so that ChatBot can listen to user continuously.
5. Realized how to manage change request when there are significant changes in the project requirements.
6. Learned from experience of sharing work among team mates and manage project with small team
7. By using Github version control system for managing different iterations, we were able to share code among sponsor, team and also saved different versions successfully with our losing code.

Next Steps:

1. Improve the speed of the entire text analytics process as it consumed almost 8 hours to run on the entire data set. Pyspark could be a perfect solution to increase the speed. Another way to increase the speed is by multi threading. There is one loop in the program which took lot of time. Implementing multi threading might be help even if the speed is reduced by 50%.
2. The keywords the has been extracted are tested manually if they mean something or not. So, there is a scope for improvement there. By building a testing frame work, one can save time by getting rid of the manual testing process. Basic idea of testing framework that has been thought of is, comparing different issues and calculating how different the keywords are to each other across issues while how similar they are among themselves in one particular issue. Along with this, by using word2vec, one could understand how many English phrases are extracted which could also be an indicator. In order to improve RAKE's performance, we can run a script that cycles through runs with different sets of parameters and evaluates the quality of keywords for each run if testing is created. It could then return the parameters that performed best on this dataset.

3. There are few algorithms which can perform a better job. Deep learning algorithms such as Recurrent Neural Networks are well known for this type of keyword extraction. This could be improving the quality of keywords extracted.
4. Slack bot can be made smarter by using machine learning concepts and Natural language processing.