CAPSTONE PROJECT

Topic Modelling on
Amazon Help tweet

Presented by: Peggy Man

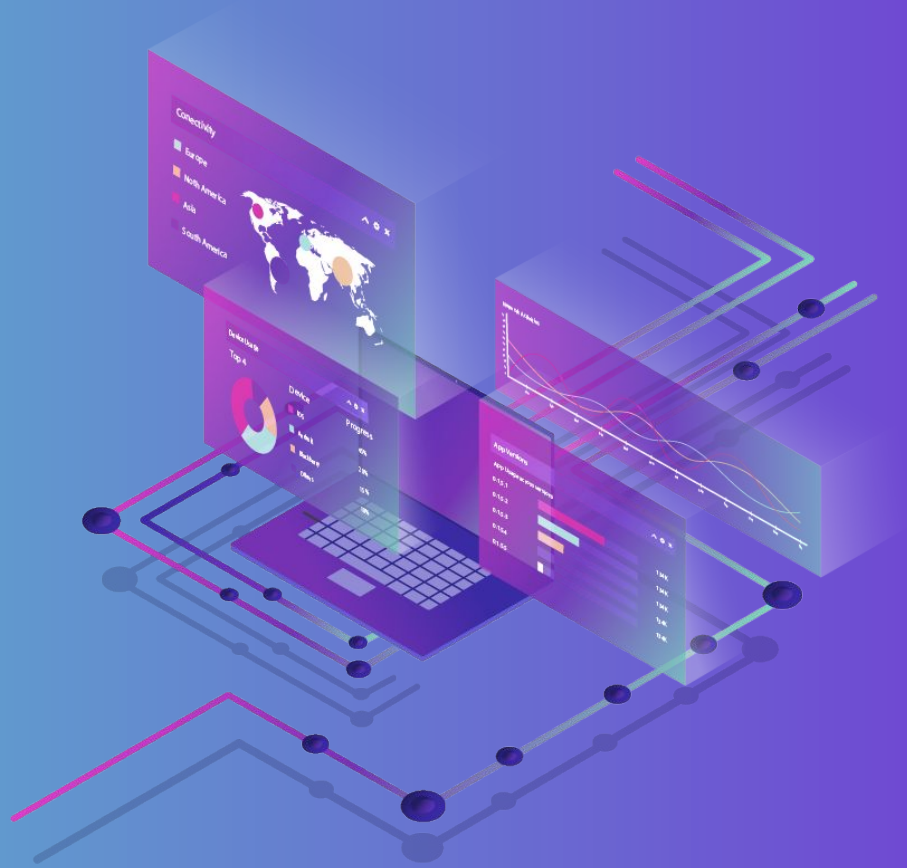


TABLE OF CONTENTS

PROBLEM
01 STATEMENT

WHAT IS THE PROBLEM?

- Focus on Amazon Customer Service twitter account - @AmazonHelp
- Customer satisfaction on is a challenge due to **HIGH** volume of tweets
- Qualities and efficiency of customer service

TACKLE THE PROBLEM

TOPIC

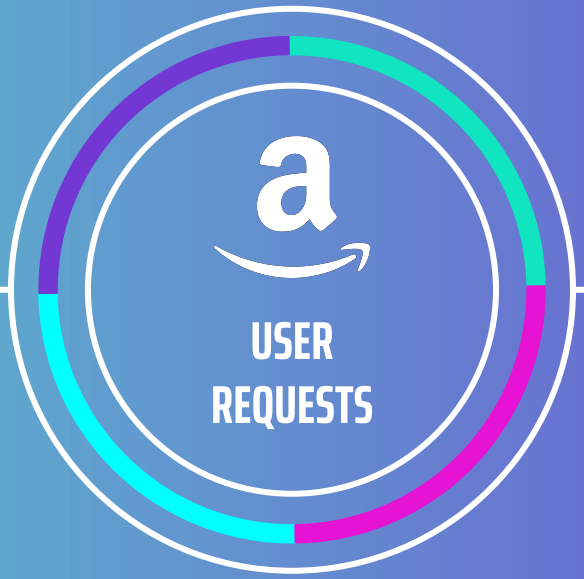
SUB CS TEAM

0

TOPIC

1

SUB CS TEAM



SUB CS TEAM

2

TOPIC

3

SUB CS TEAM

TOPIC

@AmazonHelp

DATASETS 02

UNDERSTANDING THE DATASETS

INBOUND TWEETS

787,346
First inbound tweets (not a reply tweet)

COMPANIES

108
Number of companies tweets

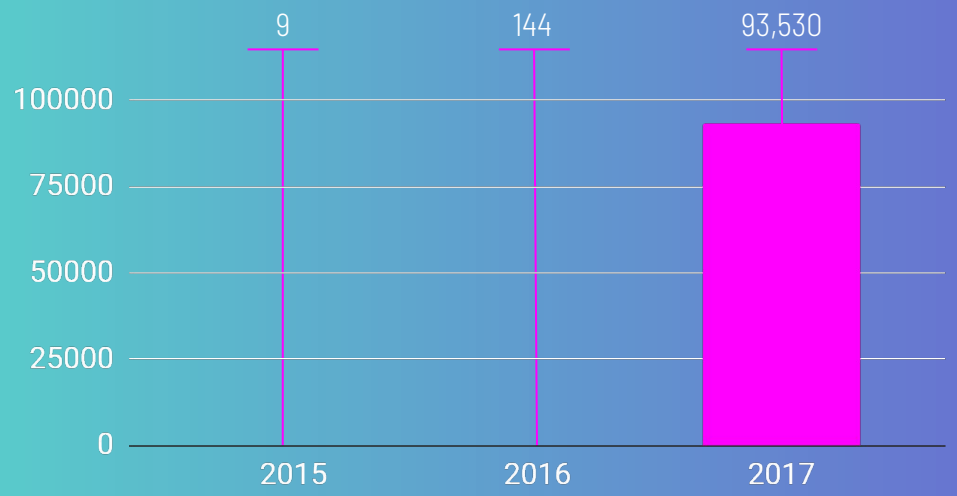
SIZE OF DATAFRAME

2,811,774 x 7
Dataframe shape

RESPONSE TWEETS

794,299
Number of tweets response by companies

SELECTION WITHIN DATASETS



Number of Tweets by Years of @AmazonHelp

@AmazonHelp
As selected company

93,530
Total Number of Tweets

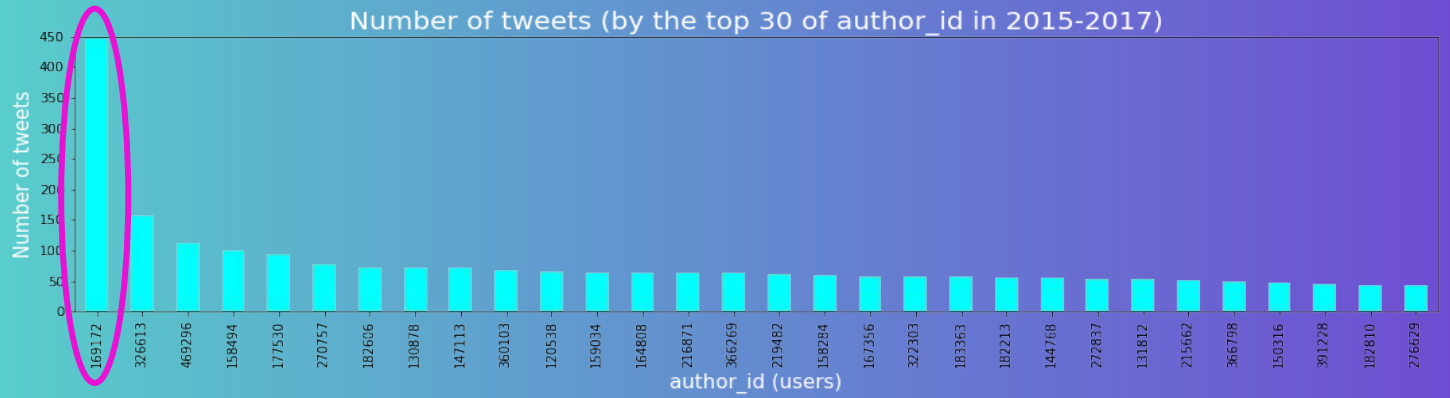
13 Jun 2015
First Tweet

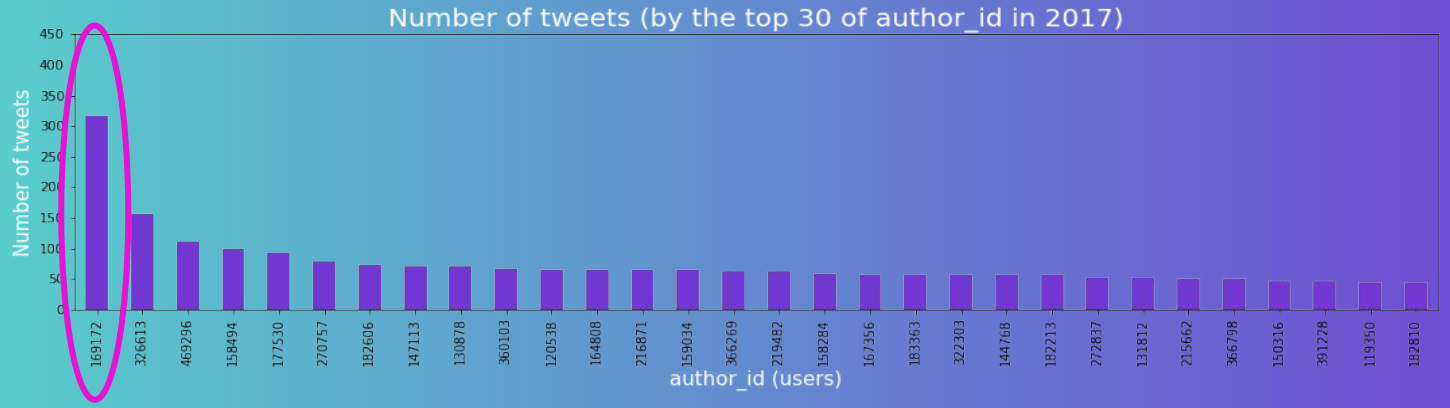
03 Dec 2017
Last Tweet

DATA CLEANING

03 & EDA

TOP 30 TWEET USERS



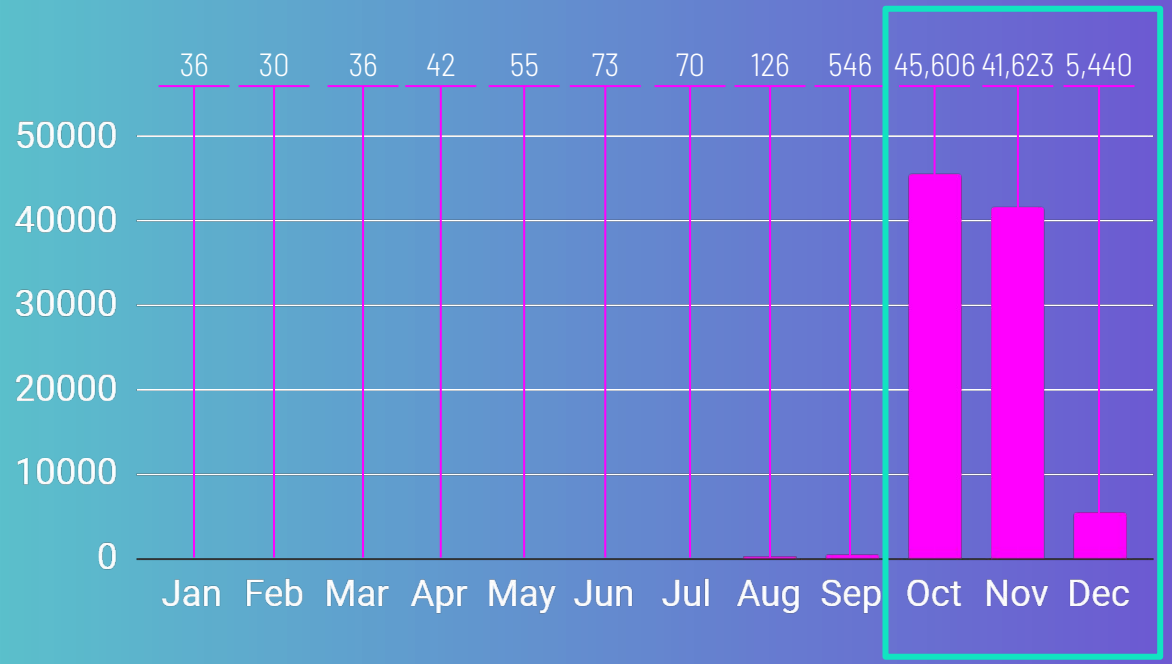


TWEET FROM THE USER ID: @169172

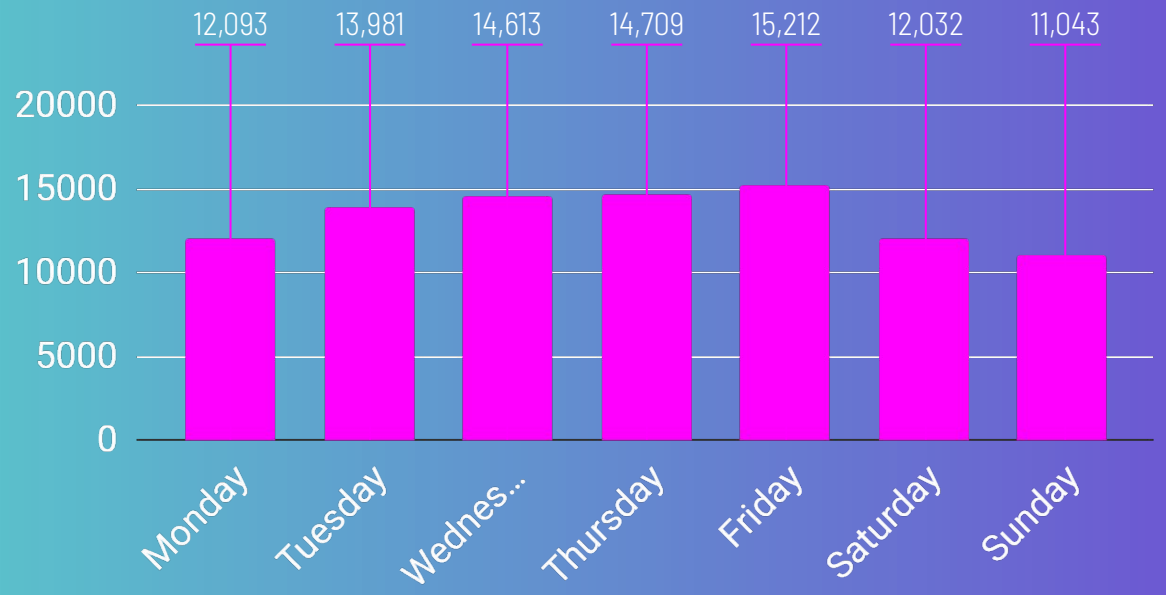
```
created_at
2016-08-17 15:33:17+00:00    Very unhappy with @115830 @AmazonHelp item I ordered sold by Amazon dropped 15~ % 1/2 da
ys after ordering and won't honour price drop
2016-08-18 01:57:28+00:00    @115830 @AmazonHelp Can you help please? Haven't received response
2016-08-19 19:14:02+00:00    @115830 @AmazonHelp Can I get some sort of response please, third time asking?
2016-08-20 20:10:03+00:00    @115830 @AmazonHelp 4th day requesting any kind of response
2016-08-21 16:16:38+00:00    @115830 @AmazonHelp 5th day requesting any kind of response
                                                    ...
2017-11-29 21:00:27+00:00    @AmazonHelp @115830 @115851 463rd day requesting any kind of response
2017-11-30 19:00:12+00:00    @AmazonHelp @115830 @115851 464th day requesting any kind of response
2017-12-01 19:15:52+00:00    @AmazonHelp @115830 @115851 465th day requesting any kind of response
2017-12-02 18:50:53+00:00    @AmazonHelp @115830 @115851 466th day requesting any kind of response
2017-12-03 19:18:04+00:00    @AmazonHelp @115830 @115851 467th day requesting any kind of response
Name: text, Length: 447, dtype: object
```

Total **447** spam tweets
Requesting @AmazonHelp to response

NUMBER OF TWEETS BY MONTH



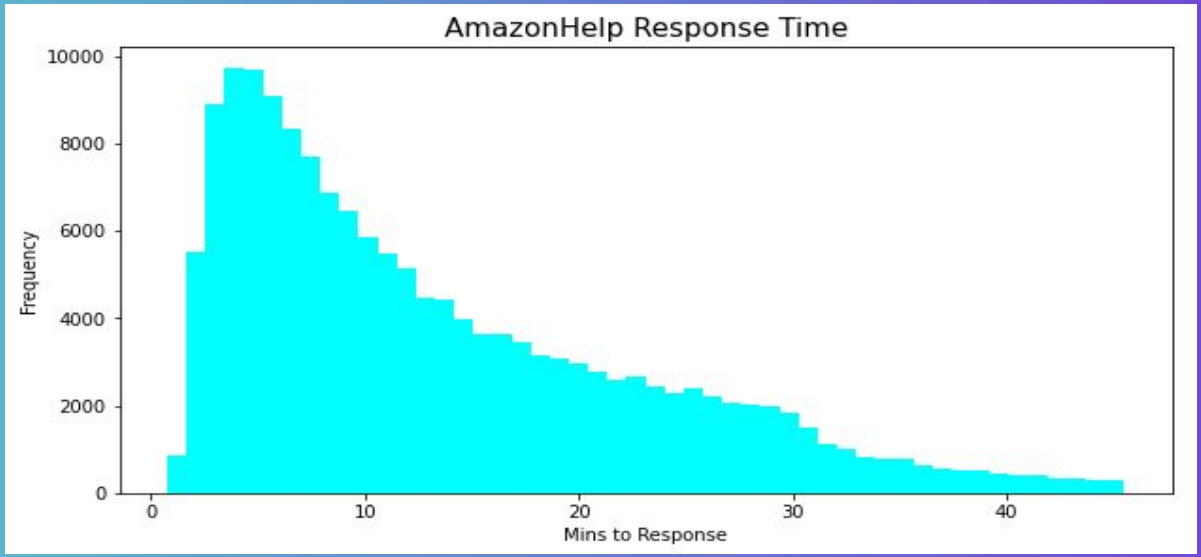
NUMBER OF TWEETS BY DAY



SERVICE LEVEL AGREEMENT

13.44
Minutes

AVERAGE
RESPONSE TIME



PROCEED WITH ONLY 2017 TWEETS

Number of Tweets

93,530
Tweets

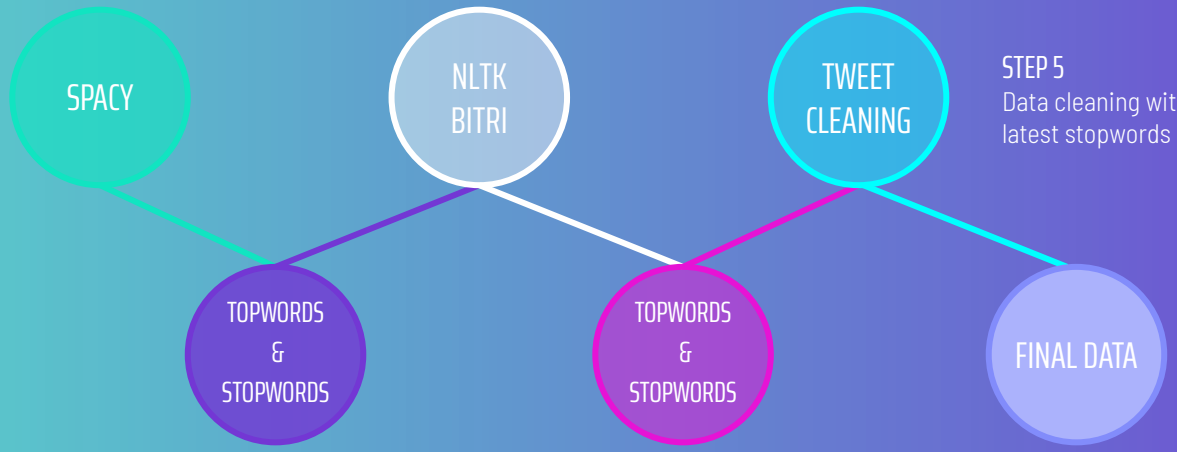
Duration

336
Days
(01-Jan-2017 to 03-Dec-2017)

Average tweet per hour

11.6
Tweets

DATA PRE-PROCESSING



STEP 1
Run SPACY library

STEP 3
Run NLTK BITRI with latest stopwords

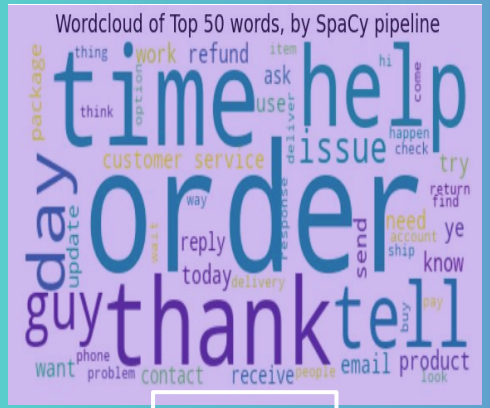
STEP 5
Data cleaning with latest stopwords

STEP 2
Generated topwords & stopwords from Spacy (keep the result)

STEP 4
Generated new topwords & stopwords from NLTK Bitri (keep the result)

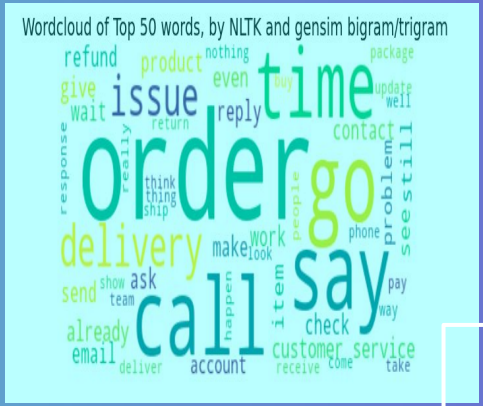
STEP 6
3 output will be use for modelling.
SPACY, NLTK BITRI, and TWEET CLEANING

TOPWORDS CLOUD



SPACY

TWEET CLEANING



NLTK BITRI

DATA CLEANING CHALLENGES

NON-ENGLISH FILTERING

1
Filtered non-ASCII

2
Used language detection library filter non-English (12,360 rows)

3
Check on non-English list, found out it included English within

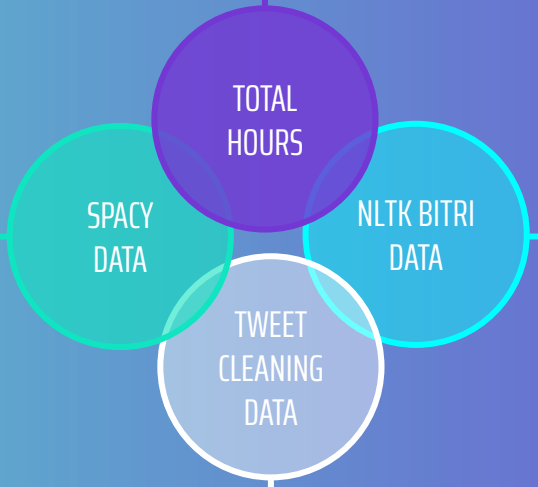
FINAL TRADE OFF

Lose some of the English data due to mis-detection of language detector

TOPIC MODELLING 04

LATENT DIRICHLET ALLOCATION MODEL (LDA)

TOTAL ~3 HOURS SPENT
In order to run all collected data



HIGHEST COHERENCE SCORE: 0.3519
Number of topic is 4
Number of iteration is 10

HIGHEST COHERENCE SCORE: 0.3688
Number of topic is 3
Number of iteration is 10

HIGHEST COHERENCE SCORE: 0.3549
Number of topic is 6
Number of iteration is 10

HYPER PARAMETER TUNING

TOTAL HOURS SPENT

13:58

BEST COHERENCE SCORE

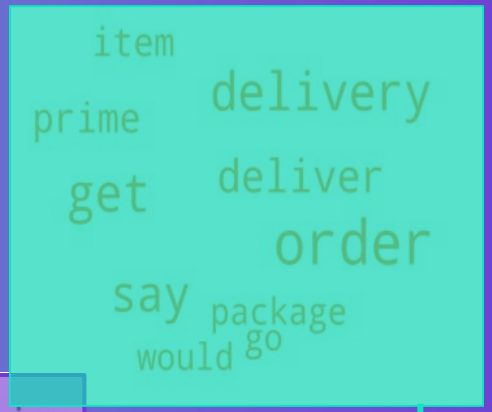
0.4400192893615461

BEST PARAMETER

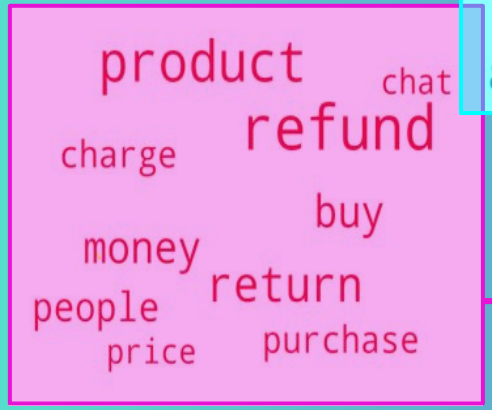
coherence_tuning(corpus,data_name,k=4, a='symmetric', b=0.7000000000000001)

RESULT OF TOPICS

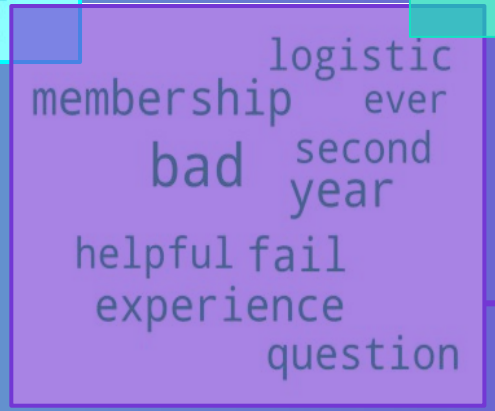
TOPIC 0
Account Issue
E-mail follow up



TOPIC 1
Delivery with PRIME account



TOPIC 3
Refund & Product return

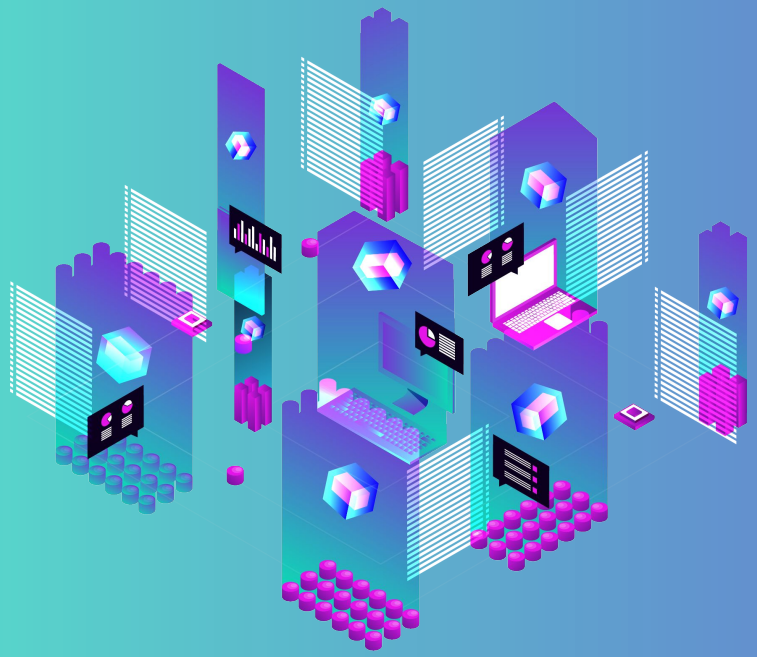


TOPIC 2
Membership & Complaints

05 FUTURE IMPROVEMENT

LIMITATION & IMPROVEMENT

1. WRONG JUDGEMENT: Kept non-English tweets such as Spanish, French & German. (As per Topic 1)
    a. IMPROVEMENT: Research on better language detector.
2. LACK OF DATA: Most of the tweets are regarding delivery order.
    a. IMPROVEMENT: Self perform data crawling to gather more data.
3. LOW COHERENCE SCORE: Due to time constraint, the highest coherence score gotten was 0.44.
    a. IMPROVEMENT: Further tuning on hyper parameters.



CONCLUSION 06

CONCLUSION & RECOMMENDATION

RESTRUCTURE

Restructure team to a sub-team based on the topic result
Sub-team A focus on account issue
Sub-team B focus on order issue

STAFF ALLOCATION

Based on the result of data analysis. The peak usually falls on Oct, Nov, and Dec.
Be prepared with enough workforce and allocate to the appropriate team. The SLA should be less than 14 mins.

THANKS