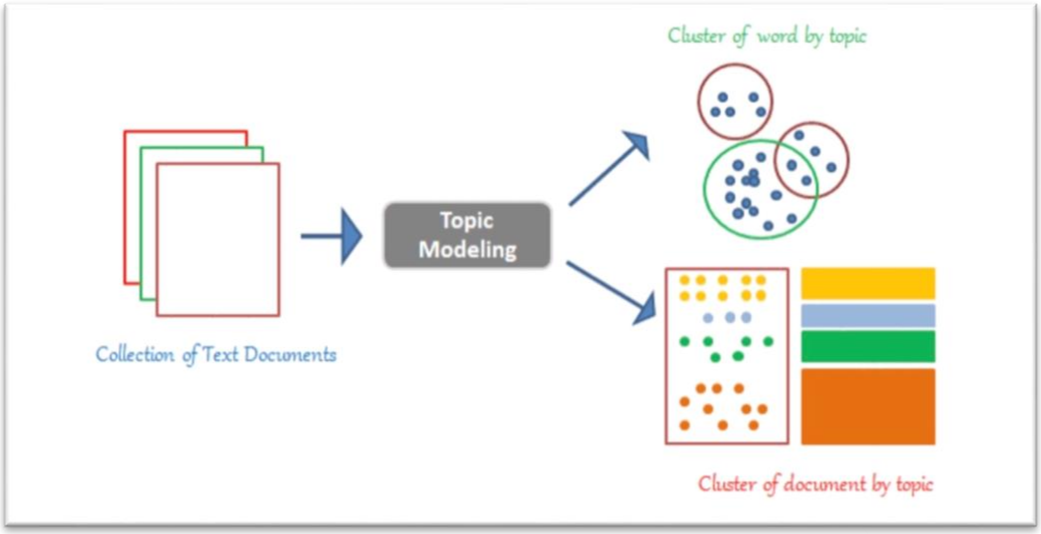
Topic Modeling using Watson NLP

Topic modeling is an unsupervised machine learning algorithm. That's used to convert unstructured content into structured format in the form of set of similar documents, detecting word and phrase patterns within them. It is automatically clustering word groups and similar expressions to best characterize a set of similar documents.

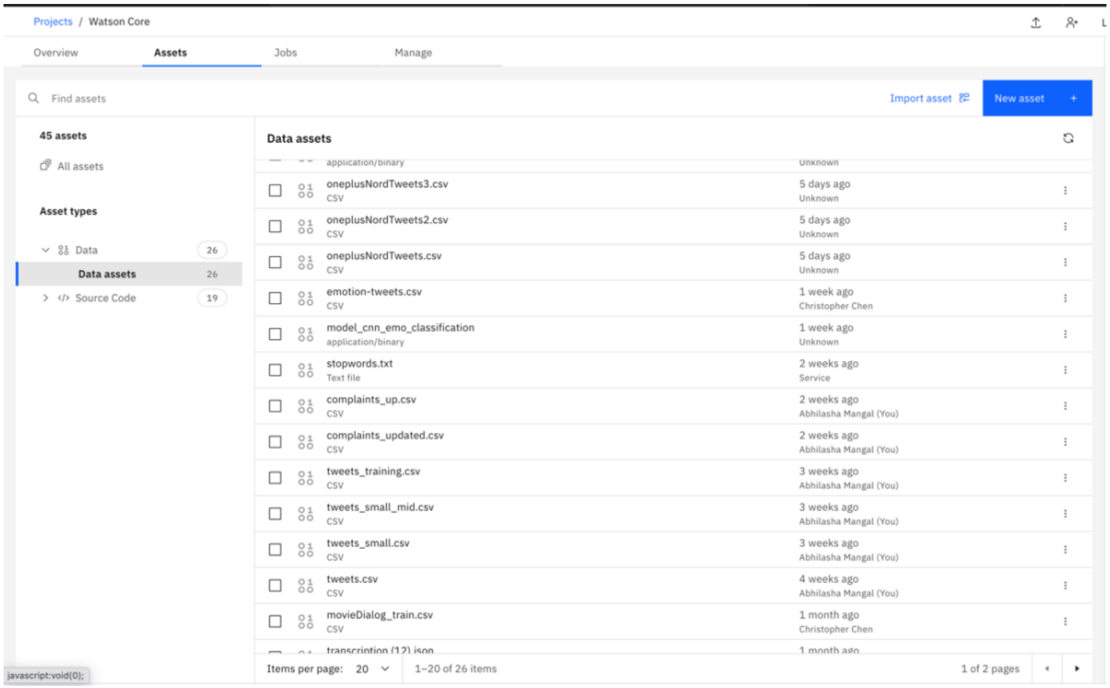


In these days unstructured content is increasing rapidly in huge amount. To handle this data and converted into structured format is very time consuming and overloading process for the employs. By using topic analysis model, A machine will be able to sort through endless lists of unstructured content into similar documents. It will save time & money for the companies.

This notebook demonstrates how to analyse consumer financial data using Watson NLP step by step.

1. Collecting the dataset

Let's an example Consumer Financial complaint database collected from Consumer complaint database . To use here we normalize this dataset by removing which rows does not have value of consumer complaints. This data set contains 999285 consumer complaints with the date received, submitted via, products, sub-products and company information. Once you have downloaded the dataset, you can upload it to the Watson Studio instance by going to the Assets tab and then dropping the data files as shown below.



Projects / Watson Core

45 assets

All assets

Asset types

- Data 26
  - Data assets 26
- Source Code 19

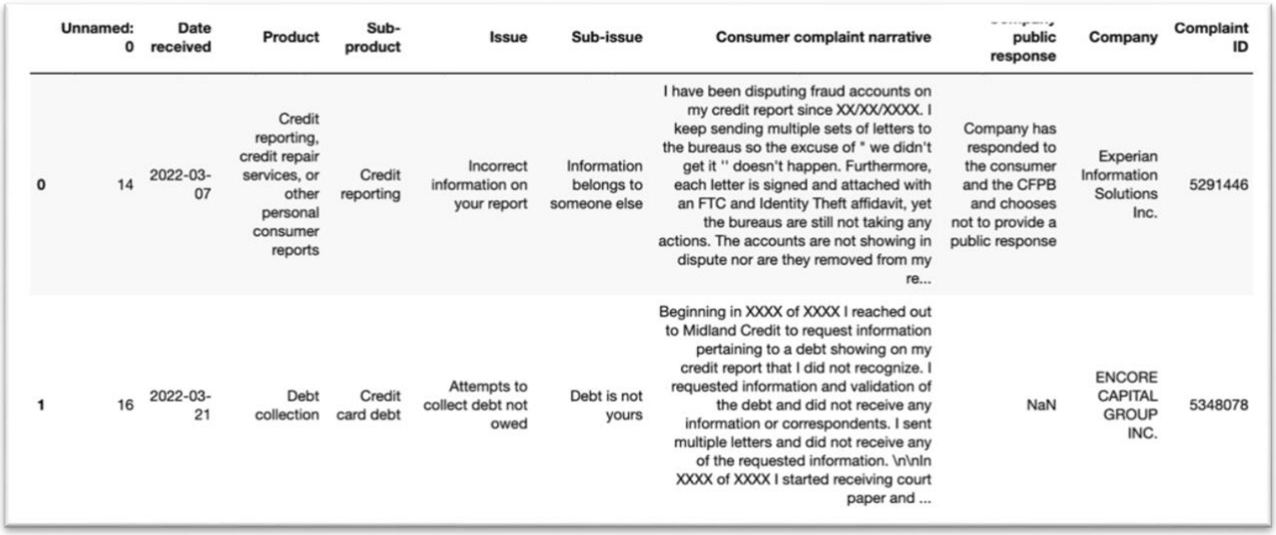
Data assets

Name	Type	Last modified	By
	application/binary		Unknown
oneplusNordTweets3.csv	CSV	5 days ago	Unknown
oneplusNordTweets2.csv	CSV	5 days ago	Unknown
oneplusNordTweets.csv	CSV	5 days ago	Unknown
emotion-tweets.csv	CSV	1 week ago	Christopher Chen
model_cnn_emo_classification	application/binary	1 week ago	Unknown
stopwords.txt	Text file	2 weeks ago	Service
complaints_up.csv	CSV	2 weeks ago	Abhilasha Mangal (You)
complaints_updated.csv	CSV	2 weeks ago	Abhilasha Mangal (You)
tweets_training.csv	CSV	3 weeks ago	Abhilasha Mangal (You)
tweets_small_mid.csv	CSV	3 weeks ago	Abhilasha Mangal (You)
tweets_small.csv	CSV	3 weeks ago	Abhilasha Mangal (You)
tweets.csv	CSV	4 weeks ago	Abhilasha Mangal (You)
movieDialog_train.csv	CSV	1 month ago	Christopher Chen
transcription (12).json		1 month ago	

Items per page: 20 1–20 of 26 items 1 of 2 pages

javascript:void(0);

Once the dataset is added to the project, you can access it from the notebook and read the csv file into panda's data frame.



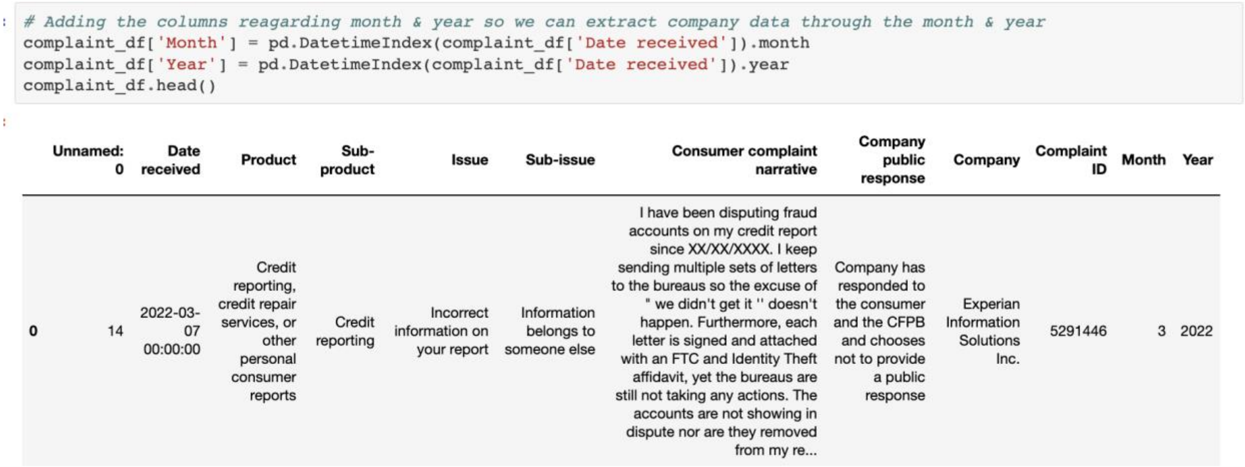
	Unnamed: 0	Date received	Product	Sub-product	Issue	Sub-issue	Consumer complaint narrative	Company public response	Company	Complaint ID
0	14	2022-03-07	Credit reporting, credit repair services, or other personal consumer reports	Credit reporting	Incorrect information on your report	Information belongs to someone else	I have been disputing fraud accounts on my credit report since XX/XX/XXXX. I keep sending multiple sets of letters to the bureaus so the excuse of " we didn't get it '' doesn't happen. Furthermore, each letter is signed and attached with an FTC and Identity Theft affidavit, yet the bureaus are still not taking any actions. The accounts are not showing in dispute nor are they removed from my re...	Company has responded to the consumer and the CFPB and chooses not to provide a public response	Experian Information Solutions Inc.	5291446
1	16	2022-03-21	Debt collection	Credit card debt	Attempts to collect debt not owed	Debt is not yours	Beginning in XXXX of XXXX I reached out to Midland Credit to request information pertaining to a debt showing on my credit report that I did not recognize. I requested information and validation of the debt and did not receive any information or correspondents. I sent multiple letters and did not receive any of the requested information. \n\nIn XXXX of XXXX I started receiving court paper and ...	NaN	ENCORE CAPITAL GROUP INC.	5348078

2. Data processing & Exploratory Data Analysis:

To process this data for Topic Modeling categories by month & company wise.

2.1 Month & Year Wise Data

To collect the data by using months so converted date received into time frame and added 'Month' & 'Year' two more columns.

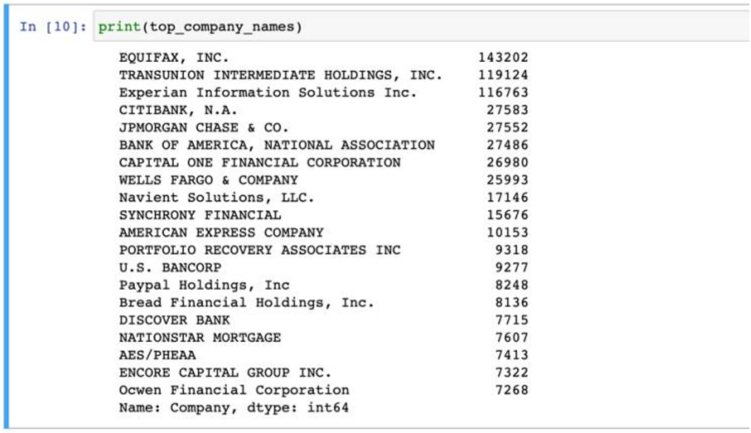


```
# Adding the columns reagarding month & year so we can extract company data through the month & year
complaint_df['Month'] = pd.DatetimeIndex(complaint_df['Date received']).month
complaint_df['Year'] = pd.DatetimeIndex(complaint_df['Date received']).year
complaint_df.head()
```

	Unnamed: 0	Date received	Product	Sub-product	Issue	Sub-issue	Consumer complaint narrative	Company public response	Company	Complaint ID	Month	Year
0	14	2022-03-07 00:00:00	Credit reporting, credit repair services, or other personal consumer reports	Credit reporting	Incorrect information on your report	Information belongs to someone else	I have been disputing fraud accounts on my credit report since XX/XX/XXXX. I keep sending multiple sets of letters to the bureaus so the excuse of " we didn't get it '' doesn't happen. Furthermore, each letter is signed and attached with an FTC and Identity Theft affidavit, yet the bureaus are still not taking any actions. The accounts are not showing in dispute nor are they removed from my re...	Company has responded to the consumer and the CFPB and chooses not to provide a public response	Experian Information Solutions Inc.	5291446	3	2022

2.2 Company Wise Data

To collect the data company wise took “Top 20 Companies” which have more complaints.



```
In [10]: print(top_company_names)

    EQUIFAX, INC.                                  143202
    TRANSUNION INTERMEDIATE HOLDINGS, INC.         119124
    Experian Information Solutions Inc.            116763
    CITIBANK, N.A.                                  27583
    JPMORGAN CHASE & CO.                            27552
    BANK OF AMERICA, NATIONAL ASSOCIATION           27486
    CAPITAL ONE FINANCIAL CORPORATION               26980
    WELLS FARGO & COMPANY                           25993
    Navient Solutions, LLC.                         17146
    SYNCHRONY FINANCIAL                             15676
    AMERICAN EXPRESS COMPANY                        10153
    PORTFOLIO RECOVERY ASSOCIATES INC                9318
    U.S. BANCORP                                     9277
    Paypal Holdings, Inc                             8248
    Bread Financial Holdings, Inc.                   8136
    DISCOVER BANK                                    7715
    NATIONSTAR MORTGAGE                              7607
    AES/PHEAA                                        7413
    ENCORE CAPITAL GROUP INC.                        7322
    Ocwen Financial Corporation                      7268
    Name: Company, dtype: int64
```

2.3 Text Pre-processing

Our first step is to pre-process the documents in a way that cleans distracting signals and makes them easier to process and analyse. This is a standard step in many NLP pipelines. Here we perform three types of pre-processing:

1. Stop-words filtering:
   To remove stop-words we used **Watson NLP pre- defined list**. Which we can remove & extend this stop words list. You can download this list by using ‘downlaod_and_load’method of Watson NLP library.

```
wnlp_stop_words = watson_nlp.download_and_load('text_stopwords_classi
fication_ensemble_en_stock').stopwords
```

2. Remove some Patterns:
   This dataset has consumer personal information. That is hide by Pattern of 'XX/XX/XXXX' or another format. Replaced all these patterns by blank.

3. Lemmatization:
   Variability in surface forms, as in derivation (drive, driving, drives) and inflection (am, is, are), creates a challenge for a clustering algorithm that works on top of term frequency (TF) representations since in this setup the algorithm is not aware of the strong semantic relation between them.

After applying all these steps to documents and clean the data for further process to extract Keyword - Noun Phrase Extraction & Topic Modeling.

3. Keyword & Noun Phrase Extraction:

To extract the noun phrase & key word extraction we are using pre-defined model **noun-phrases_rbr_en_stock & 'keywords_text-rank_en_stock'** of Watson NLP for English language**.** You can download & load these models by using below steps :

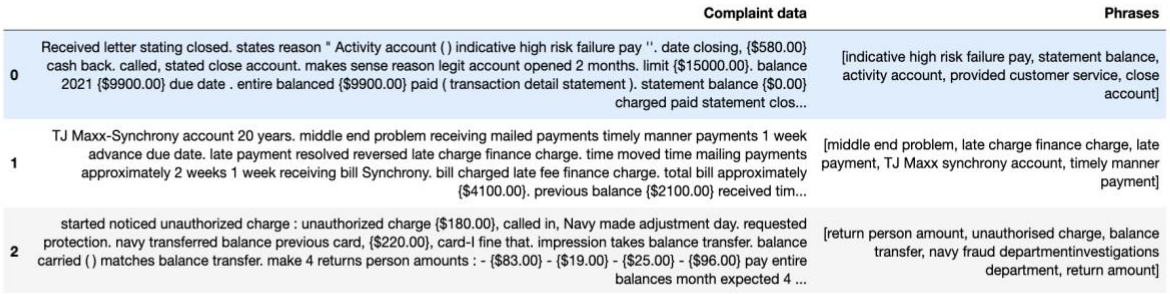
```
noun_phrases_model = watson_nlp.load(watson_nlp.download('noun-phrases_rbr_en_stock'))

keywords_model = watson_nlp.load(watson_nlp.download('keywords_text-rank_en_stock'))
```

By using above model, we are collecting all Noun & Keyword -phrases for each document. To collect all these, you can use below steps:

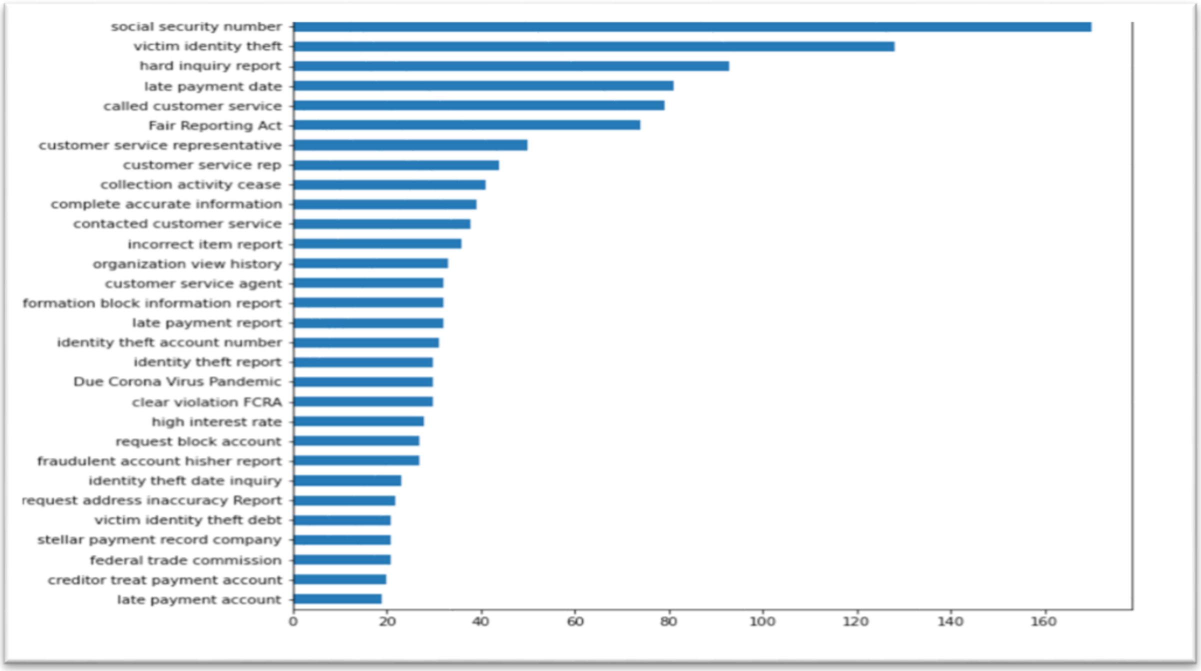
```
# Run the Syntax and Noun Phrases models
    syntax_prediction = syntax_model.run(text, parsers=('token', 'lemma',
'part_of_speech'))
    noun_phrases = noun_phrases_model.run(text)
    # Run the keywords model
    keywords = keywords_model.run(syntax_prediction, noun_phrases, limit=
5)
    keywords_list =keywords.to_dict()['keywords']
```

After collecting all Noun-phrases & Keywords create the data frame to see which are most talks coming form consumer at document level & as well company level

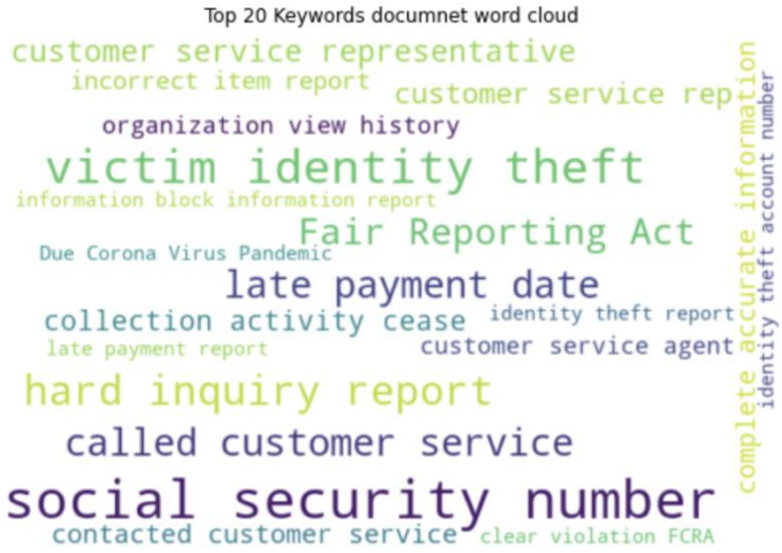


	Complaint data	Phrases
0	Received letter stating closed. states reason " Activity account ( ) indicative high risk failure pay ''. date closing, {$580.00} cash back. called, stated close account. makes sense reason legit account opened 2 months. limit {$15000.00}. balance 2021 {$9900.00} due date . entire balanced {$9900.00} paid (transaction detail statement). statement balance {$0.00} charged paid statement clos...	[indicative high risk failure pay, statement balance, activity account, provided customer service, close account]
1	TJ Maxx-Synchrony account 20 years. middle end problem receiving mailed payments timely manner payments 1 week advance due date. late payment resolved reversed late charge finance charge. time moved time mailing payments approximately 2 weeks 1 week receiving bill Synchrony. bill charged late fee finance charge. total bill approximately {$4100.00}. previous balance {$2100.00} received tim...	[middle end problem, late charge finance charge, late payment, TJ Maxx synchrony account, timely manner payment]
2	started noticed unauthorized charge : unauthorized charge {$180.00}, called in, Navy made adjustment day. requested protection. navy transferred balance previous card, {$220.00}, card-I fine that. impression takes balance transfer. balance carried ( ) matches balance transfer. make 4 returns person amounts : - {$83.00} - {$19.00} - {$25.00} - {$96.00} pay entire balances month expected 4 ...	[return person amount, unauthorised charge, balance transfer, navy fraud departmentinvestigations department, return amount]

By using above data-frame we applied some pre -processing steps to calculate the length of each phrase & Keyword. We removed 1-gram & bi-gram from this data frame and we had collected Top 30 Keywords & Phrases to see which most frequent complaints are coming. we can say here mostly consumer talked about "social_security_number" and "victim_identify_theft".



Phrase & Keywords Word cloud:



People are mostly talking about **social security number and victim identify theft**

4. Model Building

Topic Modelling Watson NLP uses two types of modeling:

1. Summary Model
2. Hierarchal Clustering Model

To use above functionality, we have downloaded some pre-required libs.

A summary model consists of a mapping of words to their occurrences over all of the documents Additionally, the summary model can be provided a dictionary of parameters to modify the summary model that is trained.

Summary Model Train Params :

```
{ 'min_words_per_utterance': 5, 'num_turns_to_remove': 0, 'beginning_ratio': 1,
'beginning_weighting_factor': 1, 'min_ngram_size': 5, 'max_ngram_size': 8, 'max_ngrams': 10, }
```

```
summary_model = NGramSummary.train(train_data=syntax_data,train_params=train_params)
```

By using above method, we can train documents in summary model. Here train data is provided in form of Syntax data which is processed by Syntax model in above step.

This Summary model output will be passed as input into Hierarchical Clustering Modeling.

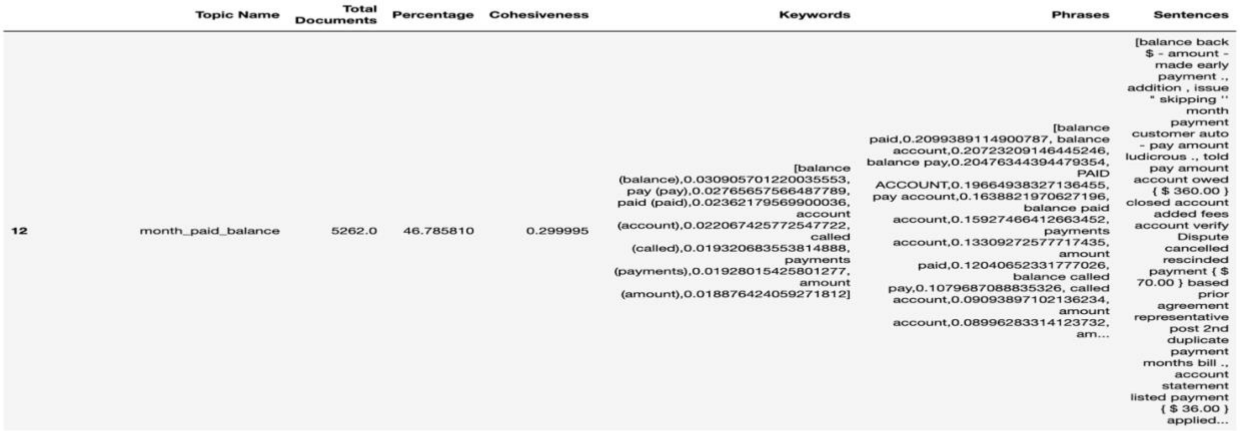
Train Params:

```
train_params = { 'king_cluster_min_ratio': .5, 'min_records_per_king_cluster': 10,
'num_topics_per_iteration': 40, 'max_num_iters_per_model': 4, 'min_word_support': 0.01,
'max_word_support': 0.7, 'max_ngrams_per_topic': 10, }
```

By using below code, we can train topic modeling

```
topic_model = HierarchicalClustering.train(train_data=syntax_data,
summary_model=summary_model, train_params = {'king_cluster_min_ratio': .5,
'min_records_per_king_cluster': 4000,  'num_topics_per_iteration': 40,'max_num_iters_per_model':
4,'max_ngrams_per_topic':10})
```

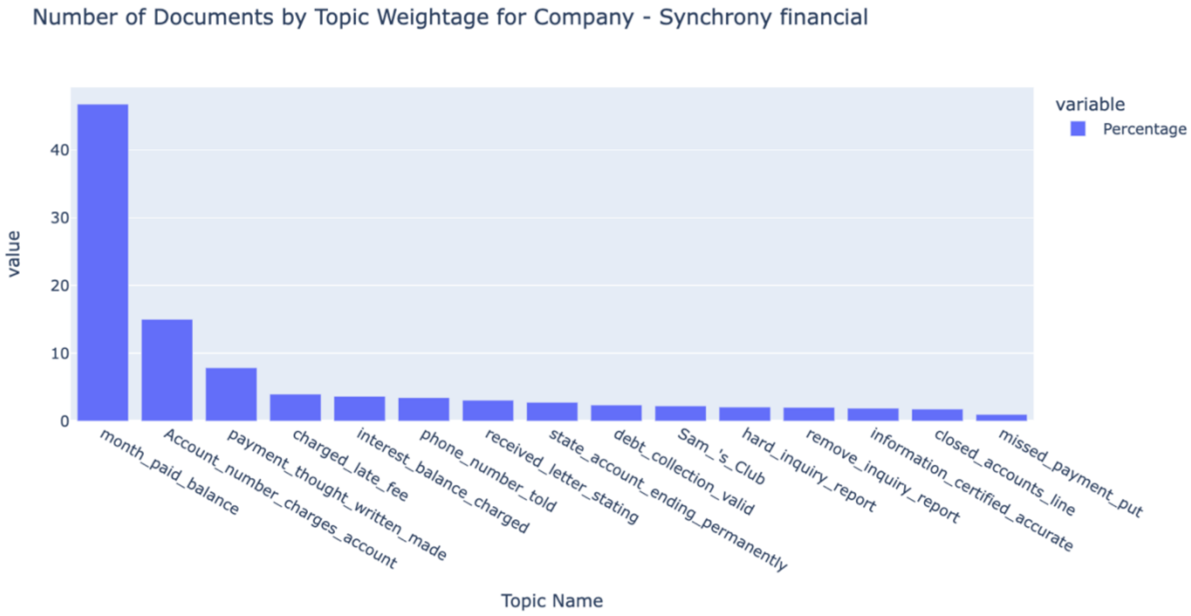
This Topic model provides as output in JSON format. Which can consist of the Topic Name, No of documents, percentage, most important keywords, phrases, sentences and etc.



	Topic Name	Total Documents	Percentage	Cohesiveness	Keywords	Phrases	Sentences
12	month_paid_balance	5262.0	46.785810	0.299995	[balance (balance),0.030905701220035553, pay (pay),0.02765657566487789, paid (paid),0.02362179569900036, account (account),0.022067425772547722, called (called),0.019320683553814888, payments (payments),0.01928015425801277, amount (amount),0.018876424059271812]	[balance paid,0.2099389114900787, balance account,0.20723209146445246, balance pay,0.20476344394479354, PAID ACCOUNT,0.19664938327136455, pay account,0.1638821970627196, balance paid account,0.15927466412663452, payments account,0.13309272577717435, amount paid,0.12040652331777026, balance called pay,0.1079687088835326, called account,0.09093897102136234, amount account,0.08996283314123732, am...	[balance back $ - amount - made early payment ., addition , issue " skipping '' month payment customer auto - pay amount ludicrous ., told pay amount account owed { $ 360.00 } closed account added fees account verify Dispute cancelled rescinded payment { $ 70.00 } based prior agreement representative post 2nd duplicate payment months bill ., account statement listed payment { $ 36.00 } applied...

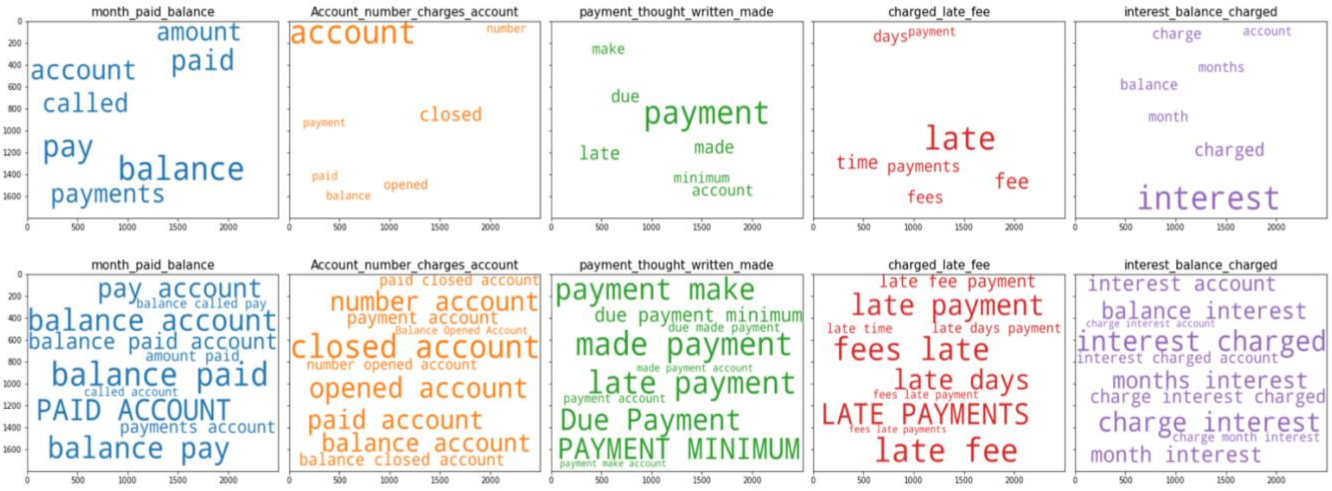
After analysis of this trained topic model we can see these top 15 topics which is talked most by the consumer regrading company.

Top 15 topic documents of Synchrony Financial company.



We can say here most of consumer complaining regarding 'month_paid_balance' & 'account_number_ charges so employ can transfer these types of issue directly to bank account field.

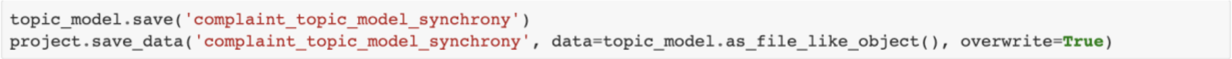
Top 5 Topics Keywords & Phrases of Synchrony Financial company:



Here we can see Top -5 Topics keywords & phrases.

4.1 Save trained model

Once the model has been trained, you can easily save it using the save() method from watson_nlp library or using the project.save_data() function from Watson Studio as shown below.



```
topic_model.save('complaint_topic_model_synchrony')
project.save_data('complaint_topic_model_synchrony', data=topic_model.as_file_like_object(), overwrite=True)
```

4.2 Load trained model

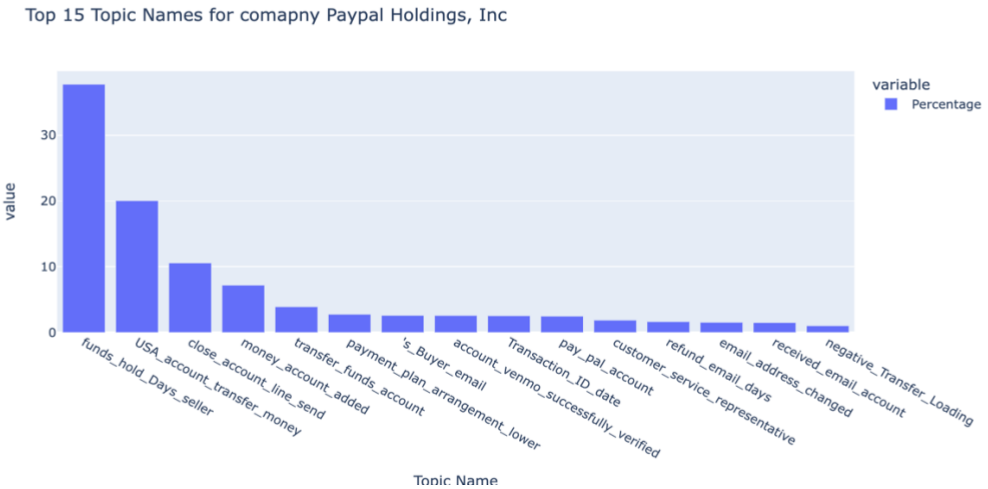
Once the model has been saved, you can easily load it using the load() method from watson_nlp library as shown below . So we can use further this model to predict the topics & keywords form the text.

```
topic_model = watson_nlp.load('complaint_topic_model_synchrony')
```

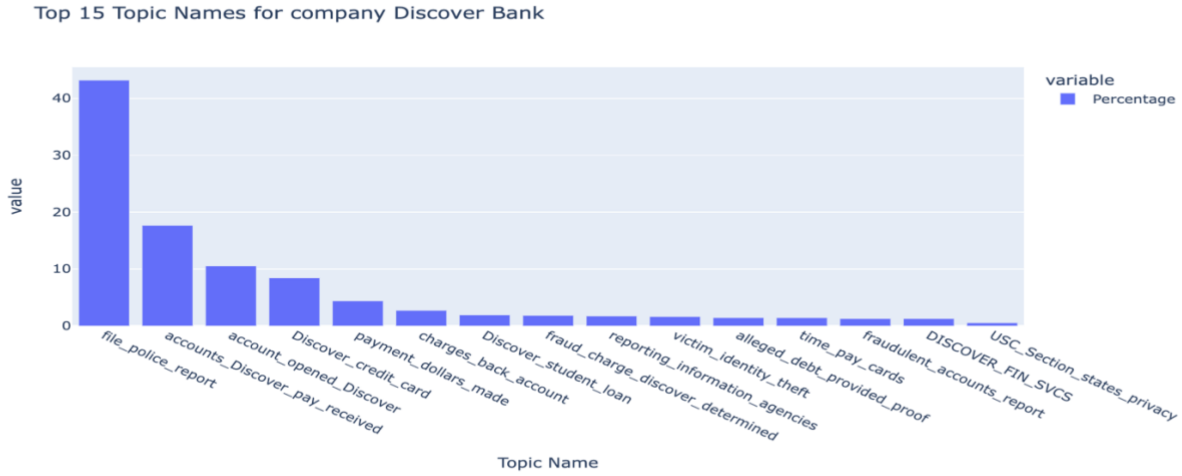
Analysis on 2 Companies Topics & Keywords:

By using this dataset, we had collected two company's data & trained the model as previous shown above steps to see how the different topics our occur company to company.

1. Paypal Holdings, Inc Company:



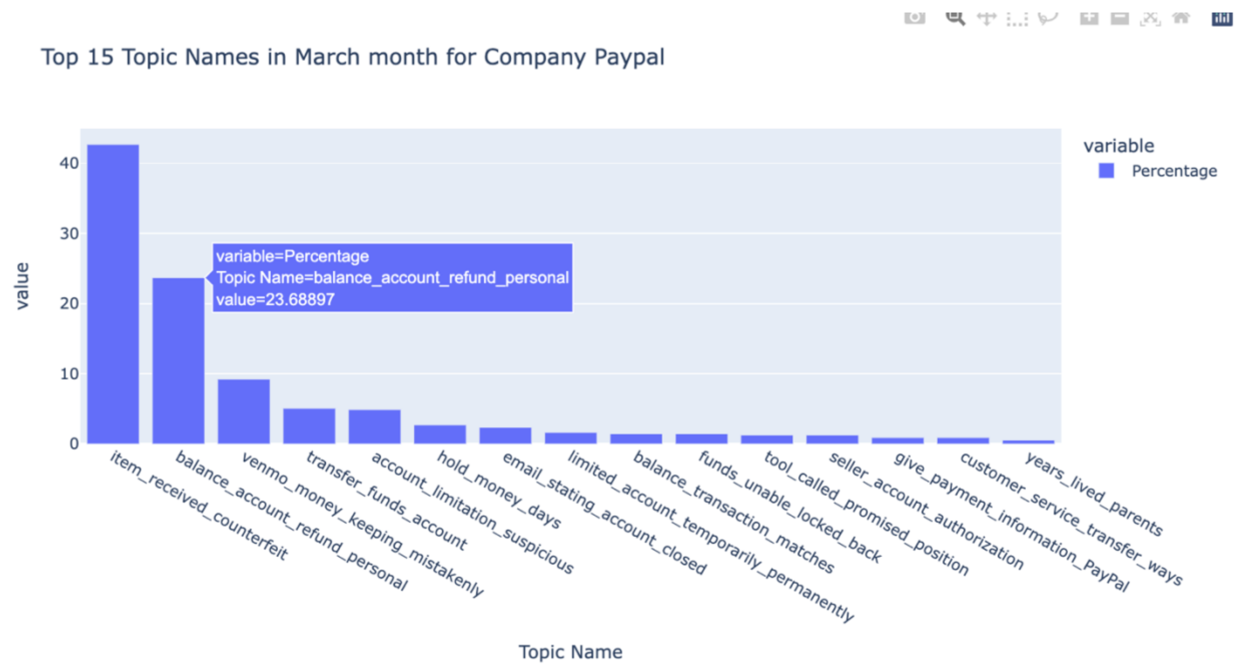
2. Discover Bank:



After analysis of both companies, we can easily find out both companies have different talk & issues and pain points of their customers. Like : Paypal Holdings company has most occur topics "Funds_hold_days" & "USA_account_transfer_money" but in Discover Bank has different topics like "File_police_report" &"account_discover_pay_recieved".

Analysis month wise Companies Topics & Keywords:

By using this dataset, we had collected month wise company data & trained the model as previous shown above steps to see how the different topics our occur in a company day by day. In which month which issues/Topics occur so much & how we can reduce in upcoming month.



The most frequent issue for the PayPal in the month of March is the item received counterfeit and balance account refund personal. This information can be used for either the subsequent month or for the next April (or any other month). they can directly resolve this issue rather than doing more investigation again and again.

5. Conclusion

We have seen how easily we are able to identify the topics from consumer financial dataset .This topic model can be used to understand the pain points and major areas of improvement day by day/ Weekly/Monthly or Yearly basis. Based on this information, they can create self-service content or direct support to help customers. Rather than trying to work out who needs to speak to the customer, a topic modelling tool can tag conversations and then, using workflows, route them to the most appropriate team.