AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

1) **Who** (company, agency, organization) collected the data?
   a) Who they are, what do they do?
      The data is taken from data.gov it is a government website that provides data from various genre's like health, crime etc. The dataset consists of data from customers of various financial organizations like banks. The complaints are collected about various financial products and services like mortgage, loans, debit/credit card issues the customer had to face. The Product, Sub-product, Issues, Complaint, location, company name have been provided in the dataset.
   b) What is their role/purpose?
      The purpose of any financial organization is to provide reliability and trustworthy service to their customer since a very crucial thing like money is involved here, no customer would like to compromise on the quality of service. Thus, by analyzing these complaints the organizations can take necessary actions to gain more customers and make sure the existing ones are here to stay.
2) **Need**
   a) *Why* did they collect this data?
      Any organization wants to keep track of what their customers think about them and if they are doing good in the market, this data helps to do just that. They would be able to recognize their weak areas and work on them to ensure quality service to their customers.
3) What potential *questions* could be answered by studying this data?
   a) List some *specific questions*, and *plan* to *answer them* in your analysis
      Questions like on a scale of 0 to -5 how unhappy are their customers with them, -5 being the most unhappy/dissatisfied. What products are customers repeatedly complaining about for particular financial companies. Example, Wells Fargo is repeatedly facing customer complaints for incorrect information of credit score. The analysis will provide answers to these sort of questions for the financial organizations.

Description of dataset:

The dataset is a 620 mb csv file. It consists of 18 columns among which Product, Sub-product, Zipcode, Issue, Sub-issue, Complaint, Company are some of them. 1,048,575 number of rows.

There are Nominal/Categorical datatypes for most attributes.

Requirements, resources needed

- What software and hardware resources will you use to study this data?
  The project was based on a csv dataset. Softwares used were Microsoft Excel, Anaconda Jupyter and R studio.
  On Microsoft Windows 8, 8 Gb RAM, 64-bit operating system for hardware.

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

Present the Results/Findings

1) Explore the dataset using relevant tools discussed in the course (R, SQL, Python, Tableau, etc)

Since the dataset is rich with information for analyses, I decided to implement some NLP techniques, on the columns, Product, Issue, State, Company and Consumer Complaint narrative. Among the issues the top 3 products were Mortgage, Student Loan and Credit Card.

So to help the financial service providing companies an analysis for the top 10 issues with above products was done.

The following results were found:

Top 10 issues with product mortgage according to count.

	Issue	Count
1	Loan modification,collection,foreclosure	108541
2	Loan servicing, payments, escrow account	72184
3	Application, originator, mortgage broker	15898
4	Trouble during payment process	15334
5	Struggling to pay mortgage	13526
6	Settlement process and costs	8150
7	Other	5425
8	Credit decision / Underwriting	5218
9	Applying for a mortgage or refinancing an existing mort...	2782
10	Closing on a mortgage	2321

Among which loan modification, collection, foreclosure. Loan servicing, payments, escrow account and Application, originator, mortgage broker were the top 3 based on their count. So, financial companies providing
Mortgage as a service can have a look at these top 10 issues and plan their service accordingly to provide quality service to the customer.
Similarly for companies providing student loan and credit card services can have a look at the results below.

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

Top 10 issues with student loan based on count

	Issue	Count
1	Dealing with my lender or servicer	14840
2	Dealing with your lender or servicer	7896
3	Can't repay my loan	7641
4	Repaying your loan	3820
5	Struggling to repay your loan	3260
6	Problems when you are unable to pay	1697
7	Getting a loan	842
8	Incorrect information on your report	666
9	Problem with a credit reporting company's investigation...	201
10	Credit monitoring or identity theft protection services	36

Top 10 issues with credit card based on count

	Issue	Count
1	Billing disputes	14074
2	Other	8698
3	Identity theft / Fraud / Embezzlement	7864
4	Closing/Cancelling account	5888
5	APR or interest rate	5229
6	Late fee	3396
7	Customer service / Customer relations	3201
8	Delinquent account	2924
9	Credit determination	2865
10	Advertising and marketing	2635

We can do this analysis for any product as per the requirements.

After the top 10 issue with product, some analysis was done to fetch commonly used top 10 words by the consumer for banks Wells Fargo and Bank of America. It can be done for any company in the dataset.

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

Results for most used words for wells fargo.

Top 10 most common words in consumer complaint narrative using NLP for companies

1. Wells Fargo

```
# A tibble: 10 x 2
       word     n
     <fctr> <int>
 1     xxxx 95938
 2    fargo 26652
 3       xx 24429
 4  account 13936
 5     loan 10299
 6   credit  8485
 7     bank  8014
 8 mortgage  7979
 9  payment  7940
10     told  6582
```

Similarly for BOFA 8,502 records were fetched

	X	Product	Issue	State	Consumer.complaint.narrative	Company
1	23	Debt collection	Communication tactics	TN	Bank of America has called 50 times in the past 30 minut...	BANK OF AMERICA, NAT
2	24	Checking or savings account	Closing an account	FL	I LIVE AT THE HOMELESS SHELTER DOWNTOWN, A FEW D...	BANK OF AMERICA, NAT
3	30	Checking or savings account	Problem with a lender or other company charging your ...	CA	FACT : My name is XXXX XXXX. On XX/XX/XXXX, Bank of Am...	BANK OF AMERICA, NAT
4	37	Checking or savings account	Managing an account	CA	On XX/XX/XXXX I deposited a check in the amount of XXX...	BANK OF AMERICA, NAT
5	48	Debt collection	Attempts to collect debt not owed	SC	On XX/XX/XXXX at XXXX XXXX I received a call from XXXX. T...	BANK OF AMERICA, NAT
6	57	Bank account or service	Account opening, closing, or management	NY	Bank of America let me apply for a mortgage for a cond...	BANK OF AMERICA, NAT
7	65	Mortgage	Application, originator, mortgage broker	CA	Bank of America processed a predatory loan modificatio...	BANK OF AMERICA, NAT
8	70	Checking or savings account	Opening an account	GA	My purse and car were stolen a few years back. As a res...	BANK OF AMERICA, NAT
9	97	Credit card or prepaid card	Fees or interest	TX	Bank of America has charged interest on my account im...	BANK OF AMERICA, NAT
10	98	Bank account or service	Account opening, closing, or management	FL	Good Afternoon. My wife and I have multiple accounts ...	BANK OF AMERICA, NAT
11	107	Bank account or service	Deposits and withdrawals	WA	On XXXX/XXXX/XX/XX/2015 A customer paid a bill at my re...	BANK OF AMERICA, NAT
12	112	Mortgage	Settlement process and costs	FL	I began a loan modification with boa for a heloc XX/XX/X	BANK OF AMERICA, NAT

After this, sentiment score was calculated, since this is a complaint dataset, ona scale of -0.1 to -5 top 6 negative comments were fetched for Wells fargo and Bank Of America.

Sentiment score for bofa top 6 negative comments

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

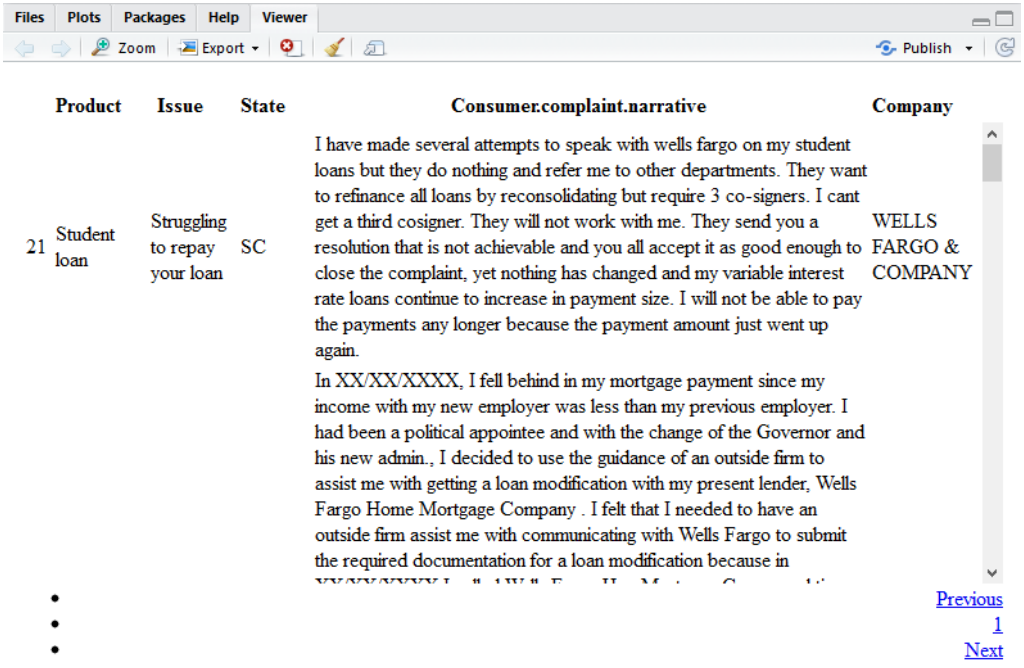
-Radhika Mittal

```
> sentiment_lines = calculate_sentiment(bofa)
> head(sentiment_lines)
# A tibble: 6 x 3
      X  sentiment words
  <int>      <dbl> <int>
1    23 -1.0000000     8
2    30 -0.7741935    31
3    37  0.0625000    32
4    48 -1.1538462    13
5    57 -0.8000000     5
6    65 -0.1250000     8
>
```

For Wells Fargo:

```
> sentiment_lines = calculate_sentiment(wellsfargo)
> head(sentiment_lines)
# A tibble: 6 x 3
      X  sentiment words
  <int>      <dbl> <int>
1    21  1.0000000     7
2    29 -0.2340426    47
3    59 -0.7647059    17
4    64 -0.7647059    17
5    95 -1.4000000     5
6   107 -1.8000000    10
```

Unique integer id's of the complaints was also fetched to read the complaint using the id.



AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

	Product	Issue	State	Consumer.complaint.narrative	Company
29	Mortgage	Struggling to pay mortgage	MD	rep. at XXXX (in a separate conversation) who also felt that XXXX XXXX was very combative, that I was behind in my property taxes that are by the way escrowed and paid by Wells Fargo which is why the {$5500.00}. was due. This was far from the truth. In fact, XXXX XXXX and I spoke and she stated that she researched it with the taxed dept. and it was concluded that the burden of responsibility was on Wells Fargo and that they had to pay the $ XXXX.in order to get the quit claim deed recorded. XXXX XXXX stated that I was absolutely correct in stating that I was not behind in my property taxes like XXXX XXXX had raised his voice and said and that Wells Fargo 's position was that they would have to make a special request to XXXX XXXX XXXX County court to send them an invoice earlier than normal and that was what they were moving towards. XXXX XXXX did state that she was not a home preservation specialist but she was trying to assist me. Another week passed and suddenly Wells Fargo decided to move forward with reviewing my file and made a decision to reopen my file for a decision. Something in the back of my mind told me not to trust Wells Fargo because after several weeks of demanding a quit claim	WELLS FARGO & COMPANY

Previous 1 Next

To do this, text mining, tidytext libraries were used in R. The logic used behind this analysis is almost like SQL queries.

Metadata Information:

The attributes in the dataset with their respective datatypes:

Date: Ordinal

Product: Categorical/Nominal

Sub-Product: Categorical/Nominal

Issue: Categorical/Nominal

Sub-Issue: Categorical/Nominal

Consumer Complaint Narrative: None Plaintext

Company public response: None Plain Text

Company: Categorical/Nominal

State: Ordinal

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

Zip Code: Ordinal

These were some of the important ones.

Are there any privacy, quality, or other issues with this data?

It had a lot of null values, making it difficult to use for analysis with accurate results, thus that wa taken care of by removing the null values. Data was manipulated and pre-processed for applying NLP techniques.

Visualization:

Summary Statistics:

```
> summary(custcomplaint)
   Date.received                                                                                  Product
 9/8/2017 :   2911   Mortgage                                                                       :250743
 9/9/2017 :   2246   Debt collection                                                                :198129
 1/19/2017:   1697   Credit reporting, credit repair services, or other personal consumer reports:145477
 1/20/2017:   1340   Credit reporting                                                               :129628
 9/13/2017:   1274   Credit card                                                                    : 82708
 4/5/2018 :   1138   Bank account or service                                                        : 80940
 (Other)  :1038037   (Other)                                                                        :161018
                     Sub.product                                          Issue
                           :217360   Loan modification,collection,foreclosure:108541
 Credit reporting          :142676   Incorrect information on credit report  : 95529
 Other mortgage            : 85541   Incorrect information on your report    : 86044
 Checking account          : 75440   Loan servicing, payments, escrow account: 72184
 Conventional fixed mortgage: 65495   Cont'd attempts collect debt not owed   : 54661
 I do not know             : 40620   Account opening, closing, or management : 35721
 (Other)                   :421511   (Other)                                 :595963
                                                Sub.issue
                                                     :478611
 Account status                                      : 35241
 Information belongs to someone else                 : 34898
 Debt is not mine                                    : 33220
 Information is not mine                             : 30010
 Their investigation did not fix an error on your report: 22725
 (Other)                                             :413938
```

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

```
ncated>
                                                                                Company.public.response
                                                                                               :715217
Company has responded to the consumer and the CFPB and chooses not to provide a public response:223313
Company believes it acted appropriately as authorized by contract or law                      : 46601
Company chooses not to provide a public response                                              : 45612
Company believes the complaint is the result of a misunderstanding                            :  4311
Company disputes the facts presented in the complaint                                         :  3974
(Other)                                                                                       :  9615
                                Company           State           ZIP.code                                Tags
EQUIFAX, INC.                       : 88678   CA     :145492          : 67285                                 :908963
Experian Information Solutions Inc.   : 79741   FL     :101204   300XX  :  4160   Older American               : 65377
BANK OF AMERICA, NATIONAL ASSOCIATION : 74964   TX     : 85765   770XX  :  3695   Older American, Servicemember: 11104
TRANSUNION INTERMEDIATE HOLDINGS, INC.: 72720   NY     : 71703   331XX  :  3076   Servicemember                : 63199
WELLS FARGO & COMPANY               : 62825   GA     : 53456   606XX  :  2980
JPMORGAN CHASE & CO.                : 52250   IL     : 40497   334XX  :  2940
(Other)                             :617465   (Other):550526   (Other):964507
       Consumer.consent.provided.     Submitted.via    Date.sent.to.company                Company.response.to.consumer Timely.response.
            : 16121                        :    68   9/8/2017 :   2779    Closed with explanation       :806972         :    68
Consent not provided:244821       Email      :   375   9/9/2017 :   2204    Closed with non-monetary relief:129464        No :  26760
Consent provided    :229018       Fax        : 17936   1/19/2017:   1323    Closed with monetary relief    : 62260        Yes:1021815
Consent withdrawn   :  1090       Phone      : 69861   1/20/2017:   1211    Closed without relief          : 17868
N/A                 :542027       Postal mail: 64213   9/13/2017:   1211    Closed                         : 16304
Other               : 15566       Referral   :164491   4/10/2018:   1059    In progress                    :  5358
                                  Web        :731699   (Other)  :1038856    (Other)                        : 10417
Consumer.disputed.  Complaint.ID
   :    68         Min.   :      1
N/A:336816         1st Qu.: 921200
No :575882         Median :1859747
Yes:135877         Mean   :1746258
                   3rd Qu.:2662855
                   Max.   :3070683
                   NA's   :68
```

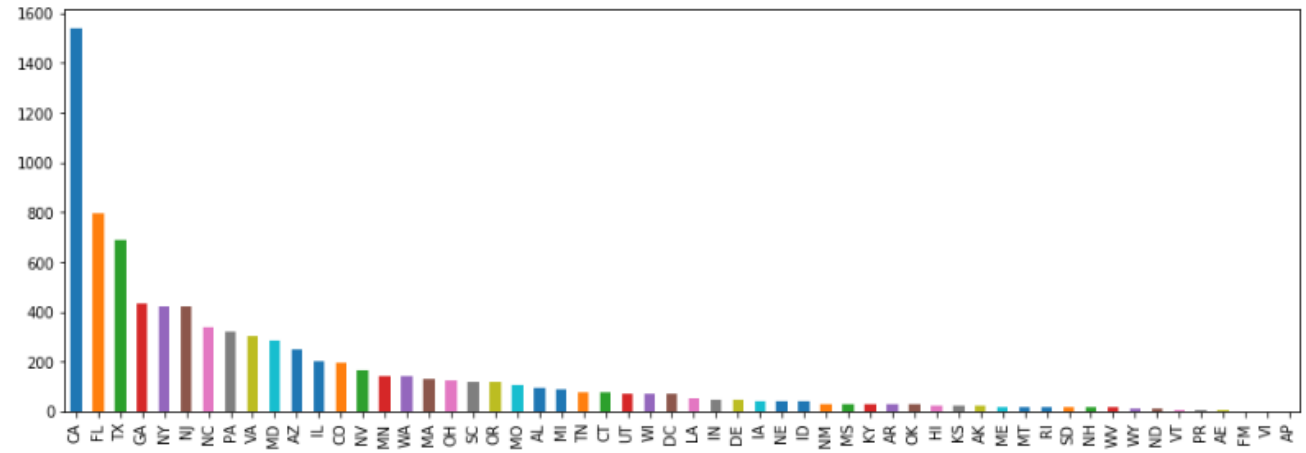
Summary statistics for this kind of data is just for the record.

Bar plot:

Bar plot were plotted according to state for particular companies, customer complaints was calculated to see that California has the most complaints for Bank of America, Wells Fargo and Equifax. These 3 were chosen for comparison purpose. We can run the code for any other company names from the dataset too. This was done in Python using Pandas and Matplotlib, following were the results obtained.

```
import matplotlib
pandas.value_counts(wellsfargo['State']).plot.bar(figsize=(15, 5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xefa799e860>
```

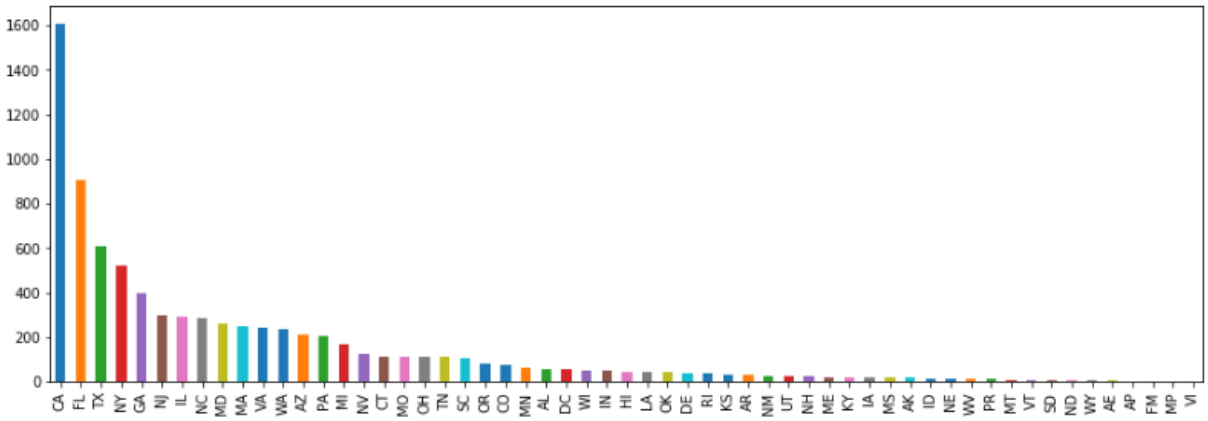


AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

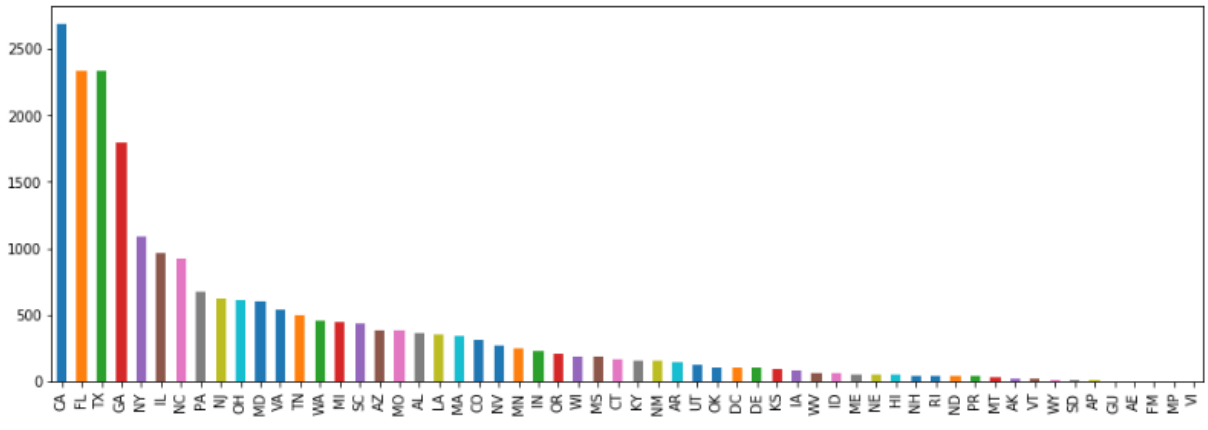
In [63]:
```
bofa = pandas.read_csv('C:/Users/HP/Desktop/bofacustcomplaint.csv', encoding='utf-8')
pandas.value_counts(bofa['State']).plot.bar(figsize=(15, 5))
```

Out[63]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xef90c6e748>
```



In [65]:
```
equifax = pandas.read_csv('C:/Users/HP/Desktop/equifaxcustcomplaint.csv', encoding='utf-8')
pandas.value_counts(equifax['State']).plot.bar(figsize=(15, 5))
```

Out[65]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xefa7afb668>
```



We can see the Y-axes scale change is significant for Equifax as compared to Wells Fargo and Bankof America. This is a good result to see the location and the population as a factor to improve services.

AIT580 Project: Consumer Complaint database for NLP(Sentiment Analysis & Topic Modeling)

-Radhika Mittal

Explain/define terms

- Include *explanation* of any *technical terms* relevant to the project domain

NLP: Natural Language Processing is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.[3]

NLTK: It is a Natural Language Tool Kit used to process natural language and implement NLP.

Libraries: These are used in Python and R to use functions related to data analysis as required. Some of the libraries used here were dplyr, ggplot2, tm, tidytext, textcat etc. in R and Pandas and Matplotlib in Python.

Sentiment Analysis: Also known as opinion mining, as to if the customer has a positive, negative or neutral opinion about the company or a particular service it received.

Topic Modeling: It is a technique to determine topics from document to make it easy to infer what the document or a large text is all about.

References:

[1] “Consumer complaint database”, https://catalog.data.gov/dataset/consumer-complaint-database

[2] “Stack overflow”, https://stackoverflow.com/

[3] “Wikipedia”, https://en.wikipedia.org/wiki/Natural_language_processing