CS 410 Final Project Report

Team Member:

Shengqi Zhou (shengqi7)

Project Overview

The goal of this project is to build an interactive application using CFPB consumer complaints narratives data, which allows users to explore complaints issues (e.g. Identity theft) and complaint monthly trends by combining text retrieval and phrase level topic mining.

This application has two main functions:

1. Part 1: Query based word association discovery

   The user would enter a query (e.g. Identity theft), and the system will retrieve top 10 relevant complaints using BM25+RM3 model and mines the most characteristic phrases that co-occur with the query.

2. Part 2: Monthly trending topics discovery

   The user would enter a date range and an optional company/bank filter, and the system will output frequent multi-word phrases and tracks how often they appear in complaints narratives over the time user entered, and then the system will visualize that as a monthly trend table and chart.

Dataset and Preprocessing

The system uses the CFPB Consumer complaints dataset stored as complaints_data.csv. Preprocessing is designed both to clean the data and to normalize it for retrieval:

Link to dataset:

https://www.consumerfinance.gov/data-research/consumer-complaints/search/?chartType=line&dateInterval=Month&dateRange=3y&date_received_max=2025-12-10&date_received_min=2022-12-10&lens=Product&searchField=all&subLens=sub_product&tab=Trends

I further filtered the data by below criteria to get the final dataset I use for this project:

- Date received complaints: Nov 2022 to Nov 2025
- Company name: American Express, Bank of America, JPMorgan Chase, Navy Federal Credit Union

Data preprocessing:

1. Data normalization: The Date received column is stripped, changed into datetime format. The Consumer complaint narrative field is converted to lowercase, and rows with missing complaint narratives are dropped.

2. Censored data handling: Censored characters (e.g., “XXXX”) are replaced with the token “[REDACTED]” to explicitly mark censored data while keeping their presence.

Text Retrieval and Text Mining

1. Indexing: This project uses Pyserini (Lucene) package to build a search index and writes each document as a separate JSON file containing id and narratives. A Lucene index is built with pyserini.index.lucene and configured to store positions, docvectors, and narratives. Indexing is skipped if an existing index is detected.

2. Retrieval Model: The primary model used in this application for retrieving relevant consumer narratives related to query is BM25 with RM3 pseudo-relevance feedback
3. Mining Association Phrases: Based on the retrieved result, the system would mine co-occurring multi-word phrases (Bigrams/Trigrams). After filtering of function words, bank names, generic nouns, and phrases that are essentially just the same as the query. The system would count each phrase and rank them by document frequency.

4. Top Phrases/Topic mining: The system also identifies “topic-like” phrases over the full corpus or over a date range and company user entered. It aggregates document-level counts and selects the most frequent, informative bigrams/trigrams as topics. These phrases are later used to build monthly trend tables and plots that show how major complaint themes evolve over time.

Interactive Application

Below shows how the user would be able to interact with the system:

Part 1 – Query based word association discovery

- Inputs:
    - Free-text query.
    - Optional year filter.
- Outputs:
    - Filter summary: query, year filter, and number of complaints retrieved by RM3.
    - A list of top associated phrases with their document counts.
    - A list of top 10 retrieved complaints, showing:
        - date, company, issue label, RM3 score, narrative.

Part 2 – Monthly trending topics discovery

- Inputs:
    - Optional start date and end date.
    - Optional company filter.
    - Number of phrases to plot in the trend chart.
- Outputs:
    - Filter Summary showing selected filters and number of topic phrases tracked.
    - Table of topic phrases with document counts.
    - Monthly trend table for the top phrases (rows = months, columns = phrases).
    - Trend chart showing how complaint topics change over time by month.

Evaluation of retrieval and mining results

- Part 1 Retrieval results evaluation

For the example query “identity theft”, I manually inspected the top 10 complaints returned by the RM3 searcher.

Below screenshot shows top 3 retrieved results as an example of output:

```
Rank 1
  RM3 score: 2.3142
  Date:      2024-09-14 00:00:00
  Company:   BANK OF AMERICA, NATIONAL ASSOCIATION
  Issue:     Took or threatened to take negative or legal action
  Narrative:
    i am a victim of identity theft. an identity thief used my personal information to execute a loan with [redacted]
[redacted] [redacted] on [redacted]/[redacted]/[redacted], although the fraudulent loan was removed from all 3 credit
reports by [redacted]/[redacted]/[redacted] [redacted] [redacted] [redacted] threatened to steal my home for payment.
the fbi defines housr stealing as indentity theft and mortgage fraud. i've attached my theft affidavit abd ftc identity
theft report.

Rank 2
  RM3 score: 2.2683
  Date:      2024-03-04 00:00:00
  Company:   BANK OF AMERICA, NATIONAL ASSOCIATION
  Issue:     Improper use of your report
  Narrative:
    i recently noticed that i am a victim of identity theft, and i am requesting an extended fraud alert to be immediately
placed in my credit file so that no new credit information or applications will be approved or issued until the lender
first verified my identity.fortunately, in accordance with the fcra sections 605b,615 ( f ) and 623 ( a ) ( 6 ), an
identity theft report can be used to permanent block fraudulent information that results from identity theft, such as
accounts or addresses from appearing on a victims credit report. identity theft reports can prevent a company from
continuing to collect debts that result from identity theft. pursuant to the fcra 605b ( 15 u.s.c. 1681c-2 ), please
block all information resulting from idenitiy theft, except as otherwise provided in this section a, in which a consumer
reporting agency shall block the reporting of any information in the file of a consumer identifies as information
resulted from an alleged identity theft, no later than [redacted] business days

Rank 3
  RM3 score: 2.2475
  Date:      2024-12-02 00:00:00
  Company:   NAVY FEDERAL CREDIT UNION
  Issue:     Incorrect information on your report
  Narrative:
    i am filing this cfpb complaint to request pursuant to fcra and [redacted] [redacted] u.s.c and [redacted], that all
credit report agencies, block information showing on my consumer credit report that show identity theft and fraud within
4 business days of receiving this complaint. i am filing the ftc identity theft report to a fraudulent account,
appearing on my consumer file credit report. i do not recognize this account, which results in identity theft and fraud.
this fraudulent account was open between the years of [redacted] through [redacted] on the day of
[redacted]/[redacted]/[redacted]. this account has resulted in a fraudulent inquiry appearing on my consumer fico
credible report. i demand the aforementioned fraudulent account to be removed from each other one of my consumer credit
report and all activities that was being said to item cease pursuant to my rights with the fcra and fdcpa law. the
aforementioned account that was listed is a result of identity theft and fraud. i have also attached a copy of my ftc
```

Using a simple 3-level relevance scale

- Highly relevant – identity theft is the central topic of the complaint
- Relevant – identity theft is present but secondary to another main issue
- Not relevant

Rank	Issue label	Relevance	Justification
1	Took or threatened to take negative/legal action	Highly relevant	Explicitly describes a fraudulent loan and "identity theft" multiple times.
2	Improper use of your report	Highly relevant	Detailed discussion of credit report fraud alerts and blocking identity-theft items.
3	Incorrect information on your report	Highly relevant	Focused on removing fraudulent accounts caused by identity theft.
4	Credit monitoring or identity theft services	Highly relevant	Account taken during identity theft; discussion of restoration and credit impact.
5	Incorrect information on your report	Highly relevant	Formal complaint about a fraudulent account opened due to identity theft.
6	Incorrect information on your report	Highly relevant	Direct statement "I am a victim of identity theft…" and credit report harm.
7	Incorrect information on your report	Highly relevant	Multiple fraudulent accounts opened, police report filed for identity theft.
8	Improper use of your report	Highly relevant	Requests deletion of items because "I am a victim of identity theft."
9	Unexpected or other fees	Highly relevant	Benefits fraudulently used after identity theft; dispute with bank about reimbursement.
10	Credit monitoring or identity theft services	Highly relevant	Fraudulent Navy Federal account; cites identity-theft laws and affidavits.

All 10 retrieved complaints describe identity theft and its consequences (fraudulent accounts, incorrect information on reports, or stolen benefits). Under this relevance definition:

- Precision@10 = 10/10 = 1.0
- Precision@5 = 5/5 = 1.0

Even though the Issue labels vary ("Incorrect information on your report", "Unexpected or other fees", "Took or threatened to take legal action", etc.), the narratives themselves are consistently about identity theft. This confirms that narrative-based retrieval is crucial in this dataset, because relying only on the issue labels would underestimate how many complaints are truly identity-theft-related.

- Part 1 Word association results evaluation

```
--------------------------------------------------------------------
PART 1 - TOP ASSOCIATED PHRASES
--------------------------------------------------------------------
 1. personal information  (appears in 33 complaints)
 2. trade commission  (appears in 23 complaints)
 3. social security  (appears in 18 complaints)
 4. block information  (appears in 16 complaints)
 5. fraudulent accounts  (appears in 14 complaints)
 6. made result  (appears in 13 complaints)
 7. copy ftc identity  (appears in 12 complaints)
 8. remove fraudulent  (appears in 12 complaints)
 9. agency block information  (appears in 11 complaints)
10. made result identity  (appears in 11 complaints)
11. violation rights  (appears in 11 complaints)
12. filed police  (appears in 11 complaints)
13. relate transaction made  (appears in 10 complaints)
14. information resulted  (appears in 10 complaints)
15. current address  (appears in 10 complaints)
16. fraudulent also  (appears in 10 complaints)
17. hard inquiries  (appears in 10 complaints)
18. jpmcb card  (appears in 10 complaints)
19. transaction made result  (appears in 9 complaints)
20. fraudulent information  (appears in 9 complaints)
```

I conducted a manual evaluation of the top 20 associated phrases returned for the query “identity theft”. A phrase was decided to be accurate if it was directly related to identity-theft (e.g., personal information, trade commission, social security, fraudulent accounts, block information, filed police, hard inquiries, fraudulent information).

- Number of phrases evaluated: 20
- Number of phrases judged accurate: ≈ 14
- Phrase Precision@20: 14 / 20 ≈ 70%

In summary, about 70% of the highest-ranked phrases are semantically meaningful and useful for analysts, while the remaining ~30% are more noisy or awkward constructions (e.g., made result, relate transaction made).

- Part 2 Topic phrases results evaluation

```
--------------------------------------------------------------------
PART 2 - FILTER SUMMARY
--------------------------------------------------------------------
Date range:   2024-01-01  ->  2024-12-31
Company:      All companies
Number of topic phrases tracked: 20

--------------------------------------------------------------------
PART 2 - SELECTED TOPIC PHRASES
--------------------------------------------------------------------
 1. debit card  (appears in 1279 complaints)
 2. identity theft  (appears in 1091 complaints)
 3. phone number  (appears in 1061 complaints)
 4. never received  (appears in 1052 complaints)
 5. money back  (appears in 1020 complaints)
 6. personal information  (appears in 991 complaints)
 7. reached out  (appears in 851 complaints)
 8. financial bureau  (appears in 803 complaints)
 9. social security  (appears in 803 complaints)
10. fraud department  (appears in 794 complaints)
11. resolve issue  (appears in 769 complaints)
12. gift card  (appears in 750 complaints)
13. received letter  (appears in 727 complaints)
14. same day  (appears in 725 complaints)
15. third party  (appears in 689 complaints)
16. received email  (appears in 688 complaints)
17. next day  (appears in 681 complaints)
18. call back  (appears in 679 complaints)
19. there nothing  (appears in 669 complaints)
20. several times  (appears in 645 complaints)
```

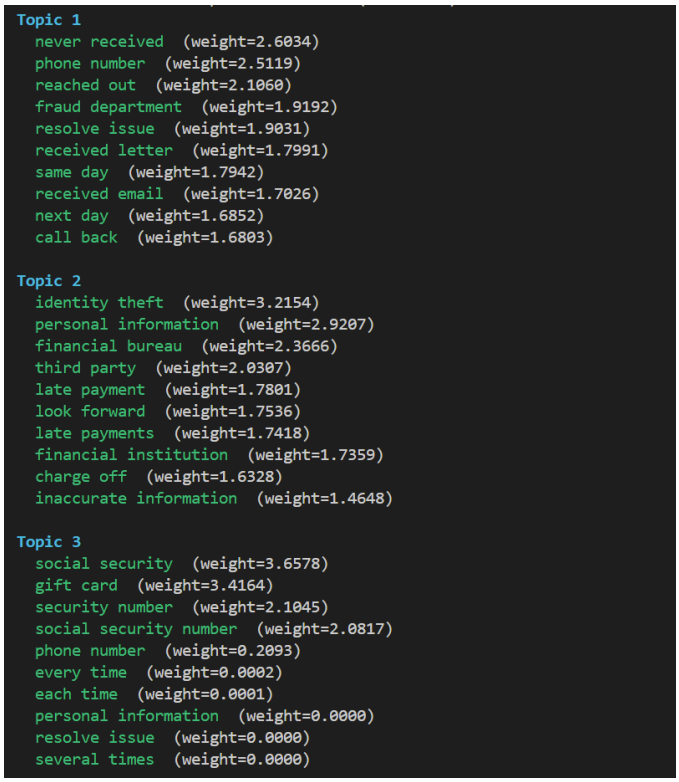
In terms of topic phrases results, a phrase was considered as on-topic if it captured a meaningful complaint topic (card/fraud, identity theft, contact channel, or dispute resolution). Examples of strong phrases include “debit card,” “identity theft,” “money back,” “personal information,” “social security,” “fraud department,” “gift card,” and “third party.” Phrases like “same day,” “next day,” “several times,” and “there nothing” were considered to be noisy or too generic to be useful topics, even though they frequently occur in narratives.

- Number of phrases evaluated: 20
- Number of phrases judged meaningful topics: ≈ 14
- Topic Precision@20: 14 / 20 ≈ 70%

Overall, the topic-mining component produces many domain-relevant phrases that summarize key complaint topics (cards, identity theft, government agencies, contact attempts), but still shows ~30% generic phrases that could be further filtered with stricter lexical thresholds.

Additional Experiment

As an additional experiment, I applied a PLSA-style NMF model to the top 50 mined topics/phrases from function Part 2. The model was configured with 5 latent topics and produced groups such as:



While some clusters are interpretable, many phrases still overlap across topics and several generic phrases (e.g., “same day”, “several times”) appear with non-zero weights in multiple topics. Overall, the PLSA layer provides only modest additional structure compared to the direct n-gram frequency method, and topic interpretability remains mixed. I therefore treat this as an exploratory extension rather than the primary topic-mining approach.

Conclusion

In this project, I built a complaint exploration tool that combines text retrieval with phrase-level topic mining on CFPB consumer complaint narratives data. The system successfully supports two workflows: (1) starting from a specific concern such as identity theft and surfacing both highly relevant complaints and their characteristic co-occurring phrases, and (2) scanning the corpus over a chosen time window to see which complaint themes are most prevalent and how they evolve month by month.

The BM25+RM3 retrieval model performed very strongly for the identity-theft use case: manual inspection of the top 10 results yielded Precision@10 = 1.0, with all retrieved complaints describing identity theft and its consequences, even when the issue labels differed. This confirms that narrative-based retrieval is essential for this dataset and that pseudo-relevance feedback is effective at amplifying identity-theft signals. The phrase-mining components provide useful higher-level structure: both the query-based association phrases and the global topic phrases achieved roughly 70% precision@20, indicating that most top phrases are interpretable, domain-relevant descriptors of fraud, identity theft, card issues, or contact attempts, while a smaller portion remain generic or noisy.

Overall, the system demonstrates how relatively simple IR techniques can be combined into an interactive tool that gives analysts both access to detailed analysis of complaint narratives and high-level overview of complaint trends.