Understanding Complaints Against Financial Institutions

**Abstract**

The goal of this project was to create a topic modeling system using natural language process and unsupervised learning model. The dataset was acquired from Consumer Financial Protection Bureau open data which contained consumer complaints against financial institutions’ operations, services, products, etc.

**Design**

Consumer Financial Protection Bureau is an agency of the United States government responsible for consumer protection in the financial sector. It works to prevent unfair, deceptive and abusive practices from financial companies by taking action against those that break the law. Complaints about financial products and services – and any documents a consumer provides – are sent directly to financial firms that generally must respond within 15 days. The CFPB also refer some complaints to other federal agencies. Although, complaints may not necessary against every financial institution, knowing common themes of those complaints and enhance policies and procedures of some complaints can help financial institution lower the number of future similar complaints and retain customers.

This project demonstrates natural language processing can be used to identify common complaint topics to towards different financial institutions.

**Data**

The dataset used in the project is the [Consumer Complaint Narratives](https://www.consumerfinance.gov/data-research/consumer-complaints/search/?dataNormalization=None&dateRange=3y&date_received_max=2021-12-13&date_received_min=2018-12-13&searchField=all&state=IL&tab=Map) from Consumer Financial Protection Bureau. The raw dataset contains 7846 for analysis.

**Algorithms**

Data: The dataset was generated including most complaints from its source by eliminating “Credit Report” or anything similar from Product/Sub-product.

EDA: Rows without Consumer Complaints Narrative was removed from the dataset.

Text Preprocessing: After numbers, punctuations, white spaces were removed from the corpus, it was tokenized with parts of speech. Only Noun, Adj, Verb, Adv were retained for topic modeling.

Topic Modeling: Baseline model was created using TF-IDF vectorizer and LDA model since TF-IDF can be used to visualize topics.

**Tools**

* NumPy and Pandas for data manipulation
* NLTK package for text processing
* Matplotlib for plotting

**Communication**

A page of writeup and a presentation pdf are accessible in my Github repository.