Feature Engineering and Exploratory Data Analysis

Authors

1) Bhavya Haridas
2) Harshitha S Gadadhar
3) Saumil Shah
4) Nikitha Gawde

DATA PREP AND PRE-PROCESSING

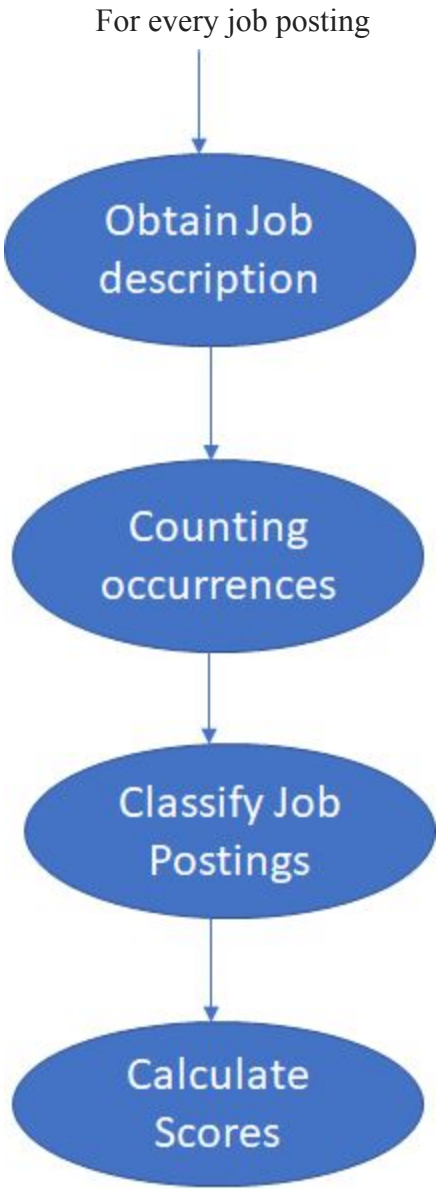
- From each of the 12 teams, the data sets containing top 100 words based on TF-IDF, TextRank, and Wordcount algorithms are compiled into a single file.
- The single data set containing the words from each algorithm are again cleaned, by removing, all the duplicates and handling missing words.

FORMING CLUSTERS(Categories) FOR DIFFERENT AREAS IN FINTECH

- A single dataset of top 100 words is obtained by manually reviewing and cleaning of the data obtained from the 12 teams.
- The 8 clusters chosen are Data analytics and Machine Learning, Network and Cybersecurity, Software development, Audit and Finance, Marketing, sales and accounts, Business intelligence and analysis, Investment and risk management, Admin/clerical/HR
- The 8 clusters are formed manually. The words bucketed into these clusters are keywords related to the Cluster domain obtained from job postings related to the cluster and WEF articles.

FEATURE ENGINEERING

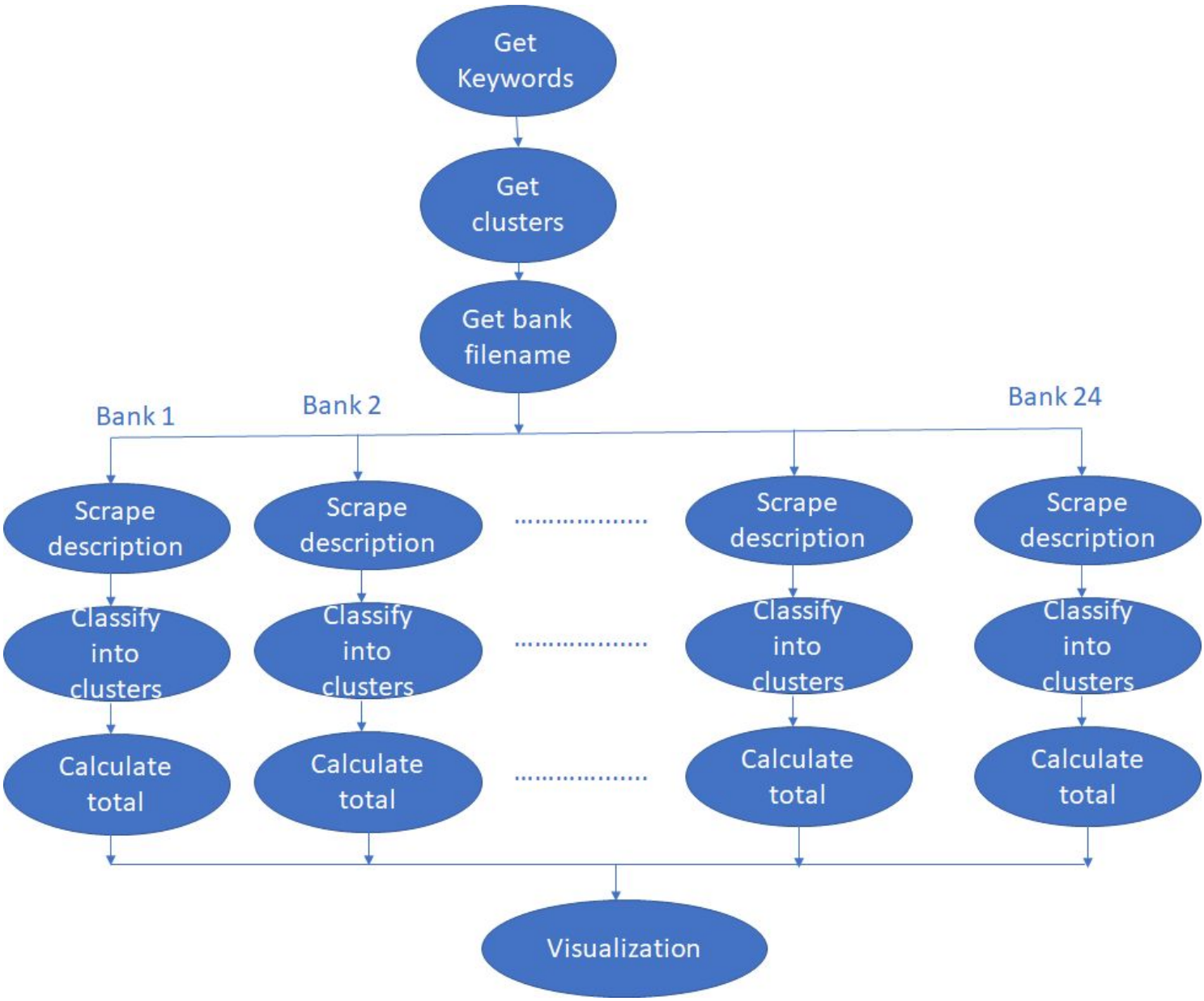
- Feature Engineering is the process of transforming the training data and augment it with additional features so that ML algorithms more effective.
- The keywords are matched with Clusters and bucketed.
- Below is the Flow chart to describing the methodology used



For every job posting in a Bank
Step 1 : First the job description for the particular Job Listing is obtained.
Step 2 : The occurrence of each keyword in the Job listing is counted.
Step 3 : Based on the number of occurrences of the keyword the job bucketed to a cluster.
Step 4 : Once clusters are assigned, based on the scores obtained for each cluster by the job posting, it is tagged either as a Fintech or Non Fintech job.

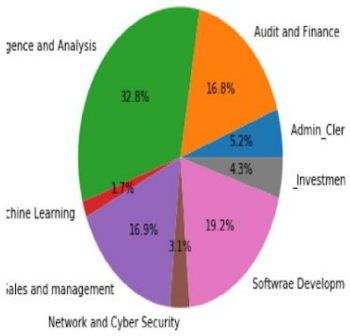
BUILDING A PIPELINE AND AUTOMATING IT

- Python library used - Dask
- A pipeline was built to:
    - Automate data transfer between various stages.
    - Process chunks of data in parallel multi-threads.
- After getting the metadata required for analyzing features, the job description of each posting was matched against the selected keywords in parallel and assigned clusters.

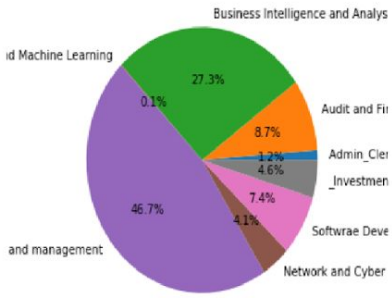


ANALYZING THE DATA AND GAINING INSIGHTS

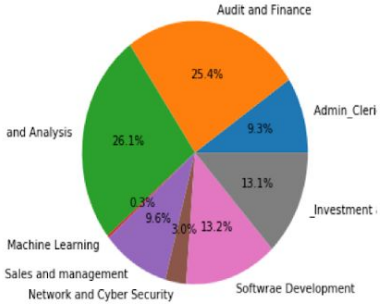
American Express



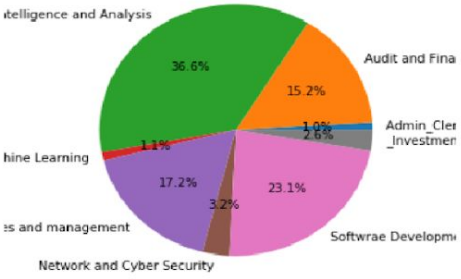
Bank of America



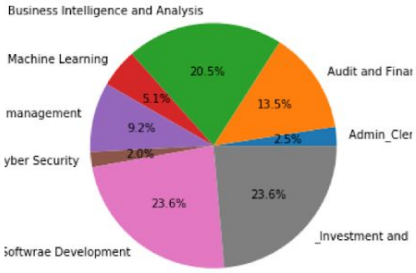
BBT & Corp



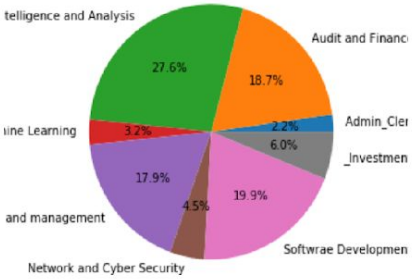
BNY Mellon



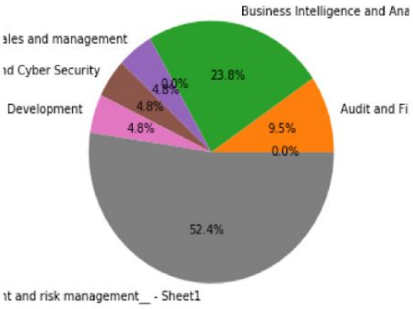
Capital One



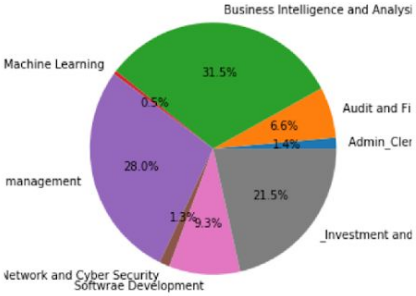
Charles Schwab



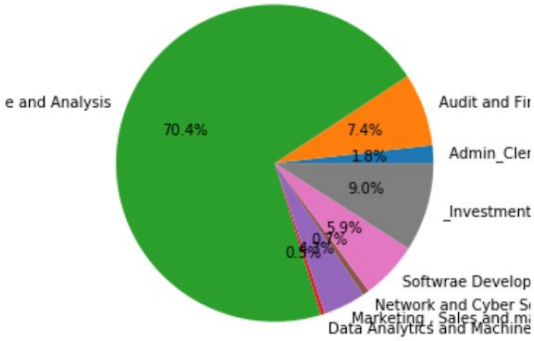
CitiGroup



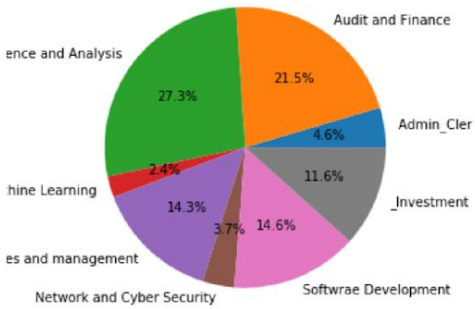
Citizens Bank



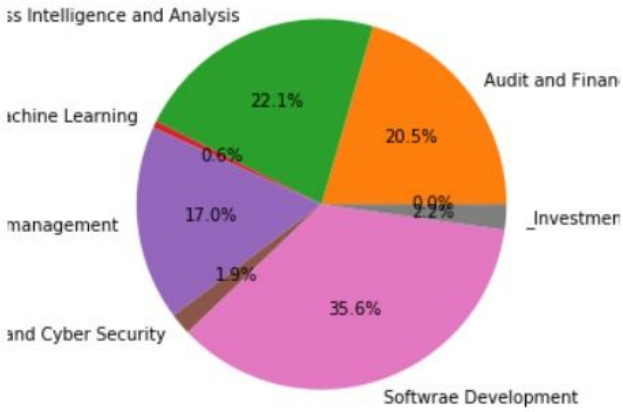
CoAmerica



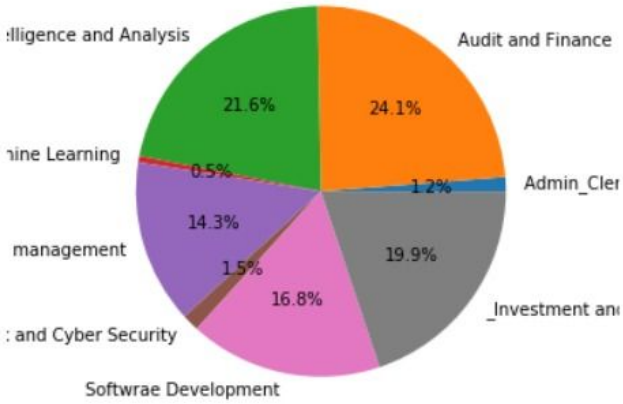
DISCOVER



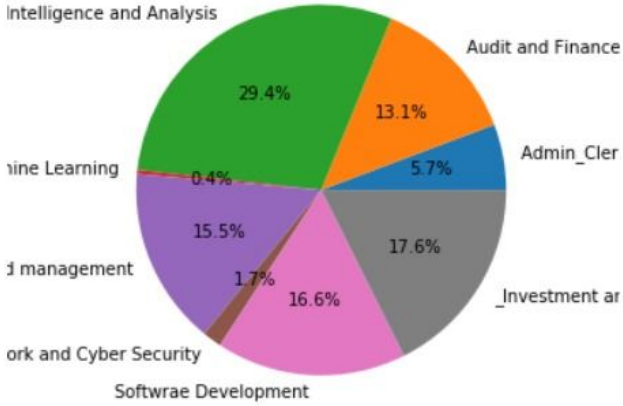
STATE STREET



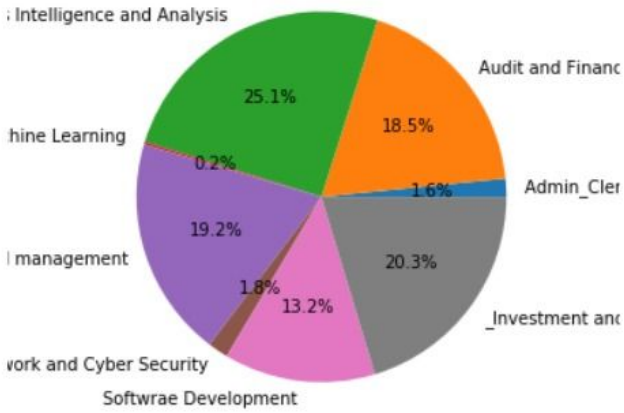
SUNTRUST



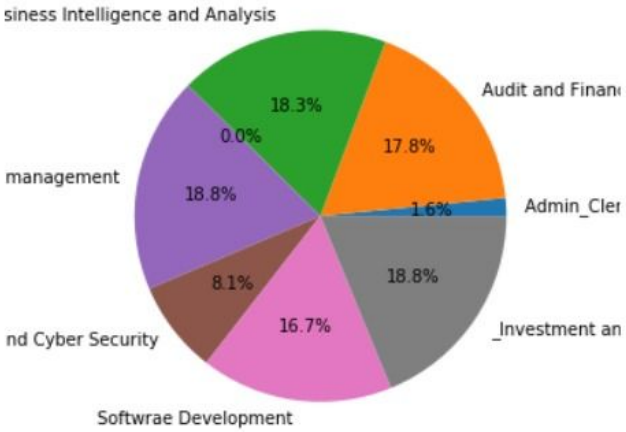
US BANK



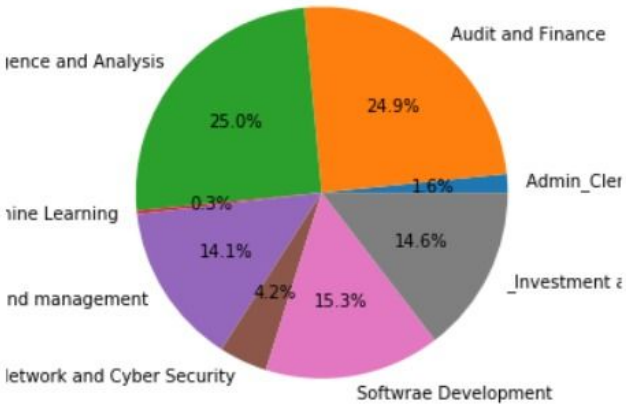
WELLS FARGO



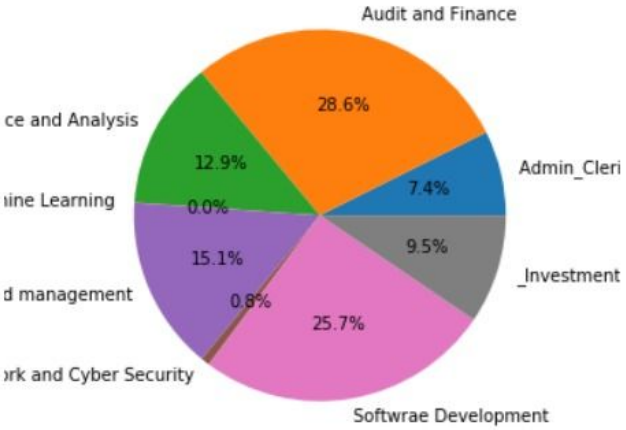
KEYCORP



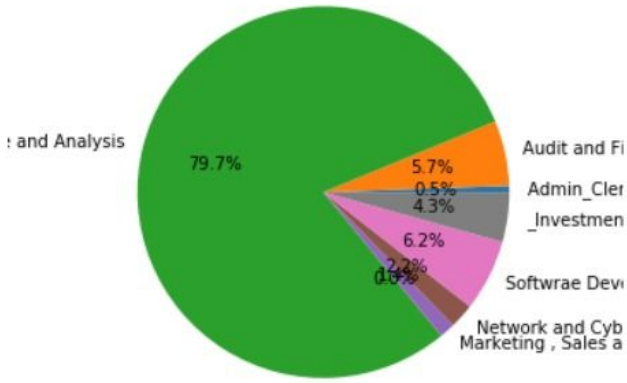
M & T BANK



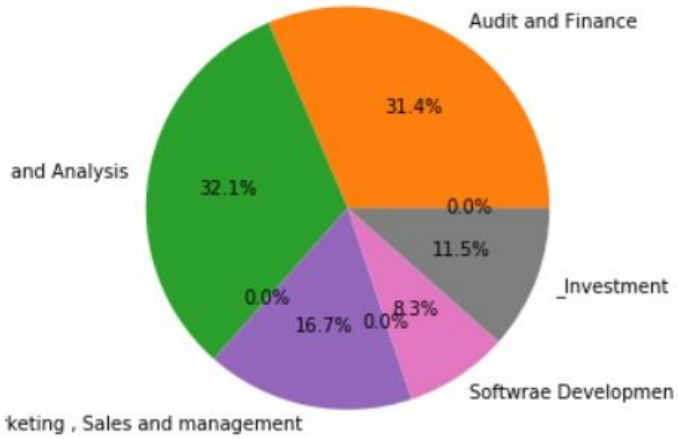
MORGAN STANLEY



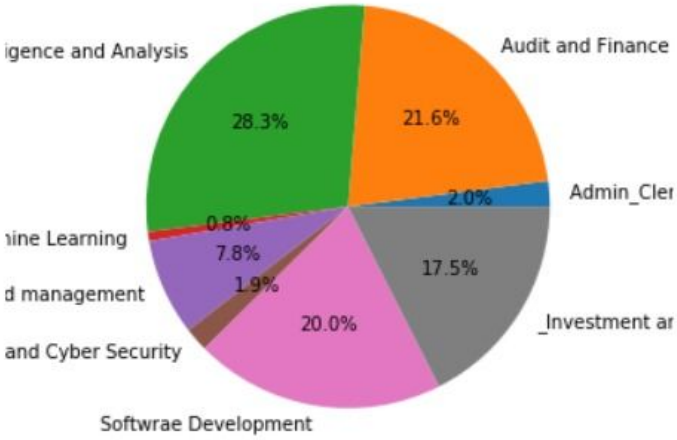
REGIONS BANK



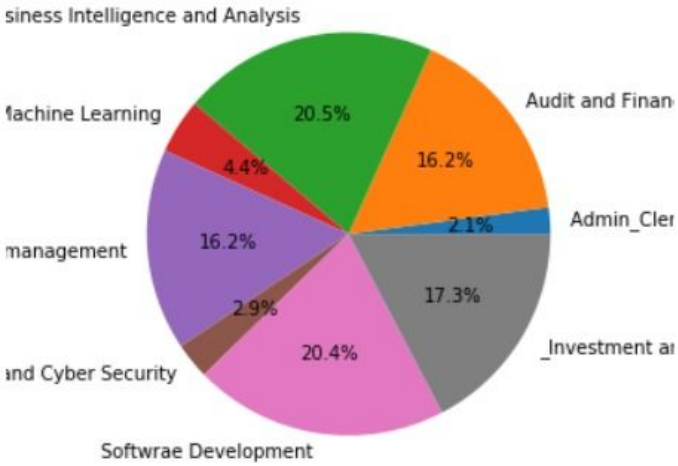
FIFTH THIRD BANK



HUNTINGTON BANK



JP MORGAN



FINTECH AND NON-FINTECH SCORES