Feature Engineering and Exploratory Data Analysis

Summary	Build a dataset with features that can be used to analyze hiring trends in Fintech in the 24 largest banks by market cap in the United States
URL	https://github.com/pratikshsawant5293/ADS_Spring_2019/tree/master/Assignment2
Category	Web
Environment	NA
Status	Version 1
Feedback Link	https://github.com/pratikshsawant5293/ADS_Spring_2019/tree/master/Assignment2
Team No	7
Team Members	Maneendar Sorupaka (sorupaka.m@husky.neu.edu) Pratiksha Sawant (sawant.prat@husky.neu.edu) Rajsharavan Senthilvelan (senthilvelan.r@husky.neu.edu) Reema Mehta (mehta.r@husky.neu.edu)

Introduction

In Assignment 1 we scraped a few reports related to Fintech and gathered key terms that are associated with Fintech. We also scraped jobs from two of the assigned largest banks in US and evaluated hiring trends in these banks.
In continuation of this analysis, we further perform feature engineering and exploratory data analysis on a single dataset collected for 24 US banks.

Subject of Analysis:

Top 24 largest US Banks by market cap

- Bank Of America
- JP MORGAN
- WellsFargo
- CitiGroup
- Morgan Stanley
- Goldman Sachs
- American Express
- US Bank
- Capital One
- BNY Mellon
- BB&T Corp
- State Street
- Suntrust Bank
- Discover Financials
- M&T Bank
- Northern Trust
- Key Corp Bank
- Fifth Third Bank
- Citizen
- Regions
- Charles Schwab
- PNC Bank
- Huntington Bancshares
- Comerica Inc

Data Prep and Pre-processing

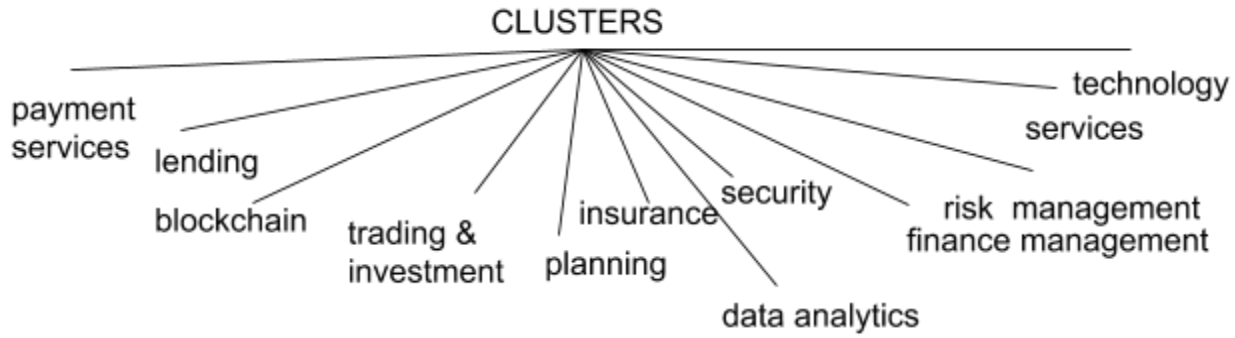
We gathered data of 24 banks from all the 12 teams and combined into a single dataset after removing duplicates/missing values from columns.

Forming Clusters for Different Areas in Fintech

We created clusters from the list of keywords derived by 3 methods in Assignment 1:

- Word count
- TF/IDF
- Text Rank

All the fintech words are grouped into 12 clusters as mentioned below:



payment services	blockchain	trading and investments	planning	lending	insurance	data analytics	security	finance management	risk management	services	technology
payment solution	smart contract	profitability	acquisition	underbanked	ledger	analysis	compliance	equity	fraud	digital banking	automation
payment schemes	dlt	share	shareholders	fund	insurtechs	analytics	fraudulent	revenue	financial risk	banking	ai
emv	blockchain	tax	agile	mortgage	life insurance	big data	regulatory governance	crowdfunding	financial crisis	AWS	cloud
global payment	cryptocurrency	tradeshift	performance	mutfund		prediction	cyber attack	audit	risk capital	saas	machine learning
micropayments		stocks	accuracy	economy		statistics	cyberinsecurity	asset	risk associated	dtcc	networks
		equity	technological advancements	debt		data source	cybersecurity	trade		software	artificial intelligence
		taxation	provisioning	lender		algorithmic	regulatory compliance	cryptocurrency		retail banking	advanced machine learning
		transaction	supply chain management	lend		metadata	authentication	sale		bitpay	devops
		investment management		banked		machine learning	securitization	regulatory		development	process automation tools
		retail	market infrastructure	lending platform			validation	legal entity		financial services	digitization
		trading	business model				security	capital		operations	.net
		funding platform	trading strategy				coco bonds	wealth management		kba	java
		investment	process externalization				identity transaction	overcapitalization		pos	api
								taxation		manufacturing	biometrics
								hedge		product service	cryptography
								monetizing		financial product	framework
								fidelity		mobile payment	sql
								finance		global payment	business intelligence
								fintech		reporting	cognitive computing
								capital raising		cloud	hacking
								financial product		fis	digitalization
								capital market		optimization	visualizations
								financial system			techcrunch
								bitcoin			financial technology
											ledger technology
											fintech
											digital
											etl
											aml
											anti money laundering

Feature Engineering

A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done and feature engineering is a process of using domain knowledge to create features for detailed analysis. To provide a complete understanding of the hiring trends of US Banks in Fintech we add two features in the dataset:

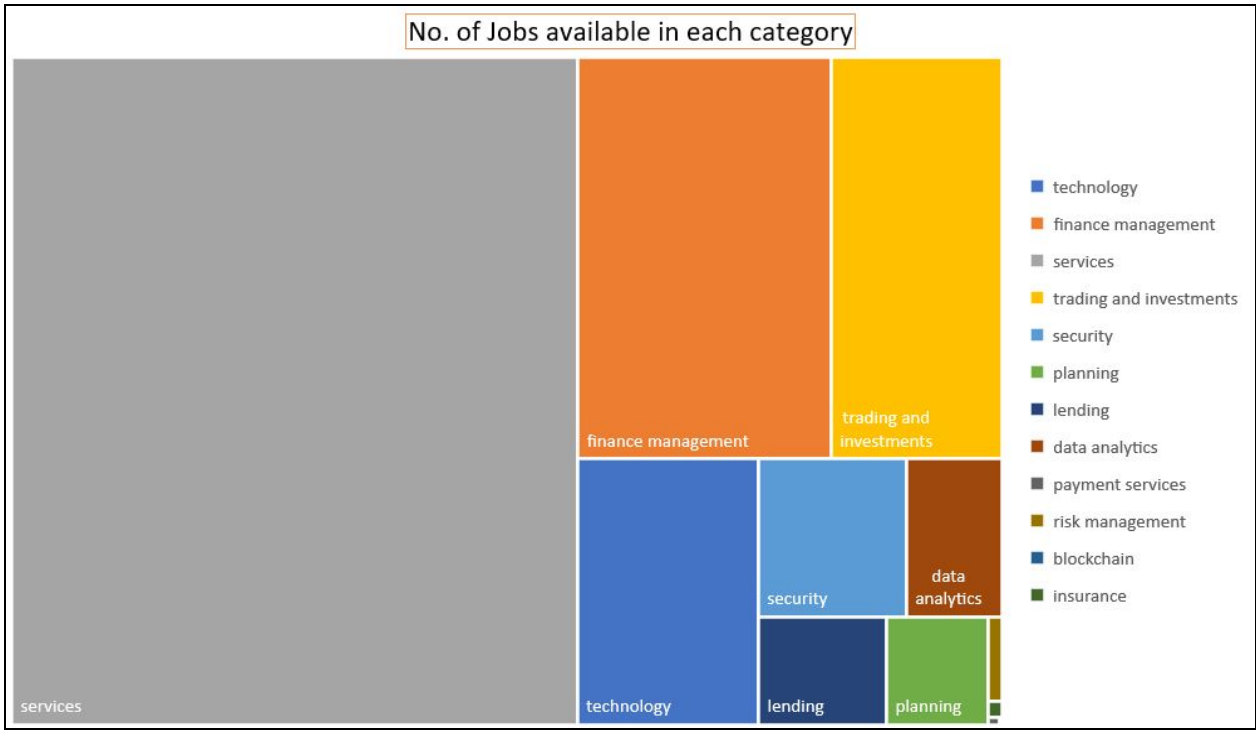
- ***Classifying each job description into different clusters***
  Each job description is assigned to a suitable cluster based on the count of that cluster. The counter for each cluster increments if any word from that cluster appears in the job description.

- ***Identifying the focused area/fintech under each cluster***
  Each job description is featured to identify the focused area/fintech under the assigned cluster. For example if Bank of America has Job J1 which is categorized into technology, we would determine the focused area under technology cluster by selecting the maximum count of Fintech words under that cluster.

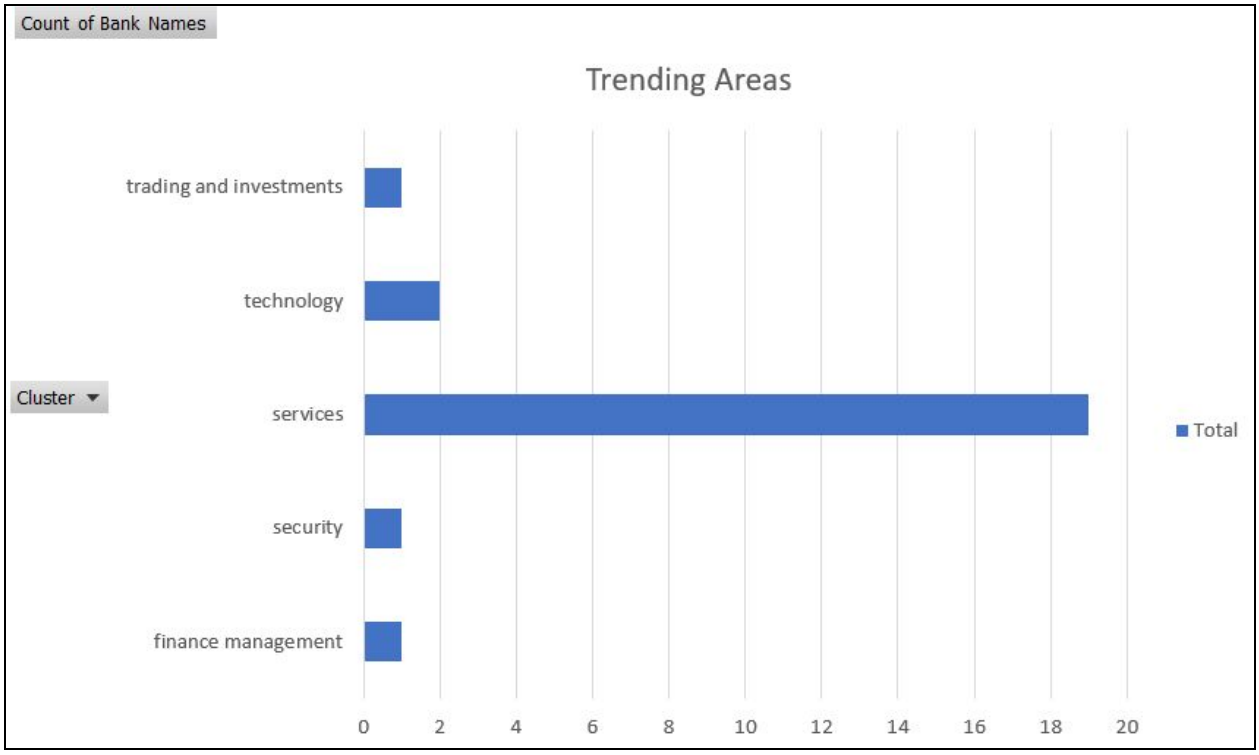
Analysis

We analyze the dataset generated above to answer a few questions listed below, which provides us a detailed study of the job hiring trends in fintech industry

- **How are the top 24 banks hiring?**
  All the 24 banks are majorly hiring for professionals skilled in services and financial management domain having skill sets in SAAS, AWS, crowdfunding monetizing etc. However, there are proportionate job opportunities in trading & investments, technology sector, data analytics, and security sector

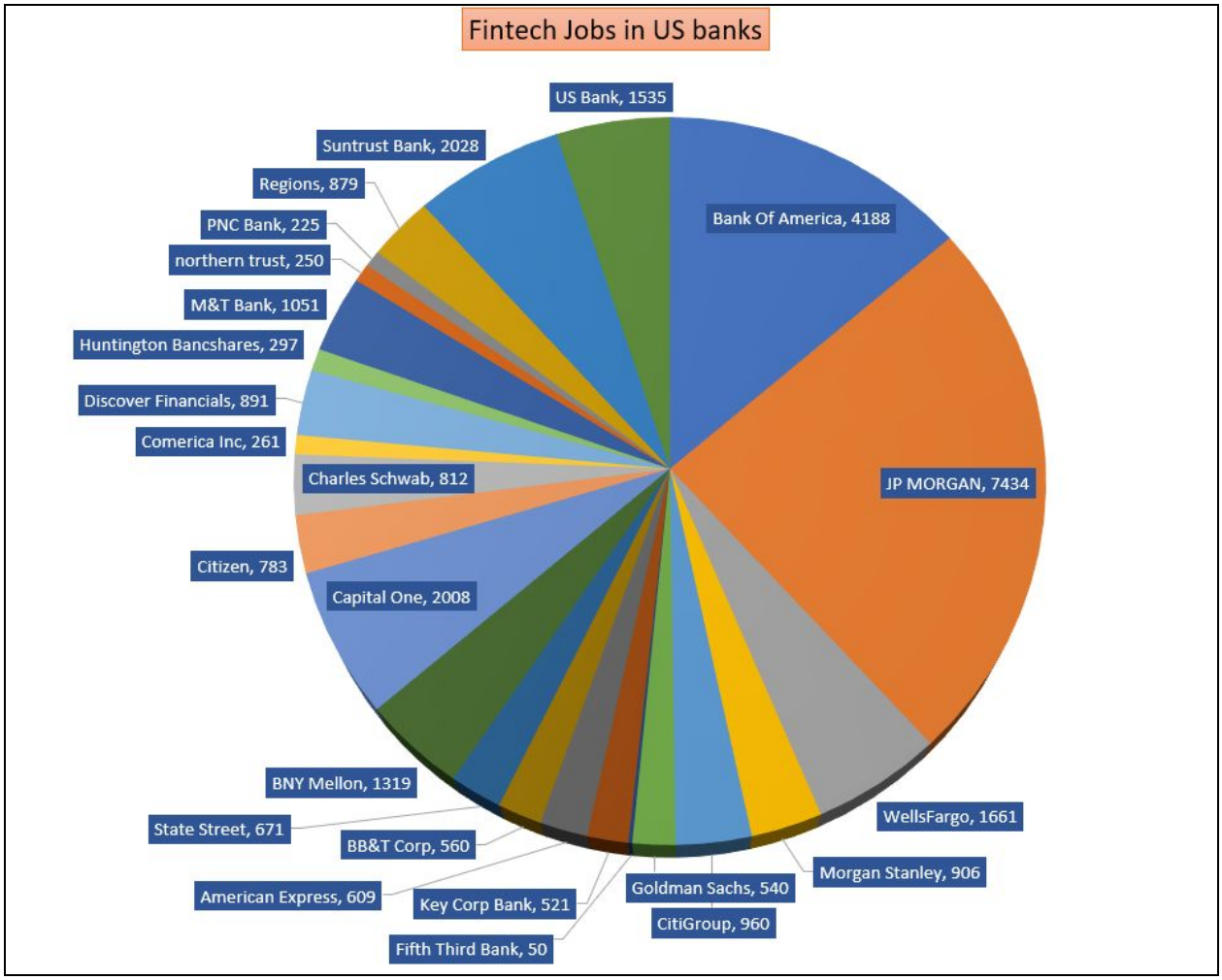


- **How are the fintech related job hiring trends?**
  The data visualization given below represents total number of banks vs top 5 hiring domains in each bank. We analyze that out of 24 banks, more than 18 banks are hiring under services category and looking for professionals who are skilled to work in development, reporting, digital banking, operations, cloud etc.

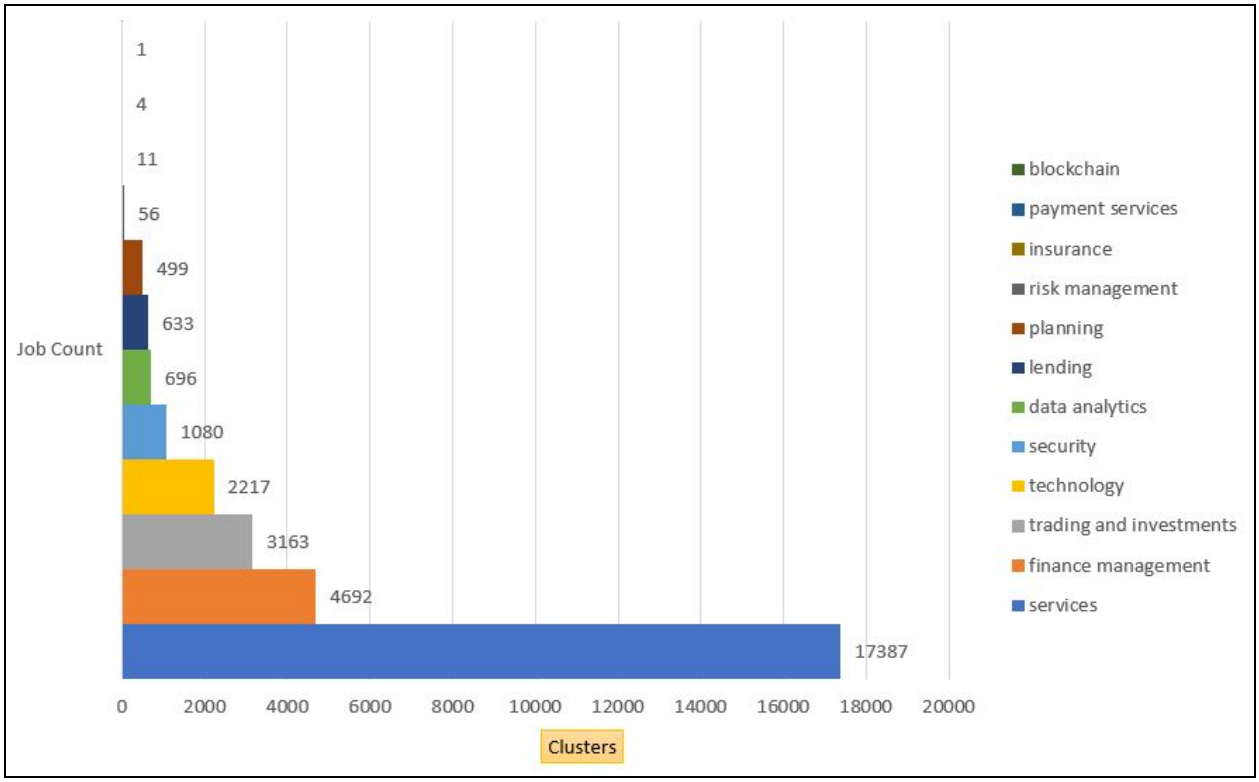


We fetched what is the focused area from the clusters for each job description and calculated count of most focusing area of the fintech world. This word cloud is showing what are the current sectors where these banks are hiring.So as per analysis, banking, software, development, financial services and operations are some area, where jobs openings heading.

- **Which companies have the most fintech related jobs and which ones least?**
  We took the data of 24 banks and segregated based on fintech sectors they are associated with and summation of data brought us in outcome that JP Morgan is leading in fintech hiring followed by Banks of America and Golman sachs.PNC and Comerica Inc. is still in phase of growth where fintech openings are limited.

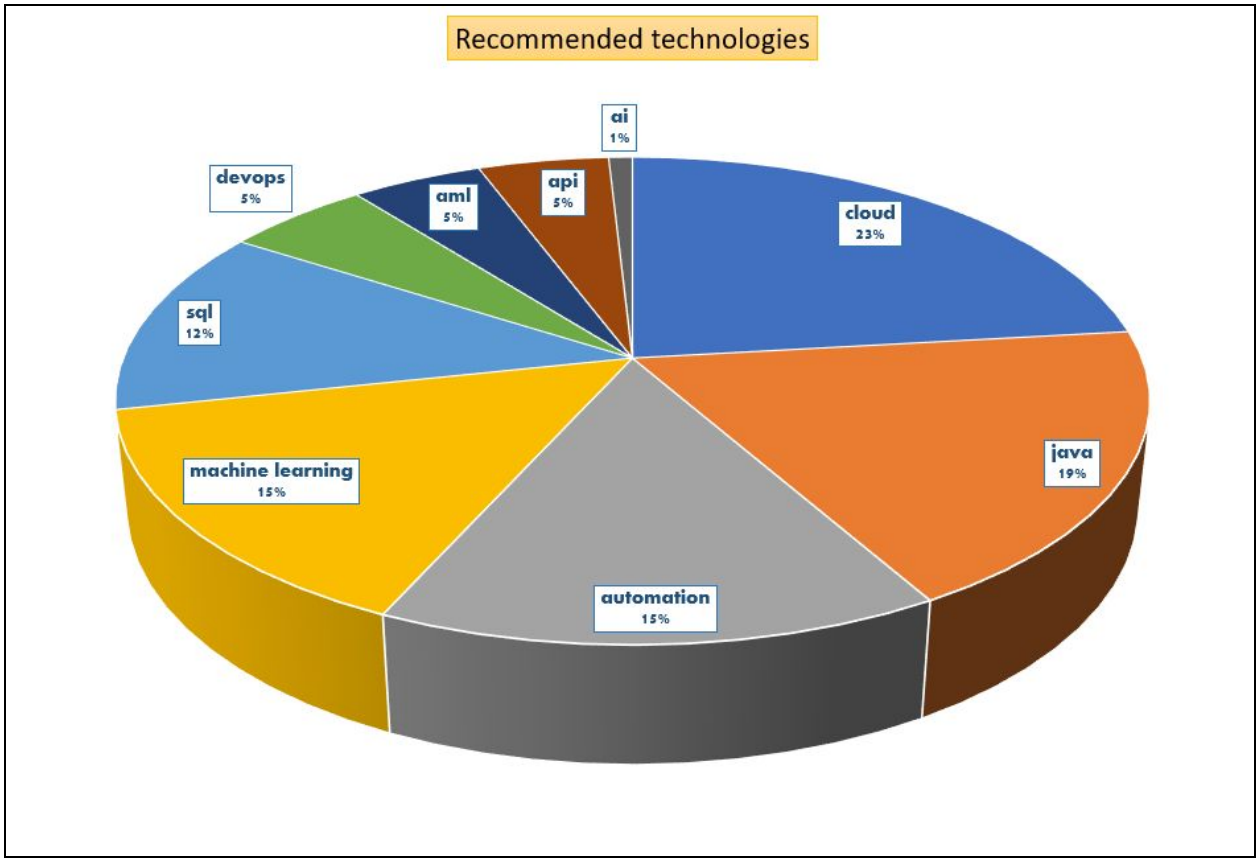


- **Which categories have the most jobs and which ones least?**
  As per the statistics provided below our analysis confirms that there are maximum job openings available under *services category* which includes opportunities in digital banking,retail banking, mobile payment etc.



- **Which areas would you recommend the job seeker to focus on based on available jobs?**

As per our analysis US Banks are focusing on technologies such as cloud, java, automation, machine learning etc and so we would recommend the job seeker to focus on these technologies.



Languages and Tools Used

Language	Python 3.7
Libraries/Tools	Pandas, nltk, Docker, Chromedriver
Visualization	Datawrapper, Excel

Citations

https://stackoverflow.com/questions/40718637/upload-csv-in-mybucket-and-read-file-in-s3-aws-using-python

https://stackoverflow.com/questions/40718637/upload-csv-in-mybucket-and-read-file-in-s3-aws-using-python