Principles of Big Data Management

Project Report:
SILLION VALLEY BANK COLLAPSE

Instructor: Dr. Praveen Rao

Sampath Kumar Arpula

KRPNS Santosh

University of Missouri

PROJECT OUTLINE

INTRODUCTION

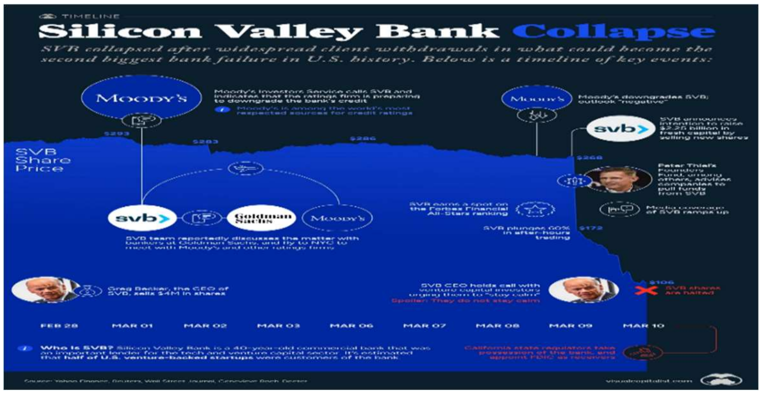
Silicon Valley Bank (SVB) focuses on serving the requirements of the technology and life science sectors. SVB is a distinctive bank that is designed especially for the technology and life science sectors. Its team of industry experts offers clients tailored financial solutions and insights to help them thrive because they have a thorough understanding of these sectors. SVB provides a variety of products and services to meet the needs of businesses at different phases of development, including project finance for specialized projects, growth financing for established businesses, and early-stage financing for new ventures. This makes it possible for SVB to support companies at every stage of their existence and help them realize their full potential. SVB, however, had a serious difficulty in 2020 that ultimately resulted in its demise. The COVID-19 epidemic, which significantly disrupted the world economy and financial markets, had a significant negative impact on the bank. The pandemic significantly reduced SVB's asset value and raised its loan default rates.

Due to these difficulties, SVB was compelled to declare bankruptcy in 2021. Given that SVB was a significant lender to numerous startups and other tech companies, this had significant effects on the tech sector as well as the overall economy. Additionally, the demise of SVB had important social and political ramifications. The bank was regarded as a prominent player in the tech ecosystem and was well known for its tight ties to the technology sector. Its demise aroused concerns about the stability of the tech sector and the pandemic's effects on the overall economy.

Increased regulation and supervision of the tech sector have been demanded in the wake of SVB's demise, particularly regarding financial services. A rising understanding of the need to create financial systems that are more durable and resilient and are better equipped to resist economic shocks and crises has also emerged.

The failure of SVB serves as a reminder of how crucial it is to comprehend the risks and difficulties that contemporary financial institutions are confronting as well as the demand for proactive measures to manage these risks. Additionally, it emphasizes how crucial data analysis and monitoring are for identifying potential dangers and weaknesses in financial systems.

Considering this, we suggest a big data research project that concentrates on Twitter data pertaining to SVB's demise. The project's goal is to find patterns and trends in Twitter data that could be a sign of the dangers and weaknesses that face financial systems and the digital sector.



MOTIVATION

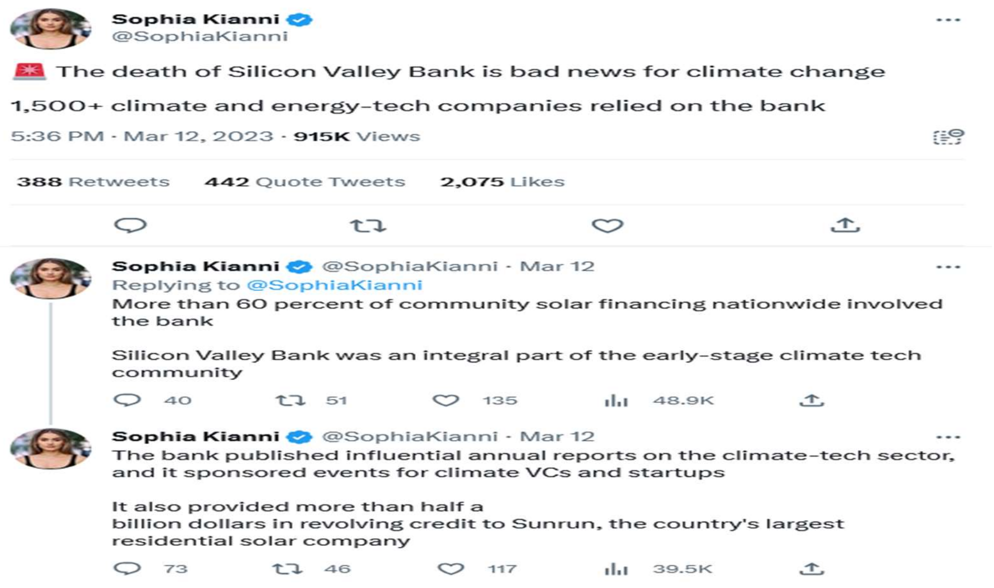
The decision to work with SVB Bank on a big data project was driven in part by the bank's distinctive position as a pioneer in the technology and life science sectors. SVB Bank is a prime choice to investigate the potential of big data analysis due to its emphasis on offering specialized financial solutions to businesses in these industries.

The team of professionals at SVB Bank has extensive experience of the technology and life science sectors, and you may take advantage of this knowledge to learn important insights about market trends and patterns. Large data collections can be analyzed to find patterns that can guide business choices and benefit organizations.

Additionally, SVB Bank's assortment of services and solutions that address various phases of a company's growth offers a wealth of data to work with. SVB Bank provides a significant amount of financial data that may be used to acquire insights into market trends and company performance, from early-stage financing for startups through growth financing for established businesses.

Big data analysis tools can be used to find patterns and insights that may not be seen from a straightforward review of financial statements. For instance, it might be feasible to spot new trends in the technology and life science sectors by evaluating social media data, which could guide investment choices.

The main reason for selecting SVB Bank for a big data project is to take advantage of its experience in the technology and life science sectors to learn important insights about market trends and patterns. Big data analysis techniques can be used to find hidden patterns and insights that can guide business choices and benefit organizations.



Sophia Kianni
@SophiaKianni

🚨 The death of Silicon Valley Bank is bad news for climate change

1,500+ climate and energy-tech companies relied on the bank

5:36 PM · Mar 12, 2023 · 915K Views

388 Retweets 442 Quote Tweets 2,075 Likes

Sophia Kianni @SophiaKianni · Mar 12
Replying to @SophiaKianni
More than 60 percent of community solar financing nationwide involved the bank

Silicon Valley Bank was an integral part of the early-stage climate tech community

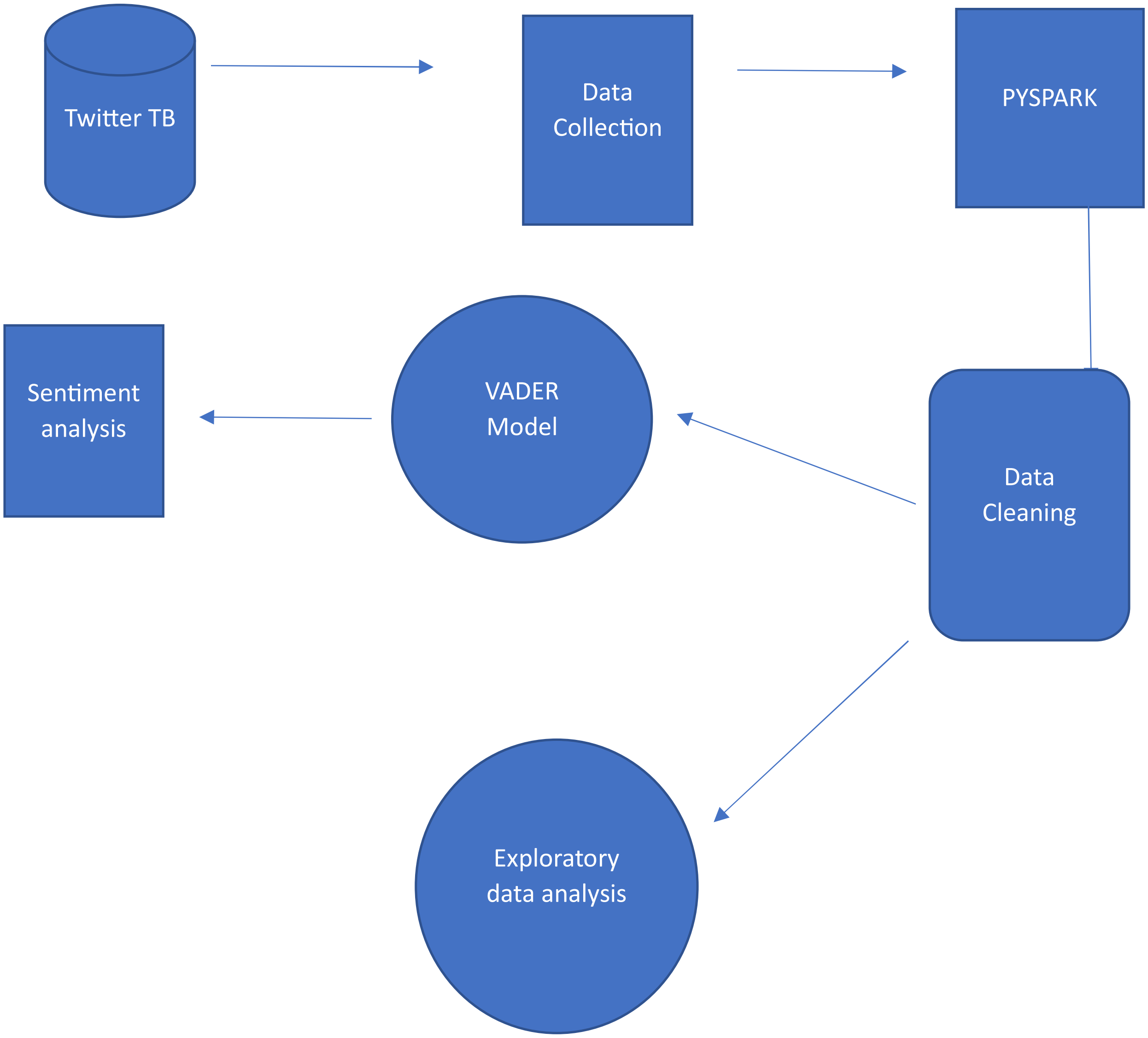
40 51 135 48.9K

Sophia Kianni @SophiaKianni · Mar 12
The bank published influential annual reports on the climate-tech sector, and it sponsored events for climate VCs and startups

It also provided more than half a
billion dollars in revolving credit to Sunrun, the country's largest residential solar company

73 46 117 39.5K

ARCHITECTURE



DATA COLLECTION

The process of collecting data is a crucial step in any data analysis project. In the context of this project, the data collection process involved collecting Twitter data related to the topic of "SVB Collapse". The data was collected using the Twitter API, which provides a way to access tweets containing specific keywords or hashtags.

After authentication, a Tweepy API object was created. This object was used to search for tweets containing the keyword "SVB Collapse" using the Cursor object. The maximum number of tweets to download was set to 10,000. The tweets were downloaded in batches using the Cursor object and stored in a list. After the tweets were collected, they were stored in a JSON file. This data can now be used for further exploratory analysis, such as sentiment analysis, topic modeling, or visualization.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a way to analyze and summarize data to better understand it and find insights that may be hidden. In my study, I analyzed a set of tweets related to the Silicon Valley Bank collapse using EDA.

We started by looking at the basic information about the dataset, such as the number of rows. Then, we checked the distribution of the favorite and retweet counts of the tweets. The highest number of favorites. Retweet count, character length and on average. This means that the tweets about the Silicon Valley Bank collapse were quite popular.

Next, we analyzed the text data in the tweets. I calculated the highest and average number of characters and words in the tweets. This shows that the tweets related to the Silicon Valley Bank collapse were relatively short and concise.

These findings give us a general understanding of the characteristics of tweets related to the Silicon Valley Bank collapse. We could further analyze the sentiment and topics of the tweets to find deeper insights and better understand the underlying themes and opinions surrounding this topic.

SENTIMENT ANALYSIS

VADER Model:

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a Python-based sentiment analysis library that is specifically designed to analyze sentiments expressed in social media. This tool is based on a rule-based, lexicon-driven approach and uses a combination of lexical and grammatical rules and machine learning techniques to analyze the sentiment of a given text.

Developed by researchers at the Georgia Institute of Technology, VADER uses a lexicon of words and phrases that are tagged with a polarity score, indicating the degree of positive or negative sentiment expressed by each term. The lexicon also includes rules for handling negation, intensifiers, and modifiers, which can affect the sentiment of a given text.

VADER calculates four scores for a given text: positivity, negativity, neutrality, and an overall sentiment score. These scores are calculated based on the number of positive, negative, and neutral words in the text, as well as the presence of certain grammatical constructs. One of VADER's unique strengths is its ability to handle sentiment analysis for text that is typically found in social media. This text often includes unconventional language and grammatical structures, and VADER is designed to handle such cases effectively. Additionally, VADER is able to detect sarcasm and irony, which are often difficult for other sentiment analysis tools to identify.

Overall, VADER is a flexible and powerful tool for analyzing sentiment in text. It can be used in various applications, including social media monitoring, market research, and customer feedback analysis.

Sentiment Analysis:

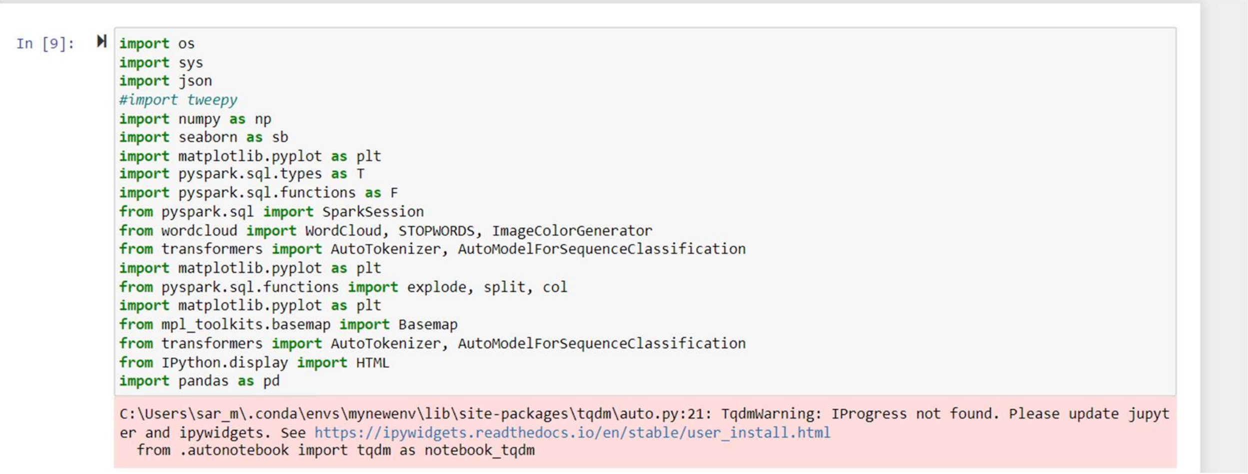
We have used a pre-trained model called 'twitter- Vader sentiment' for sentiment analysis. This model has been trained on a large dataset of tweets and is capable of classifying them into three categories i.e., positive, negative, or neutral.

We have passed the preprocessed tweets to the model and it returns the sentiment of the tweets in the form of one of the three categories. Positive, ngative and netural and used the VADER analyzer.polarity_scores() method is used to analyze the sentiment of the tweet text, producing a dictionary of sentiment scores for the tweet.

IMPLMENTATION and RESULTS:

Loading libraries

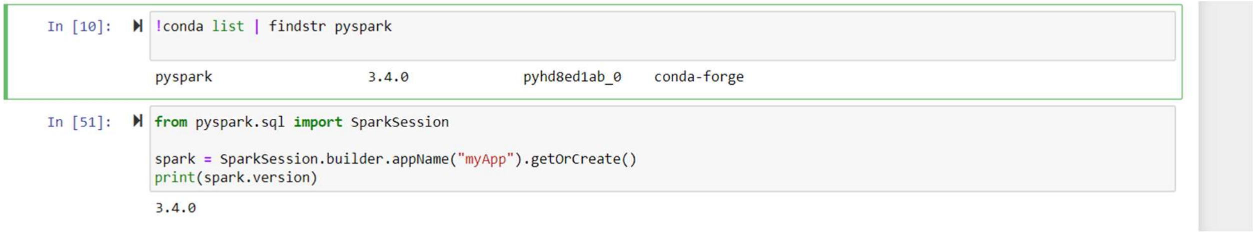
Please find the code snippet for libraries which are used.



```
In [9]:
import os
import sys
import json
#import tweepy
import numpy as np
import seaborn as sb
import matplotlib.pyplot as plt
import pyspark.sql.types as T
import pyspark.sql.functions as F
from pyspark.sql import SparkSession
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import matplotlib.pyplot as plt
from pyspark.sql.functions import explode, split, col
import matplotlib.pyplot as plt
from mpl_toolkits.basemap import Basemap
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from IPython.display import HTML
import pandas as pd
```

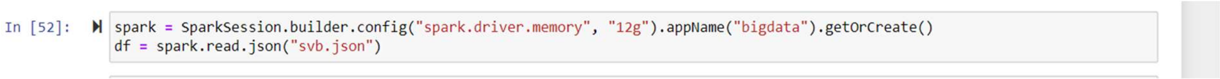
```
C:\Users\sar_m\.conda\envs\mynewenv\lib\site-packages\tqdm\auto.py:21: TqdmWarning: IProgress not found. Please update jupyt
er and ipywidgets. See https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
```

Spark is a powerful data processing system. We created a instance of a spark session object which is necessary to interact with spark. We also set the application name and have printed the version of spark.



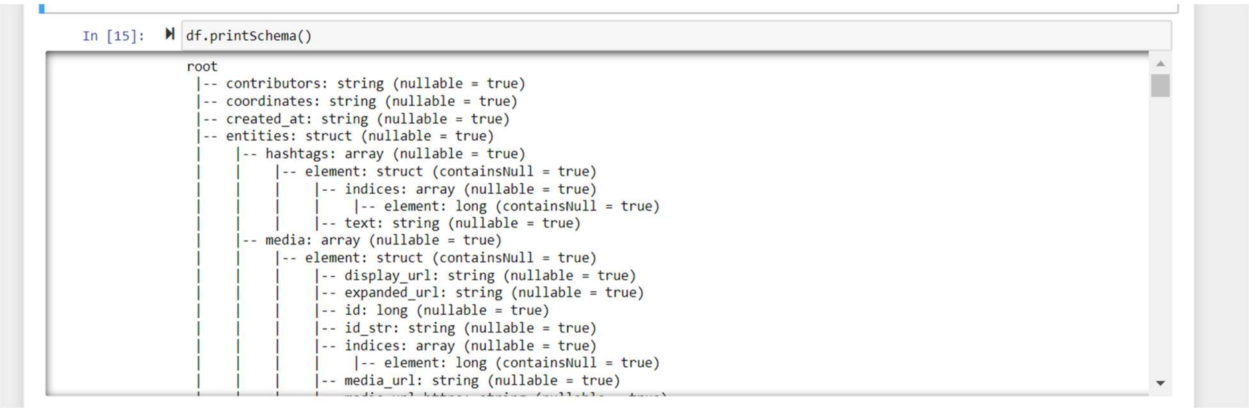
```
In [10]:  !conda list | findstr pyspark

          pyspark                   3.4.0              pyhd8ed1ab_0    conda-forge

In [51]:  from pyspark.sql import SparkSession

          spark = SparkSession.builder.appName("myApp").getOrCreate()
          print(spark.version)

          3.4.0
```

The tweets are extracted from twitter and exported into a json file named svb.json. Now we are loading that svb.json file into spark dataframes, to do exploratory data analysis and sentimental analysis on the topic Sillicon valley bank.



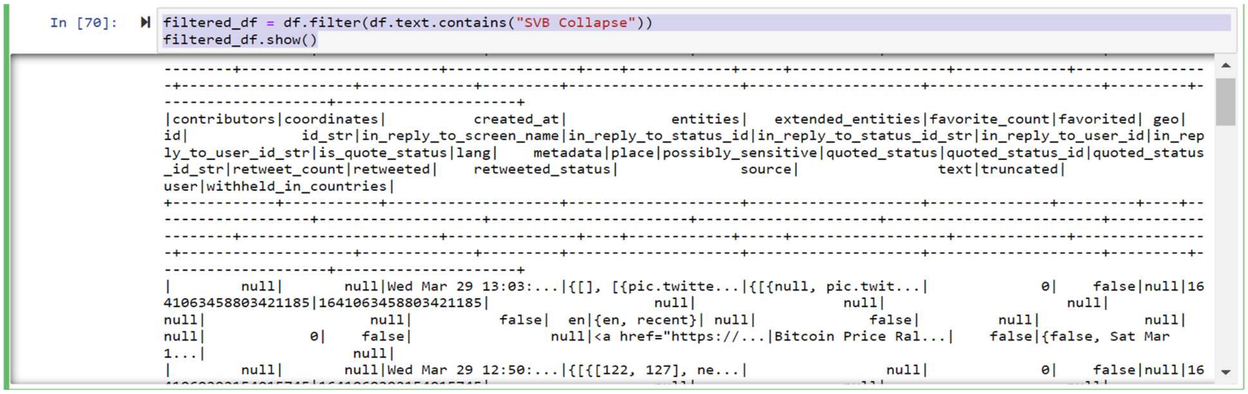
```
In [52]:  spark = SparkSession.builder.config("spark.driver.memory", "12g").appName("bigdata").getOrCreate()
          df = spark.read.json("svb.json")
```

The output shows the schema of the json file which is loaded into the spark data frames.



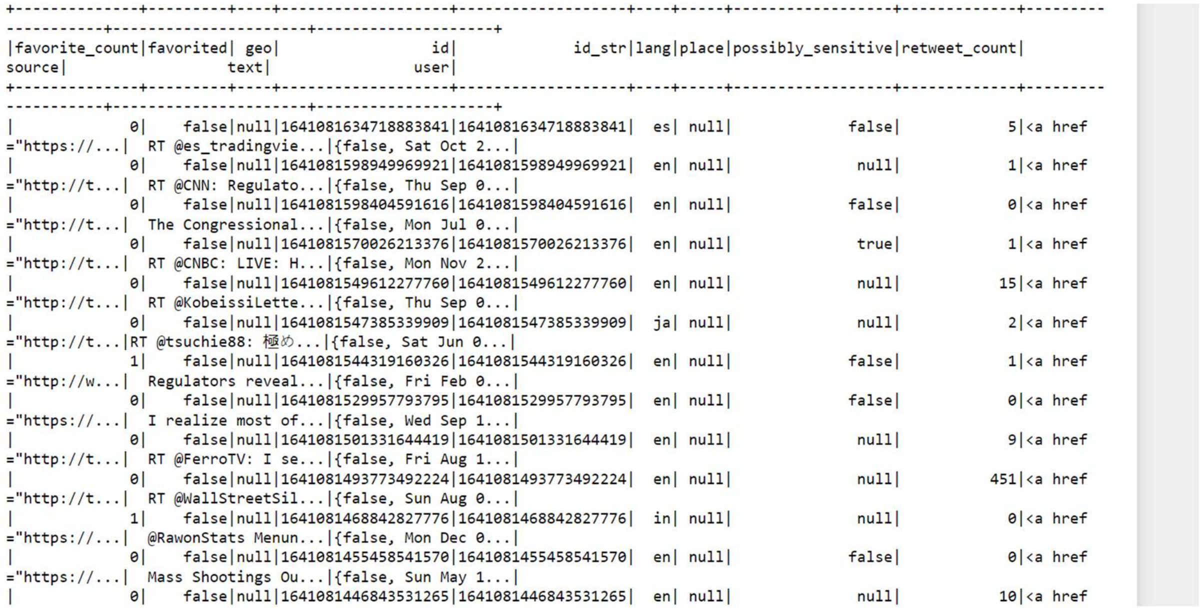
```
In [15]:  df.printSchema()

          root
           |-- contributors: string (nullable = true)
           |-- coordinates: string (nullable = true)
           |-- created_at: string (nullable = true)
           |-- entities: struct (nullable = true)
           |    |-- hashtags: array (nullable = true)
           |    |    |-- element: struct (containsNull = true)
           |    |    |    |-- indices: array (nullable = true)
           |    |    |    |    |-- element: long (containsNull = true)
           |    |    |    |-- text: string (nullable = true)
           |    |-- media: array (nullable = true)
           |    |    |-- element: struct (containsNull = true)
           |    |    |    |-- display_url: string (nullable = true)
           |    |    |    |-- expanded_url: string (nullable = true)
           |    |    |    |-- id: long (nullable = true)
           |    |    |    |-- id_str: string (nullable = true)
           |    |    |    |-- indices: array (nullable = true)
           |    |    |    |    |-- element: long (containsNull = true)
           |    |    |    |-- media_url: string (nullable = true)
```

From the file which is loaded in the spark dataframes, we have filtered the dataframes to extract the text SVBCollapse and stored in another dataframes called filtered_df. The same is also displayed.



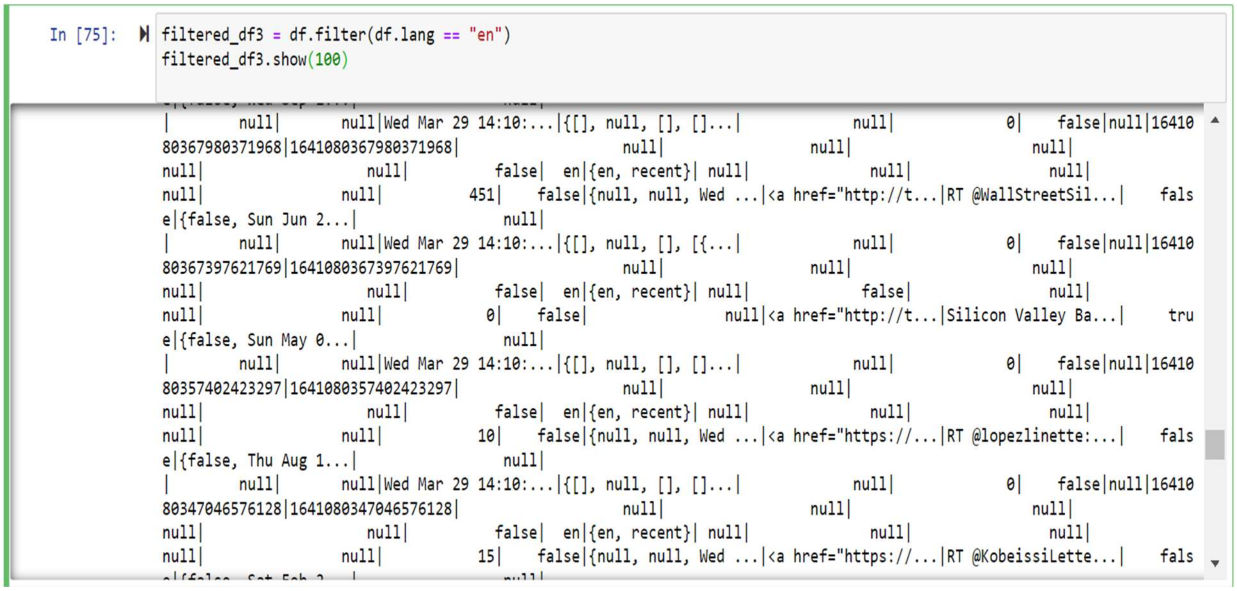
```
In [70]:  filtered_df = df.filter(df.text.contains("SVB Collapse"))
          filtered_df.show()

--------+---------------------+-------------+----+-----------+-----+-----------------+------------+-------------
-+-------------------+-----------+--------+-------------------+-------------------+-------------------+--------+-
------------------+--------------------+
|contributors|coordinates|          created_at|            entities|   extended_entities|favorite_count|favorited| geo|
id|              id_str|in_reply_to_screen_name|in_reply_to_status_id|in_reply_to_status_id_str|in_reply_to_user_id|in_rep
ly_to_user_id_str|is_quote_status|lang|     metadata|place|possibly_sensitive|quoted_status|quoted_status_id|quoted_status
_id_str|retweet_count|retweeted|     retweeted_status|                 source|                text|truncated|
user|withheld_in_countries|
+------------+-----------+--------------------+--------------------+--------------------+--------------+---------+----+--
-----------------+--------------------+-----------------------+---------------------+-------------------------+----------
--------+---------------------+-------------+----+-----------+-----+-----------------+------------+-------------
-+-------------------+-----------+--------+-------------------+-------------------+-------------------+--------+-
------------------+--------------------+
|        null|       null|Wed Mar 29 13:03:...|{[], [{pic.twitte...|{[{null, pic.twit...|             0|    false|null|16
41063458803421185|1641063458803421185|                   null|                 null|                     null|
null|              null|          false|  en|{en, recent}| null|              false|         null|            null|
null|            0|    false|             null|<a href="https://...|Bitcoin Price Ral...|    false|{false, Sat Mar
1...|              null|
|        null|       null|Wed Mar 29 12:50:...|{[{[122, 127], ne...|                null|             0|    false|null|16
```

The output shows the filtered data with regards to the columns “favorite_count", "favorited", "geo", "id", "id_str", "lang", "place", "possibly_sensitive", "retweet_count", "source", "text", and "user" for performing data analysis.



```
+--------------+---------+----+-------------------+-------------------+----+-----+------------------+-------------+---------
-----------+---------------------+--------------------+
|favorite_count|favorited| geo|                 id|             id_str|lang|place|possibly_sensitive|retweet_count|
source|                 text|                user|
+--------------+---------+----+-------------------+-------------------+----+-----+------------------+-------------+---------
-----------+---------------------+--------------------+
|             0|    false|null|1641081634718883841|1641081634718883841|  es| null|             false|            5|<a href
="https://...|  RT @es_tradingvie...|{false, Sat Oct 2...|
|             0|    false|null|1641081598949969921|1641081598949969921|  en| null|              null|            1|<a href
="http://t...|  RT @CNN: Regulato...|{false, Thu Sep 0...|
|             0|    false|null|1641081598404591616|1641081598404591616|  en| null|             false|            0|<a href
="http://t...|  The Congressional...|{false, Mon Jul 0...|
|             0|    false|null|1641081570026213376|1641081570026213376|  en| null|              true|            1|<a href
="http://t...|  RT @CNBC: LIVE: H...|{false, Mon Nov 2...|
|             0|    false|null|1641081549612277760|1641081549612277760|  en| null|              null|           15|<a href
="http://t...|  RT @KobeissiLette...|{false, Thu Sep 0...|
|             0|    false|null|1641081547385339909|1641081547385339909|  ja| null|              null|            2|<a href
="http://t...|RT @tsuchie88: 極め...|{false, Sat Jun 0...|
|             1|    false|null|1641081544319160326|1641081544319160326|  en| null|             false|            1|<a href
="http://w...|  Regulators reveal...|{false, Fri Feb 0...|
|             0|    false|null|1641081529957793795|1641081529957793795|  en| null|             false|            0|<a href
="https://...|  I realize most of...|{false, Wed Sep 1...|
|             0|    false|null|1641081501331644419|1641081501331644419|  en| null|              null|            9|<a href
="http://t...|  RT @FerroTV: I se...|{false, Fri Aug 1...|
|             0|    false|null|1641081493773492224|1641081493773492224|  en| null|              null|          451|<a href
="http://t...|  RT @WallStreetSil...|{false, Sun Aug 0...|
|             1|    false|null|1641081468842827776|1641081468842827776|  in| null|              null|            0|<a href
="https://...|  @RawonStats Menun...|{false, Mon Dec 0...|
|             0|    false|null|1641081455458541570|1641081455458541570|  en| null|             false|            0|<a href
="https://...|  Mass Shootings Ou...|{false, Sun May 1...|
|             0|    false|null|1641081446843531265|1641081446843531265|  en| null|              null|           10|<a href
```

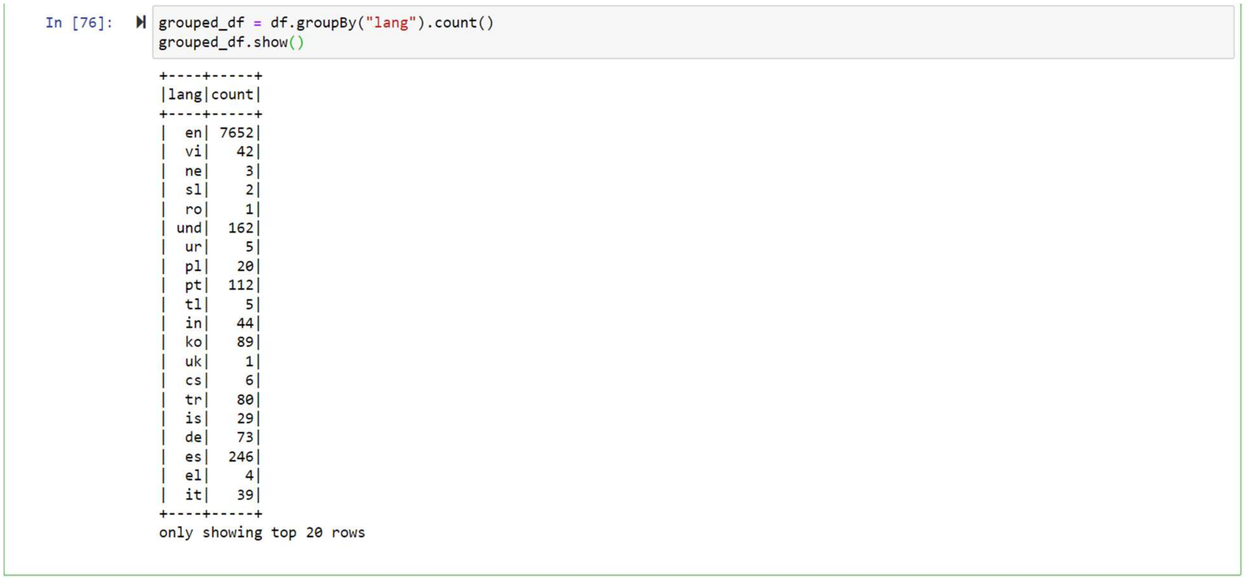
The data is further filtered to check tweets only from the language which is english.



```
In [75]:  filtered_df3 = df.filter(df.lang == "en")
          filtered_df3.show(100)
```

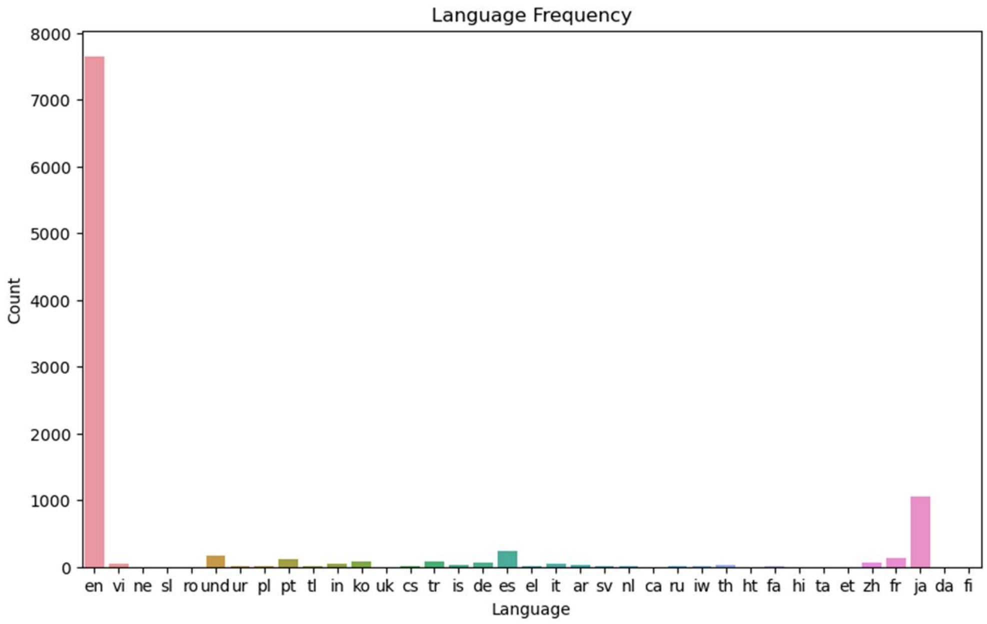
```
e|{false, Wed Sep 1...|                 null|
|        null|       null|Wed Mar 29 14:10:...|{[], null, [], []...|             null|             0|    false|null|16410
80367980371968|1641080367980371968|                        null|                 null|                  null|
null|                  null|          false|  en|{en, recent}| null|             null|                 null|
null|                null|            451|    false|{null, null, Wed ...|<a href="http://t...|RT @WallStreetSil...|    fals
e|{false, Sun Jun 2...|                 null|
|        null|       null|Wed Mar 29 14:10:...|{[], null, [], [{...|             null|             0|    false|null|16410
80367397621769|1641080367397621769|                        null|                 null|                  null|
null|                  null|          false|  en|{en, recent}| null|            false|                 null|
null|                null|              0|    false|                 null|<a href="http://t...|Silicon Valley Ba...|     tru
e|{false, Sun May 0...|                 null|
|        null|       null|Wed Mar 29 14:10:...|{[], null, [], []...|             null|             0|    false|null|16410
80357402423297|1641080357402423297|                        null|                 null|                  null|
null|                  null|          false|  en|{en, recent}| null|             null|                 null|
null|                null|             10|    false|{null, null, Wed ...|<a href="https://...|RT @lopezlinette:...|    fals
e|{false, Thu Aug 1...|                 null|
|        null|       null|Wed Mar 29 14:10:...|{[], null, [], []...|             null|             0|    false|null|16410
80347046576128|1641080347046576128|                        null|                 null|                  null|
null|                  null|          false|  en|{en, recent}| null|             null|                 null|
null|                null|             15|    false|{null, null, Wed ...|<a href="https://...|RT @KobeissiLette...|    fals
e|{false, Sat Feb 2...|                 null|
```

The output shows the grouped data by language in which the tweets are tweeted after counting the number of tweets in each language.

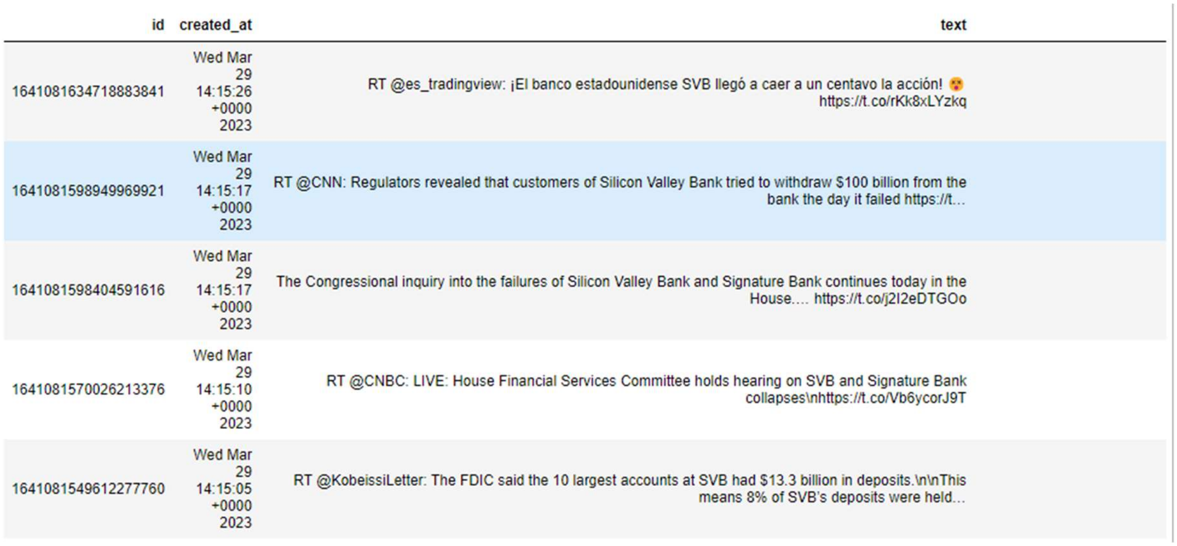


```
In [76]:  grouped_df = df.groupBy("lang").count()
          grouped_df.show()

          +----+-----+
          |lang|count|
          +----+-----+
          |  en| 7652|
          |  vi|   42|
          |  ne|    3|
          |  sl|    2|
          |  ro|    1|
          | und|  162|
          |  ur|    5|
          |  pl|   20|
          |  pt|  112|
          |  tl|    5|
          |  in|   44|
          |  ko|   89|
          |  uk|    1|
          |  cs|    6|
          |  tr|   80|
          |  is|   29|
          |  de|   73|
          |  es|  246|
          |  el|    4|
          |  it|   39|
          +----+-----+
          only showing top 20 rows
```

This analysis has provided valuable insights into the distribution of languages used on Twitter regarding the SVB collapse topic. By examining the frequency of tweets in different languages and presenting the results in a bar chart, we were able to gain a deeper understanding of the languages in which users have tweeted about the SVB collapse.



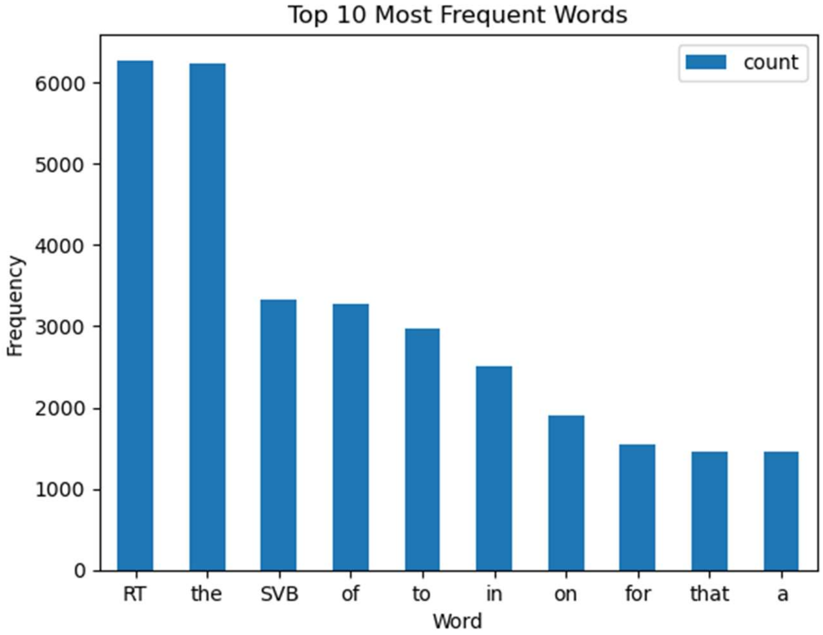
Just to display the tweets in a table, we have loaded the data in pandas and used html table to display the output in a html table.



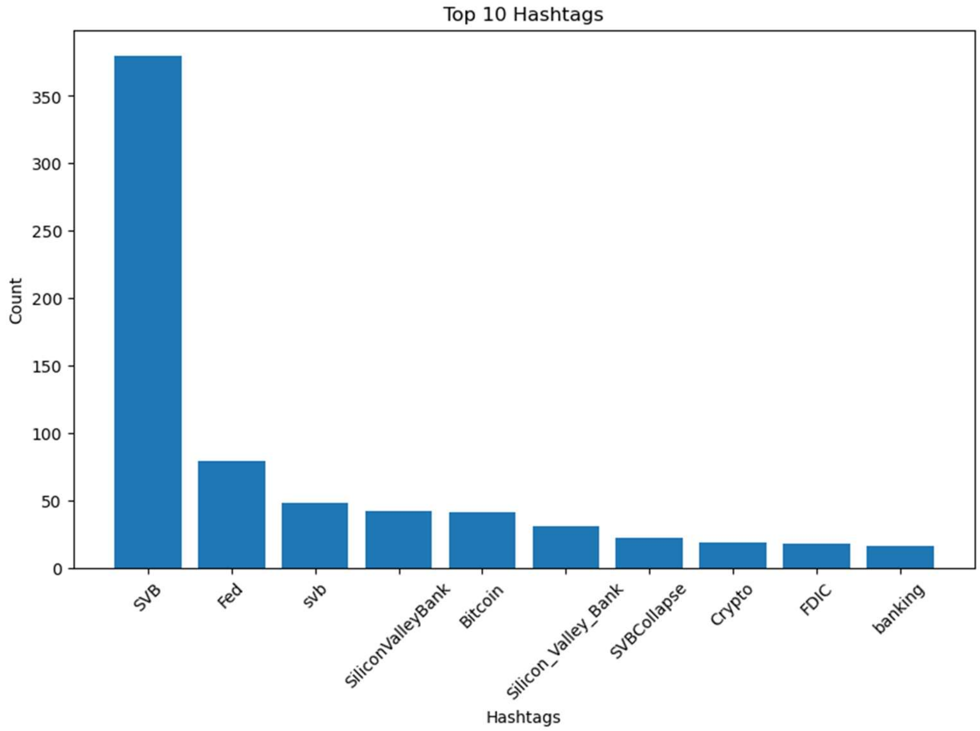
id	created_at	text
1641081634718883841	Wed Mar 29 14:15:26 +0000 2023	RT @es_tradingview: ¡El banco estadounidense SVB llegó a caer a un centavo la acción! 😲 https://t.co/rKk8xLYzkq
1641081598949969921	Wed Mar 29 14:15:17 +0000 2023	RT @CNN: Regulators revealed that customers of Silicon Valley Bank tried to withdraw $100 billion from the bank the day it failed https://t…
1641081598404591616	Wed Mar 29 14:15:17 +0000 2023	The Congressional inquiry into the failures of Silicon Valley Bank and Signature Bank continues today in the House.… https://t.co/j2I2eDTGOo
1641081570026213376	Wed Mar 29 14:15:10 +0000 2023	RT @CNBC: LIVE: House Financial Services Committee holds hearing on SVB and Signature Bank collapses\nhttps://t.co/Vb6ycorJ9T
1641081549612277760	Wed Mar 29 14:15:05 +0000 2023	RT @KobeissiLetter: The FDIC said the 10 largest accounts at SVB had $13.3 billion in deposits.\n\nThis means 8% of SVB's deposits were held…

To identify patterns and trends, we have performed text analysis on pyspark dataframes by splitting the text column into individual words and counting how often each word appears and then identifying the top 10 most frequent words.

We also displayed the ouput in a bar graph that displays the frequency of the top 10 most frequently used words.



To identify most popular topics in the tweets dataset which is extracted, we have performed hashtag analysis on data which is loaded in pyspark dataframes. We did it by counting number of times each hashtag appears tweets. The output is displayed in a bar chart with top 10 most commonly used hashtags.



In process of performing data analysis on the data, i.e; loaded in spark dataframes. We have done below:

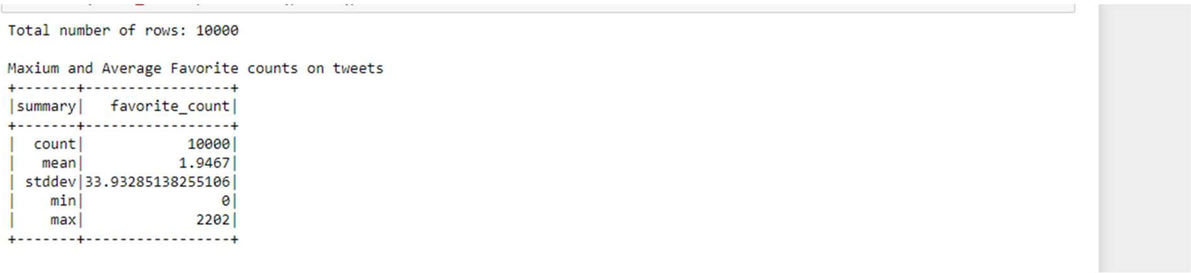
Printed total number of rows in the Dataframe.

Calculated the maximum and average number of like and retweets for the posts.

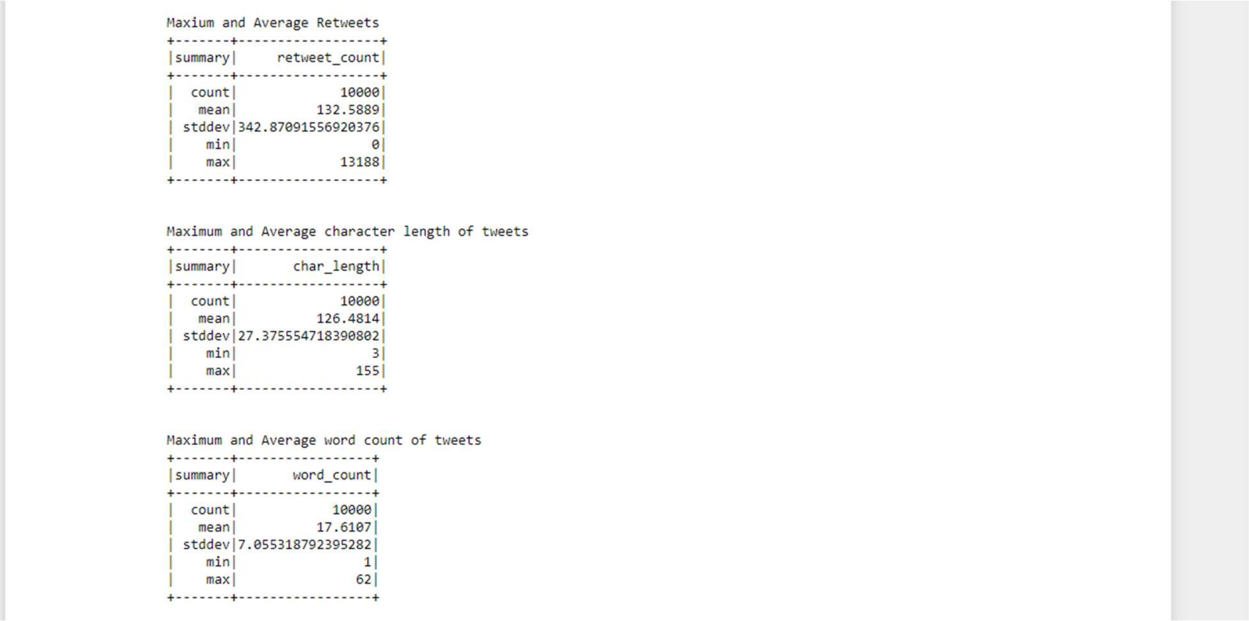
Calculated the maximum and average character length of the posts.

Calculated the maximum and average word count of the posts.

All these values are printed. These analsys provides us insights into how people are engaging and reactinbg on the topic Sillicon valley bank.



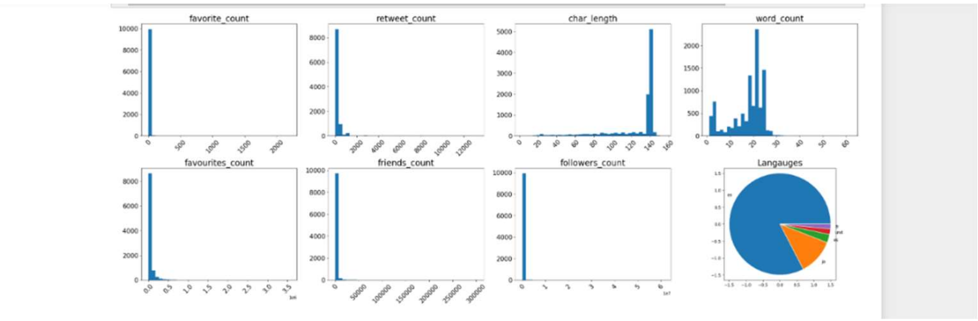
```
Total number of rows: 10000

Maxium and Average Favorite counts on tweets
+-------+-----------------+
|summary|   favorite_count|
+-------+-----------------+
|  count|            10000|
|   mean|           1.9467|
| stddev|33.93285138255106|
|    min|                0|
|    max|             2202|
+-------+-----------------+
```



```
Maxium and Average Retweets
+-------+------------------+
|summary|     retweet_count|
+-------+------------------+
|  count|             10000|
|   mean|          132.5889|
| stddev|342.87091556920376|
|    min|                 0|
|    max|             13188|
+-------+------------------+


Maximum and Average character length of tweets
+-------+------------------+
|summary|       char_length|
+-------+------------------+
|  count|             10000|
|   mean|          126.4814|
| stddev|27.375554718390802|
|    min|                 3|
|    max|               155|
+-------+------------------+


Maximum and Average word count of tweets
+-------+-----------------+
|summary|       word_count|
+-------+-----------------+
|  count|            10000|
|   mean|          17.6107|
| stddev|7.055318792395282|
|    min|                1|
|    max|               62|
+-------+-----------------+
```

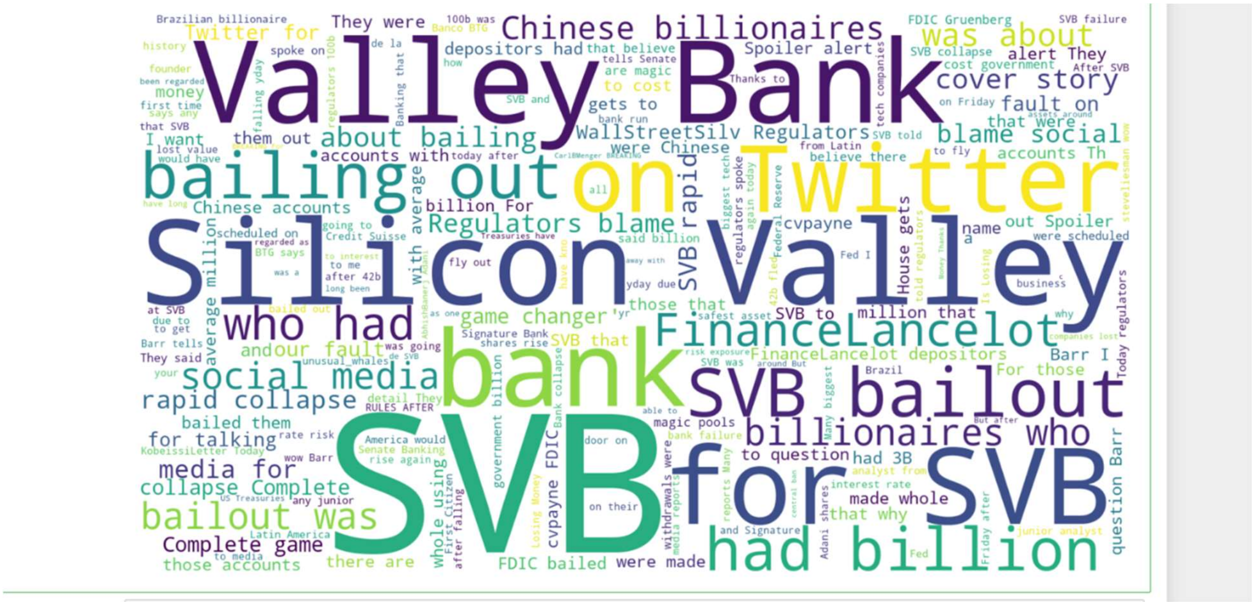
We have performed interesting data visualizations on the data as part of exploratory data analysis. The results depicting diffrerent variables of interest in the data for favorite count, retweet count, character length of tweets, word count of tweets, number of favorites of users who tweeted. and number of followers of user who tweeted are shown and distribution of values for each variable are showin in histogram charts. This helps us to identify pattern in the data.

A pie chart also displayed shows the top five languages in which tweets are tweeted. This visualization helps us to understand the composition of dataset with regards to language in which the tweets are present in the dataset.

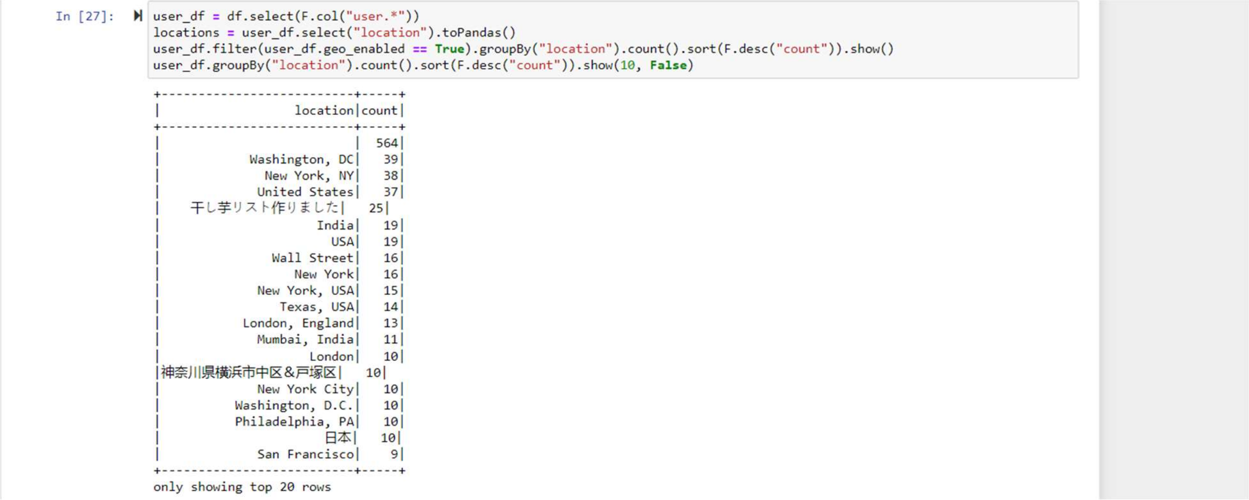


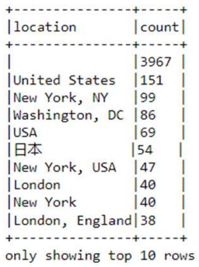
Word Cloud: A word cloud is a visual representation of text in the data, where the size of the word shows the frequency of that word, in simple terms we can say that if the word is larger it means that its appeared more times than the word which is not.

So as part of interesting data visualization, we have used word cloud to display most frequent words from the tweets. It’s a interesting visualization and provides insights like overall sentiment of tweets and trends that are most commonly discussed in the tweets.



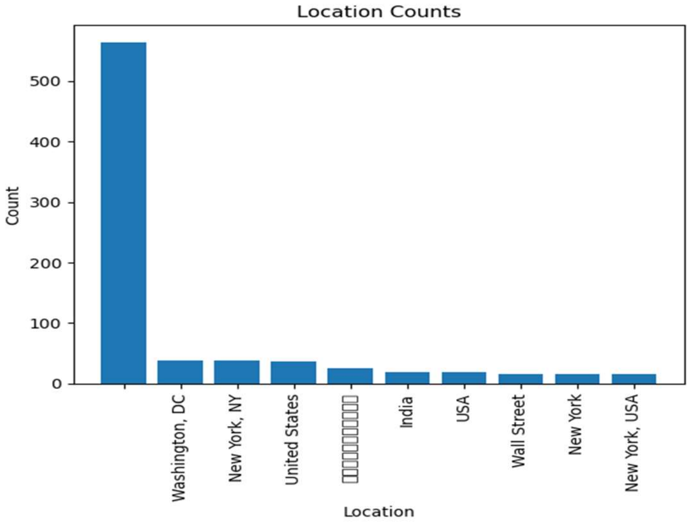
The output shows the extracted data from users location from the tweets after performing a group by operation to count the number of tweets from each location.


```
In [27]:  user_df = df.select(F.col("user.*"))
          locations = user_df.select("location").toPandas()
          user_df.filter(user_df.geo_enabled == True).groupBy("location").count().sort(F.desc("count")).show()
          user_df.groupBy("location").count().sort(F.desc("count")).show(10, False)

          +--------------------------+-----+
          |                  location|count|
          +--------------------------+-----+
          |                          |  564|
          |            Washington, DC|   39|
          |              New York, NY|   38|
          |             United States|   37|
          |     干し芋リスト作りました|   25|
          |                     India|   19|
          |                       USA|   19|
          |               Wall Street|   16|
          |                  New York|   16|
          |             New York, USA|   15|
          |                Texas, USA|   14|
          |           London, England|   13|
          |             Mumbai, India|   11|
          |                    London|   10|
          |神奈川県横浜市中区&戸塚区|   10|
          |             New York City|   10|
          |          Washington, D.C.|   10|
          |          Philadelphia, PA|   10|
          |                      日本|   10|
          |             San Francisco|    9|
          +--------------------------+-----+
          only showing top 20 rows
```

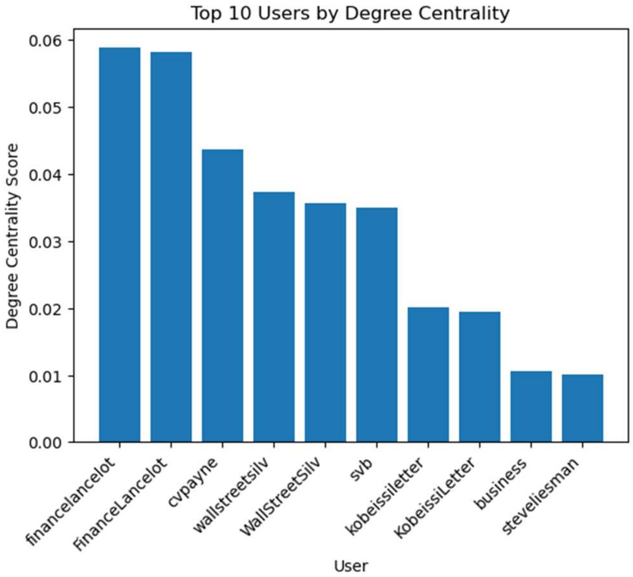


```
+---------------+-----+
|location       |count|
+---------------+-----+
|               |3967 |
|United States  |151  |
|New York, NY   |99   |
|Washington, DC |86   |
|USA            |69   |
|日本            |54   |
|New York, USA  |47   |
|London         |40   |
|New York       |40   |
|London, England|38   |
+---------------+-----+
only showing top 10 rows
```

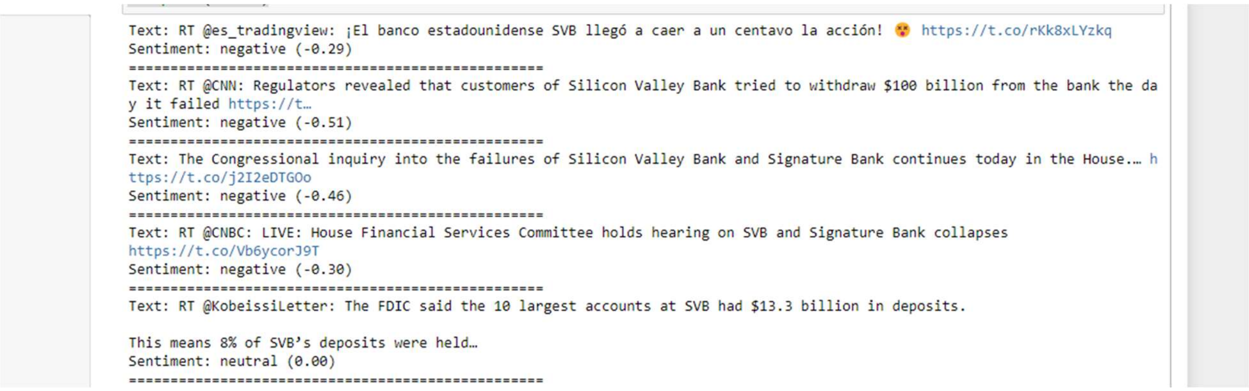
The same is displayed in a bar chart.



The output shows the degree centrality in a bar chart of twitter users and their degree centrality score.

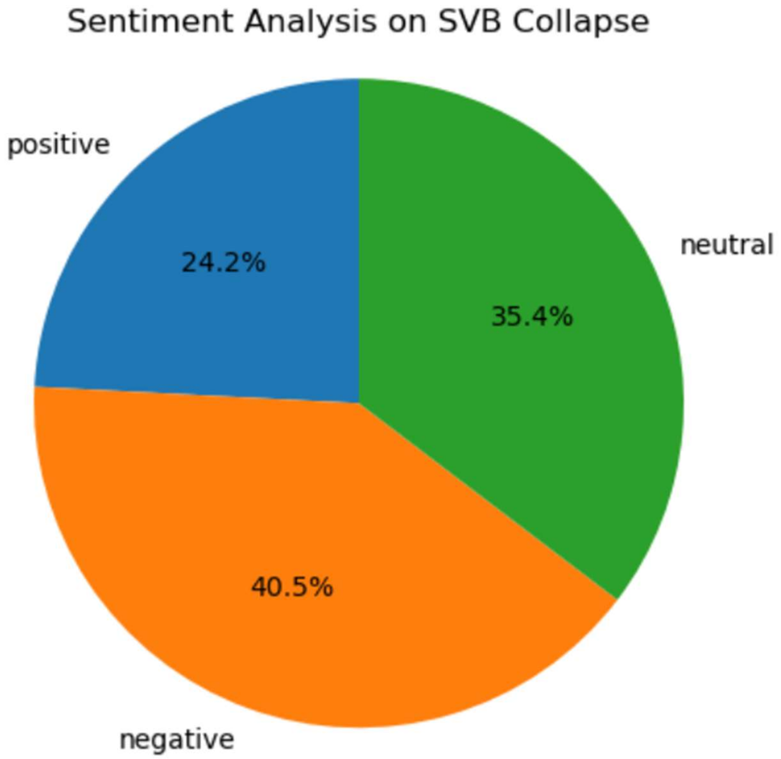


The results show the compound score of sentiment analysis that is performed through VADER model. The text in the tweets is extracted, each text is analyzed in the VADER model and given a score. Tweets with compound score > 0.05 are considered as positive, scores less that 0.05 are considered as negative and in between scores are considered as neutral.

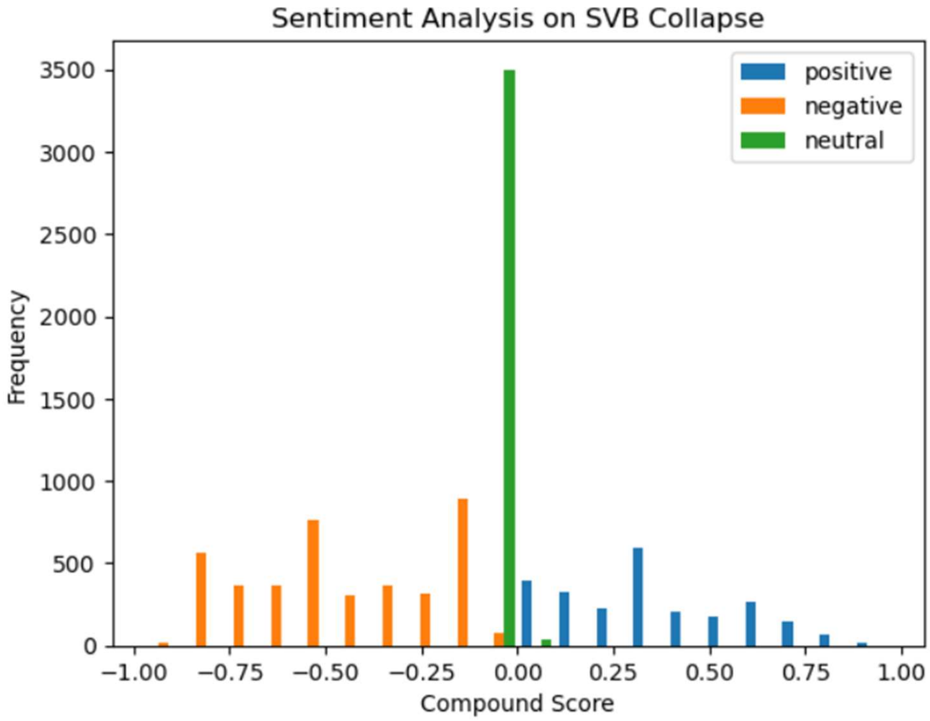


```
Text: RT @es_tradingview: ¡El banco estadounidense SVB llegó a caer a un centavo la acción! 😲 https://t.co/rKk8xLYzkq
Sentiment: negative (-0.29)
==================================================
Text: RT @CNN: Regulators revealed that customers of Silicon Valley Bank tried to withdraw $100 billion from the bank the da
y it failed https://t…
Sentiment: negative (-0.51)
==================================================
Text: The Congressional inquiry into the failures of Silicon Valley Bank and Signature Bank continues today in the House.… h
ttps://t.co/j2I2eDTGOo
Sentiment: negative (-0.46)
==================================================
Text: RT @CNBC: LIVE: House Financial Services Committee holds hearing on SVB and Signature Bank collapses
https://t.co/Vb6ycorJ9T
Sentiment: negative (-0.30)
==================================================
Text: RT @KobeissiLetter: The FDIC said the 10 largest accounts at SVB had $13.3 billion in deposits.

This means 8% of SVB’s deposits were held…
Sentiment: neutral (0.00)
==================================================
```

The same scores are used to plot a pie chart to show the overall Sentiment on tweets of SVB Bank.



Below results show the polarity scores of sentiment analysis of positive negative and netural are plotted in a histogram.



Results:

We obtained the sentiment of each tweet by passing it through the pre-trained model. The sentiments were then classified into three categories i.e., positive, negative, or neutral. The results were analyzed to determine the overall sentiment of Twitter users towards the topic of SVB bank collapse.

CONCLUSION

As this study on analyzing Twitter data related to the collapse of SVB Bank demonstrates, the use of big data and sophisticated data management technologies like Hadoop and Spark has significantly impacted the field of data analysis. We successfully used PySpark, PySpark SQL, and a variety of libraries, including Plot and Pyspark, to edit, query, and analyze the data. This project has effectively demonstrated the ability of these technologies to extract valuable insights from large datasets.

Moreover, by utilizing machine learning functions like sentiment analysis and enhancing data visualization with Pandas, we were able to gain a better understanding of the data and uncover new information about SVB's collapse. Overall, this study emphasizes the critical importance of utilizing various tools and methodologies when analyzing data, particularly when working with vast datasets.

Through effective application of sentiment analysis and exploratory data analysis, we derived essential insights that could guide decision-making and promote corporate success. The need for efficient data management and analysis methods is ever-increasing as the volume of data generated continues to grow. Therefore, big data and advanced data management technologies have the potential to help organizations gain insightful information that can guide decision-making and promote business success.

Frameworks:

Tools/Libraries	Description
IDE	Anaconda – Jupyter Notebook
Hadoop	A distributed processing framework that provides high availability and fault tolerance for big data storage and processing
Spark	An open-source distributed computing system that processes big data in-memory and provides faster processing than Hadoop
PySpark	A Python API for Apache Spark that allows users to write Spark applications in Python
Pandas	A library for data manipulation and analysis in Python that provides easy-to-use data structures and data analysis tools
Seaborn	A Python data visualization library based on matplotlib that provides an interface for drawing attractive and informative statistical graphics
Matplotlib	A Python data visualization library that provides tools for creating static, animated, and interactive visualizations in Python
Wordcloud	A Python library for generating word clouds from text data
Transformers	A Python-based library for natural language processing (NLP) tasks, including pre-trained models for sentiment analysis and other NLP tasks

References:

https://www.semanticscholar.org/paper/VADER%3A-A-Parsimonious-Rule-Based-Model-for-Analysis-Hutto-Gilbert/bcdc102c04fb0e7d4652e8bcc7edd2983bb9576d

https://www.visualcapitalist.com/wp-content/uploads/2023/03/silicon-valley-bank-collapse.jpg

https://i.kym-cdn.com/photos/images/original/002/549/751/74e.png

https://realpython.com/python-nltk-sentiment-analysis/