Big Data Programming

Project Increment 2 Report

Project Description:

1. Project Title and Team Members:

Project title:

Analysis of COVID-19 Impact on the Job Sectors

Team Members:

1) Chaturvedi Devna
2) Muppalla Gireesh Kumar
3) Yarava Yugandhar
4) Gundumogula Mani Sai

2. Introduction:

The coronavirus outbreak accelerated during March 2020 and continued to spread globally. Certain cities and the communities turned out to be more vulnerable, which critically impacted the workforce. The pandemic has crumpled specific industries. A countries economy has its roots in different sectors. Primarily the cities and various metropolitan areas specialize as they are the hub for multiple industries, for example, Finance in New York, Information Technology in Seattle, and so on.

The industries heavily impacted by COVID-19 encompass leisure and hospitality. There were heavy closures in the initial 8 – 10 months, and many workers in various sectors faced the highest unemployment rate. As the economies are still reviving from the pandemic's ripples, there are still many concerns amongst employers, contractors, students, part-time employees about their future in the current market scenario and their place in it.

3. Background

Since the pandemic began in March 2020, apart from being a Health crisis, COVID-19 has turned out to be a financial one. Closing of businesses, job losses, reduction in working hours, lowering the pay scales, layoffs are unfolding into a recession situation. About 25 – 26 million Americans have already filed for unemployment insurance by June 2020. The job losses seem to concentrate on the areas that are directly getting impacted by COVID-19 restrictions: the social media platforms, namely Facebook, LinkedIn, and Twitter. Twitter is one of the most reliable platforms providing real-time information on critical topics and ensuring that people are aware of the recent market trends catering to various sectors.

Unemployment is one of the most critical issues in the present scenario. Tracking the wave of unemployment through social media and analyzing the results can help the workforce make them aware and help them make informed decisions. The project aims at addressing the unemployment problem through the project we are implementing by data analysis and visualization.

4. Goals and Objectives

- **Motivation**

The COVID-19 situation has impacted the workforce causing many professionals to lose their jobs globally in many sectors, namely, IT, Hardware, Banking, Finance, Telecom. The concept of big data and its services have been implemented in various sectors to analyze this impact.

We streamlined the process of analyzing social media data using the existing techniques, visualize the data in a feasible way.

- **Significance**

As the workforce is globally getting impacted by COVID-19, the government should track the people getting unemployed. The project analysis helps to answer the following questions

1) How is the economy of the country getting impacted?

2) To design policies on how to tackle the situation.

Big data supports the analysis of abundant amounts of data with great efficiency. Using the data extracted from Twitter, we can perform the analysis to deduce the job losses in various sectors, look at the sectors most affected by the pandemic. Also, this analysis can help the workforce to make their decision in an informed manner.

- **Objectives**
  - To analyze the number of people worldwide who suffered from job losses, people searching for jobs in the current market situation about various sectors.
  - To analyze the data based on job sectors, job losses, gender, age group, etc.
  - We will be using a combination of tools using Hadoop Ecosystem, RDBMS, SNScrape, Twitter API, and python to collect the data, store, analyze and visualize the data.

- **Features**
  - Analysis of real-time tweets from Twitter using SNScrape and Twitter API.
  - Pre-processing the text data using NLP libraries.
  - Extracting data from Twitter and loading it to SQL
  - Performing ETL using Sqoop to import and export data between RDBMS and HDFS.
  - Querying using HQL

5. Design, Implementation, Testing and Reporting

- **Dataset**

The data we are collecting focuses on job losses and unemployment. The social media platform we have selected for data extraction is Twitter. We are collecting Twitter data using Twitter batch API using Twitter developer account credentials. All tweets are in CSV format. The collected data has an array of information about a tweet like Datetime, Tweet Id, Text, Username, Language, conversation Id, retweet Count, quoted Tweet, like Count, retweeted Tweet. The hashtags used to scrape the data

are Covid-19, CovidatWork, unemployment, coronavirus, pandemic, career, working poor, poverty.
Detail design of features with workflow

- **Analysis (Details about Data)**

For the data extraction from Twitter, an account is created in Twitter Developers API. Then API tokens and credentials received are used to download tweets using tweepy and Twitter streaming API in python and stored tweets into CSV file. 7.8 million records of Twitter extracted from January to December 2020 using SNScrape. We are performing real-time data extraction using Twitter API.

- **Implementation**

- **Phase 2 Extract, Transform and Load the data**

    - **Part 1 – Extracting Text using SNScrape and Twitter API from Twitter**

    The first step includes extracting text data from Twitter using SNScrape and Twitter API: Hashtags used: Covid-19, CovidatWork, unemployment, coronavirus, pandemic, career, working poor, poverty. Saving the data to a shared drive

In [18]:
```
import tweepy,json
```

In [2]:
```
consumer_key="HdOI177rKuZrD1k6MGACqKbD6"
consumer_secret="hgVsDcRNV7hwheoBbVhYymvzBMTPoNwvRoqydmRWzlKcfl64lJ"
access_token="1359235244524785669-oDvB2pv5wDHAXZpUllJhbya3PUAbCB"
access_token_secret="ewU7Cg2V6ava1LTkTo5kYVb2T0zOJ2fhuJvbJF2MNXZMo"
```

In [3]:
```
auth= tweepy.OAuthHandler(consumer_key,consumer_secret)
auth.set_access_token(access_token,access_token_secret)
```

In [5]:
```
tweet_list=[]
class MyStreamListener(tweepy.StreamListener):
    def __init__(self,api=None):
        super(MyStreamListener,self).__init__()
        self.num_tweets=0
        self.file=open("tweet.txt","w")
    def on_status(self,status):
        tweet=status._json
        self.file.write(json.dumps(tweet)+ '\n')
        tweet_list.append(status)
        self.num_tweets+=1
        if self.num_tweets<20000:
            return True
        else:
            return False
        self.file.close()
```

```
l = MyStreamListener()
stream =tweepy.Stream(auth,l)

stream.filter(track=['Covid19', 'CovidAtWork', 'Covid-19', 'CoronaVirus', 'CoronaVirus19', 'CovidatWork', 'unemployment', 'ca
stream.filter(languages=["en"])
```

In [8]:

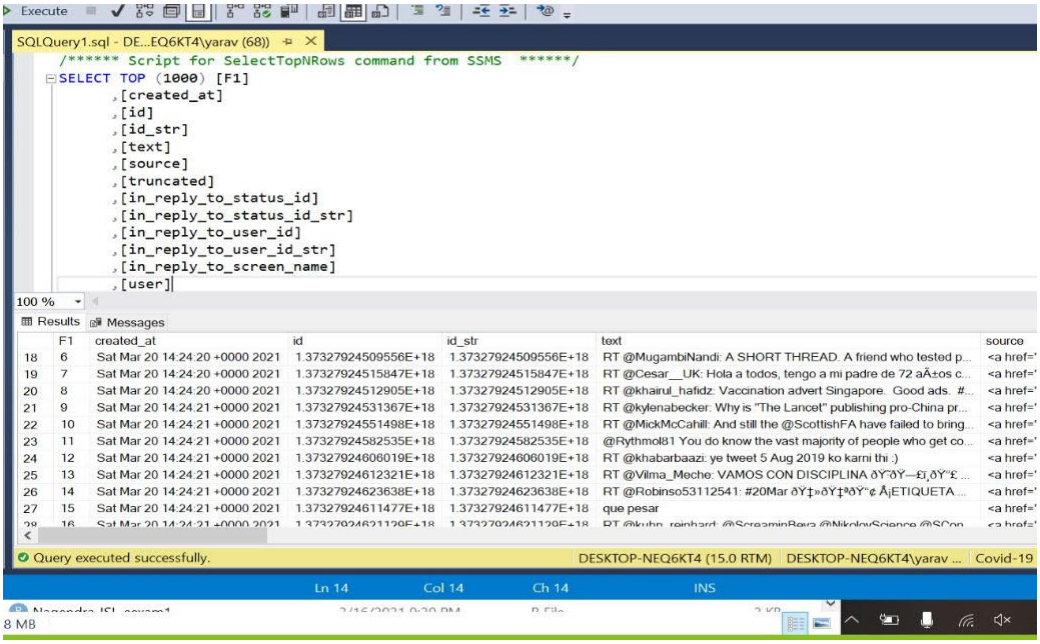
```
tweets_data_path='tweet.txt'
tweets_data=[]
tweets_file=open(tweets_data_path,"r")
#read in tweets and store on list
for line in tweets_file:
    tweet=json.loads(line)
    tweets_data.append(tweet)
tweets_file.close()
print(tweets_data[0])
```

{'created_at': 'Fri Mar 26 16:22:19 +0000 2021', 'id': 1375483261326135303, 'id_str': '1375483261326135303', 'text': 'RT @lesleyanngreen: 50 NOT OUT!!! https://t.co/KibAzagQ0k via @lesleyanngreen @MaryEGater first published Feb 2019 even more relevant now w…', 'source': 'Twitter Web App', 'truncated': False, 'in_reply_to_status_id': None, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'id': 232617366, 'id_str': '232617366', 'name': 'Cheshire Girl', 'screen_name': 'lesleyanngreen', 'location': 'Cheshire. UK GB ', 'url': 'http://Www.psychologytoday.com/profile/412570', 'description': 'HEALTH is WEALTH\nPsychotherapist, writer, health advocate for over 50's women \nBSc Psych (hons), MSc Clinical Science\n\nBlog -: https://t.co/Y3M4l0AlfX', 'translator_type': 'none', 'protected': False, 'verified': False, 'followers_count': 2734, 'friends_count': 4854, 'listed_count': 59, 'favourites_count': 47463, 'statuses_count': 23180, 'created_at': 'Fri Dec 31 17:55:16 +0000 2010', 'utc_offset': None, 'time_zone': None, 'geo_enabled': True, 'lang': None, 'contributors_enabled': False, 'is_translator': False, 'profile_background_color': '000000', 'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_tile': False, 'profile_link_color': '19CF86', 'profile_sidebar_border_color': '000000', 'profile_sidebar_fill_color': '000

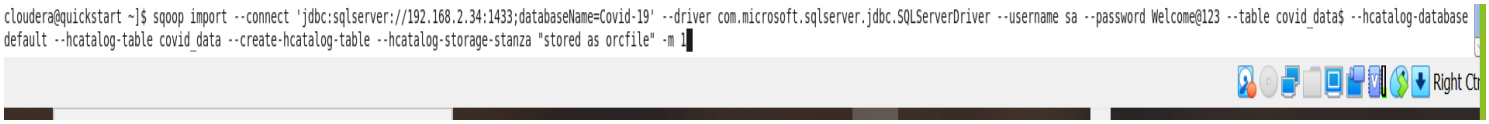
In [43]:

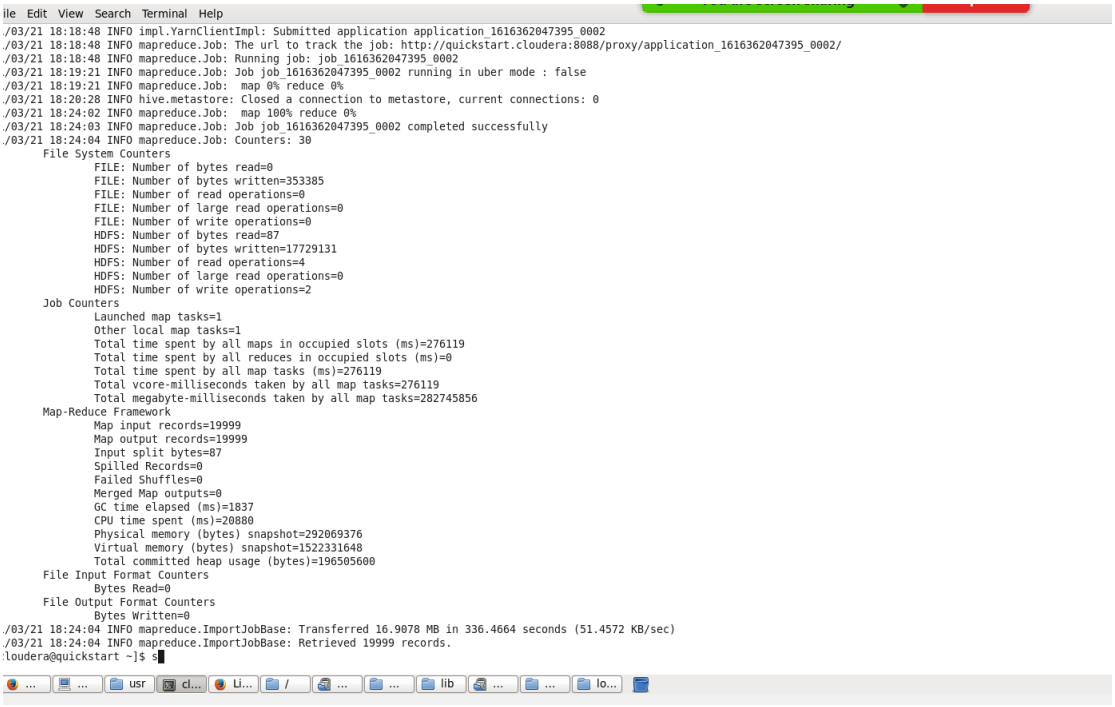
```
import snscrape.modules.twitter as sntwitter
# Creating list to append tweet data to
tweets_list2 = []
for n,j in enumerate(dates):
#     print(j)
    if int(j[5:7])==12:
        #print(j)
        query='(#Covid19 OR #CovidAtWork OR #Covid-19 OR
        print(query)
        maxTweets = 20000
        print(str(query))
        # Using TwitterSearchScraper to scrape data and ap
        for i,tweet in enumerate(sntwitter.TwitterSearchS
            print(i,'value')
            if i>maxTweets:
                break

            tweets_list2.append([tweet.date, tweet.id, twe
```

- **Part 2 - Loaded the data into SQL server using data import and export tool.**



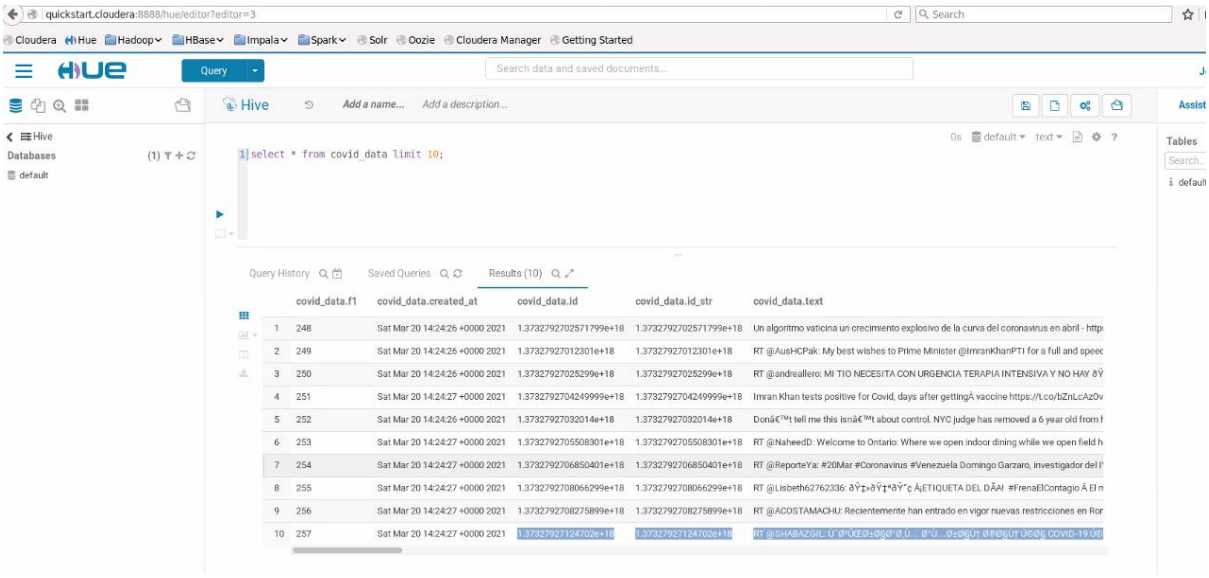
- **Part 3 - Use the SQOOP command to connect SQL server and import the data from SQL to Hive**



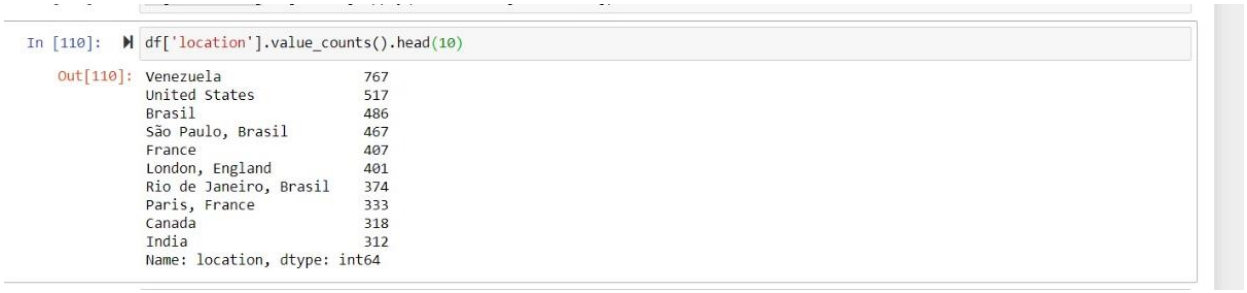


- **Part 4 - Viewed the data using by retrieving top 5 records.**

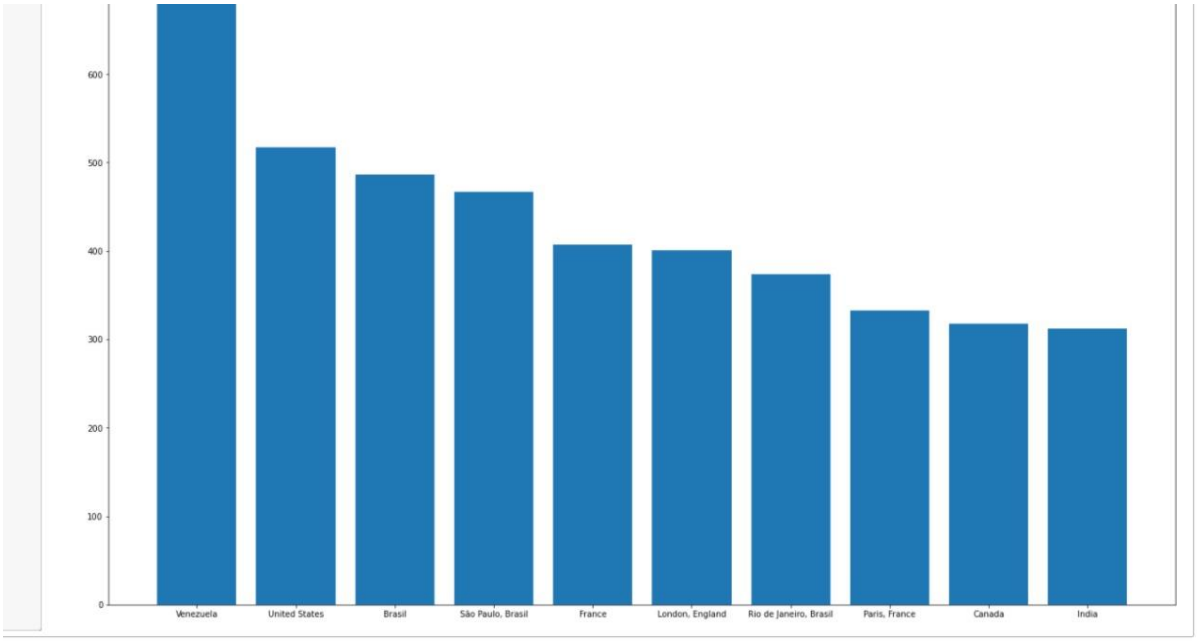
Visualizing the data using Hue



- ## Part 5 - Top 10 location counts based on the tweets



```
In [110]: df['location'].value_counts().head(10)

Out[110]: Venezuela                767
          United States            517
          Brasil                   486
          São Paulo, Brasil        467
          France                   407
          London, England          401
          Rio de Janeiro, Brasil   374
          Paris, France            333
          Canada                   318
          India                    312
          Name: location, dtype: int64
```



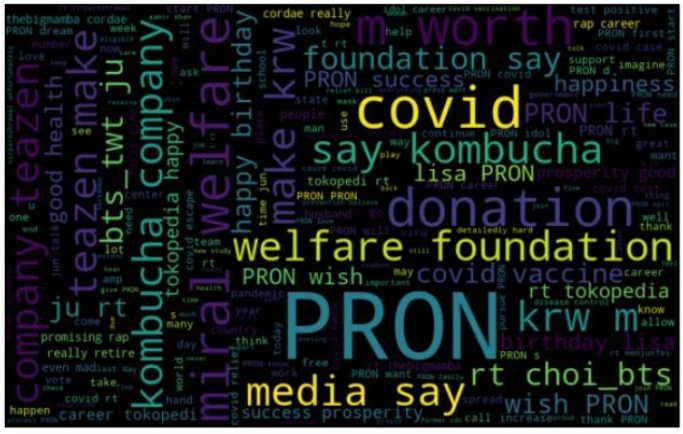
- **Part 6: Sentiment Analysis**

A) Word Cloud on whole data

In [25]:
```
# create text from all tweets
all_words = ' '.join([text for text in df['new_Text']])
import matplotlib.pyplot as plt
from wordcloud import WordCloud
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(all_words)

plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

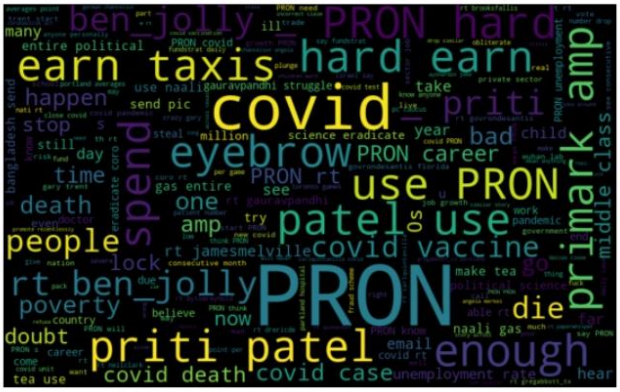
B) Word Cloud on positive data

In [26]:
```
# create text from just positive tweets
normal_words = ' '.join([text for text in df['new_Text'][df['sentiment'] == 'Positive']])

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```



C) Word Cloud on negative data

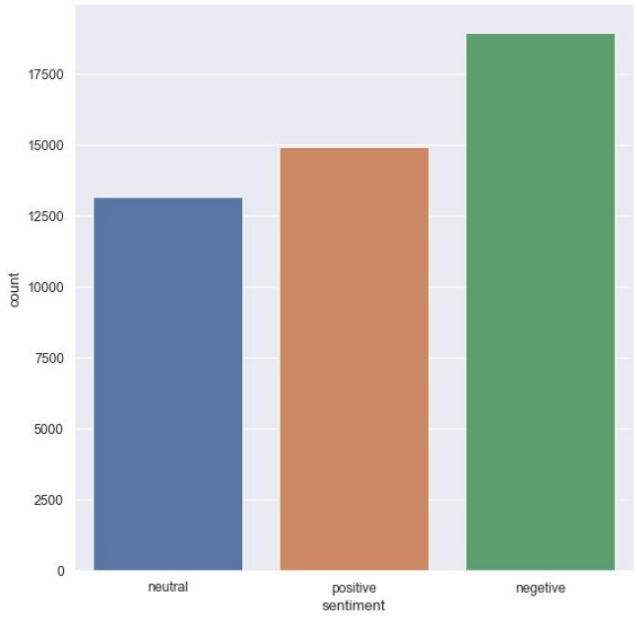
```
normal_words = ' '.join([text for text in df['new_Text'][df['sentiment'] == 'Negative']])

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```



D) Sentiment Analysis Graph

In [24]:
```
import seaborn as sns
sns.set(font_scale=1.1)
sns.catplot("sentiment", data=temp, kind="count", height=8)
```

Out[24]: <seaborn.axisgrid.FacetGrid at 0x210d72ab070>



- **Project Management**
  - **Implementation Status Report**
    - **Work Completed**
      - Description
        - ✓ Phase 1 and Phase 2 of the project focused on Designing the project, investigating the datasets, flow of the project, and tools required for implementation.
      - **Responsibility (Task, Person)**
        - ✓ Phase 1- Design and Analysis
          - Design of the project – Yugandhar Yarava
          - Investigating the datasets – Mani Sai Gundumogula
          - Designing the flow of the project – Devna Chaturvedi
          - Decide the tools to be used for implementation - Gireesh Kumar Muppalla
        - ✓ Phase 2 – Extract, Transform and Loading of Data
          - Extracting Text using SNScrape and Twitter API from Twitter – Mani Sai Gundumogula
          - Pre-processing of Text Data in Python – Gireesh Kumar
          - Imported the data to SQL server – Yugandhar Yarava
          - Export the data from SQL to HADOOP - HIVE using Sqoop – Devna Chaturvedi
      - **Contributions**
          - Gireesh Kumar Muppalla – 25%
          - Mani Sai Gundumogula – 25%
          - Devna Chaturvedi – 25%
          - Yugandhar Yarava – 25%

    - **Work to be Completed**
      - **Description**
        - ✓ Phase 3 of the project composes of Data Analysis and Visualization of data through sentiment Analysis and performing data analysis. Loading data into Solr and perform queries.
      - **Responsibilities**
        - ✓ **Phase 3 – Visualization and Analysis**
          - Loading Data into Spark and use pyspark for data analysis – Devna Chaturvedi
          - Spark queries for data analysis and statistics using SPARK SQL – Gireesh Kumar Muppalla
          - Sentiment Analysis using Matplotlib and Seaborn - Yugandhar Yarava
          - Loading Data into Solr and implement queries on Solr Mani Sai Gundumogula

Assignment 2 – Story Telling

Who?

The data is related to the story's main characters, the people who are facing the brunt of unemployment in various sectors. The data extracted comes from the tweets of people who have disclosed their issues and concerns on unemployment. The use of hashtags like unemployment in combination with COVID-19, covidatwork, poverty, working poor shows how people are getting subjected to the loss of employment hours and layoffs.

What?

The event happening at workplaces like unemployment, the closing of businesses, loss in total working hours, less compensation for work are rampant due to the pandemic. This data analysis can help prospective people looking for jobs in the current market scenario and focus on specific employment opportunities making them more aware of the real-time situation.

When?

The event started in March 2020 and has been spreading rapidly. People react to the unemployment scenario on social media platforms, especially on Twitter, where the hashtags covid-19 in conjugation with unemployment hashtags help filter the data focused on unemployed people. The data is collected in real-time from twitter. 7.8 million tweets are collected till now. The data collected from Twitter using SNScrape is from January 2020 to December 2020.

Where?

Globally, from all over the world the data is extracted using Twitter. As a significant proportion of the society is composed of the workforce, the data is relevant to various age groups and multiple sectors. Thus, we are working on this dataset considering all the factors. The data is collected from Twitter to explore the global scenario and filter the top 10 countries who are getting impacted.

Why?

Most people initially believed that the unemployment wave would halt, and the situation will normalize, but this doesn’t seem to happen in the present scenario. Industries and the workforce are striving hard to restore normality. The concern of unemployment surrounding the people is our motivation behind doing the tweet data collection and perform the data analysis.

Calibration 1

The problem project and explained in project increment 1 describes the global scenario on unemployment and our story's characters getting impacted by it. Assignment 2 emphasizes how the data collected relates to the problem and the characters mentioned in assignment 1. The conditions of unemployment, job losses, closing of businesses, loss in compensations the needs of the story in assignment 1 align with the assignment 2 data extracted from Twitter.

References:

1. https://abcnews.go.com/Business/52-million-americans-file-unemployment-amid-covid-19/story?id=70180549
2. https://www.brookings.edu/research/explaining-the-economic-impact-of-covid-19-core-industries-and-the-hispanic-workforce/
3. https://blog.twitter.com/en_us/topics/company/2020/covid-19.html
4. https://www.npr.org/sections/money/2020/10/27/927842540/the-dark-side-of-the-recovery-revealed-in-big-data
5. https://www.hdfstutorial.com/sqoop-architecture/
6. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
7. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
8. https://hadoop.apache.org/docs/r1.2.1/commands_manual.pdf
9. https://linoxide.com/images/hadoop-hdfs-commands-cheatsheet-900x1500.png
10. https://www.thegeekstuff.com/2015/02/hadoop-command-reference/#comments
11. https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/documents/briefingnote/wcms_767028.pdf
12. https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/documents/briefingnote/wcms_767028.pdf
13. https://www.pwc.com/us/en.html
14. https://www.washingtonpost.com/news/wonk/wp/2014/04/07/twitter-is-surprisingly-accurate-at-predicting-unemployment/