3. Analysis of Emojis, Emoticons, and Hashtags

This notebook covers the frequency distributions of emojis, emoticons, and hashtags across positive, neutral, and negative sentiments. The intention is to analyze and determine if there is are patterns amongst the use of emojis, emoticons, and hashtags that could possibly further help determine the sentiment of the tweet.

Import Libraries and Read the Cleaned Data

In [1]:
```
import pandas as pd
from collections import Counter
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
```

Read 'Tweets_cleaned.csv'

In [2]:
```
df_tweets_cleaned = pd.read_csv('..\data\Tweets_cleaned.csv', encoding='utf-8'
)
```

Separate the dataframe into positive, neutral, and negative classified sentiment.

In [3]:
```
df_tweets_positive = df_tweets_cleaned.loc[df_tweets_cleaned['airline_sentimen
t'] == 'positive']
df_tweets_neutral = df_tweets_cleaned.loc[df_tweets_cleaned['airline_sentimen
t'] == 'neutral']
df_tweets_negative = df_tweets_cleaned.loc[df_tweets_cleaned['airline_sentimen
t'] == 'negative']
```

Distribution of emojis grouped by positive, neutral, and negative sentiments

Positive tweets with Emojis

In [4]:
```
#create a list of tokens from the tweets
list_of_emojis_positive = []
df_tweets_positive_emojis = df_tweets_positive.loc[df_tweets_positive['emojis_
flag'] == True]
```

Number of tweets of positive sentiment that contained emojis:

In [5]:
```
len(df_tweets_positive_emojis)
```

Out[5]: 189

In [6]:
```
#create a list
positive_emojis = []

for i, row in df_tweets_positive_emojis.iterrows():
    tweet_emojis = df_tweets_positive_emojis.at[i, 'emojis']
    #print(type(tweet_emojis))

    tweet_emojis_list = list(tweet_emojis.split(","))

    for emoji in tweet_emojis_list:
        #print('emoji: ' + emoji)

        #strip brackets, quote, and spaces
        emoji = emoji.strip('[]')
        emoji = emoji.replace("\'", "")
        emoji = emoji.strip()

        #print('emoji_strip: ' + emoji)
        positive_emojis.append(emoji)
```

Number of emojis used in positive tweets:

In [7]:
```
len(positive_emojis)
```

Out[7]: 457

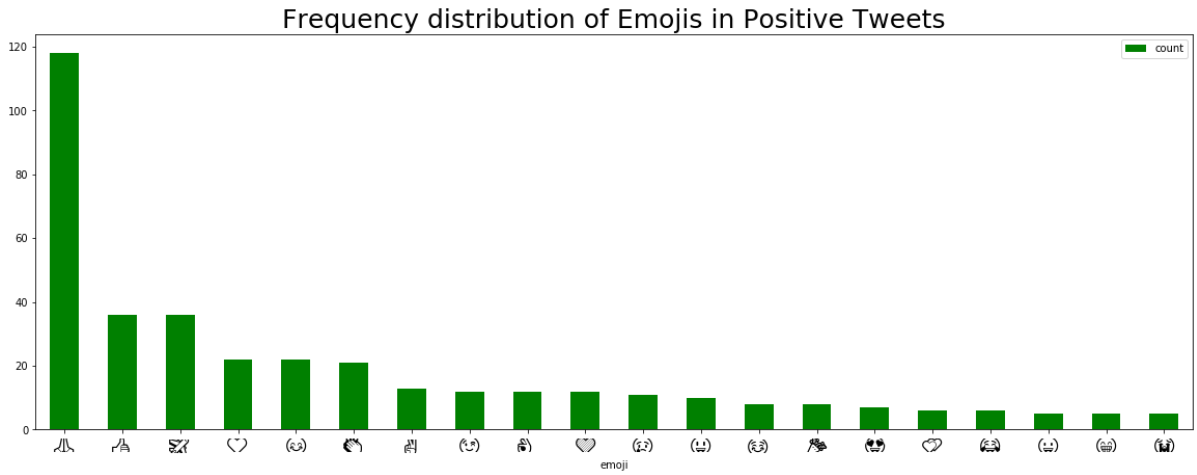
In [8]:
```
positive_emoji_counter = Counter(positive_emojis)
positive_emoji_counter.most_common(20)
```

Out[8]:
```
[('🙏', 118),
 ('👍', 36),
 ('✈', 36),
 ('❤', 22),
 ('😊', 22),
 ('👏', 21),
 ('☝', 13),
 ('😉', 12),
 ('👌', 12),
 ('🤍', 12),
 ('😢', 11),
 ('😃', 10),
 ('☺', 8),
 ('🎉', 8),
 ('😍', 7),
 ('💞', 6),
 ('😂', 6),
 ('😀', 5),
 ('😁', 5),
 ('🤳', 5)]
```

In [9]:
```
# convert list of tuples into data frame
freq_df_positive_emoji = pd.DataFrame.from_records(positive_emoji_counter.most
_common(20),
                                   columns=['emoji', 'count'])

# create bar plot
distribution_bar_positive_emojis = freq_df_positive_emoji.plot(
    kind='bar',
    x='emoji',
    color='green',
    figsize=(20,7)
)
distribution_bar_positive_emojis.set_title(
    'Frequency distribution of Emojis in Positive Tweets',
    fontsize=25
)
distribution_bar_positive_emojis.set_xticklabels(
    freq_df_positive_emoji.emoji.tolist(),
    rotation = 0,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

```
Out[9]: [Text(0, 0, '🙏'),
         Text(0, 0, '👍'),
         Text(0, 0, '✈'),
         Text(0, 0, '❤'),
         Text(0, 0, '😊'),
         Text(0, 0, '👌'),
         Text(0, 0, '✌'),
         Text(0, 0, '😉'),
         Text(0, 0, '👌'),
         Text(0, 0, '♡'),
         Text(0, 0, '😘'),
         Text(0, 0, '😃'),
         Text(0, 0, '☺'),
         Text(0, 0, '🎉'),
         Text(0, 0, '😍'),
         Text(0, 0, '💕'),
         Text(0, 0, '😂'),
         Text(0, 0, '😀'),
         Text(0, 0, '😁'),
         Text(0, 0, '😭')]
```



Analysis of Frequency distribution of Emojis in Positive Tweets

Interestingly, the `Prayer Hands` emoji topped the frequency distribution. This emoji is commonly used as a way of saying "please" or "thank you", and therefore it makes sense it should be identified as an emoji with positive sentiment. All other emojis were not as commonly used. Another interesting emoji related to positive sentiment was the `Airplane` emoji; upon manual examination the text of the tweets, users sometimes tweeted this emoji as a way of expressing excitement on their trip. Surprisingly at first glance, the `Happy Face` emoji and its several variations were not used quite as often; in retrospect, this made sense simply because there is a large variety of `Happy Face` emojis available for users to use. It may be worth manually clustering the `Happy Face` emojis, as there are a large number of such categories in the lower frequency distributions.

Neutral tweets with Emojis

In [10]:
```
#create a list of tokens from the tweets
list_of_emojis_neutral = []
df_tweets_neutral_emojis = df_tweets_neutral.loc[df_tweets_neutral['emojis_fla
g'] == True]
```

Number of tweets of neutral sentiment that contained emojis:

In [11]:
```
len(df_tweets_neutral_emojis)
```

Out[11]: 127

In [12]:
```
#create a list
neutral_emojis = []

for i, row in df_tweets_neutral_emojis.iterrows():
    tweet_emojis = df_tweets_neutral_emojis.at[i, 'emojis']
    #print(type(tweet_emojis))

    tweet_emojis_list = list(tweet_emojis.split(","))

    for emoji in tweet_emojis_list:
        #print('emoji: ' + emoji)

        #strip brackets, quote, and spaces
        emoji = emoji.strip('[]')
        emoji = emoji.replace("\'", "")
        emoji = emoji.strip()

        #print('emoji_strip: ' + emoji)
        neutral_emojis.append(emoji)
```

Number of emojis used in neutral tweets:

In [13]:
```
len(neutral_emojis)
```

Out[13]: 239

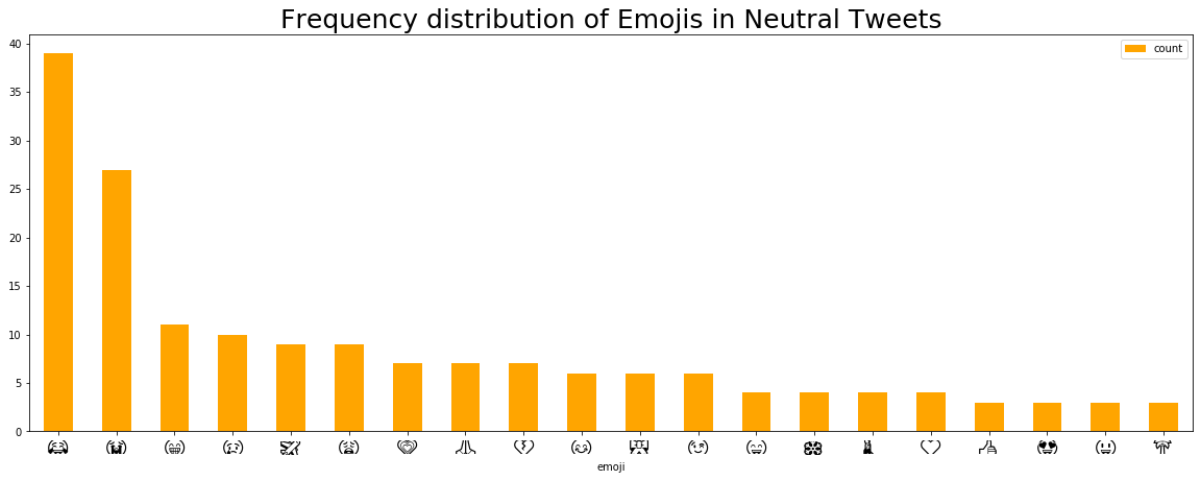
In [14]:
```
neutral_emoji_counter = Counter(neutral_emojis)
neutral_emoji_counter.most_common(20)
```

Out[14]:
```
[('😂', 39),
 ('😭', 27),
 ('😁', 11),
 ('😢', 10),
 ('✈', 9),
 ('😩', 9),
 ('💝', 7),
 ('🙏', 7),
 ('💔', 7),
 ('😋', 6),
 ('🎀', 6),
 ('😉', 6),
 ('😄', 4),
 ('❄', 4),
 ('🎂', 4),
 ('❤', 4),
 ('👍', 3),
 ('😍', 3),
 ('😃', 3),
 ('🌴', 3)]
```

In [15]:
```
# convert list of tuples into data frame
freq_df_neutral_emoji = pd.DataFrame.from_records(neutral_emoji_counter.most_c
ommon(20),
                                        columns=['emoji', 'count'])

# create bar plot
distribution_bar_neutral_emojis = freq_df_neutral_emoji.plot(
    kind='bar',
    x='emoji',
    color='orange',
    figsize=(20,7)
)
distribution_bar_neutral_emojis.set_title(
    'Frequency distribution of Emojis in Neutral Tweets',
    fontsize=25
)
distribution_bar_neutral_emojis.set_xticklabels(
    freq_df_neutral_emoji.emoji.tolist(),
    rotation = 0,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

Out[15]:
```
[Text(0, 0, '😂'),
 Text(0, 0, '😭'),
 Text(0, 0, '😁'),
 Text(0, 0, '😢'),
 Text(0, 0, '✈'),
 Text(0, 0, '😩'),
 Text(0, 0, '💝'),
 Text(0, 0, '🙏'),
 Text(0, 0, '💔'),
 Text(0, 0, '😊'),
 Text(0, 0, '🎀'),
 Text(0, 0, '😉'),
 Text(0, 0, '😄'),
 Text(0, 0, '❄'),
 Text(0, 0, '🎅'),
 Text(0, 0, '❤'),
 Text(0, 0, '👍'),
 Text(0, 0, '😍'),
 Text(0, 0, '😐'),
 Text(0, 0, '👼')]
```



Analysis of Frequency distribution of Emojis in Neutral Tweets

There is much overlap of emojis used in positive and neutral tweets. For example, several of the various `Happy Face` emojis are used in both positive ane neutral tweets. On the flip side, there are only two emojis which overlap in neutral and negative sentiments. Overall,the number of emojis used in neutral tweets were far lower than positive tweets, even if there were overlaps in the emoji usage.

Negative tweets with Emojis

In [16]:
```
#create a list of tokens from the tweets
list_of_emojis_negative = []
df_tweets_negative_emojis = df_tweets_negative.loc[df_tweets_negative['emojis_
flag'] == True]
```

Number of tweets of negative sentiment that contained emojis:

In [17]:
```
len(df_tweets_negative_emojis)
```

Out[17]: 173

In [18]:
```
#create a list
negative_emojis = []

for i, row in df_tweets_negative_emojis.iterrows():
    tweet_emojis = df_tweets_negative_emojis.at[i, 'emojis']
    #print(type(tweet_emojis))

    tweet_emojis_list = list(tweet_emojis.split(","))

    for emoji in tweet_emojis_list:
        #print('emoji: ' + emoji)

        #strip brackets, quote, and spaces
        emoji = emoji.strip('[]')
        emoji = emoji.replace("\'", "")
        emoji = emoji.strip()

        #print('emoji_strip: ' + emoji)
        negative_emojis.append(emoji)
```

Number of emojis used in negative tweets:

In [19]:
```
len(negative_emojis)
```

Out[19]: 280

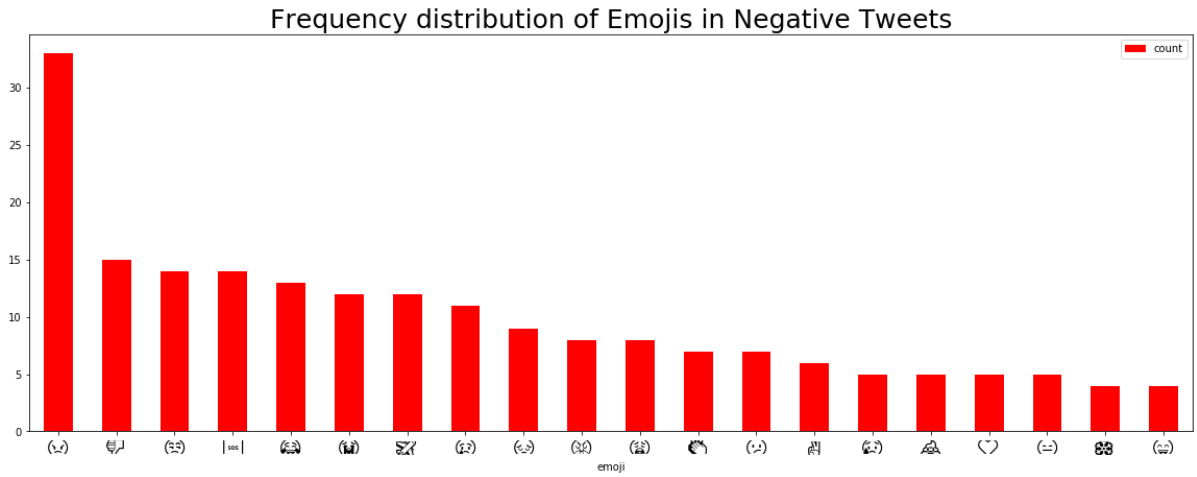
In [20]:
```
negative_emojis_counter = Counter(negative_emojis)
negative_emojis_counter.most_common(20)
```

Out[20]:
```
[('😠', 33),
 ('👎', 15),
 ('😣', 14),
 ('🆘', 14),
 ('😂', 13),
 ('😭', 12),
 ('✈', 12),
 ('😢', 11),
 ('😔', 9),
 ('😵', 8),
 ('😩', 8),
 ('👏', 7),
 ('😕', 7),
 ('☝', 6),
 ('😥', 5),
 ('💩', 5),
 ('❤', 5),
 ('😑', 5),
 ('❄', 4),
 ('😁', 4)]
```

In [21]:
```
# convert list of tuples into data frame
freq_df_negative_emojis = pd.DataFrame.from_records(negative_emojis_counter.mo
st_common(20),
                                    columns=['emoji', 'count'])

# create bar plot
distribution_bar_negative_emojis = freq_df_negative_emojis.plot(
    kind='bar',
    x='emoji',
    color='red',
    figsize=(20,7)
)
distribution_bar_negative_emojis.set_title(
    'Frequency distribution of Emojis in Negative Tweets',
    fontsize=25
)
distribution_bar_negative_emojis.set_xticklabels(
    freq_df_negative_emojis.emoji.tolist(),
    rotation = 0,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

```
Out[21]: [Text(0, 0, '😠'),
          Text(0, 0, '👎'),
          Text(0, 0, '😤'),
          Text(0, 0, '🆘'),
          Text(0, 0, '😂'),
          Text(0, 0, '😭'),
          Text(0, 0, '✈'),
          Text(0, 0, '😢'),
          Text(0, 0, '😔'),
          Text(0, 0, '😵'),
          Text(0, 0, '😩'),
          Text(0, 0, '👏'),
          Text(0, 0, '😕'),
          Text(0, 0, '✌'),
          Text(0, 0, '😥'),
          Text(0, 0, '💩'),
          Text(0, 0, '❤'),
          Text(0, 0, '😑'),
          Text(0, 0, '❄'),
          Text(0, 0, '😁')]
```



Analysis of Frequency distribution of Emojis in Negative Tweets

The `Angry Face` emoji was by far the most used to express negative sentiment. The usage of the emojis indicated that users were mostly angry, sad, or dissatisfied. There is very little overlap of the emojis between negative and positive tweets, and negative and neutral tweets, but the general variety of emojis used is consistent to their usage of negative expression of sentiment. Overall number of emojis used in negative tweets was slightly higher than neutral tweets.

Distribution of emoticons grouped by positive, neutral, and negative sentiments

Positive tweets with Emoticons

In [22]:
```
#create a list of tokens from the tweets
list_of_emoticons_positive = []
df_tweets_positive_emoticons = df_tweets_positive.loc[df_tweets_positive['emot
icons_flag'] == True]
```

Number of tweets of positive sentiment that contained emoticons:

In [23]:
```
len(df_tweets_positive_emoticons)
```

Out[23]: 146

In [24]:
```
#create a list
positive_emoticons = []

for i, row in df_tweets_positive_emoticons.iterrows():
    tweet_emoticons = df_tweets_positive_emoticons.at[i, 'emoticons']
    #print(type(tweet_emoticons))

    tweet_emoticons_list = list(tweet_emoticons.split(","))

    for emoticon in tweet_emoticons_list:
        #print('emoticon: ' + emoticon)

        #strip brackets, quote, and spaces
        emoticon = emoticon.strip('[]')
        emoticon = emoticon.replace("\'", "")
        emoticon = emoticon.strip()

        #print('emoticon_strip: ' + emoticon)
        positive_emoticons.append(emoticon)
```

Number of emoticons used in positive tweets:

In [25]:
```
len(positive_emoticons)
```

Out[25]: 150

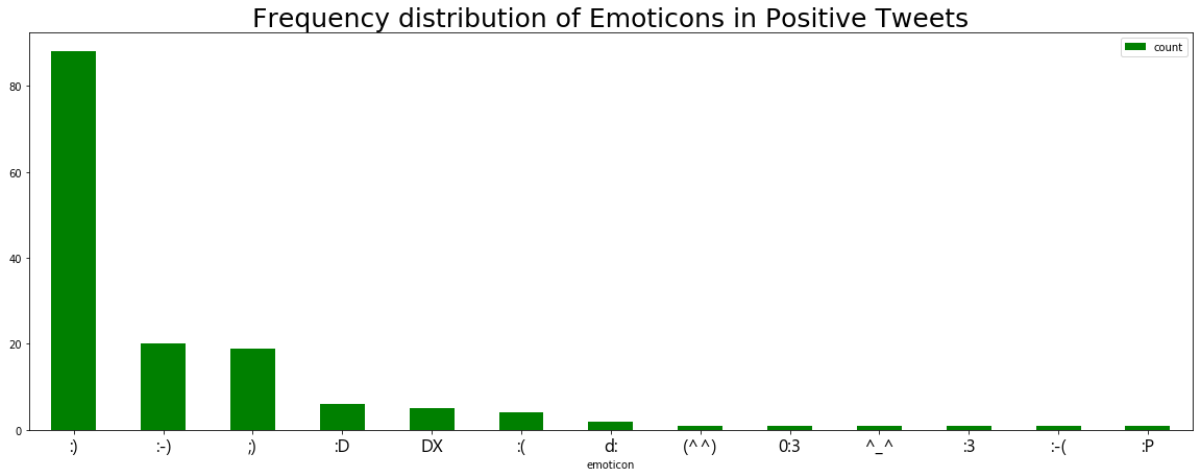
In [26]:
```
positive_emoticons_counter = Counter(positive_emoticons)
positive_emoticons_counter.most_common(20)
```

Out[26]:
```
[(':)', 88),
 (':-)', 20),
 (';)', 19),
 (':D', 6),
 ('DX', 5),
 (':(', 4),
 ('d:', 2),
 ('(^^)', 1),
 ('0:3', 1),
 ('^_^', 1),
 (':3', 1),
 (':-(', 1),
 (':P', 1)]
```

In [27]:
```
# convert list of tuples into data frame
freq_df_positive_emoticons = pd.DataFrame.from_records(positive_emoticons_coun
ter.most_common(20),
                                       columns=['emoticon', 'count'])

# create bar plot
distribution_bar_positive_emoticons = freq_df_positive_emoticons.plot(
    kind='bar',
    x='emoticon',
    color='green',
    figsize=(20,7)
)
distribution_bar_positive_emoticons.set_title(
    'Frequency distribution of Emoticons in Positive Tweets',
    fontsize=25
)
distribution_bar_positive_emoticons.set_xticklabels(
    freq_df_positive_emoticons.emoticon.tolist(),
    rotation = 0,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

Out[27]:
```
[Text(0, 0, ':)'),
 Text(0, 0, ':-)'),
 Text(0, 0, ';)'),
 Text(0, 0, ':D'),
 Text(0, 0, 'DX'),
 Text(0, 0, ':('),
 Text(0, 0, 'd:'),
 Text(0, 0, '(^^)'),
 Text(0, 0, '0:3'),
 Text(0, 0, '^_^'),
 Text(0, 0, ':3'),
 Text(0, 0, ':-('),
 Text(0, 0, ':P')]
```



Analysis of Frequency distribution of Emoticons in Positive Tweets

Compared to the use of emojis in positive tweets, emoticons were not as popular. This makes sense given the "youthful" nature of Twitter, most users would know how to use emojis, instead of typing out emoticons, which is considered the "old school" method of expression emotion via pseudo-visual imagery. Unlike emojis in positive tweets, the `Happy Face` emoticon was by far the most used. Other emoticons were not used quite as frequently.

Neutral tweets with Emoticons

In [28]:
```
#create a list of tokens from the tweets
list_of_emoticons_neutral = []
df_tweets_neutral_emoticons = df_tweets_neutral.loc[df_tweets_neutral['emoticons_flag'] == True]
```

Number of tweets of neutral sentiment that contained emoticons:

In [29]:
```
len(df_tweets_neutral_emoticons)
```

Out[29]: 71

In [30]:
```
#create a list
neutral_emoticons = []

for i, row in df_tweets_neutral_emoticons.iterrows():
    tweet_emoticons = df_tweets_neutral_emoticons.at[i, 'emoticons']
    #print(type(tweet_emoticons))

    tweet_emoticons_list = list(tweet_emoticons.split(","))

    for emoticon in tweet_emoticons_list:
        #print('emoticon: ' + emoticon)

        #strip brackets, quote, and spaces
        emoticon = emoticon.strip('[]')
        emoticon = emoticon.replace("\'", "")
        emoticon = emoticon.strip()

        #print('emoticon_strip: ' + emoticon)
        neutral_emoticons.append(emoticon)
```

Number of emoticons used in neutral tweets:

In [31]:
```
len(neutral_emoticons)
```

Out[31]: 71

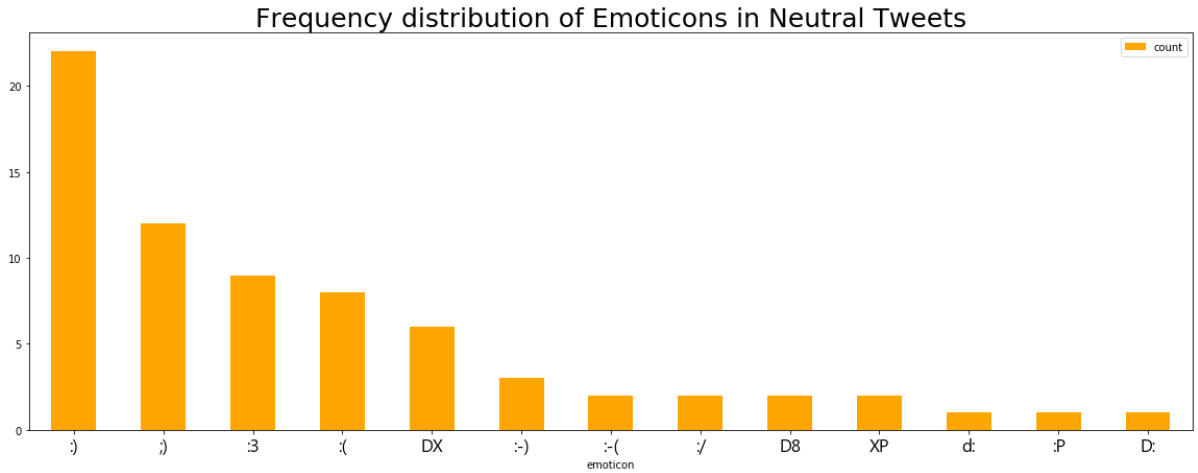
In [32]:
```
neutral_emoticons_counter = Counter(neutral_emoticons)
neutral_emoticons_counter.most_common(20)
```

Out[32]:
```
[(':)', 22),
 (';)', 12),
 (':3', 9),
 (':(', 8),
 ('DX', 6),
 (':-)', 3),
 (':-(', 2),
 (':/', 2),
 ('D8', 2),
 ('XP', 2),
 ('d:', 1),
 (':P', 1),
 ('D:', 1)]
```

In [33]:
```
# convert list of tuples into data frame
freq_df_neutral_emoticons = pd.DataFrame.from_records(neutral_emoticons_counte
r.most_common(20),
                                     columns=['emoticon', 'count'])

# create bar plot
distribution_bar_neutral_emoticons = freq_df_neutral_emoticons.plot(
    kind='bar',
    x='emoticon',
    color='orange',
    figsize=(20,7)
)
distribution_bar_neutral_emoticons.set_title(
    'Frequency distribution of Emoticons in Neutral Tweets',
    fontsize=25
)
distribution_bar_neutral_emoticons.set_xticklabels(
    freq_df_neutral_emoticons.emoticon.tolist(),
    rotation = 0,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

Out[33]:
```
[Text(0, 0, ':)'),
 Text(0, 0, ';)'),
 Text(0, 0, ':3'),
 Text(0, 0, ':('),
 Text(0, 0, 'DX'),
 Text(0, 0, ':-)'),
 Text(0, 0, ':-('),
 Text(0, 0, ':/'),
 Text(0, 0, 'D8'),
 Text(0, 0, 'XP'),
 Text(0, 0, 'd:'),
 Text(0, 0, ':P'),
 Text(0, 0, 'D:')]
```



Analysis of Frequency distribution of Emoticons in Neutral Tweets

Similar to the use of emojis in neutral tweets, there is much overlap between the use of emoticons between positive and neutral sentiments, and not as much overlap between the use of emoticons between neutral and negative sentiments. There were not a lot of neutral tweets that contained emoticons either.

Negative tweets with Emoticons

In [34]:
```
#create a list of tokens from the tweets
list_of_emoticons_negative = []
df_tweets_negative_emoticons = df_tweets_negative.loc[df_tweets_negative['emot
icons_flag'] == True]
```

Number of tweets of negative sentiment that contained emoticons:

In [35]:
```
len(df_tweets_negative_emoticons)
```

Out[35]: 156

In [36]:
```
#create a list
negative_emoticons = []

for i, row in df_tweets_negative_emoticons.iterrows():
    tweet_emoticons = df_tweets_negative_emoticons.at[i, 'emoticons']
    #print(type(tweet_emoticons))

    tweet_emoticons_list = list(tweet_emoticons.split(","))

    for emoticon in tweet_emoticons_list:
        #print('emoticon: ' + emoticon)

        #strip brackets, quote, and spaces
        emoticon = emoticon.strip('[]')
        emoticon = emoticon.replace("\'", "")
        emoticon = emoticon.strip()

        #print('emoticon_strip: ' + emoticon)
        negative_emoticons.append(emoticon)
```

Number of emoticons used in negative tweets:

In [37]:
```
len(negative_emoticons)
```

Out[37]: 165

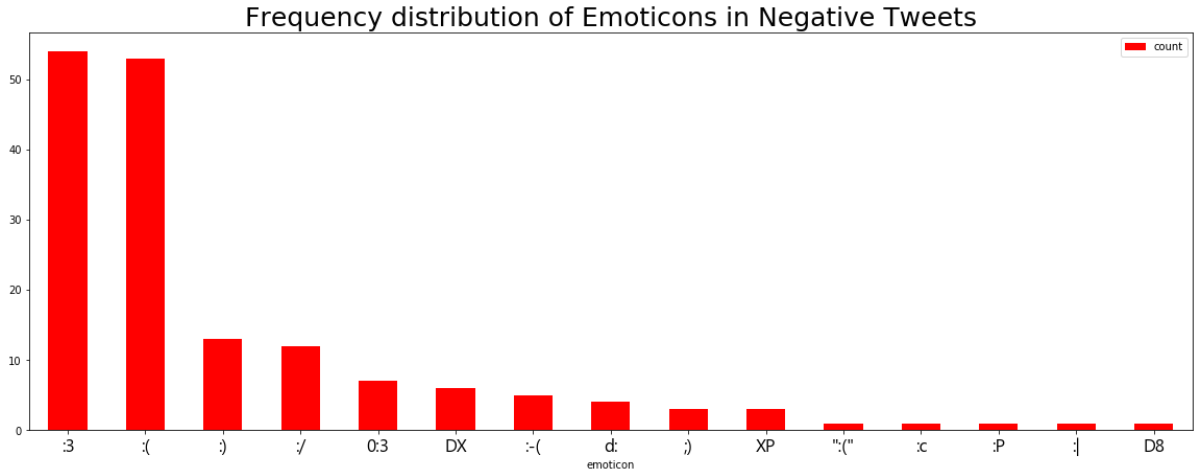
In [38]:
```
negative_emoticons_counter = Counter(negative_emoticons)
negative_emoticons_counter.most_common(20)
```

Out[38]:
```
[(':3', 54),
 (':(', 53),
 (':)', 13),
 (':/', 12),
 ('0:3', 7),
 ('DX', 6),
 (':-(', 5),
 ('d:', 4),
 (';)', 3),
 ('XP', 3),
 ('":("', 1),
 (':c', 1),
 (':P', 1),
 (':|', 1),
 ('D8', 1)]
```

In [39]:
```
# convert list of tuples into data frame
freq_df_negative_emoticons = pd.DataFrame.from_records(negative_emoticons_coun
ter.most_common(20),
                                       columns=['emoticon', 'count'])

# create bar plot
distribution_bar_negative_emoticons = freq_df_negative_emoticons.plot(
    kind='bar',
    x='emoticon',
    color='red',
    figsize=(20,7)
)
distribution_bar_negative_emoticons.set_title(
    'Frequency distribution of Emoticons in Negative Tweets',
    fontsize=25
)
distribution_bar_negative_emoticons.set_xticklabels(
    freq_df_negative_emoticons.emoticon.tolist(),
    rotation = 0,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

Out[39]:
```
[Text(0, 0, ':3'),
 Text(0, 0, ':('),
 Text(0, 0, ':)'),
 Text(0, 0, ':/'),
 Text(0, 0, '0:3'),
 Text(0, 0, 'DX'),
 Text(0, 0, ':-('),
 Text(0, 0, 'd:'),
 Text(0, 0, ';)'),
 Text(0, 0, 'XP'),
 Text(0, 0, '":("'),
 Text(0, 0, ':c'),
 Text(0, 0, ':P'),
 Text(0, 0, ':|'),
 Text(0, 0, 'D8')]
```



Analysis of Frequency distribution of Emoticons in Negative Tweets

It must be stated that it was initially surprising to see the `Cat Face` , or `:3` emoticon top the frequency distribution. However, upon closer inspection of the data, there is a strong possibility that there was a mistake in our data cleaning steps; for some reason, tweets contained times (e.g. "12:30") would be extracted and interpreted as emoticons. This represents a data quality issue that will need to be fixed. Other emoticons with similar data quality issue include `0:3` , `DX` , and `d:` . Despite these data quality issues, it can be noted that some of the emoticons, such as the typical `Sad Face` , or `:(` emoticon do represent what is considered a normal way of expressing negative sentiment.

Distribution of hashtags grouped by positive, neutral, and negative sentiments

Positive tweets with Hashtags

In [40]:
```
#create a list of tokens from the tweets
list_of_hashtags_positive = []
df_tweets_positive_hashtags = df_tweets_positive.loc[df_tweets_positive['hasht
ags_flag'] == True]
```

Number of tweets of positive sentiment that contained hashtags:

In [41]:
```
len(df_tweets_positive_hashtags)
```

Out[41]: 435

In [42]:
```
#create a list
positive_hashtags = []

for i, row in df_tweets_positive_hashtags.iterrows():
    tweet_hashtags = df_tweets_positive_hashtags.at[i, 'hashtags']
    #print(type(tweet_hashtags))

    tweet_hashtags_list = list(tweet_hashtags.split(","))

    for hashtag in tweet_hashtags_list:
        #print('hashtag: ' + hashtag)

        #strip brackets, quote, and spaces
        hashtag = hashtag.strip('[]')
        hashtag = hashtag.replace("\'", "")
        hashtag = hashtag.strip()

        #print('hashtag_strip: ' + hashtag)
        positive_hashtags.append(hashtag)
```

Number of hashtags used in positive tweets:

In [43]:
```
len(positive_hashtags)
```

Out[43]:
```
699
```

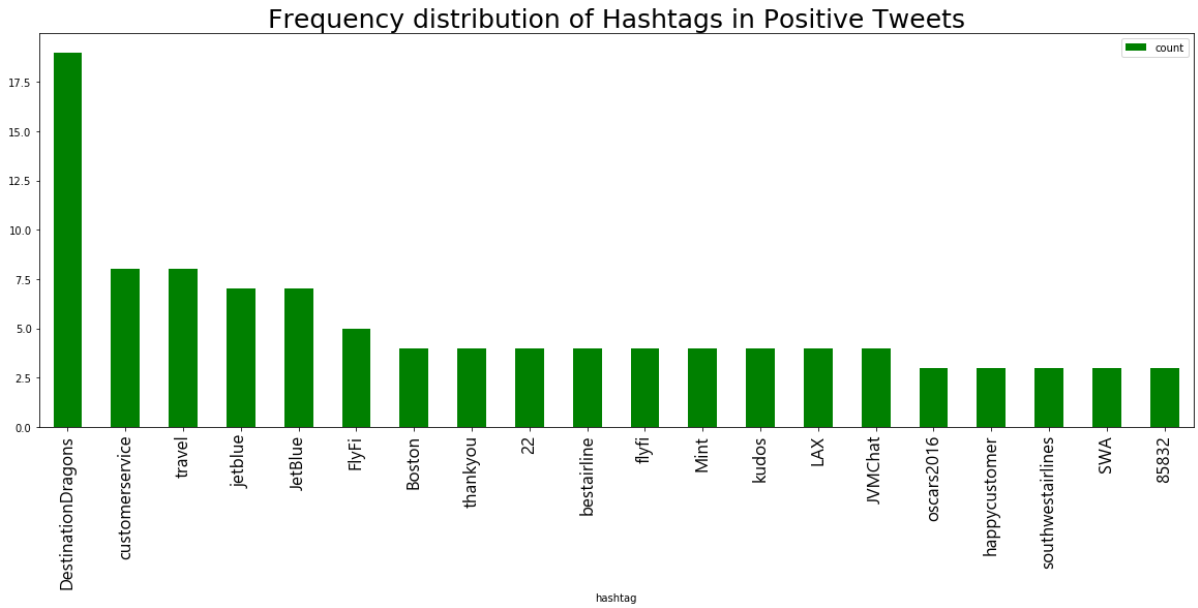
In [44]:
```
positive_hashtags_counter = Counter(positive_hashtags)
positive_hashtags_counter.most_common(20)
```

Out[44]:
```
[('DestinationDragons', 19),
 ('customerservice', 8),
 ('travel', 8),
 ('jetblue', 7),
 ('JetBlue', 7),
 ('FlyFi', 5),
 ('Boston', 4),
 ('thankyou', 4),
 ('22', 4),
 ('bestairline', 4),
 ('flyfi', 4),
 ('Mint', 4),
 ('kudos', 4),
 ('LAX', 4),
 ('JVMChat', 4),
 ('oscars2016', 3),
 ('happycustomer', 3),
 ('southwestairlines', 3),
 ('SWA', 3),
 ('85832', 3)]
```

In [45]:
```
# convert list of tuples into data frame
freq_df_positive_hashtags = pd.DataFrame.from_records(positive_hashtags_counte
r.most_common(20),
                                      columns=['hashtag', 'count'])

# create bar plot
distribution_bar_positive_hashtags = freq_df_positive_hashtags.plot(
    kind='bar',
    x='hashtag',
    color='green',
    figsize=(20,7)
)
distribution_bar_positive_hashtags.set_title(
    'Frequency distribution of Hashtags in Positive Tweets',
    fontsize=25
)
distribution_bar_positive_hashtags.set_xticklabels(
    freq_df_positive_hashtags.hashtag.tolist(),
    rotation = 90,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

```
Out[45]: [Text(0, 0, 'DestinationDragons'),
          Text(0, 0, 'customerservice'),
          Text(0, 0, 'travel'),
          Text(0, 0, 'jetblue'),
          Text(0, 0, 'JetBlue'),
          Text(0, 0, 'FlyFi'),
          Text(0, 0, 'Boston'),
          Text(0, 0, 'thankyou'),
          Text(0, 0, '22'),
          Text(0, 0, 'bestairline'),
          Text(0, 0, 'flyfi'),
          Text(0, 0, 'Mint'),
          Text(0, 0, 'kudos'),
          Text(0, 0, 'LAX'),
          Text(0, 0, 'JVMChat'),
          Text(0, 0, 'oscars2016'),
          Text(0, 0, 'happycustomer'),
          Text(0, 0, 'southwestairlines'),
          Text(0, 0, 'SWA'),
          Text(0, 0, '85832')]
```



Analysis of Frequency distribution of Hashtags in Positive Tweets

Despite the large number of tweets with positive hashtags, the hashtags themselves were split across a large variety of hashtags, with hashtag `DestinationDragons` only appearing 19 times in 699 hashtags in 435 tweets. The hashtag `DestinationDragons` refers to the promotional event hosted by Southwest Airlines for the Imagine Dragons tour: http://destinationdragons.com/ (http://destinationdragons.com/). However, upon closer inspection, there is the possibility of another data quality issue, namely that the there were tweets with the same letters, but different case; examples include `jetblue` vs `JetBlue` , and `FlyFi` vs `flyfi` . This could possibly explain why the distribution of hashtags was seemingly distributed so equally.

Neutral tweets with Hashtags

In [46]:
```
#create a list of tokens from the tweets
list_of_hashtags_neutral = []
df_tweets_neutral_hashtags = df_tweets_neutral.loc[df_tweets_neutral['hashtags
_flag'] == True]
```

Number of tweets of neutral sentiment that contained hashtags:

In [47]:
```
len(df_tweets_neutral_hashtags)
```

Out[47]: 410

In [48]:
```
#create a list
neutral_hashtags = []

for i, row in df_tweets_neutral_hashtags.iterrows():
    tweet_hashtags = df_tweets_neutral_hashtags.at[i, 'hashtags']
    #print(type(tweet_hashtags))

    tweet_hashtags_list = list(tweet_hashtags.split(","))

    for hashtag in tweet_hashtags_list:
        #print('hashtag: ' + hashtag)

        #strip brackets, quote, and spaces
        hashtag = hashtag.strip('[]')
        hashtag = hashtag.replace("\'", "")
        hashtag = hashtag.strip()

        #print('hashtag_strip: ' + hashtag)
        neutral_hashtags.append(hashtag)
```

Number of hashtags used in neutral tweets:

In [49]:
```
len(neutral_hashtags)
```

Out[49]: 653

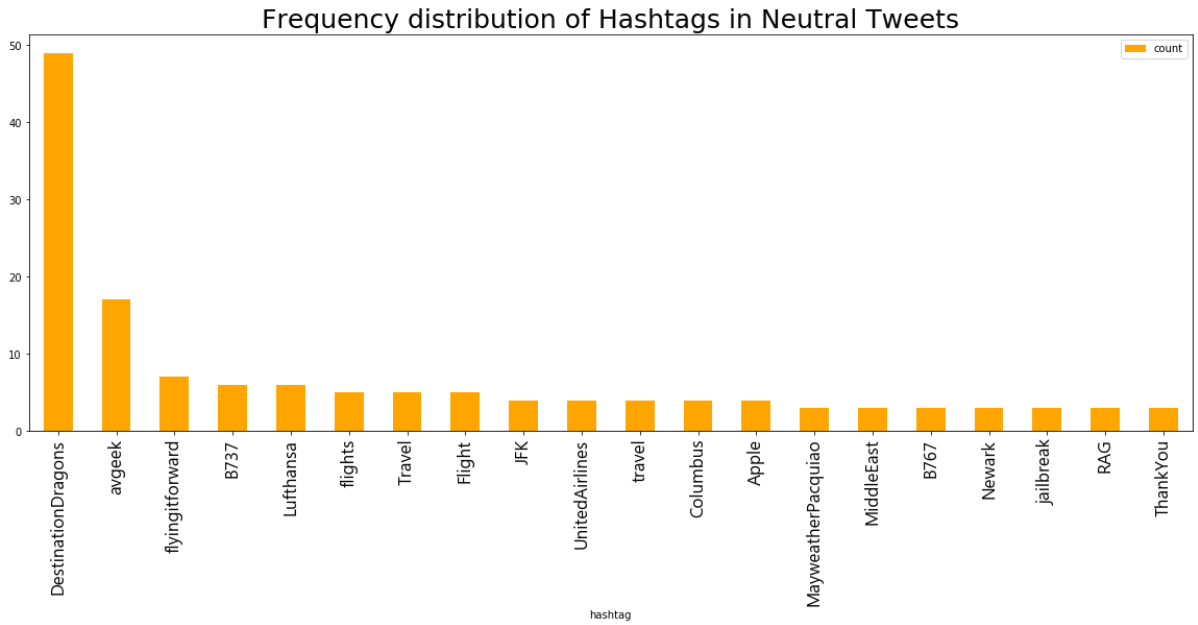
In [50]:
```
neutral_hashtags_counter = Counter(neutral_hashtags)
neutral_hashtags_counter.most_common(20)
```

Out[50]:
```
[('DestinationDragons', 49),
 ('avgeek', 17),
 ('flyingitforward', 7),
 ('B737', 6),
 ('Lufthansa', 6),
 ('flights', 5),
 ('Travel', 5),
 ('Flight', 5),
 ('JFK', 4),
 ('UnitedAirlines', 4),
 ('travel', 4),
 ('Columbus', 4),
 ('Apple', 4),
 ('MayweatherPacquiao', 3),
 ('MiddleEast', 3),
 ('B767', 3),
 ('Newark', 3),
 ('jailbreak', 3),
 ('RAG', 3),
 ('ThankYou', 3)]
```

In [51]:
```
# convert list of tuples into data frame
freq_df_neutral_hashtags = pd.DataFrame.from_records(neutral_hashtags_counter.
most_common(20),
                                     columns=['hashtag', 'count'])

# create bar plot
distribution_bar_neutral_hashtags = freq_df_neutral_hashtags.plot(
    kind='bar',
    x='hashtag',
    color='orange',
    figsize=(20,7)
)
distribution_bar_neutral_hashtags.set_title(
    'Frequency distribution of Hashtags in Neutral Tweets',
    fontsize=25
)
distribution_bar_neutral_hashtags.set_xticklabels(
    freq_df_neutral_hashtags.hashtag.tolist(),
    rotation = 90,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

```
Out[51]: [Text(0, 0, 'DestinationDragons'),
          Text(0, 0, 'avgeek'),
          Text(0, 0, 'flyingitforward'),
          Text(0, 0, 'B737'),
          Text(0, 0, 'Lufthansa'),
          Text(0, 0, 'flights'),
          Text(0, 0, 'Travel'),
          Text(0, 0, 'Flight'),
          Text(0, 0, 'JFK'),
          Text(0, 0, 'UnitedAirlines'),
          Text(0, 0, 'travel'),
          Text(0, 0, 'Columbus'),
          Text(0, 0, 'Apple'),
          Text(0, 0, 'MayweatherPacquiao'),
          Text(0, 0, 'MiddleEast'),
          Text(0, 0, 'B767'),
          Text(0, 0, 'Newark'),
          Text(0, 0, 'jailbreak'),
          Text(0, 0, 'RAG'),
          Text(0, 0, 'ThankYou')]
```



Analysis of Frequency distribution of Hashtags in Neutral Tweets

For some reason, neutral tweets featured `DestinationDragons` more frequently at 49 occurences, compared to positive tweets featuring the same hashtag. Unfortunately, the data quality issue with hashtags mentioned earlier with positive tweets likely has affected neutral tweets as well: `travel` vs `Travel`, `flights` vs `Flight` (although this is also a case of not lemmatizing the hashtags). Something that is curious is that neutral tweets had hashtags that referenced a specific airplane model, such as `B737` and `B767`, which I find personally quite odd.

Negative tweets with Hashtags

In [52]:
```
#create a list of tokens from the tweets
list_of_hashtags_negative = []
df_tweets_negative_hashtags = df_tweets_negative.loc[df_tweets_negative['hasht
ags_flag'] == True]
```

Number of tweets of negative sentiment that contained hashtags:

In [53]:
```
len(df_tweets_negative_hashtags)
```

Out[53]: 1523

In [54]:
```
#create a list
negative_hashtags = []

for i, row in df_tweets_negative_hashtags.iterrows():
    tweet_hashtags = df_tweets_negative_hashtags.at[i, 'hashtags']
    #print(type(tweet_hashtags))

    tweet_hashtags_list = list(tweet_hashtags.split(","))

    for hashtag in tweet_hashtags_list:
        #print('hashtag: ' + hashtag)

        #strip brackets, quote, and spaces
        hashtag = hashtag.strip('[]')
        hashtag = hashtag.replace("\'", "")
        hashtag = hashtag.strip()

        #print('hashtag_strip: ' + hashtag)
        negative_hashtags.append(hashtag)
```

Number of hashtags used in negative tweets:

In [55]:
```
len(negative_hashtags)
```

Out[55]: 2140

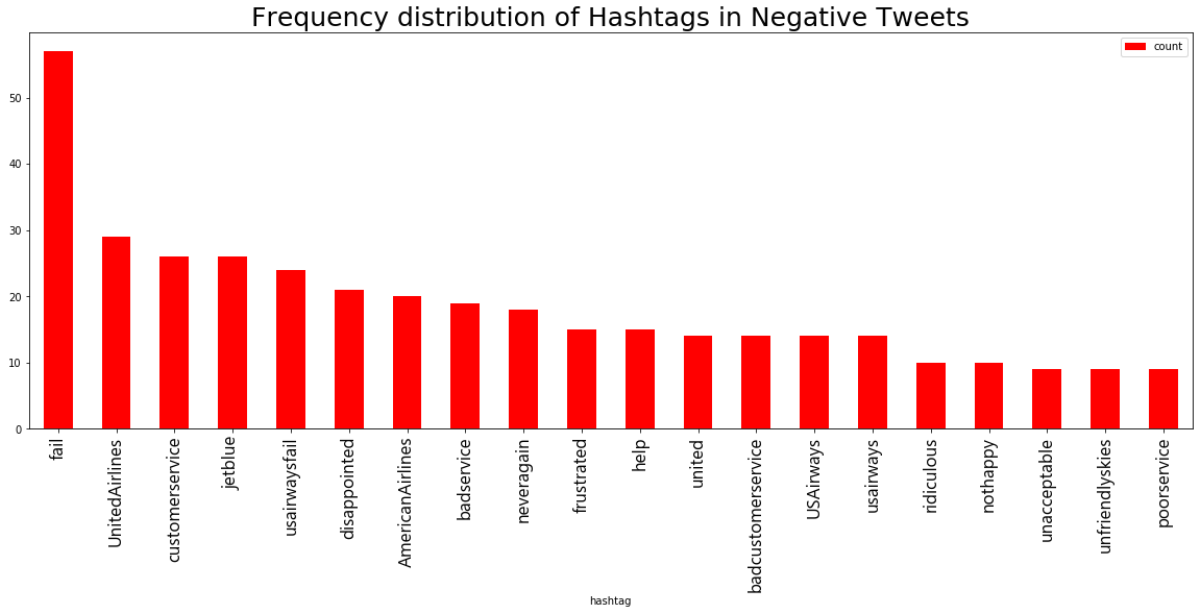
In [56]:
```
negative_hashtags_counter = Counter(negative_hashtags)
negative_hashtags_counter.most_common(20)
```

Out[56]:
```
[('fail', 57),
 ('UnitedAirlines', 29),
 ('customerservice', 26),
 ('jetblue', 26),
 ('usairwaysfail', 24),
 ('disappointed', 21),
 ('AmericanAirlines', 20),
 ('badservice', 19),
 ('neveragain', 18),
 ('frustrated', 15),
 ('help', 15),
 ('united', 14),
 ('badcustomerservice', 14),
 ('USAirways', 14),
 ('usairways', 14),
 ('ridiculous', 10),
 ('nothappy', 10),
 ('unacceptable', 9),
 ('unfriendlyskies', 9),
 ('poorservice', 9)]
```

In [57]:
```
# convert list of tuples into data frame
freq_df_negative_hashtags = pd.DataFrame.from_records(negative_hashtags_counte
r.most_common(20),
                                      columns=['hashtag', 'count'])

# create bar plot
distribution_bar_negative_hashtags = freq_df_negative_hashtags.plot(
    kind='bar',
    x='hashtag',
    color='red',
    figsize=(20,7)
)
distribution_bar_negative_hashtags.set_title(
    'Frequency distribution of Hashtags in Negative Tweets',
    fontsize=25
)
distribution_bar_negative_hashtags.set_xticklabels(
    freq_df_negative_hashtags.hashtag.tolist(),
    rotation = 90,
    fontsize = 17,
    #fontproperties=prop
    fontname='Segoe UI Emoji'
)
```

Out[57]:
```
[Text(0, 0, 'fail'),
 Text(0, 0, 'UnitedAirlines'),
 Text(0, 0, 'customerservice'),
 Text(0, 0, 'jetblue'),
 Text(0, 0, 'usairwaysfail'),
 Text(0, 0, 'disappointed'),
 Text(0, 0, 'AmericanAirlines'),
 Text(0, 0, 'badservice'),
 Text(0, 0, 'neveragain'),
 Text(0, 0, 'frustrated'),
 Text(0, 0, 'help'),
 Text(0, 0, 'united'),
 Text(0, 0, 'badcustomerservice'),
 Text(0, 0, 'USAirways'),
 Text(0, 0, 'usairways'),
 Text(0, 0, 'ridiculous'),
 Text(0, 0, 'nothappy'),
 Text(0, 0, 'unacceptable'),
 Text(0, 0, 'unfriendlyskies'),
 Text(0, 0, 'poorservice')]
```



Analysis of Frequency distribution of Hashtags in Negative Tweets

The number of negative tweets with hashtags was approximately double the number of positive and negative tweets combined. This itself is interesting because it seems to imply a casuality that users would use hashtags more often to express negative sentiments. There is some credence to this theory, despite the warnings of the oft-repeated phrase "Correlation does not imply causation", as users who are likely upset or sad with the airlines will want to raise this issue as quickly as possible in social media so that they are likelier to be rewarded/compensated; in Twitter, a likely way of getting a viral tweet is using as many hashtags as possible so it appears more often in multiple user's Twitter feeds. Nevertheless, this is just a theory: what was suggested is just a hypothesis, and for a hypothesis to be proven statistically significant, it must be tested.

The `fail` hashtag was used quite frequently, occuring a total of 57, but like the other most frequent hashtags in neutral and positive sentiments, it still is a small percentage of the total number of hashtags in negative tweets. Like the other positive and neutral tweets, the distribution of hashtags in negative tweets is affected by the same data quality issue: `USAirways` vs `usairways`. On top of that, another issue we can see from the frequency distribution that the same airline companies have different hashtags due to different naming conventions (for example, `united` vs `UnitedAirlines`). Solving this data quality issue will be quite challenging. Despite the aforementioned data quality issues, there are some positives to be taken away. Compared to the positive and neutral tweets, tt appears that many negative tweets contained the hashtags of the name of the airline company. Also, negative tweets also contained words normally associated with negative sentiment, such as `disappointed`, `frustrated`, `ridiculous`, `nothappy`, etc.

Final Thoughts

As discussed earlier, there were several data quality issues, particularly with the extraction of emoticons and hashtags. The issue with the extraction of emoticons is caused by a step in the cleaning process that confuses the emoticon with what is actually a non-emoticon string of characters. For example, the time and airports are misrepresented as emoticons. The data quality with the hashtags is one that consists of several interlinked problems ranging from simple to complex. These problems include making the hashtags to a single case (preferrably lower), reducing the words in the hashtag to its roots (via lemmatization, for example), and figuring out a way to group different hashtags that have the same entity, but different word.

Despite the aforementioned drawbacks, there are some positives to be gained from this analysis. There were some visible patterns in the counts and frequency distributions of emojis, emoticons, and hashtags. Some emojis and emoticons correlated well with positive or negative sentiments, with a little overlap between those sentiments with the neutral sentiment. Negative hashtags showed a clear pattern where many of the top frequent hashtags were either airline entities, or words commonly related to negative sentiment or emotion. These patterns show that emojis, emoticons, and hashtags may provide some value in classifying sentiment.

There are some additional explorations and analysis one could do further with emojis, emoticons, and hashtags:

- Clustering emojis and emoticons (e.g., see https://hal-amu.archives-ouvertes.fr/hal-01871045/document (https://hal-amu.archives-ouvertes.fr/hal-01871045/document)). A lot of emojis and emoticons share the same sentiment.
- Clustering could also work for hashtags, but is would be harder to implement because the lexicon of English words is significantly more than the lexicon of emojis/emoticons. It is not enough to just calculate string similarity between the hashtags as well using Levenshtein distance , since some entities have wildly different words.
- Exploring the median/average number of emojis/emoticons/hashtags. In particular, it appears that negative tweets seem to have more hashtags than positive or neutral tweets. We could potentially explore the number of hashtags, or even the length of a tweet to see if there is a statistically signficant correlation between them and sentiment.

In [ ]: