sentiment analysis

June 27, 2023

[5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import string
import nltk
import warnings
%matplotlib inline
warnings.filterwarnings('ignore')
```

[6]:
```
df = pd.read_csv('Tweets.csv')
df.head()
```

[6]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t…	NaN
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0		
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0		
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0		

```
2  @VirginAmerica I didn't today… Must mean I n…         NaN
3  @VirginAmerica it's really aggressive to blast…         NaN
4  @VirginAmerica and it's a really big bad thing…         NaN

              tweet_created tweet_location               user_timezone
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)
```

[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   tweet_id                      14640 non-null  int64
 1   airline_sentiment             14640 non-null  object
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object
 6   airline_sentiment_gold        40 non-null     object
 7   name                          14640 non-null  object
 8   negativereason_gold           32 non-null     object
 9   retweet_count                 14640 non-null  int64
 10  text                          14640 non-null  object
 11  tweet_coord                   1019 non-null   object
 12  tweet_created                 14640 non-null  object
 13  tweet_location                9907 non-null   object
 14  user_timezone                 9820 non-null   object
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

[8]:
```
def  remove_pattern(input_txt,pattern):
    r = re.findall(pattern,input_txt)
    for word in r:
        input_txt = re.sub(word,"",input_txt)
    return input_txt
```

[9]:
```
df['clean_text']=np.vectorize(remove_pattern)(df['text'],"@[\w]*]")
```

[10]:
```
df.head()
```

[10]:
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
```

```
1  570301130888122368          positive                       0.3486
2  570301083672813571           neutral                       0.6837
3  570301031407624196          negative                       1.0000
4  570300817074462722          negative                       1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                         NaN  Virgin America
1            NaN                      0.0000  Virgin America
2            NaN                         NaN  Virgin America
3     Bad Flight                      0.7033  Virgin America
4     Can't Tell                      1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                                text tweet_coord  \
0                @VirginAmerica What @dhepburn said.         NaN
1  @VirginAmerica plus you've added commercials t…         NaN
2  @VirginAmerica I didn't today… Must mean I n…         NaN
3  @VirginAmerica it's really aggressive to blast…         NaN
4  @VirginAmerica and it's a really big bad thing…         NaN

              tweet_created tweet_location               user_timezone  \
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)

                                      clean_text
0                @VirginAmerica What @dhepburn said.
1  @VirginAmerica plus you've added commercials t…
2  @VirginAmerica I didn't today… Must mean I n…
3  @VirginAmerica it's really aggressive to blast…
4  @VirginAmerica and it's a really big bad thing…
```

[11]:
```
df['text']=df['text'].str.replace("[^a-zA-Z#]"," ")
df.head()
```

[11]:
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
```

```
3  570301031407624196          negative                         1.0000
4  570300817074462722          negative                         1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                              text tweet_coord  \
0                VirginAmerica What  dhepburn said          NaN
1  VirginAmerica plus you ve added commercials t…         NaN
2  VirginAmerica I didn t today    Must mean I n…         NaN
3  VirginAmerica it s really aggressive to blast…         NaN
4  VirginAmerica and it s a really big bad thing…         NaN

              tweet_created tweet_location               user_timezone  \
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)

                                      clean_text
0               @VirginAmerica What @dhepburn said.
1  @VirginAmerica plus you've added commercials t…
2  @VirginAmerica I didn't today… Must mean I n…
3  @VirginAmerica it's really aggressive to blast…
4  @VirginAmerica and it's a really big bad thing…
```

[12]:
```
df['text']=df['text'].apply(lambda x:" ".join([w for w in x.split() if␣
 ↪len(w)>3]))
df.head()
```

[12]:
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
```

```
4  570300817074462722          negative                       1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                              text tweet_coord  \
0                  VirginAmerica What dhepburn said         NaN
1  VirginAmerica plus added commercials experienc…         NaN
2  VirginAmerica didn today Must mean need take a…         NaN
3  VirginAmerica really aggressive blast obnoxiou…         NaN
4                  VirginAmerica really thing about         NaN

              tweet_created tweet_location               user_timezone  \
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)

                                        clean_text
0                @VirginAmerica What @dhepburn said.
1  @VirginAmerica plus you've added commercials t…
2  @VirginAmerica I didn't today… Must mean I n…
3  @VirginAmerica it's really aggressive to blast…
4  @VirginAmerica and it's a really big bad thing…
```

[13]:
```
tokenized_text=df['text'].apply(lambda x:x.split())
tokenized_text.head()
```

[13]:
```
0                [VirginAmerica, What, dhepburn, said]
1    [VirginAmerica, plus, added, commercials, expe…
2    [VirginAmerica, didn, today, Must, mean, need,…
3    [VirginAmerica, really, aggressive, blast, obn…
4                [VirginAmerica, really, thing, about]
Name: text, dtype: object
```

[14]:
```
from nltk.stem.porter import PorterStemmer
stemmer = PorterStemmer()
tokenized_text = tokenized_text.apply(lambda sentence: [stemmer.stem(word) for␣
 ↪word in sentence])
tokenized_text.head()
```

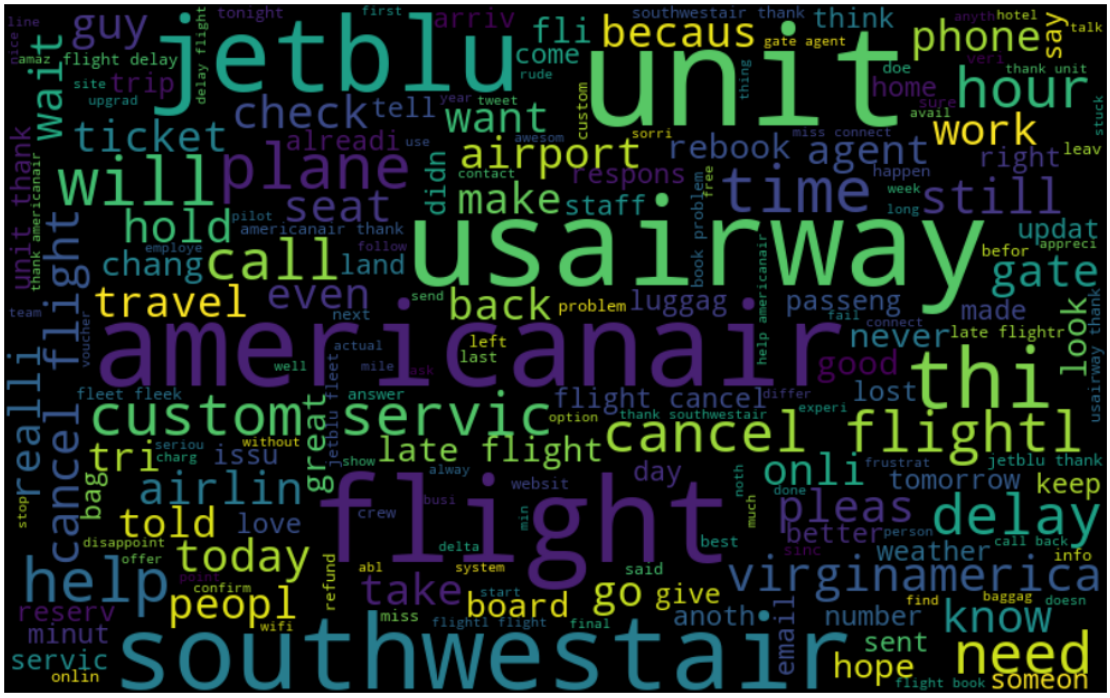
[14]:
```
0                [virginamerica, what, dhepburn, said]
1     [virginamerica, plu, ad, commerci, experi, tacki]
2     [virginamerica, didn, today, must, mean, need,…
3     [virginamerica, realli, aggress, blast, obnoxi…
4                [virginamerica, realli, thing, about]
Name: text, dtype: object
```

[15]:
```
for i in range(len(tokenized_text)):
    tokenized_text[i]=" ".join(tokenized_text[i])
df['clean_text']= tokenized_text
df.head()
```

[15]:
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                              text tweet_coord  \
0                  VirginAmerica What dhepburn said         NaN
1  VirginAmerica plus added commercials experienc…         NaN
2  VirginAmerica didn today Must mean need take a…         NaN
3  VirginAmerica really aggressive blast obnoxiou…         NaN
4                  VirginAmerica really thing about         NaN

               tweet_created tweet_location               user_timezone  \
```

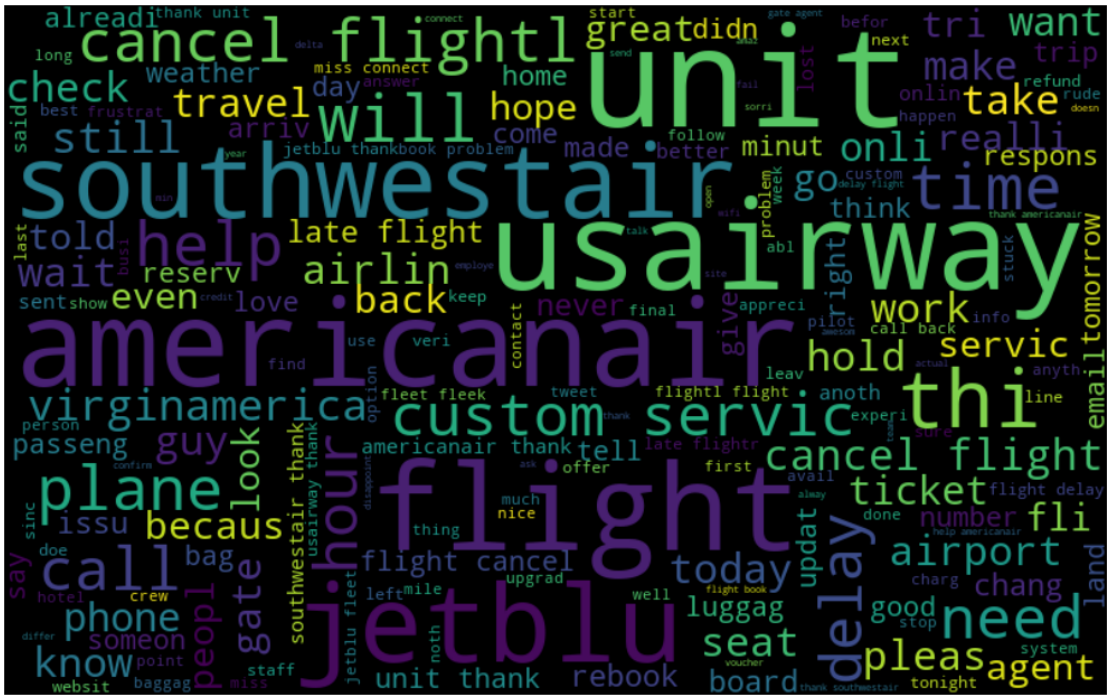
```
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)

                                      clean_text
0                   virginamerica what dhepburn said
1         virginamerica plu ad commerci experi tacki
2  virginamerica didn today must mean need take a…
3  virginamerica realli aggress blast obnoxi ente…
4                   virginamerica realli thing about
```

[16]:
```
all_words = " ".join([sentence for sentence in df['clean_text']])
from wordcloud import WordCloud
wordcloud = WordCloud(width=800,height=500,random_state=42,max_font_size=100).
 ↪generate(all_words)
plt.figure(figsize=(15,8))
plt.imshow(wordcloud, interpolation ='bilinear')
plt.axis('off')
plt.show()
```



[18]:
```
all_words = " ".join([sentence for sentence in␣
 ↪df['clean_text'][df['retweet_count']==0]])
```

```
wordcloud = WordCloud(width=800,height=500,random_state=42,max_font_size=100).
 ↪generate(all_words)
plt.figure(figsize=(15,8))
plt.imshow(wordcloud, interpolation ='bilinear')
plt.axis('off')
plt.show()
```



[19]:
```
def hashtag_extract(tweet):
    hashtags = []
    for word in tweet:
        ht = re.findall(r"#(\w+)", word)
        hashtags.append(ht)
    return hashtags
```

[20]:
```
ht_positive = hashtag_extract(df['clean_text'][df['retweet_count']==0])
ht_negative = hashtag_extract(df['clean_text'][df['retweet_count']==1])
```

[21]:
```
ht_positive[:5]
```

[21]:
```
[[], [], [], [], []]
```

[22]:
```
ht_positive = sum(ht_positive, [])
ht_negative = sum(ht_negative, [])
```

[23]:
```
ht_positive[:5]
```

[23]:
```
['fabul', 'seduct', 'stress', 'fail', 'noair']
```

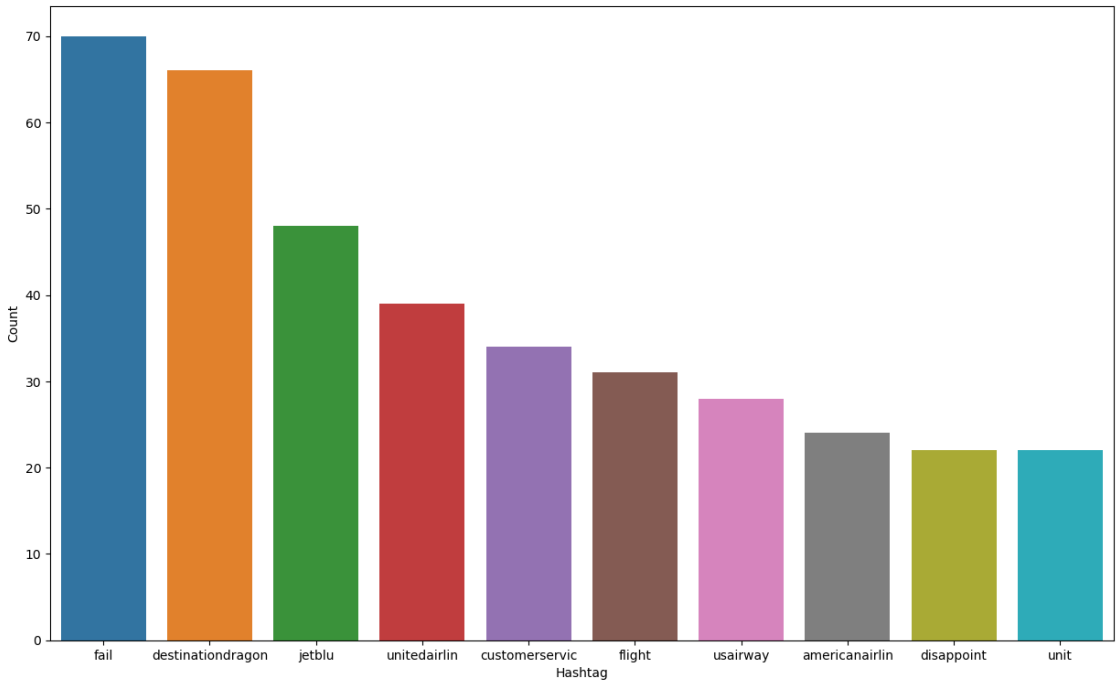
[24]:
```
freq=nltk.FreqDist(ht_positive)
d = pd.DataFrame({'Hashtag': list(freq.keys()),
                  'Count':list(freq.values())})
d.head()
```

[24]:

	Hashtag	Count
0	fabul	1
1	seduct	1
2	stress	1
3	fail	70
4	noair	1

[25]:
```
d = d.nlargest(columns='Count',n=10)
plt.figure(figsize=(15,9))
sns.barplot(data=d, x='Hashtag', y='Count')
```

[25]:
```
<AxesSubplot:xlabel='Hashtag', ylabel='Count'>
```



[26]:
```
freq=nltk.FreqDist(ht_negative)
d = pd.DataFrame({'Hashtag': list(freq.keys()),
                  'Count':list(freq.values())})
```

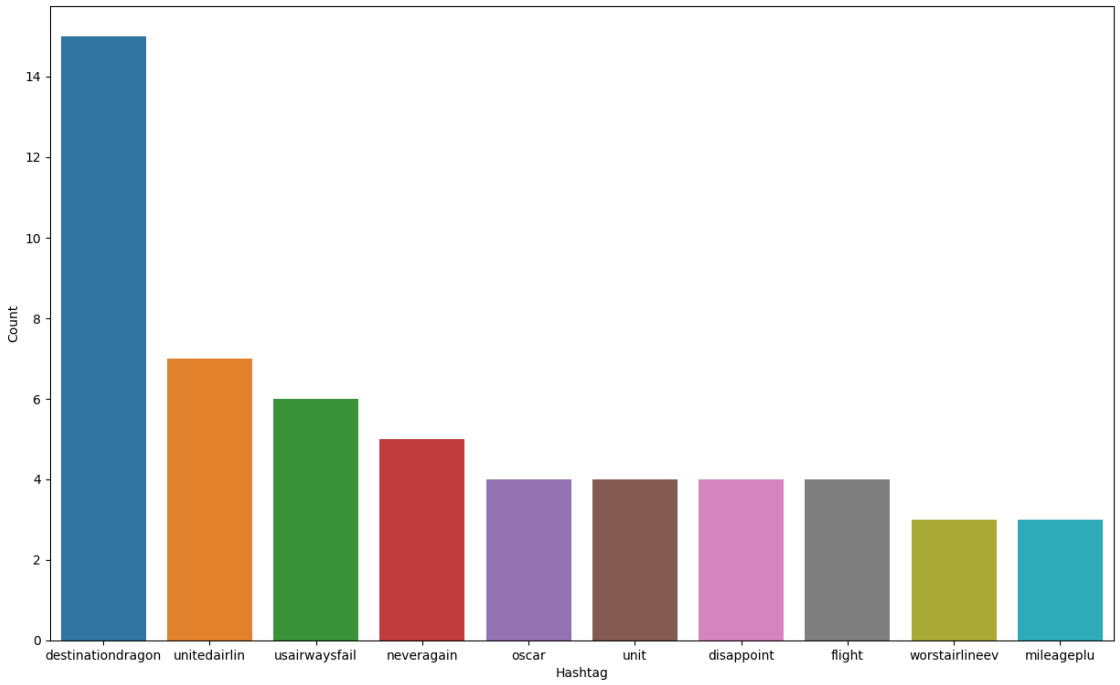
```
d.head()
```

[26]:

	Hashtag	Count
0	oscar	4
1	redcarpet	1
2	sweet	1
3	ourprincess	2
4	freyasfund	1

[27]:
```
d = d.nlargest(columns='Count',n=10)
plt.figure(figsize=(15,9))
sns.barplot(data=d, x='Hashtag', y='Count')
```

[27]: <AxesSubplot:xlabel='Hashtag', ylabel='Count'>



[28]:
```
from sklearn.feature_extraction.text import CountVectorizer
bow_vectorizer = CountVectorizer(max_df=0.90,min_df=2,␣
 ↪max_features=1000,stop_words='english')
bow=bow_vectorizer.fit_transform(df['clean_text'])
```

[29]:
```
#bow[0].toarray()
```

[30]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train ,y_test = train_test_split(bow,df['retweet_count'],␣
 ↪random_state=42, test_size =0.25)
```

[31]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score,accuracy_score
```

[32]:
```
model = LogisticRegression()
model.fit(x_train,y_train)
```

[32]:
```
LogisticRegression()
```

[38]:
```
accuracy_score(y_test,pred)
```

[38]:
```
0.940983606557377
```

[ ]:
```

```