Chapter 10: Spark Streaming

Ex1: Pre-processing Data from Tweets

Requirement:

- Read data from file (Tweets)
- Pre-process data
- Save data after pre-processing to new file.

In [1]:
```
# pip install textblob
import csv
from textblob import TextBlob
```

In [2]:
```
import pandas as pd
```

In [3]:
```
tweetdata = 'tweets_christmas.txt'
sentences = []
sentiment_polarity = []
sentiment_subjectivity = []
```

In [4]:
```
with open(tweetdata, 'r') as csvfile:
    rows = csv.reader(csvfile)
    for row in rows:
        sentence = row[0]
        blob = TextBlob(sentence)
        if ("Error on_data" not in sentence):
            print (sentence)
            print (blob.sentiment.polarity, blob.sentiment.subjectivity)
            sentences.append(sentence)
            sentiment_polarity.append(blob.sentiment.polarity)
            sentiment_subjectivity.append(blob.sentiment.subjectivity)
```

```
b'Christmas is here \xf0\x9f\x8e\x84\xf0\x9f\x8e\x81\xf0\x9f\x8e\x84 Beautifu
l Christmas tree \xf0\x9f\x8e\x84'
0.85 1.0
b'RT @VrglSullano: All I want for Christmas is cashhhhhhhhhhhhhhhhhhhhhhhhhhh
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh\xe2\x80\xa6'
0.0 0.0
b'RT @_mymusictaste: \xf0\x9f\x93\xa3ATINY! Surprise! \xf0\x9f\x8e\x86\nChris
tmas is starting earlier this year! \xf0\x9f\x8e\x81\n\nWelcome ATEEZ GLOBAL
 FANSIGN EVENT IN LOS ANGELES. \xf0\x9f\x8e\xb6\nSta\xe2\x80\xa6'
0.0 0.19999999999999998
b'RT @GrowLevel: \xf0\x9f\x8e\x84Merry Christmas\xf0\x9f\x8e\x84\n\xe3\x83\xa
1\xe3\x83\xad\xe3\x83\xb3\xe3\x82\xb1\xe3\x83\xbc\xe3\x82\xad\xe3\x82\x92\xe6
\x8a\xbd\xe9\x81\xb8\xe3\x81\xa71\xe5\x90\x8d\xe6\xa7\x98\xe3\x81\xab\xe3\x83
\x97\xe3\x83\xac\xe3\x82\xbc\xe3\x83\xb3\xe3\x83\x88\xf0\x9f\x8e\x81\xf0\x9f
\x8d\x88\n\n\xe5\x95\x86\xe5\x93\x81\xe3\x81\xaf\xe3\x82\xaf\xe3\x83\xbc\xe3
\x83\xab\xe4\xbe\xbf\xe3\x81\xa7\xe3\x81\x8a\xe5\xb1\x8a\xe3\x81\x91\xe3\x81
\x97\xe3\x81\xbe\xe3\x81\x99\xe3\x80\x82\n\xe2\x80\xbb\xe5\xa4\xa7\xe9\x98\xa
a\xe3\x81\x8b\xe3\x82\x892\xe6\x97\xa5\xe4\xbb\xa5\xe4\xb8\x8a\xe3\x81\x8b\xe
3\x81\x8b\xe3\x82\x8b\xe5\x9c\xb0\xe5\x9f\x9f\xe3\x81\xaf\xe4\xb8\x8d\xe5\x8f
\xaf\xe3\x80\x82\n\n\xe5\xbf\x9c\xe5\x8b\x9f\xe7\xb7\xa0\xe5\x88\x87\xe3\x82
```

In [5]:
```
data = pd.DataFrame({"sentence": sentences,
                     "sentiment_polarity":sentiment_polarity,
                     "sentiment_subjectivity":sentiment_subjectivity
                    })
```

In [6]:
```
data = data.drop([0, 1])
```

In [7]:
```
data.sentence = data.sentence.str.replace("b'", "")
```

In [8]:
```
data.head()
```

Out[8]:

	sentence	sentiment_polarity	sentiment_subjectivity
2	Decorating Sebastians Grave| Christmas 2019 Vl...	0.0	0.0
3	RT @_mymusictaste: \xf0\x9f\x93\xa3ATINY! Surp...	0.0	0.2
4	Christmas partyyyy'	0.0	0.0
5	Working on a maaaybe Christmas-y themed piece....	0.5	0.6
6	Coastes shenanigans pre-Christmas celebrations...	0.0	0.0

In [9]:
```
data.to_csv("tweets_christmas.csv")
```

Another solution: Build function to read txt file and convert to csv file

In [10]:
```
def read_and_pre_pro(file_in, file_out):
    sentences = []
    sentiment_polarity = []
    sentiment_subjectivity = []
    with open(file_in, 'r') as csvfile:
        rows = csv.reader(csvfile)
        for row in rows:
            sentence = row[0]
            blob = TextBlob(sentence)
            if ("Error on_data" not in sentence):
                #print (sentence)
                #print (blob.sentiment.polarity, blob.sentiment.subjectivity)
                sentences.append(sentence)
                sentiment_polarity.append(blob.sentiment.polarity)
                sentiment_subjectivity.append(blob.sentiment.subjectivity)
        data = pd.DataFrame({"sentence": sentences,
                             "sentiment_polarity":sentiment_polarity,
                             "sentiment_subjectivity":sentiment_subjectivity
                            })
        data.sentence = data.sentence.str.replace("b'", "")

        data.to_csv(file_out)
```

In [11]:
```
file_in = "tweets_football.txt"
file_out = "tweets_football.csv"
read_and_pre_pro(file_in, file_out)
```

In [12]:
```
df = pd.read_csv("tweets_football.csv", index_col=0)
```

In [13]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1333 entries, 0 to 1332
Data columns (total 3 columns):
sentence                  1333 non-null object
sentiment_polarity        1333 non-null float64
sentiment_subjectivity    1333 non-null float64
dtypes: float64(2), object(1)
memory usage: 41.7+ KB
```

In [14]:
```
df.head()
```

Out[14]:

	sentence	sentiment_polarity	sentiment_subjectivity
0	Listening on port: 5555	0.00	0.0
1	Received request from: ('127.0.0.1'	-0.75	1.0
2	Listening on port: 5555	0.00	0.0
3	Received request from: ('127.0.0.1'	-0.75	1.0
4	EVA SOCCER memenuhi keperluan football dan fut...	0.00	0.0

In [15]:
```
indexNames = df[df['sentence'].str.contains("Listening on port")].index
# Delete these row indexes from dataFrame
df = df.drop(indexNames)
```

In [16]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1331 entries, 1 to 1332
Data columns (total 3 columns):
sentence                1331 non-null object
sentiment_polarity      1331 non-null float64
sentiment_subjectivity  1331 non-null float64
dtypes: float64(2), object(1)
memory usage: 41.6+ KB
```

In [17]:
```
df.head()
```

Out[17]:

	sentence	sentiment_polarity	sentiment_subjectivity
1	Received request from: ('127.0.0.1'	-0.75	1.000000
3	Received request from: ('127.0.0.1'	-0.75	1.000000
4	EVA SOCCER memenuhi keperluan football dan fut...	0.00	0.000000
5	RT @AndrewMLind: Per ESPN	0.00	0.000000
6	RT @ANNMediaSports: USC football Head coach Cl...	0.00	0.333333