In [1]:
```
import pandas as pd
import re
import numpy as np
import scipy
import itertools
import matplotlib
import matplotlib.pyplot as plt
from scipy.spatial.distance import cdist
from collections import Counter
from random import choice
```

In [2]:
```
# Storing the training and test datasets into their respective dataframes
trained = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
```

In [3]:
```
# Parsing the stop_words.txt file and storing all the words in a list.
stopwords = []
with open('stop_words.txt','r') as file:
    for line in file:
        for word in line.split():
            stopwords.append(word)
```

In [4]:
```
# Removing all stopwords from all the tweets in training data.
trained["Tweet"] = trained["Tweet"].apply(lambda func: ' '.join(sw
                                          for sw in func.split()
                                          if sw not in stopwords))

# Removing all stopwords from all the tweets in test data.
test["Tweet"] = test["Tweet"].apply(lambda func: ' '.join(sw
                                          for sw in func.split()
                                          if sw not in stopwords))
trained.head()
```

Out[4]:

| | Sentiment | Tweet |
|---|---|---|
| 0 | neutral | @united 877 amsterdam ewr, 02.27.2015, 737-300. |
| 1 | negative | @united IT-problems link? #3thparty |
| 2 | positive | @united -today staff @ MSP took customer servi... |
| 3 | negative | @AmericanAir yet receive assistance one agents... |
| 4 | negative | @SouthwestAir let change reservation online I'... |

In [5]:
```
#Training Data
train_unique = (list(set(trained['Tweet'].str.findall("\w+").sum()))) # Finding d
train_unique_words = len(train_unique)

#Test Data
test_unique = (list(set(test['Tweet'].str.findall("\w+").sum()))) # Finding all t
test_unique_words = len(test_unique)

print("Unique words in Training Data: {}".format(train_unique_words))
print("Unique words in Test Data: {}".format(test_unique_words))
```

```
Unique words in Training Data: 15823
Unique words in Test Data: 6973
```

In [6]:
```
#List of all special characters that are to be removed.
special_chars = ["!",'"',"%","&","amp","'","(",")", "*","+",",","-",".",
                 "/",":",";","<","=",">","?","[","\\","]","^","_",
                 "`","{","|","}","~","–","@","#","$"]
```

In [7]:
```
#Training Data
trained['Tweet'] = trained['Tweet'].str.replace(r'http?://[^\s<>"]+|www\.[^\s<>"]
trained['Tweet'] = trained['Tweet'].str.replace('@[A-Za-z0-9]+', '') # Removing u
trained['Tweet'] = trained['Tweet'].str.replace(r'\B#\w*[a-zA-Z]+\w*', '') # Remo
trained['Tweet'] = trained['Tweet'].str.replace('\d+', '') # Removing numbers fro

for c in special_chars:
    trained['Tweet'] = trained['Tweet'].str.replace(c,'') # Removing all special

#Test Data
test['Tweet'] = test['Tweet'].str.replace(r'http?://[^\s<>"]+|www\.[^\s<>"]+', ''
test['Tweet'] = test['Tweet'].str.replace('@[A-Za-z0-9]+', '') # Removing usernam
test['Tweet'] = test['Tweet'].str.replace(r'\B#\w*[a-zA-Z]+\w*', '') # Removing h
test['Tweet'] = test['Tweet'].str.replace('\d+', '') # Removing numbers from all

for c in special_chars:
    test['Tweet'] = test['Tweet'].str.replace(c,'') # Removing all special charac
```

...

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

Out[8]:

| | Sentiment | Tweet |
|---|---|---|
| 0 | neutral | amsterdam ewr |
| 1 | negative | ITproblems link |
| 2 | positive | today staff MSP took customer service new le... |
| 3 | negative | yet receive assistance one agents securing ne... |
| 4 | negative | let change reservation online Im wasting time |

In [9]:
```
test.head()
```

Out[9]:

| | Sentiment | Tweet |
|---|---|---|
| 0 | neutral | jump DallasAustin market News |
| 1 | positive | Chicago seen seat A AA So far great ride On ... |
| 2 | negative | need bag bouncer Get together |
| 3 | negative | Hey Jetblue stranded entire plane supposed go... |
| 4 | negative | Big fail curbside baggage Pittsburgh charge ... |

In [10]:
```
#Training Data
train_unique = (list(set(trained['Tweet'].str.findall("\w+").sum()))) # Finding a
train_unique_words = len(train_unique)

#Test Data
test_unique = (list(set(test['Tweet'].str.findall("\w+").sum()))) # Finding all t
test_unique_words = len(test_unique)

print("Unique words in Training Data: {}".format(train_unique_words))
print("Unique words in Test Data: {}".format(test_unique_words))
```

```
Unique words in Training Data: 12416
Unique words in Test Data: 5814
```

In [12]:
```
trained.to_csv('train_clean.csv')
test.to_csv('test_clean.csv')
```

In [ ]:
```
```