#CHAPTER 6 - HANDLING TEXT

**#6.1 - Cleaning Text**

#create text

text\_data = [" Interrobang. By Aishwarya Henriette ",

"Parking and Going. By Karl Gautier",

" Today Is The night. By Jarek Prakash "]

print(text\_data)

print("original data")

#strip whitespaces

strip\_whitespace = [string.strip() for string in text\_data]

print(strip\_whitespace)

print("white spaces is gone")

#remove periods

remove\_periods = [string.replace(".", "") for string in strip\_whitespace]

print(remove\_periods)

print("periods removed")

**#6.9 Weighting Word Importance - be sure to read the discussion**

import numpy as np

from sklearn.feature\_extraction.text import TfidfVectorizer

text\_data = np.array(['I love Brazil. Brazil!]',

'Sweden is best.',

'Germany beats both'])

#create the tf.idf feture matrix

tfidf=TfidfVectorizer()

feature\_matrix = tfidf.fit\_transform(text\_data)

print(feature\_matrix)

print("feature matrix")

print(tfidf.vocabulary\_)

print ("word matrix")