**Source Codes**

**1) NLK using NLKT:**

Remove Stop Words Using NLTK

from nltk.corpus import stopwords

stopwords.words('english')

from bs4 import BeautifulSoup

import urllib.request

import nltk

from nltk.corpus import stopwords

response = urllib.request.urlopen('http://php.net/')

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

tokens = [t for t in text.split()]

clean\_tokens = tokens[:]

sr = stopwords.words('english')

for token in tokens:

if token in stopwords.words('english'):

clean\_tokens.remove(token)

freq = nltk.FreqDist(clean\_tokens)

for key,val in freq.items():

print (str(key) + ':' + str(val))

**2) NLP using Genism:**

To see similarity on some adjectives

# imports needed and logging

import gzip

import gensim

import logging

logging.basicConfig(format=’%(asctime)s : %(levelname)s : %(message)s’, level=logging.INFO)

def read\_input(input\_file):

    """This method reads the input file which is in gzip format"""

 logging.info("reading file {0}...this may take a while".format(input\_file))

    with gzip.open(input\_file, 'rb') as f:

        for i, line in enumerate(f):

  if (i % 10000 == 0):

                logging.info("read {0} reviews".format(i))

            # do some pre-processing and return list of words for each review

            # text

            yield gensim.utils.simple\_preprocess(line)

# build vocabulary and train model

    model = gensim.models.Word2Vec(

        documents,

        size=150,

        window=10,

        min\_count=2,

        workers=10,

        iter=10)