Name : Saood Ur Rahman

Roll No : 17P-6092

Section: “A”

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.feature\_extraction.text import TfidfVectorizer

corpus = open("Movies\_TV.txt").read()

import re

corpus = re.sub(r'Domain.\*\n', '', corpus)

rows = corpus.split('\n')

rows.remove(rows[-1])

def data(rows):

domain, lable, rating, reviews, = [], [], [], []

for row in rows:

dom, lab, rat, rev = row.split('\t')

domain.append(dom)

lable.append(lab)

rating.append(rat)

reviews.append(rev)

return reviews

reviews = data(rows)

def structure(reviews):

vec\_binary = CountVectorizer(ngram\_range = (1, 3), min\_df=10, max\_df=100, max\_features = 1000, binary=True)

X = vec\_binary.fit\_transform(reviews)

print("binary",X)

vec\_frequency = CountVectorizer(ngram\_range = (1, 3), min\_df=10, max\_df=100, max\_features = 1000)

Y = vec\_frequency.fit\_transform(reviews)

print("frequency",Y)

tfidf = TfidfVectorizer(ngram\_range=(1,3),min\_df=10,max\_df=100,max\_features=1000)

Z = tfidf.fit\_transform(reviews)

print("tfidf",Z)

structure(reviews)