Steps followed to achieve the result in assignment 2:

Question: Write a code to find at least 20 frequently asked questions for the subjects and topics mentioned in the dataset ([http://bit.ly/epo\_FAQ\_sample\_topics](https://www.google.com/url?q=http://bit.ly/epo_FAQ_sample_topics&sa=D&ust=1544962804005000&usg=AFQjCNHxifZ0w_haVJcXLTNnyC262o9TSQ)) from the questions in the dataset ([http://bit.ly/epo\_FAQ\_superset\_Sample](https://www.google.com/url?q=http://bit.ly/epo_FAQ_superset_Sample&sa=D&ust=1544962804005000&usg=AFQjCNGRv0GV_hVZQuWDFFj8Bqyq8cu_dQ)). The second dataset contains list of 500 questions that was asked in the interview.

**Solution:** I built my solution using python. Libraries required are Numpy, pandas, nltk.

Steps followed:

1. Reading the dataset from csv file.
2. Removing the unnecessary columns from dataset.
3. Split the dataset in questions having one or more sentences.
4. Tokenize each question into words.
5. Remove the words which are general in use also known as stop words. Also remove the question words.
6. Lemmatize each token means deriving the base word like base word of “better” is “good”.
7. Count the number of occurrence of each token in other question.
8. Question having more tokens which have occurred maximum in the dataset is the most frequently asked question.
9. Rank every question using step 4 to 8.
10. Now define keywords for identifying topics.
11. Match those keywords from words in FAQs and decide the topic.

In this way, I solved this problem.

**Python Code:**

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

def match(w1,w2):

c=0

x=min(len(w1),len(w2))

for i in range(x):

if w1[i]==w2[i]:

c+=1

if c>x//2:

return True

else:

return False

fo=open("500\_que.csv","r")

data=fo.read()

data=data.split("\n")

data=data[1:-1]

rank={}

for i in range(1,len(data)+1):

rank[i]=0

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

stop\_words = set(stopwords.words('english'))

d1=[]

que=['question','questions','how','when','what','which','why']

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

for sen in data:

word\_tokens = word\_tokenize(sen)

filtered\_sentence = []

for w in word\_tokens:

w=w.lower()

if w not in stop\_words and len(w)>2 and w not in que:

w=lemmatizer.lemmatize(w)

w=lemmatizer.lemmatize(w,pos='a')

w=lemmatizer.lemmatize(w,pos='v')

filtered\_sentence.append(w)

d1.append(filtered\_sentence)

print(d1)

for i in range(len(d1)):

for j in range(len(d1)):

for k in range(len(d1[i])):

for l in range(len(d1[j])):

if match(d1[i][k],d1[j][l]):

rank[i+1]+=1

z=0

topic={}

topic["fcfs, Operating System"]=['fcfs','first serve']

topic["dml, Database Management System"]=['dml','insert','update','delete']

topic["object, Object\_Oriented\_Programming"]=['class','object','method']

topic["inheritance, Object\_Oriented\_Programming"]=['inherit','extend']

topic["ploymorphism, Object\_Oriented\_Programming"]=['polymorphism','overloading','overwritting']

topic["encapsulation, Object\_Oriented\_Programming"]=['abstract','abstract class','interface']

topic["jdbc, Object\_Oriented\_Programming"]=['jdbc']

topic['Sql, Database management system']=['sql','query']

appendFile = open('filteredtext.csv','a')

appendFile.write(" Topic, Subject, Question,\n")

appendFile.close()

for k,v in sorted(rank.items(),reverse=True, key =lambda kv:(kv[1], kv[0])):

if z>25:

break

flag=False

for wrd in data[k-1].split():

if len(wrd)>2:

for ke,va in topic.items():

for wrd2 in va:

if match(wrd,wrd2):

print(data[k-1])

print()

print()

z+=1

appendFile = open('filteredtext.csv','a')

appendFile.write(" "+ke+", "+data[k-1]+",\n")

appendFile.close()

flag=True

break

if flag:

break

if flag:

break