In [1]:
```
import sqlite3
import pandas as pd
from nltk.tokenize import word_tokenize
from sqlalchemy import create_engine # database connection
import os
import datetime as dt
from datetime import datetime
```

In [ ]:
```
#Creating db file from csv
#Learn SQL: https://www.w3schools.com/sql/default.asp
if not os.path.isfile('trainy.db'):
    start = datetime.now()
    disk_engine = create_engine('sqlite:///trainy.db')
    start = dt.datetime.now()
    chunksize = 10000
    j = 0
    index_start = 1
    for df in pd.read_csv('Train.csv', names=['Id', 'Title', 'Body', 'Tags'], ch
        df.index += index_start
        j+=1
        print('{} rows'.format(j*chunksize))
        df.to_sql('no_dup_train', disk_engine, if_exists='append')
        index_start = df.index[-1] + 1
    print("Time taken to run this cell :", datetime.now() - start)
```

In [2]:
```
#Learn SQL: https://www.w3schools.com/sql/default.asp
if os.path.isfile('trainy.db'):
    start = datetime.now()
    con = sqlite3.connect('trainy.db')
    data = pd.read_sql_query('SELECT * FROM no_dup_train LIMIT 50000', con)
    con.close()#SELECT * FROM `your_table` LIMIT 1001, 5000
    print("Time taken to run this cell :", datetime.now() - start)
else:
    print("Please download the train.db file from drive or run the first to genar
```

```
Time taken to run this cell : 0:00:00.893951
```

In [83]:
```
data.head(5)
```

Out[83]:

	index	Id	Title	Body	Tags	tagcount
0	1	Id	Title	Body	Tags	1
1	2	1	How to check if an uploaded file is an image w...	<p>I'd like to check if an uploaded file is an...	php image-processing file-upload upload mime-t...	5
2	3	2	How can I prevent firefox from closing when I ...	<p>In my favorite editor (vim), I regularly us...	firefox	1
3	4	3	R Error Invalid type (list) for variable	<p>I am import matlab file and construct a dat...	r matlab machine-learning	3
4	5	4	How do I replace special characters in a URL?	<p>This is probably very simple, but I simply ...	c# url encoding	3

In [ ]:
```
```

In [84]:
```
data.describe()
```

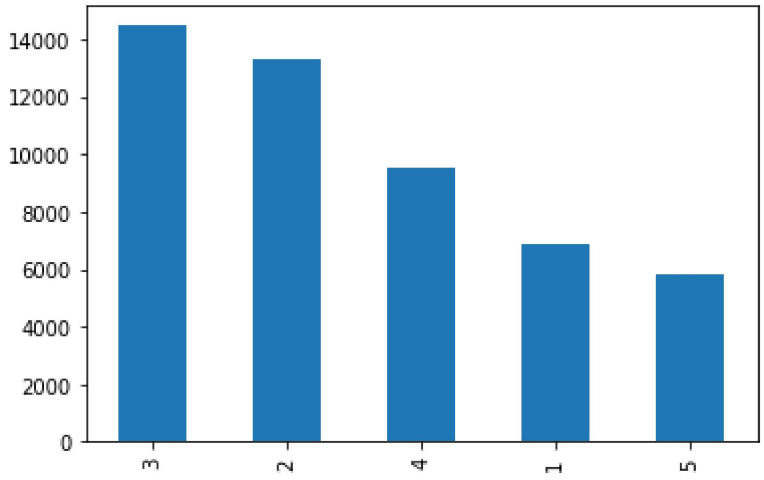
Out[84]:

	index	tagcount
count	49851.000000	49851.000000
mean	64912.698481	2.885599
std	50277.996082	1.207786
min	1.000000	1.000000
25%	22470.500000	2.000000
50%	54954.000000	3.000000
75%	97473.500000	4.000000
max	150000.000000	5.000000

In [3]:
```
#very important thing to learn here is that for list we should not use
#list.len(). we should use len(list)
#iam just specifying number of tags occuring in each question
data['tagcount']=data['Tags'].apply(lambda text: len(text.split()))
```

In [4]:
```
#if u want to draw bar graph for the tags youn can specify
#value_counts().plot(kind='bar')
%matplotlib inline
import matplotlib.pyplot as plt
data['tagcount'].value_counts().plot(kind='bar')
```

Out[4]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2242af02358>
```



In [5]:
```
data=data.drop_duplicates(subset={'Title','Tags','tagcount'},keep='first')
```

In [6]:
```
#we have already seen one important thing in countvectorizer
#that is get_feature_names()
from sklearn.feature_extraction.text import CountVectorizer
#In this only i got to know giving tokenizer is much important
vect=CountVectorizer(tokenizer=lambda x: x.split())
tagdatamatrix=vect.fit_transform(data['Tags'])
```

In [7]:
```
#this will tell me no. of datapoints
print(tagdatamatrix.shape[0])
#to know number of unique tags
print(tagdatamatrix.shape[1])
#to know what are the tags present
tagy=vect.get_feature_names()
tagy[:10]
#see tsgy is a type of listso we cant give tagy.head()
#we can mention like this that the number of elements i require
#but like data[:10] we can give for dataframe
```

```
49851
13892
```

Out[7]:
```
['.class-file',
 '.each',
 '.emf',
 '.hgtags',
 '.htaccess',
 '.htpasswd',
 '.mov',
 '.net',
 '.net-1.1',
 '.net-2.0']
```

In [8]:
```
#now i want to know each tag occuring number of times
freqs=tagdatamatrix.sum(axis=0).A1
res=dict(zip(tagy,freqs))
```

In [9]:
```
# this is the important invention i have a set i want it to
#convert it into dataframe so i first convert it into
#list wit setname.items() and i use pd.DataFrame specifying
#column names
s=pd.DataFrame(list(res.items()),columns=['tags','counts'])
```

In [10]:
```
s.head()
```

Out[10]:

	tags	counts
0	.class-file	1
1	.each	6
2	.emf	1
3	.hgtags	1
4	.htaccess	175

In [11]:
```
#now we wnt dataframe to be sorted so i can use sort.values() function
ssorted=s.sort_values(['counts'],ascending=False)
```

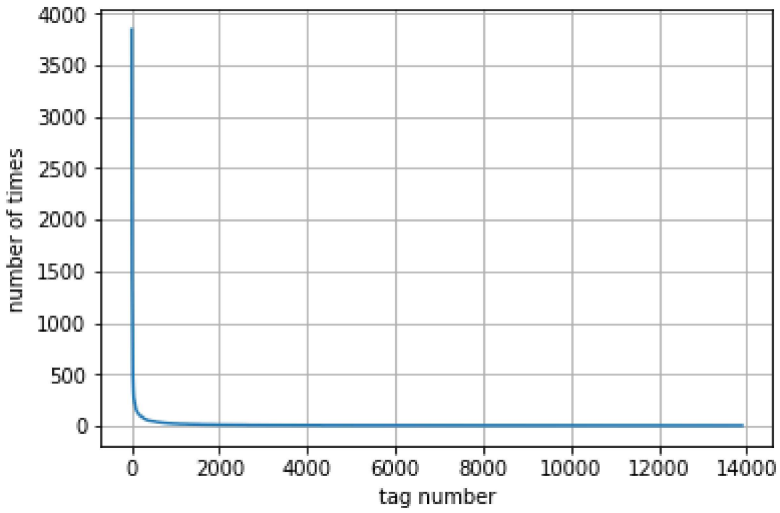
In [12]:
```
ssorted.head()
```

Out[12]:

	tags	counts
1520	c#	3845
6021	java	3399
8927	php	3231
6056	javascript	3089
419	android	2673

In [13]:
```
#now i take an array i take values of ssorted counts
scounts=ssorted['counts'].values
```

In [14]:
```
#now i plot a grid plot since it is continous values
plt.plot(scounts)
plt.grid()
plt.xlabel("tag number")
plt.ylabel("number of times")
plt.show()
```



In [15]:
```
data.head()
```

Out[15]:

	index	Id	Title	Body	Tags	tagcount
0	1	Id	Title	Body	Tags	1
1	2	1	How to check if an uploaded file is an image w...	<p>I'd like to check if an uploaded file is an...	php image-processing file-upload upload mime-t...	5
2	3	2	How can I prevent firefox from closing when I ...	<p>In my favorite editor (vim), I regularly us...	firefox	1
3	4	3	R Error Invalid type (list) for variable	<p>I am import matlab file and construct a dat...	r matlab machine-learning	3
4	5	4	How do I replace special characters in a URL?	<p>This is probably very simple, but I simply ...	c# url encoding	3

In [16]:
```
data['Body'].head()
```

Out[16]:
```
0                                                Body
1    <p>I'd like to check if an uploaded file is an...
2    <p>In my favorite editor (vim), I regularly us...
3    <p>I am import matlab file and construct a dat...
4    <p>This is probably very simple, but I simply ...
Name: Body, dtype: object
```

In [17]:
```
a=str(data['Body'][:1])
print(a)
data['Body'][2]
```

```
0    Body
Name: Body, dtype: object
```

Out[17]:
```
'<p>In my favorite editor (vim), I regularly use ctrl-w to execute a certain ac
tion. Now, it quite often happens to me that firefox is the active window (on w
indows) while I still look at vim (thinking vim is the active window) and press
ctrl-w which closes firefox. This is not what I want. Is there a way to stop ct
rl-w from closing firefox?</p>\n\n<p>Rene</p>\n'
```

In [18]:
```
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
def striphtml(data):
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', str(data))
    return cleantext
stop_words=set(stopwords.words('english'))
stemmer=SnowballStemmer("english")
```

In [19]:

```
def create_connection(db_file):
    """ create a database connection to the SQLite database
        specified by db_file
    :param db_file: database file
    :return: Connection object or None
    """
    try:
        conn = sqlite3.connect(db_file)
        return conn
    except Error as e:
        print(e)

    return None

def create_table(conn, create_table_sql):
    """ create a table from the create_table_sql statement
    :param conn: Connection object
    :param create_table_sql: a CREATE TABLE statement
    :return:
    """
    try:
        c = conn.cursor()
        c.execute(create_table_sql)
    except Error as e:
        print(e)

def checkTableExists(dbcon):
    cursr = dbcon.cursor()
    str = "select name from sqlite_master where type='table'"
    table_names = cursr.execute(str)
    print("Tables in the databse:")
    tables =table_names.fetchall()
    print(tables[0][0])
    return(len(tables))

def create_database_table(database, query):
    conn = create_connection(database)
    if conn is not None:
        create_table(conn, query)
        checkTableExists(conn)
    else:
        print("Error! cannot create the database connection.")
    conn.close()

sql_create_table = """CREATE TABLE IF NOT EXISTS QuestionsProcessed (question te
create_database_table("Processed.db", sql_create_table)
```

```
Tables in the databse:
QuestionsProcessed
```

In [20]:
```
# http://www.sqlitetutorial.net/sqlite-delete/
# https://stackoverflow.com/questions/2279706/select-random-row-from-a-sqlite-tal
start = datetime.now()
read_db = 'trainy.db'
#vv.v.vimportant thing i am giving 'Processed.db'as my write_db
write_db = 'Processed.db'
if os.path.isfile(read_db):
    conn_r = create_connection(read_db)
    if conn_r is not None:
        reader =conn_r.cursor()
        reader.execute("SELECT Title, Body, Tags From no_dup_train ORDER BY RANDO

if os.path.isfile(write_db):
    conn_w = create_connection(write_db)
    if conn_w is not None:
        tables = checkTableExists(conn_w)
        writer =conn_w.cursor()
        if tables != 0:
            writer.execute("DELETE FROM QuestionsProcessed WHERE 1")
            print("Cleared All the rows")
print("Time taken to run this cell :", datetime.now() - start)
```

```
Tables in the databse:
QuestionsProcessed
Cleared All the rows
Time taken to run this cell : 0:00:41.373159
```

In [21]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\uib25207\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[21]:
```
True
```

In [22]:
```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\uib25207\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[22]:
```
True
```

In [23]:
```
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
def striphtml(data):
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', str(data))
    return cleantext
stop_words=set(stopwords.words('english'))
stemmer=SnowballStemmer("english")
```

In [24]:

```
start = datetime.now()
preprocessed_data_list=[]
reader.fetchone()
#fetchone() is used to fetch the records
# it fetches int the form of rows
#<code> if u see it presents in all all the title and body
#i want to remove <code from this title and body
# i have assigned rows of header to columns
#ro[0] to title,row[1] to questionwhich is body,row[2] to tags
questions_with_code=0
len_pre=0
len_post=0
questions_proccesed = 0
for row in reader:
    is_code = 0
    title, question, tags = row[0], row[1], row[2]
    if '<code>' in question:
        questions_with_code+=1
        is_code = 1
    x = len(question)+len(title)
    len_pre+=x
# iam storing adding the lengths before preprocessing so that i can find average
#i want to substitute the code using re.DOTALL
    code = str(re.findall(r'<code>(.*?)</code>', question, flags=re.DOTALL))
# i will substitute this <code>woth question  with flags=re.MULTILINE|re.DOTALL
    question=re.sub('<code>(.*?)</code>', '', question, flags=re.MULTILINE|re.DO
    question=striphtml(question.encode('utf-8'))
#at same time iam making preprocessing of title and question which is body
    title=title.encode('utf-8')
# this is the important step whereiam mixing question and title to make a questi
    question=str(title)+" "+str(question)
    question=re.sub(r'[^A-Za-z]+',' ',question)
#there there is two types of tokenize one is word_tokenize
#another one is sent_tokenize
#word_tokenize will tokenize based on the tokenize based on the number of words
#sent_tokenize will tokenize sentences based on sentences based on fullstop
#word_tokenize works based on space
    words=word_tokenize(str(question))
#there is reason we have not tokenized till now is till now we had htmlletters
#Removing all single letter and and stopwords from question exceptt for the lette
    question=' '.join(str(stemmer.stem(j)) for j in words if j not in stop_words

    len_post+=len(question)
    tup = (question,code,tags,x,len(question),is_code)
    questions_proccesed += 1
    #now i am putting all my sentences in writer
    writer.execute("insert into QuestionsProcessed(question,code,tags,words_pre,w
    if (questions_proccesed%100000==0):
        print("number of questions completed=",questions_proccesed)

no_dup_avg_len_pre=(len_pre*1.0)/questions_proccesed
no_dup_avg_len_post=(len_post*1.0)/questions_proccesed

print( "Avg. length of questions(Title+Body) before processing: %d"%no_dup_avg_l
print( "Avg. length of questions(Title+Body) after processing: %d"%no_dup_avg_le
print ("Percent of questions containing code: %d"%((questions_with_code*100.0)/q
```

```
print("Time taken to run this cell :", datetime.now() - start)
```

```
Avg. length of questions(Title+Body) before processing: 1165
Avg. length of questions(Title+Body) after processing: 341
Percent of questions containing code: 56
Time taken to run this cell : 0:00:31.127748
```

In [25]:
```
# dont forget to close the connections, or else you will end up with locks
conn_r.commit()
conn_w.commit()
conn_r.close()
conn_w.close()
```

In [26]:
```
if os.path.isfile(write_db):
    conn_r = create_connection(write_db)
    if conn_r is not None:
        reader =conn_r.cursor()
        reader.execute("SELECT question From QuestionsProcessed LIMIT 10")
        print("Questions after preprocessed")
        print('='*100)
        reader.fetchone()
        for row in reader:
            print(row)
            print('-'*100)
conn_r.commit()
conn_r.close()
```

```
Questions after preprocessed
====================================================================================================
('differ packag version suit recent got surpris result that impli differ versio
n experiment thought possibl what happen is bug mayb',)
----------------------------------------------------------------------------------------------------
('silverlight open childwindow time frame pass hope titl make sens discrib issu
use childwindow silverlight display process messag rotat imag ui work onc compl
et event call window close problem look littl ugli ui perform quick task child
window open close second what want abl child window open second process pass cl
ose complet ad section xaml call child search find anyth might possibl void edi
t close object sender eventarg editchanneldetail edit sender editchanneldetail
etc',)
----------------------------------------------------------------------------------------------------
('ms powerpoint window microsoft powerpoint mac ppt with comment slide made ms
powerpoint window save ppt ppt extens sent friend mac microsoft powerpoint he u
nabl see comment is way see comment',)
----------------------------------------------------------------------------------------------------
('help need problem regard diffcult manipul binomi coeffici what coeffici dot p
osit integ is close form answer',)
----------------------------------------------------------------------------------------------------
('jqueri titl attribut textbox new jqueri list link app want abl take titl attr
ibut link put text box page link click this got far work ani help would great a
ppreci thank',)
----------------------------------------------------------------------------------------------------
('can instal build essenti ubuntu jaunti here etc apt sourc list tri result cou
ldn find packag build essenti result sever err http us archiv ubuntu com jaunti
updat multivers sourc not found ip fail fetch http secur ubuntu com ubuntu dist
jaunti secur main binari packag not found ip error ani idea',)
----------------------------------------------------------------------------------------------------
('parameter straight line use polar coordin without angl parameter straight lin
e start point endpoint my idea use equat line goe two point that frac quad quad
quad quad frac quad quad quad quad when solv ny quad quad left frac quad quad r
ight which plot end straight line pass given point exercic happi teacher advis
easi way alreadi know find posit line use polar coordin read subject far unders
tand need two thing length angl said line axi right find bit confus would like
```

```
know equat realli bad way case polar coordin best way parameter exercis underst
and oper find angl assum alreadi know start endpoint ab thank help one whole da
y think sure miss someth obvious',)
--------------------------------------------------------------------------------
--------------------
('read vector array xml plot xml file creat facetrack softwar give rotat vector
well scale translat origin point look like what would like know could read vect
or xml file plot video file use use xcode os thank',)
--------------------------------------------------------------------------------
--------------------
('vb net datagridview an item key alreadi ad use uniqu index vb net net hello l
ook around seem find solut problem eventlog object inherit datagridview public
variabl eventlist list of eventlogitem eventlogitem seven properti describ even
t includ index set eventlist count time entri ad uniqu everyth work fine tri ad
d entri serial port datareceiv event handler upon receiv follow except an error
occur creat form see except innerexcept detail the error an item key alreadi ad
click view detail expand innerexcept yield inform here relev code the eventlog
class eventlist the eventlogitem class code insert item event log the code caus
problem ch cp namespac code resid class main form class ani idea thank lot adva
nc',)
--------------------------------------------------------------------------------
--------------------
```

In [27]:
```
#finally iam taking now the data into database
#after this i train my machine learning models
#Taking 0.5 Million entries to a dataframe.
write_db = 'Processed.db'
if os.path.isfile(write_db):
    conn_r = create_connection(write_db)
    if conn_r is not None:
        preprocessed_data = pd.read_sql_query("""SELECT question, Tags FROM Ques

conn_r.commit()
conn_r.close()
```

In [28]:
```
preprocessed_data.head()
```

Out[28]:

	question	tags
0	configur mb vps runnabl jar applic want run ja...	java jar hosting vps
1	differ packag version suit recent got surpris ...	debian package-management
2	silverlight open childwindow time frame pass h...	silverlight childwindow
3	ms powerpoint window microsoft powerpoint mac ...	osx microsoft-powerpoint microsoft-powerpoint-...
4	help need problem regard diffcult manipul bino...	binomial-coefficients

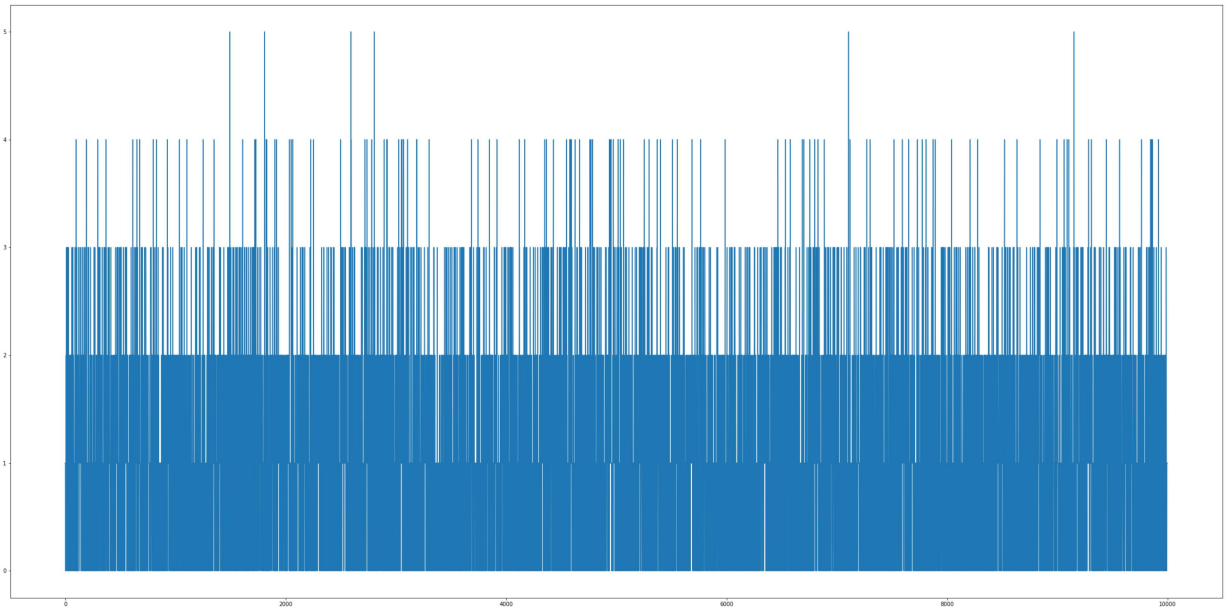
In [150]:
```
ste=[]

str1=' '
for i in preprocessed_data['tags']:
    stemy=[]
    for w in i.split():
        stemy.append(stemmer.stem(w))
    #print(stemy)
    str1=' '.join(stemy)
    #print(str1)
    ste.append(str1)
```

In [172]:
```
commonlength=[]
taglength=[]
difflength=[]
for i in range(9999):
    awords=ste[i].split(' ')
    bwords=preprocessed_data['tags'][i].split()
    common =len(set(awords).intersection(set(bwords)))
    commonlength.append(common)
    taglength.append(len(bwords))
s1=np.array(commonlength)
s2=np.array(taglength)
print(s2-s1)
```

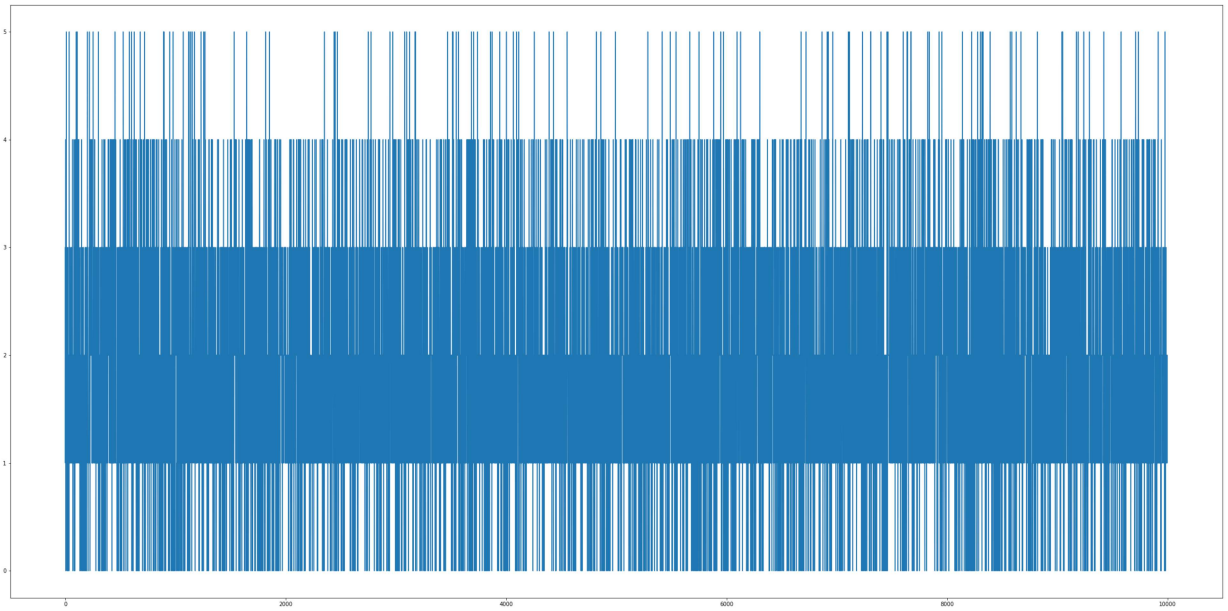
```
[1 1 0 ... 1 1 1]
```

In [179]:
```
fig,ax=plt.subplots(figsize=(40,20))
ax.plot(s2-s1)
plt.show()
```



AS MENTIONED IN COMMENT THIS IS THE EDA I PERFORMED .BASICALLY WE ARE DOING ANALYSIS GIVING QUESTION AS INPUT AND PREDICTING THE TAGS AS OUTPUT. IF WE SEE FOR MOST OF THE QUESTIONS 2 TAGS ARE NOT PRESENT IN QUESTION ITSELF . IN SOME CASE 3TAGS THAT PRESENT IN TAGS NOT PRESENT IN QUESTION.THIS IS THE MAJOR ANAYSIS STEP WHICH SHOWS US THE LENGTH OF UNCOMMON TAGS.

In [180]:
```
fig,ax=plt.subplots(figsize=(40,20))
ax.plot(commonlength)
plt.show()
```



THE ABOVE VISUALISATION SHOWS THE COMMON TAGS LENGTH PRESENT BETWEEN BOTH THE QUESTION AND TAGS.

In [ ]:
```
#converting tags for multilabel problems
from sklearn.feature_extraction.text import CountVectorizer
vec=CountVectorizer(tokenizer=lambda x: x.split())
multilabely=vec.fit_transform(preprocessed_data['tags'])
```

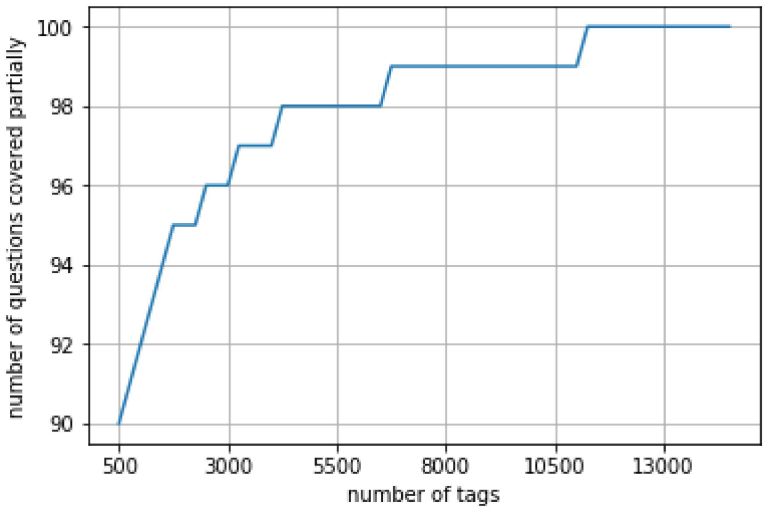
In [30]:
```
#due to low computation power
#we have to limit the number of tags we gonna choose
#we will use some criterrion to do this
```

In [31]:
```
def tagstochoose(n):
    t=multilabely.sum(axis=0).tolist()[0]
    #From this we will get the which tag is occuring more frequently
    #this is the criterrion we gonna consider
    #so we will take the important tags like this
    sortedtags=sorted(range(len(t)),key=lambda i :t[i],reverse=True)
    multilabely1=multilabely[:,sortedtags[:n]]

    return multilabely1
```

In [32]:
```
def questionsexplained1(n):
    multilabelyn=tagstochoose(n)
    x=multilabelyn.sum(axis=1)
    return (np.count_nonzero(x==0))
```

In [33]:
```
import numpy as np
questionsexplained=[]
totaltags=multilabely.shape[1]
totalquestions=preprocessed_data.shape[0]
for i in range(500,totaltags,100):
    questionsexplained.append(np.round(((totalquestions-questionsexplained1(i))/
fig,ax=plt.subplots()
ax.plot(questionsexplained)
xlabel=list(500+np.array(range(-50,450,50))*50)
ax.set_xticklabels(xlabel)
plt.xlabel('number of tags')
plt.ylabel('number of questions covered partially')
plt.grid()
plt.show()
```



In [34]:
```
multilabelyx=tagstochoose(5500)
totalsize=preprocessed_data.shape[0]
trainsize=int(0.8*totalsize)
xtrain=preprocessed_data.head(trainsize)
xtest=preprocessed_data.tail(totalsize-trainsize)
ytrain=multilabelyx[0:trainsize]
ytest=multilabelyx[trainsize:totalsize,:]
```

In [35]:
```
print(ytrain.shape)
print(xtrain.shape)
print(xtest.shape)
print(ytest.shape)
```

```
(7999, 5500)
(7999, 2)
(2000, 2)
(2000, 5500)
```

In [36]:
```
print(xtrain[:2])
```

```
                                              question  \
0  configur mb vps runnabl jar applic want run ja...
1  differ packag version suit recent got surpris ...

                         tags
0       java jar hosting vps
1  debian package-management
```

In [37]:
```
print(xtrain.iloc[0,:])
```

```
question    configur mb vps runnabl jar applic want run ja...
tags                                       java jar hosting vps
Name: 0, dtype: object
```

In [38]:
```
print(ytrain[:2])
```

```
  (0, 1)        1
  (0, 536)      1
  (0, 464)      1
  (0, 959)      1
  (1, 109)      1
  (1, 4856)     1
```

In [39]:
```
print(xtest[:2])
```

```
                                                question  \
7999  instal ubuntu netbook remix exist karmic insta...
8000  not abl connect internet use chrome abl connec...

                                          tags
7999  ubuntu ubuntu-9.10 ubuntu-netbook-remix
8000           windows-7 google-chrome firefox
```

actually it is multi class label we are predicting the tags so we have to use multi class classification algortihm we can implement multiclass using the logistic regression and svm using the normal sgd classfier we normally use one vs rest classifier along with sgd classifier how we use the use the sgd classifier is we change the losses generally we know that log loss is for logistic regressiion hinge loss for svm square loss for regression generally when we are using the cross validation with calibratedclassifier cv we se this along eith sgd classifier when we are applying regression using logistic and svm we use the same method

we can use onevsrest classifier for multi class and multilabel also

In [40]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
vecty=TfidfVectorizer(min_df=0.00009,tokenizer= lambda x:x.split())
xtrainmultilabel=vecty.fit_transform(xtrain['question'])
xtestmultilabel=vecty.fit_transform(xtest['question'])
```

In [41]:
```
print(xtrainmultilabel.shape)
```

```
(7999, 18298)
```

In [42]:
```
print(xtestmultilabel.shape)
```

```
(2000, 7975)
```

this is the methiodn we gonna use for multiclass classification using knn actually due to memory constraints we are not writing this in code other\ wise it is powerful to use\ from sklearn.multilearn.adapt import MLKNN classifier=MLKNN(k=21) classifier.fit(xtrainmultilabel,ytrain) prdictions=classifier.predict(xtestmultilabel) print(accuracy_score(ytest,predictions)

actually there are two types of f1score

print(metrics.f1score(ytest,predictions,average='micro') print(metrics.f1score(ytest,predictions,averge='macro') print(metrics.hamming_loss(ytest,predictions))

this is the method we gonna use for multiclass uisng the logistic regression

classifierlogistic=OneVsRestCalssifier(SGDClassifier(loss='log',alpha=0.00001,penalty='l1'),n_jobs=- classifierlogistic.fit(xtrainmultilabel,ytrain) predictions=classifierlogistic.predict(xtestmultilabel) print("accuracy",accuracy_score(ytest,predictions) print("f1score",metrics.f1_score(ytest,predictions,average='micro') print('macro f1_score',metrics.f1_score(ytest,predictions,average='macro') print(hamming_loss(ytest,predictions) print('precision reccall report',metrics.classification_report(ytest,predictions))\

actually this one vs rest classifier used for

both multi class and multi label

In [43]:
```
sql_create_table = """CREATE TABLE IF NOT EXISTS QuestionsProcessed (question te
create_database_table("Titlemoreweight.db", sql_create_table)
```

```
Tables in the databse:
QuestionsProcessed
```

In [44]:
```
preprocessed_data.shape
```

Out[44]:
```
(9999, 2)
```

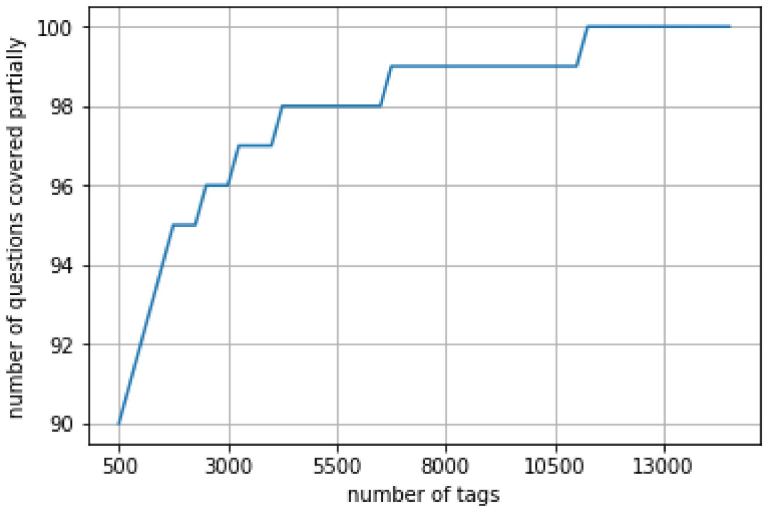
In [45]:
```
print("number of data points in sample :", preprocessed_data.shape[0])
print("number of dimensions :", preprocessed_data.shape[1])
```

```
number of data points in sample : 9999
number of dimensions : 2
```

In [46]:
```
vectorizer = CountVectorizer(tokenizer = lambda x: x.split(), binary='true')
multilabel_y = vectorizer.fit_transform(preprocessed_data['tags'])
```

In [47]:
```
questions_explained = []
total_tags=multilabel_y.shape[1]
total_qs=preprocessed_data.shape[0]
for i in range(500, total_tags, 100):
    questions_explained.append(np.round(((totalquestions-questionsexplained1(i)
```

In [48]:
```
fig,ax=plt.subplots()
ax.plot(questions_explained)
xlabel=list(500+np.array(range(-50,450,50))*50)
ax.set_xticklabels(xlabel)
plt.xlabel('number of tags')
plt.ylabel('number of questions covered partially')
plt.grid()
plt.show()
```



In [49]:
```
multilabel_yx = tagstochoose(500)
#print("number of questions that are not covered :", questionsexplained(500),"out
```

In [50]:
```
preprocessed_data.shape
```

Out[50]:
```
(9999, 2)
```

In [51]:
```
vectory=TfidfVectorizer(tokenizer=lambda x:x.split())
xtrainymultilabel=vectory.fit_transform(preprocessed_data['question'])
```

In [52]:
```
print(xtrainymultilabel.shape)
```

```
(9999, 20840)
```

In [53]:
```
print(multilabel_yx.shape)
```

```
(9999, 500)
```

In [54]:
```
from sklearn.model_selection import train_test_split
xtrain1,xtest1,ytrain1,ytest1=train_test_split(xtrainymultilabel,multilabel_yx,t
```

In [55]:
```
print(xtrain1.shape)
print(xtest1.shape)
print(ytrain1.shape)
print(ytest1.shape)
```

```
(6999, 20840)
(3000, 20840)
(6999, 500)
(3000, 500)
```

In [56]:
```
print(xtrain[:2])
```

```
                                              question  \
0  configur mb vps runnabl jar applic want run ja...
1  differ packag version suit recent got surpris ...

                        tags
0       java jar hosting vps
1  debian package-management
```

In [57]:
```
print(ytrain[:2])
```

```
  (0, 1)        1
  (0, 536)      1
  (0, 464)      1
  (0, 959)      1
  (1, 109)      1
  (1, 4856)     1
```

In [58]:
```
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import SGDClassifier
from sklearn.metrics import accuracy_score
from sklearn import metrics
```

In [57]:
```
classifiery=OneVsRestClassifier(SGDClassifier(loss='log',alpha=0.00001,penalty='
classifiery.fit(xtrain1,ytrain1)
predictions2=classifiery.predict(xtest1)
print('accuray',accuracy_score(predictions2,ytest1))
print('f1score',metrics.f1_score(predictions2,ytest1,average='micro'))
print('hamming loss',metrics.hamming_loss(predictions2,ytest1))
```

```
accuray 0.16866666666666666
f1score 0.378419452887538
hamming loss 0.003272
```

In [58]:
```
print('accuray',accuracy_score(predictions2,ytest1)*100)
print('f1score',metrics.f1_score(predictions2,ytest1,average='micro'))
print('hamming loss',metrics.hamming_loss(predictions2,ytest1))
```

```
accuray 16.866666666666667
f1score 0.378419452887538
hamming loss 0.003272
```

In [59]:
```
from sklearn.linear_model import LogisticRegression
classifierfinal=OneVsRestClassifier(LogisticRegression(penalty='l1'),n_jobs=-1)
classifierfinal.fit(xtrain1,ytrain1)
predicttionsfinal=classifierfinal.predict(xtest1)
print('accuracy',accuracy_score(predicttionsfinal,ytest1))
print('f1score',metrics.f1_score(predicttionsfinal,ytest1,average='micro'))
print('macrof1score',metrics.f1_score(predicttionsfinal,ytest1,average='macro'))
print('hammingloss',metrics.hamming_loss(predicttionsfinal,ytest1))
```

```
accuracy 0.171
f1score 0.331195985503206
macrof1score 0.1808513690582869
hammingloss 0.0031986666666666665

c:\legacyapp\python36\lib\site-packages\sklearn\metrics\classification.py:1143:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels w
ith no predicted samples.
  'precision', 'predicted', average, warn_for)
c:\legacyapp\python36\lib\site-packages\sklearn\metrics\classification.py:1145:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels w
ith no true samples.
  'recall', 'true', average, warn_for)
```

1.bag of words upto 4 grams

#due to computation limitation iam limiting to trigrams

#computing the f1 score

In [59]:
```
from sklearn.feature_extraction.text import CountVectorizer
vectory=CountVectorizer(min_df=3,ngram_range=(1,3))
xtrainymultilabel=vectory.fit_transform(preprocessed_data['question'])
xtrain1,xtest1,ytrain1,ytest1=train_test_split(xtrainymultilabel,multilabel_yx,t
```

In [60]:
```
print(xtrain1.shape)
print(xtest1.shape)
print(ytrain1.shape)
print(ytest1.shape)
```

```
(6999, 38883)
(3000, 38883)
(6999, 500)
(3000, 500)
```

In [185]:
```
from sklearn.linear_model import LogisticRegression
classifierfinal=OneVsRestClassifier(LogisticRegression(penalty='l1',C=10))
```

In [186]:
```
classifierfinal.fit(xtrain1,ytrain1)
```

Out[186]:
```
OneVsRestClassifier(estimator=LogisticRegression(C=10, class_weight=None, dual=
False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l1', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False),
          n_jobs=None)
```

In [187]:
```
predicttionsfinal=classifierfinal.predict(xtest1)
```

In [184]:
```
import warnings
warnings.filterwarnings('ignore')
print('accuracy',accuracy_score(predicttionsfinal,ytest1))
print('f1score',metrics.f1_score(predicttionsfinal,ytest1,average='micro'))
print('macrof1score',metrics.f1_score(predicttionsfinal,ytest1,average='macro'))
print('hammingloss',metrics.hamming_loss(predicttionsfinal,ytest1))
```

```
accuracy 0.16833333333333333
f1score 0.3910874625872963
macrof1score 0.2719146526613718
hammingloss 0.003662
```

In [0]:
```
#2.performing hyperparameter tunung for logistic regression to improve performan
```

In [65]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
estimate = [(10**i) for i in range(-4,4,1)]
pen=['l1','l2']
params = {'estimator__C':estimate}
clf = OneVsRestClassifier(LogisticRegression(penalty='l1'))
classifierfinal= GridSearchCV(clf, params, cv=3)
classifierfinal.fit(xtrain1, ytrain1)
```

Out[65]:
```
GridSearchCV(cv=3, error_score='raise-deprecating',
       estimator=OneVsRestClassifier(estimator=LogisticRegression(C=1.0, class_
weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l1', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False),
          n_jobs=None),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'estimator__C': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 100
0]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=0)
```

In [66]:
```
print(classifierfinal.best_params_)
```

```
{'estimator__C': 1}
```

In [67]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.linear_model import LogisticRegression
classifierfinal=OneVsRestClassifier(LogisticRegression(C=1,penalty='l1'))
classifierfinal.fit(xtrain1,ytrain1)
predicttionsfinal=classifierfinal.predict(xtest1)
print('accuracy',accuracy_score(predicttionsfinal,ytest1))
print('f1score',metrics.f1_score(predicttionsfinal,ytest1,average='micro'))
print('macrof1score',metrics.f1_score(predicttionsfinal,ytest1,average='macro'))
print('hammingloss',metrics.hamming_loss(predicttionsfinal,ytest1))
```

```
accuracy 0.17666666666666667
f1score 0.39777120764157387
macrof1score 0.27156845868104307
hammingloss 0.0035306666666666668
```

In [0]:
```
#3.one vs rest classifier with sgd classifier with hinge loss
```

In [0]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.linear_model import SGDClassifier
classifierfinal=OneVsRestClassifier(SGDClassifier(loss='hinge'))
classifierfinal.fit(xtrain1,ytrain1)
predicttionsfinal=classifierfinal.predict(xtest1)
print('accuracy',accuracy_score(predicttionsfinal,ytest1))
print('f1score',metrics.f1_score(predicttionsfinal,ytest1,average='micro'))
print('macrof1score',metrics.f1_score(predicttionsfinal,ytest1,average='macro'))
print('hamming`loss',metrics.hamming_loss(predicttionsfinal,ytest1))
```

```
accuracy 0.182
f1score 0.3044510385756677
macrof1score 0.11571108860943441
hammingloss 0.0031253333333333333
```

DOCUMENTATION

#conclusions: ***introduction***

- stackoverflow tag prediction is the business probelm with problem statement to predict *the tags of the questions *the type of classififcation we are performing is multilabel calssification #we can employ our machine learning models which we use for our classification
- we are using onevsrest classifier with our machinbe learning algorithms *we are using f1 score as our metric to be used

preprocessing *then we started with the data preprocessing steps which involves *removal of duplicates *filling th null values *but we did not remove any outliers because mainly there are not many features and leass number of numerical features *but we cleaned our text of questions where we removed the unwanted threads.

feature engineering *then we started with the feature engineering *we analysed the number of tags *number off tags that are occuring frequntly *number of data points and number of tags *number of tags per question *after viewing all this we have taken that are occuring more frequently. based on the vocabulary

machine learning models

- we can using multiple techinques in bulidong a machine learning model such as .binary relevance .classifier chains .label powerset
- we have selected top tags based on their presence in the tags top 500 are selected

- we are using tfidf model and bag of words model due to computatition limitation we were not able to perform the multilabel knn actually... *the accuaracy we obtained in this models are less since it is a multi label classification *we have selected only top 500 tags not all

*we have performed algorithms like 1.logistic regression model with onevs restr on tfidf vectorizer

2.logisticregression model with sgd classifier of log loss (difference between both is the logistic regression model uses gradient descent where as sgde classifier uses stochastic radient descent for optimisation)

In [0]:
```
s=[[0.18,0.369,0.003],[0.179,0.32,0.003]]
s1=pd.DataFrame(s,columns=['accuracy','f1score','hammingloss'],index=['sgd class
s1
```

Out[92]:

	accuracy	f1score	hammingloss
sgd classifier with log loss	0.180	0.369	0.003
logistic regression	0.179	0.320	0.003

*FOR ASSIGNMENTS *

1.for logistic regression 2.logistiv regression after hyper parameter training 3.sgd regressor with hinge loss for svm

In [188]:
```
s=[[0.165,0.395,0.27,0.0035],[0.176,0.383,0.272,0.0035],[0.182,0.30,0.11,0.003]]
s1=pd.DataFrame(s,columns=['accuracy','f1score','macrof1score','hammingloss'],in
s1
```

Out[188]:

	accuracy	f1score	macrof1score	hammingloss
logistic regression	0.165	0.395	0.270	0.0035
logistic regression after hyperparameter tuning	0.176	0.383	0.272	0.0035
sgd regressor with hinge loss	0.182	0.300	0.110	0.0030

179,180,181,182 CELL NUMBERS CONTAINS THE EDA PERFORMED