ML BASED SEARCH ENGINE ON STACKOVERFLOW DATA :

Problem :

We have Stackoverflow database (https://www.kaggle.com/c/facebook-recruiting-iii-keyword-extraction/data), And we have to built a search engine.

A user will take a query related to the programming questions related to JAVA, C, C++ and C# and our search engine will return the similiar queries that are in our database by use of NLP and ML.

We here are focus mainly in the Search engine characteristics rather than the UI in which result will be given.

DATA SET :

We have mainly 3 fields

1. The title
2. The Tags
3. The Body

What We Want?
We want that after entering a title we get various results related to our query from which we can navigate to our body of title.

We will enter a query example

CONSTRAINTS :

The data we have is in CSV format and has size of 7.2 GB and my machine has only 8GB RAM, So how its gonna work out?
We are gonna use Sqlite3 to access our data.

In [2]:

```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import sqlite3
import csv
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from wordcloud import WordCloud
import re
import os
from sqlalchemy import create_engine # database connection
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import metrics
from sklearn.metrics import f1_score,precision_score,recall_score
from datetime import datetime
```

In [5]:

```python
if not os.path.isfile('train.db'):
    start = datetime.now()
    disk_engine = create_engine('sqlite:///train.db')
    start = datetime.now()
    chunksize = 180000
    j = 0
    index_start = 1
    for df in pd.read_csv('dataset/Train.csv', names=['Id', 'Title', 'Body', 'Tags'], chunksize=chunksize,
        df.index += index_start
        j+=1
        print('{} rows'.format(j*chunksize))
        df.to_sql('data', disk_engine, if_exists='append')
        index_start = df.index[-1] + 1
    print("Time taken to run this cell :", datetime.now() - start)
```

```
180000 rows
360000 rows
540000 rows
720000 rows
900000 rows
1080000 rows
1260000 rows
1440000 rows
1620000 rows
1800000 rows
1980000 rows
2160000 rows
2340000 rows
2520000 rows
2700000 rows
2880000 rows
3060000 rows
3240000 rows
3420000 rows
```

In [7]:

```
con = sqlite3.connect('train.db')
count = pd.read_sql_query("""SELECT count(*) FROM data""", con)
print("Total number of rows in our database is : ", count["count(*)"])
```

```
Total number of rows in our database is :  0    6034196
Name: count(*), dtype: int64
```

That means we have around 6M Datapoints
Lets first remove the duplicated entries

In [9]:

```
con.close()
con = sqlite3.connect('train.db')
df = pd.read_sql_query('SELECT Title, Body, Tags, COUNT(*) as cnt_dup FROM data GROUP BY Title, Body, Tags'
con.close()
df.head()
```

Out[9]:

	Title	Body	Tags	cnt_dup
0	Implementing Boundary Value Analysis of S...	<pre><code>#include&lt;iostream&gt;\n#include&...	c++ c	1
1	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding	1
2	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding columns	1
3	java.lang.NoClassDefFoundError: javax/serv...	<p>I followed the guide in <a href="http://sta...	jsp jstl	1
4	java.sql.SQLException:[Microsoft][ODBC Dri...	<p>I use the following code</p>\n\n<pre><code>...	java jdbc	2

In [10]:

```
print("number of duplicate questions :", count['count(*)'].values[0]- df.shape[0],\
      "(",(1-((df.shape[0])/(count['count(*)'].values[0])))*100,"% )")

```

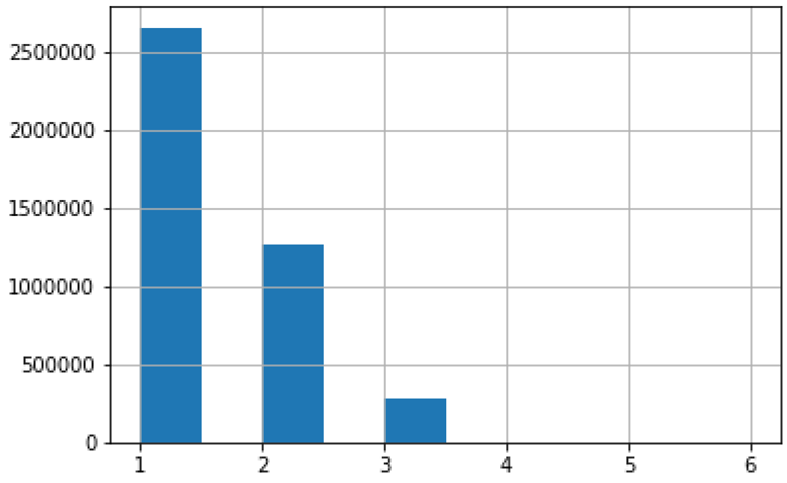
```
number of duplicate questions : 1827881 ( 30.292038906260256 % )
```

In [11]:

```
df.cnt_dup.hist()
```

Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a4c56ce10>
```



In [12]:

```
df.cnt_dup.value_counts()
```

Out[12]:

```
1    2656284
2    1272336
3     277575
4         90
5         25
6          5
Name: cnt_dup, dtype: int64
```

In [14]:

```
# So in real there are around 2.6M Data points which are our concern, Lets save them
```

In [15]:

```
del count
```

In [16]:

```python
df = df[df.cnt_dup < 2]
```

In [50]:

```python
# Now we want to make a search engine for only JAVA, C, C++, C# questions.
# So lets first find out the tags we will need
df = df.reset_index(drop=True)

```

In [56]:

```python
df.head(10)
```

Out[56]:

	Title	Body	Tags	cnt_dup
0	Implementing Boundary Value Analysis of S...	<pre><code>#include&lt;iostream&gt;\n#include&...	c++ c	1
1	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding	1
2	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding columns	1
3	java.lang.NoClassDefFoundError: javax/serv...	<p>I followed the guide in <a href="http://sta...	jsp jstl	1
4	Better way to update feed on FB with PHP SDK	<p>I am a novice with the Facebook API. I have...	facebook api facebook-php-sdk	1
5	"SQL Injection" issue preventing correct for...	<p>So I've been checking everything I can thin...	php forms	1
6	Undefined symbols for architecture i386: _OB...	<p>I have imported framework for sending email...	iphone email-integration	1
7	java.lang.NoSuchMethodError: javax.servlet.S...	<p>i want to have a servlet to process inputs ...	java servlets jboss	1
8	obtaining updated locations using gps in ser...	<p>I have app in which i have two buttons <str...	android android-widget android-service	1
9	"Specified initialization vector (IV) does no...	<p>I've had troubles using an CryptoStream for...	c# .net rijndaelmanaged cryptostream	1

In [204]:

```
from tqdm import tqdm
import multiprocessing
def separate(a, l, r):
    tags = ["c#", "java", "asp.net", "c++", "c", "ios"]
    for i in tqdm(range(l ,r)):
        for k in tags:
            if (df.iloc[i].Tags is not None) and k in df.iloc[i].Tags.split():
                a.append(i)
                break
    return a



start = datetime.now()
# p1 = multiprocessing.Process(target=print_square, args=(10, ))
# p2 = multiprocessing.Process(target=print_cube, args=(10, ))

print(datetime.now() - start)
```

```
0:00:00.001608
```

In [213]:

```
# Now we have only those rows on which we really want to search i.e JAVA, C, C++ and C#, ios

from multiprocessing import Pool

manager = multiprocessing.Manager()
shared_list = manager.list()

p = multiprocessing.Process(target=separate, args=[shared_list, 0,885427])
p2 = multiprocessing.Process(target=separate, args=[shared_list, 885427, 1770854])
p3 = multiprocessing.Process(target=separate, args=[shared_list, 1770854, df.shape[0]])

p.start()
p2.start()
p3.start()
```

```
0%|          | 0/885427 [00:00<?, ?it/s]

0%|          | 0/885427 [00:00<?, ?it/s]

0%|          | 0/885429 [00:00<?, ?it/s]

0%|          | 21/885427 [00:00<1:10:41, 208.74it/s]

0%|          | 17/885427 [00:00<1:27:29, 168.66it/s]

0%|          | 35/885427 [00:00<1:21:19, 181.43it/s]
```

In [224]:

```
b = []
b.extend(shared_list)
len(b)
```

Out[224]:

```
665656
```

In [ ]:

```
# Here by using multiprocessing we completed task of 2.5 hrs in 1hr 10 mins
```

In [228]:

```
# So now we have indices of tags lets get them
df = df.iloc[b]
```

In [231]:

```
df = df.reset_index(drop=True)
```

In [232]:

```
df.head()
```

Out[232]:

	Title	Body	Tags
0	Implementing Boundary Value Analysis of S...	<pre><code>#include&lt;iostream&gt;\n#include&...	c++ c
1	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding
2	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding columns
3	java.lang.NoSuchMethodError: javax.servlet.S...	<p>i want to have a servlet to process inputs ...	java servlets jboss
4	"Specified initialization vector (IV) does no...	<p>I've had troubles using an CryptoStream for...	c# .net rijndaelmanaged cryptostream

In [233]:

```
if not os.path.isfile('no_duplicates.db'):
    disk_dup = create_engine("sqlite:///no_duplicates.db")
    no_dup = pd.DataFrame(df, columns=['Title', 'Body', 'Tags'])
    no_dup.to_sql('no_dup_train',disk_dup)
```

In [3]:

```
con = sqlite3.connect('dataset/no_duplicates.db')
k = pd.read_sql_query("""SELECT * FROM no_dup_train""", con)
con.close()
```

In [4]:

```
k = k.drop(["index"], axis=1)
```

In [5]:

```
k.head()
```

Out[5]:

	Title	Body	Tags
0	Implementing Boundary Value Analysis of S...	<pre><code>#include&lt;iostream&gt;\n#include&...	c++ c
1	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding
2	Dynamic Datagrid Binding in Silverlight?	<p>I should do binding for datagrid dynamicall...	c# silverlight data-binding columns
3	java.lang.NoSuchMethodError: javax.servlet.S...	<p>i want to have a servlet to process inputs ...	java servlets jboss
4	"Specified initialization vector (IV) does no...	<p>I've had troubles using an CryptoStream for...	c# .net rijndaelmanaged cryptostream

In [6]:

```
f = k.Body.iloc[3]
```

In [8]:

```
f
```

Out[8]:

```
'<p>i want to have a servlet to process inputs from a standalone java program. how to deploy this servlet in
jboss. I put the servlet.class file in WEB-INF/classes and in web.xml i gave the servlet url mapping as ".d
o". From my Java client program i opened connected to the servlet using a URL object. using localhost:8080/.d
o. BUT i am getting the folowing error:</p>\n\n<pre>\n  ERROR [org.apache.catalina.connector.CoyoteAdapter] A
n exception or error occurred in the container during the request processing: \n  java.lang.NoSuchMethodErro
r: javax.servlet.ServletContext.getEffectiveSessionTrackingModes()Ljava/util/Set;\n            at\n     org.a
pache.catalina.connector.CoyoteAdapter.postParseRequest(CoyoteAdapter.java:567)\n            at org.apache.ca
talina.connector.CoyoteAdapter.service(CoyoteAdapter.java:359)\n            at org.apache.coyote.http11.Http1
1Processor.process(Http11Processor.java:877)\n            at org.apache.coyote.http11.Http11Protocol$Http11Co
nnectionHandler.process(Http11Protocol.java:654)\n            at org.apache.tomcat.util.net.JIoEndpoint$Worke
r.run(JIoEndpoint.java:951)\n</pre>\n\n<p>web.xml file contents :</p>\n\n<pre><code>&lt;?xml version="1.0" en
coding="UTF-8"?&gt; \n&lt;!DOCTYPE web-app PUBLIC "-//Sun Microsystems, Inc.//DTD Web Application 2.2//EN" "j
ava.sun.com/j2ee/dtds/web-app_2_2.dtd"&gt;; \n&lt;web-app&gt; \n     &lt;servlet&gt;\n         &lt;servlet-na
me&gt;ReverseServlet&lt;/servlet-name&gt; \n         &lt;servlet-class&gt;ReverseServlet&lt;/servlet-class&g
t; \n     &lt;/servlet&gt; \n     &lt;servlet-mapping&gt;\n          &lt;servlet-name&gt;ReverseServlet&lt;/s
ervlet-name&gt; \n          &lt;url-pattern&gt;/*.do&lt;/url-pattern&gt; \n     &lt;/servlet-mapping&gt; \n&l
t;/web-app&gt;\n</code></pre>\n'
```

In [ ]:

```

```

In [ ]:

In [ ]: