Preprocessing For Search Engine

In [1]:

```
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 [2]:

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

In [3]:

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

In [4]:
```
k.head()
```

Out[4]:

	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 [5]:
```
f = k.Body.iloc[3]
```

In [6]:
```
f
```

Out[6]:
```
'<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 [ ]:
```
# Pre Processing
```

In [ ]:

```
# Lets Clean the Title of questions
# There are Redundant Spaces in Beginning
# Removing Stop words as they are not of use
# Removing Curly Brackets
#
```

In [24]:

```
from tqdm import tqdm
from bs4 import BeautifulSoup
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(k.Title.values):
    sentance = re.sub(r"http\S+", "", sentance)
    sentance = BeautifulSoup(sentance, 'lxml').get_text()
    sentance = re.sub("\S*\d\S*", "", sentance).strip()
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not in stopwords.words('english'))
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 665656/665656 [24:18<00:00, 456.37it/s]
```

In [26]:

```
# So Our Questions are cleaned lets make them our default title
k.Title = preprocessed_reviews
```

In [31]:

```
# Lets Clean Our Tags too for This : There are multiple Tags Here so the first tag appearning will be our m
```

In [38]:

```
new_tags = []
for i in tqdm(range(k.shape[0])):
    j = k.Tags.iloc[i].split()[0]
    new_tags.append(j)
```

```
100%|██████████| 665656/665656 [00:12<00:00, 51856.45it/s]
```

In [40]:

```
k.Tags = new_tags
```

In [41]:

```
k.head()
```

Out[41]:

	Title	Body	Tags
0	implementing boundary value analysis software ...	<pre><code>#include&lt;iostream&gt;\n#include&...	c++
1	dynamic datagrid binding silverlight	<p>I should do binding for datagrid dynamicall...	c#
2	dynamic datagrid binding silverlight	<p>I should do binding for datagrid dynamicall...	c#
3	java lang nosuchmethoderror javax servlet serv...	<p>i want to have a servlet to process inputs ...	java
4	specified initialization vector iv match block...	<p>I've had troubles using an CryptoStream for...	c#

In [43]:

```
k.describe()
```

Out[43]:

	Title	Body	Tags
count	665656	665656	665656
unique	650317	659110	351
top		<p>I've now googled around and tried various m...	c#
freq	125	3	216114

In [57]:

```
k.Tags.describe()
```

Out[57]:

```
count      665656
unique        351
top            c#
freq       216114
Name: Tags, dtype: object
```

In [ ]:

```
# Lets See how many tags are there by what amount
```

In [59]:

```
k.groupby(["Tags"]).describe()
```

Out[59]:

	Title				Body			
	count	unique	top	freq	count	unique	top	freq
Tags								
.net	2044	2026	expose net object remote c client	2	2044	2042	<p>How does ASP.NET membership generate their ...	2
.net-framework	3	3	possible bind multiple datatables listview	1	3	3	<p>i'm working a module where i have to create...	1
2007	4	4	visual studio intellisense webdav davwww work ...	1	4	4	<p>We currently are running moss 2007 for an e...	1
2010	6	6	run popup window code behind web part writen c	1	6	6	<p>I have developed WCF service in Visual Stud...	1
2013	1	1	change context using rendercontext sharepoint ...	1	1	1	<p>I am trying to create an ASP workflow task ...	1
64-bit	1	1	running eclipse sdk	1	1	1	<p>I'm running the 32-bit version of Eclipse, ...	1

In [84]:

```
k[k.Tags == "2007"]
```

Out[84]:

	Title	Body	Tags
1654	default reader site group people picker retrie...	<p>I have an issue with people picker for SQL ...	2007
226767	visual studio intellisense webdav davwww work ...	<p>The page lives in a library inside SharePoi...	2007
381218	moss mvc architecture question	<p>We currently are running moss 2007 for an e...	2007
583550	set user selected date filter data view web part	<p>I have a request to set up a user based fil...	2007

In [ ]:

```
# So here we have some little problem as we can see we selected data for only C,C++, Java, ios and C#
# But as we selected the whole row so these little anomalies are there with us. We have to remove them so
# that we can have limited things to predict in our yi_s
```

In [80]:

```
tag_list = ["c#", "java", "c++", "ios", "c"]
```

In [81]:
```
new_indicies = []
for i in tqdm(range(k.shape[0])):
    for j in tag_list:
        if j == k.Tags.iloc[i]:
            new_indicies.append(i)
            break

```

```
100%|██████████| 665656/665656 [00:29<00:00, 22199.21it/s]
```

In [82]:
```
len(new_indicies)
```

Out[82]:
```
572512
```

In [85]:
```
k = k.iloc[new_indicies]
```

In [99]:
```
k.groupby(["Tags"]).describe()
```

Out[99]:

	Title				Body			
	count	unique	top	freq	count	unique	top	freq
Tags								
c	35421	34344	mean	12	35421	34765	<p>I am getting an error that says:</p>\n\n<p>...	3
c#	216114	212215		23	216114	214429	<p>I am looking for a way that would allow me ...	3
c++	90623	88325		30	90623	89547	<p>I have a project for school where we need t...	3
ios	32124	31785	uipopovercontroller dealloc reached popover st...	3	32124	31946	<p>I was reading through the documentation on ...	2
java	198230	193674		42	198230	195993	<p>I am interested in doing this C code in Jav...	3

In [ ]:
```
# Now It looks OKay and In Control
```

In [101]:
```
if not os.path.isfile('processed.db'):
    processed = create_engine("sqlite:///processed.db")
    k.to_sql('processed',processed)
```

In [104]:

```
con = sqlite3.connect('processed.db')
processed = pd.read_sql_query("""SELECT * FROM processed""", con)
con.close()
```

In [105]:

```
processed.head()
```

Out[105]:

```
   index                                           Title                                              Body  Tags
0      0  implementing boundary value analysis software ...  <pre><code>#include&lt;iostream&gt;\n#include&...   c++
1      1                 dynamic datagrid binding silverlight     <p>I should do binding for datagrid dynamicall...    c#
2      2                 dynamic datagrid binding silverlight     <p>I should do binding for datagrid dynamicall...    c#
3      3    java lang nosuchmethoderror javax servlet serv...      <p>i want to have a servlet to process inputs ...  java
4      4          specified initialization vector iv match block...   <p>I've had troubles using an CryptoStream for...    c#
```

In [106]:

```
del k
```

In [108]:

```
# Now We have Processed DB
# We also have Few duplicated rows but let them be cause they will come in top of our result in search quer
```

In [ ]:

```
# Now Lets Make Search Engine out of our Data in new Notebook
```