

Data Source : <https://www.kaggle.com/sajid576/sql-injection-dataset> (<https://www.kaggle.com/sajid576/sql-injection-dataset>)

The Primary goal of this case study to classify SQL injection attack

We have 30920 sample data and 1 class label(0/1). The 0 indicates "No SQL injection" found and 1 indicates "SQL Injection Found"

Import Necessary Libraries

In []:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In []:

```
data = pd.read_csv("Modified_SQL_Dataset.csv")
```

In []:

```
data.head()
```

Out[]:

	Query	Label
0	" or pg_sleep (__TIME__) --	1
1	create user name identified by pass123 tempora...	1
2	AND 1 = utl_inaddr.get_host_address (...	1
3	select * from users where id = '1' or @ @1 ...	1
4	select * from users where id = 1 or 1#" (...	1

In []:

```
data.shape
```

Out[]:

(30919, 2)

In []:

```
data['Label'].value_counts()
```

Out[]:

0 19537
1 11382
Name: Label, dtype: int64

We found 19537 number of rows as "NO SQL Injection attack" and 11382 found as SQL injection attack

In []:

```
data.duplicated().any()
```

Out[]:

True

In []:

```
df = data[data.duplicated()]
df
```

Out[]:

	Query	Label
866	#NAME?	1
13532	#NAME?	0
14626	1.94E+15	0
15635	#NAME?	0
17255	1.86E+15	0
17913	#NAME?	0
18233	#NAME?	0
18875	#NAME?	0
18961	7.75E+15	0
19328	26%	1
19332	29%	1
19333	28%	1

Observation : We found duplicated rows from our dataset

let's drop the duplicated entry

In []:

```
data.drop_duplicates(inplace=True)
data.head()
```

Out[]:

	Query	Label
0	" or pg_sleep (__TIME__) --	1
1	create user name identified by pass123 tempora...	1
2	AND 1 = utl_inaddr.get_host_address (...	1
3	select * from users where id = '1' or @ @1 ...	1
4	select * from users where id = 1 or 1#" (...	1

In []:

```
data.shape
```

Out[]:

(30907, 2)

In []:

```
data.duplicated().any()
```

Out[]:

False

Observation : We have removed the all duplicate rows. And Found 30907 number of sample data and 2 features value

Let' start to find Null value

In []:

```
data.isnull().any()
```

Out[]:

```
Query    False
Label    False
dtype: bool
```

Observation : We did not find any rows have null value.

Exploratory Data Analysis(EDA)

In []:

```
sns.set(rc={'figure.figsize':(30,10)})
plt.subplot(1,2,1)
ax = sns.countplot(data['Label'])
for p in ax.patches:
    ax.annotate('{:.1f}%'.format(100*p.get_height()/len(data)), (p.get_x()+0.35, p.get_height()+1))
ax.set_title('count plot of target distribution')
ax.set_xlabel('Label')
ax.set_ylabel('count')

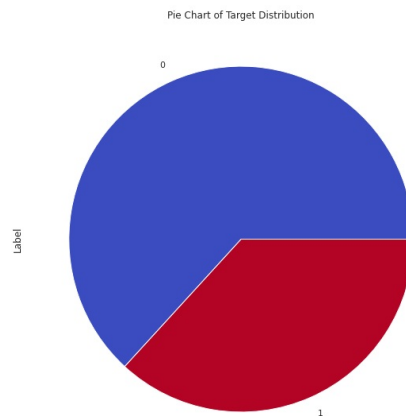
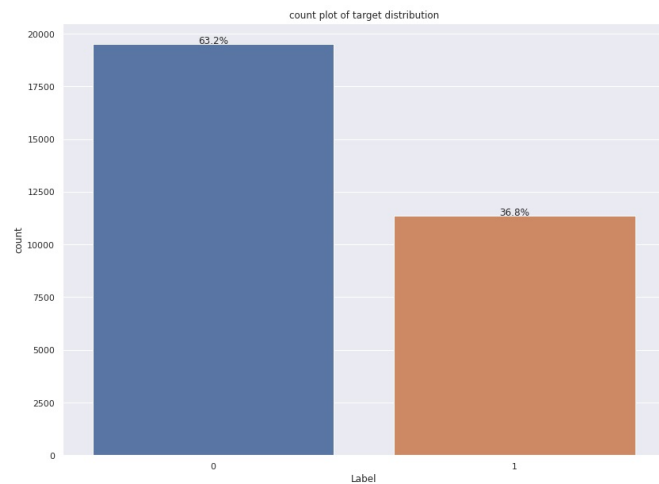
plt.subplot(1,2,2)
ax = data['Label'].value_counts().plot(kind='pie',colormap='coolwarm')
ax.set_title("Pie Chart of Target Distribution")
```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation

FutureWarning

Out[]:

Text(0.5, 1.0, 'Pie Chart of Target Distribution')



Observation : Here, it is clearly said that data points belonging to "No SQL injection" attack is 63.2% and "SQL injection attack" is 36.8%.

Let's Do Text Preprocessing

Import Neccesary Libraries

In []:

```
import re
import nltk
import string
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
from gensim.models import word2vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os
```

In []:

```
#printing some random variable
sent_0 = data['Query'].values[0]
print(sent_0)
print("="*50)

sent_2000 = data['Query'].values[2000]
print(sent_2000)
print("="*50)

sent_15000 = data['Query'].values[15000]
print(sent_15000)
print("="*50)

sent_20000 = data['Query'].values[20000]
print(sent_20000)
print("="*50)

" or pg_sleep ( __TIME__ ) --
=====
-6073" ) ) ) or 9502 = 2012#
=====
wilson@autoconstruccion.fi
=====
SELECT SUBSTR ( "SQL Tutorial", -5, 5 ) AS ExtractString;
=====
```

In []:

```
#remove the special Character : https://stackoverflow.com/a/5843547/4084039
sent_15000 = re.sub('[^A-Za-z0-9]+',' ',sent_15000)
print(sent_15000)
```

wilson autoconstruccion fi

In []:

```
sent_20000 = re.sub('[^A-Za-z0-9]+',' ',sent_20000)
print(sent_20000)
```

```
SELECT SUBSTR SQL Tutorial 5 5 AS ExtractString
```

In []:

```
stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",\
               "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', \
               'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', \
               'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',\
               'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does',\
               'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of',\
               'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'aft\
               er',\
               'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'fu\
               rther',\
               'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few',\
               'more',\
               'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
               's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', '\
               re', \
               've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn\
               ',\
               "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn',\
               "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "\
               weren't", \
               'won', "won't", 'wouldn', "wouldn't"])
```

In []:

```
# Combining all the above students
from tqdm import tqdm
preprocessed_query = []
# tqdm is for printing the status bar
for sentence in tqdm(data['Query'].values):
    sentence = re.sub('[^A-Za-z0-9]+', ' ', sentence)
    sentence = re.sub(r',', ' ', sentence)
    #https://www.machinelearningplus.com/nlp/lemmatization-examples-python/
    tokenization = nltk.word_tokenize(sentence)
    sentence = ' '.join([lemmatizer.lemmatize(w) for w in tokenization])
    # https://gist.github.com/sebleier/554280
    sentence = ' '.join(e.lower() for e in sentence.split() if e.lower() not in stopwords)
    preprocessed_query.append(sentence.strip())
```

100%|██████████| 30907/30907 [00:05<00:00, 6061.41it/s]

In []:

preprocessed_query

Out[]:

```
['pg sleep time',
 'create user name identified pass123 temporary tablespace temp default tablespace user',
 '1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 5',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 1 union select 1 version 1',
 'select name syscolumns id select id sysobjects name tablename',
 'select user id 1 1 1 1',
 '1 load file char 47 101 116 99 47 112 97 115 115 119 100 1 1 1',
 'select user id 1 1 1 1 union select 1 version 1',
 'select user id 1 union select 1 version 1',
 '1 1',
 '',
 'admin 1 1',
 'select user id 1 1 1 1',
 '1 1',
 '1 1',
 '1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 5',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'distinct',
 'select user id 1 1 1 1',
 '1 ascii lower substring select top 1 name sysobjects xtype u 1 1 116',
 'select user id 1 1 1 1',
 'admin 1 1',
 'select user id 1 1 1 1',
 'insert',
 'select user id 1 1 1 1 1 1',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 8',
```

```

'select user id 1 1 1 1',
'1 1',
'1 1',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select version 1',
'1 1',
'27 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select version 1',
'1 load file char 47 101 116 99 47 112 97 115',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select 1 1 version 1',
'select user id 1 1 union select null banner v version rownum 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'1',
'select user id 1 1 1 1',
'admin 1 1',
'select user id 1 1 union select null banner v version rownum 1 1',
'x27union select',
'print',
'select user id 1 union select version 1',
'1 1',
'1 pg sleep time',
'select user id 1 1 1 1',
'admin 1 1',
'2 1 3',
'select user id 1 1 1 1 1',
'select user id 1 1 union select null version 1',
'admin 1 1',
'not select system user sa waitfor delay 0 0 2',
'select user id 1 union select 1 banner v version rownum 1 1',
'0 0',
'select user id 1 union select version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1',
'char 4039 41 2b 40select',
'select user id 1 1 1 1 1',
'admin 1 1',
'declare varchar 200 select 0x73656c6',
'0 0',
'select user id 1 1 1 1',
'x x',
'select user id 1 union select 1 version 1',
'sleep 50',
'select user id 1 1 union select null banner v version rownum 1 1',
'1 union select 1 2 3 4 5 6 name sysobjects xtype u',
'select user id 1 1 union select version 1',
'select user id 1 1 union select null version 1',
'select user id 1 1 1 1 1',
'1 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 4',
'select user id 1 union select version 1',
'1 user name dbo',
'not substring select version 24 1 1 waitfor delay 0 0 2',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 union select null version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1 1',
'1 utl inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 6',
'1 1',
'select user id 1 1 1 1',
'text',
'1234 1 0 union select admin 81dc9bdb52d04dc20036dbd8313ed055',
'select user id 1 1 1 1',
'select user id 1 1 1 union select 1 version 1',
'admin 1 1',
'select user id 1 1 1 1 1',
'',
'1 utl inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 4',
'select information schema table',

```

```

'select user id 1 1 1 1',

'1 pg sleep time',
'select user id 1 union select version 1',
'declare q nvarchar 200 select q 0x770061',
'1 1',
'select user id 1 1 1 1 1',
'declare q nvarchar 200 select q 0x770061006900740066006f0072002000640065006c0061007900200027003000
3a0030003a0031003000270000 exec q',
'select user id 1 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'unusual unusual',
'1 select version',
'truncate',
'',
'3 3',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'select user id 1 union select 1 version 1',
'admin 1 1',
'select user id 1 1 union select version 1',
'1 benchmark 10000000 md5 1',
'select user id 1 union select 1 version 1',
'1 utl inaddr get host address select sys database name dual',
'0 0',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1',
'not substring select version 25 1 0 waitfor delay 0 0 2',
'1 sleep time',
'select user id 1 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1',
'true',
'execute immediate sel ect u er',
'hi x x',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'admin',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'union select 1 load file etc passwd 1 1 1',
'waitfor delay 0 0 time',
'unusual unusual',
'insert mysql user user host password value name localhost password pass123',
'admin 1 1',
'select user id 1 1 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'like',
'username like char 37',
'pg sleep time',
'select user id 1 1 1 1',
'3 3',
'select user id 1 1 1 1',
'select user id 1 1 1 1 1',
'exec sp',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'admin 1 1',
'waitfor delay 0 0 10',
'select user id 1 union select null version 1',
'select user id 1 union select 1 version 1',
'1 select var temp',
'hi',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 1',
'admin 1 1',
'true',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'1 pg sleep time',
'select user id 1 1 1 1',
'sleep time',
'select user id 1 1 1 1',
'admin 1 1',
'select user id 1 union select 1 version 1',
'exec xp regread',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1',
'1 1 1',
'0 0',

```

```

'union select user login char',

'select user id 1 union select 1 version 1',
'admin 1 1',
'1 select',
'select user id 1 1 1 1',
'x 1 select count tablename',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'1 benchmark 10000000 md5 1',
'procedure',
'',
'1 utl inaddr get host address select sys login user dual',
'1 utl inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 2',
'desc user',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 1 1 1',
'',
'',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'admin',
'select user id 1 1 1 1',
'utl http request',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 1',
'select user id 1 union select 1 version 1',
'benchmark 10000000 md5 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'1 uni select',
'select user id 1 union select null version 1',
'select user id 1 1 union select null version 1',
'admin 1 1',
'utl http request http 192 168 1 1',
'7659 7659',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'1 utl inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 6',
'select user id 1 1 1 1',
'true',
'select user id 1 union select 1 version 1',
'select user id 1 union select version 1',
'select user id 1 1 union select 1 version 1',
'true',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'3 3',
'1 utl inaddr get host address select global name global name',
'select user id 1 1 union select null banner v version rownum 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 union select null version 1',
'select user id 1 1 union select version version 1',
'select user id 1 1 1 1 1',
'uid like',
'select user id 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'23 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 union select version 1',
'0 1 1',
'select user id 1 1 1 1',
'admin 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1',
'1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 5',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 7',
'apos',
'sqlvuln',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
.

```



```

'',
'hi 1 1',
'select user id 1 union select null banner v version rownum 1 1',
'exec master xp cmdshell ping 172 10 1 255',
'select user id 1 1 1 union select 1 banner v version rownum 1 1',
'select user id 1 union select version 1',
'select information schema table',
'select user id 1 1 1 1',
'select top 1',
'1 1',
'',
'1 1',
'select user id 1 1 union select null version 1',
'select user id 1 union select version 1',
'',
'exec sp addlogin name password',
'select user id 1 union select null banner v version rownum 1 1',
'0x77616974666f722064656c61792027303a303a31302700 exec',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'group userid 1 1',
'select user id 1 1 union select null version 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1 1',
'1 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1',
'',
'declare varchar 22 select',
'select user id 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 1 union select version version 1',
'variable',
'exec master xp cmdshell',
'select user id 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select 1a version 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'1 select var temp',
'select user id 1 union select 1 banner v version rownum 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1',
'x x',
'',
'1 1',
'select user id 11 1 union select 1 version 1',
'x x',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 4',
'srvrolemember sysadmin 0 waitfor delay 0 0 2',
'select user id 1 1 1 1',
'select user id 1 union select version 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 union select version version 1',
'1 1',
'',
'x x',
'select user id 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'password',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 union select null version 1',
'1 utl inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 3',
'1 1',
'admin 1 1',
'select user id 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1',
'union select user login char 114 111 111 116',
'select user id 1 1 union select 1 version 1',
'1 1',
'1 utl inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 5',
'pg sleep time',
'3 3',
'select user id 1 1 union select 1 version 1',

```

```

'select user id 1 1 1 1',

'union select information schema table',
'select user id 1 1 1 1',
'benchmark 10000000 md5 1',
'1 utl inaddr get host address select count distinct column name sys tab column',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 7',
'',
'delete',
'waitfor delay 0 0 time',
'1 1',
'1 non existant table 1',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'declare varchar 200 select 0x77616974',
'x userid null',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select version 1',
'select user id 1 union select 1 version 1',
'1 utl inaddr get host address select host name v instance',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 2',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'union select',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'x full name like bob',
'x member email null',
'2 1',
'select user id 1 1 1 1 1',
'pg sleep time',
'',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'sleep time',
'sleep 50',
'select user id 1 1 1 union select 1 banner v version rownum 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'sleep time',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select version 1',
'select user id 1 union select version 1',
'select user id 1 union select null version 1',
'2 1',
'exec sel ect u er',
'anything x x',
'select user id 1 union select 1 version 1',
'select user id 1 union select 1 banner v version rownum 1 1',
'select user id 1 union select null version 1',
'',
'3 3',
'select user id 1 union select 1 version 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'admin 1 1',
'uni sel ect',
'',
'waitfor delay 0 0 time',
'select user id 1 1 1 1',
'1 waitfor delay 0 0 10',
'bfilename',
'admin 1 1',
'uef',
'password',
'1 sleep time',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 6',
'select user id 1 union select version 1',
'select user id 1 1 union select 1 version 1',
'1 1 select count tablenames',
'sleep time',
'1 utl inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 8',
'select user id 1 1 1 1 1 1',
'6',
'1',
'1 load file char 110 46 101 120 116 char 39 39 1 0',

```

```

'select user id 1 1 1 1',

'select user id 1 1 1 1 1',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'1 1',
'1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 7',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'whatever whatever',
'declare varchar 200 select 0x73656c6563742040407665727369666e exec',
'select user id 1 1 1 1',
'',
'1 utl inaddr get host address select count distinct table name sys table',
'select user id 1 union select 1 version 1',
'admin 1 1',
'select user id 1 1 union select version 1',
'select user id 1 1 1 1',
'',
'select user id 1 1 1 1',
'1 1',
'drop table temp',
'1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 2',
'benchmark 10000000 md5 1',
'x x',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'something like',
'not select serverproperty isintegratedsecurityonly 0 waitfor delay 0 0 2',
'1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 4',
'select user id 1 1 1 1',
'exec master xp cmdshell ping 10 10 1 2',
'admin 1 1',
'not select serverproperty isintegratedsecurityonly 1 waitfor delay 0 0 2',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 2',
'select user id 1 union select version version 1',
'select user id 1 1 union select null version 1',
'waitfor delay 0 0 time',
'x userid null',
'select user id 1 1 1 1 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1',
'1 1',
'select user id 1 union select 1a banner v version rownum 1 1',
'admin 1 1',
'select user id 1 1 union select 1 version 1',
'1 load file char 110 46 101 120 11',
'exec',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'1 1',
'select user id 1 1 1 1 1',
'select user id 1 union select version version 1',
'x email null',
'select user id 1 1 union select 1 version 1',
'',
'select user id 1 union select null banner v version rownum 1 1',
'select user id 11 1 union select 1 banner v version rownum 1 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'update',
'1',
'select user id 1 1 union select 1 version 1',
'pg sleep time',
'6',
'select user id 1 union select 1 version 1',
'exists',
'',
'text n text',
'select user id 1 union select null version 1',
'select user id 1 union select version 1',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'0 0',
'select user id 1 1 1 1 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 union select version version 1',
'select user id 1 1 union select version 1',
'password 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',

```

```

'select user id 1 union select 1 version 1',

'union select null select version',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'1 utl inaddr get host address select banner v version rownum 1',
'select user id 1 1 1 1',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 3',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'union select',
'asc',
'myappadmin adduser admin newpass',
'exec sp addsrvrolemember name sysadmin',
'select user id 1 1 union select 1 version 1',
'true',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 3',
'userid like',
'union select null null select version',
'',
'select user id 1 1 1 1',
'admin',
'select user id 1 union select 1 version 1',
'waitfor delay 0 0 time',
'select user id 1 union select version 1',
'select user id 1 1 1 1',
'admin 1 1',
'union select version',
'',
'sleep time',
'union select null null null null null select version',
'select user id 1 union select 1 version 1',
'select user id 1 union select version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 union select null version 1',
'select user id 1 1 1 1',
'hi',
'select name syscolumns id sele',
'select user id 1 1 1 1 1',
'not substring select version 25 1 5 waitfor delay 0 0 2',
'select user id 1 union select 1 version 1',
'2 1 3',
'union select version',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 8',
'exec master xp cmdshell nslookup www google com',
'isnull 1 0',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'declare q nvarchar 200 0x730065006c00650063',
'',
'union select version',
'select user id 1 1 1 union select 1 version 1',
'admin 1 1',
'exists',
'select user id 1 union select 1 version 1',
'declare varchar 8000 select 0x73656c',
'select user id 1 union select version 1',
'select user id 1 union select 1 banner v version rownum 1 1',
'select user id 1 1 1 1',
'select user id 1 union select 1 la version 1',
'begin declare var varchar 8000 set var select var var login password user login',
'union select null null null null select version',
'hi 1 1',
'union select',
'select user id 1 1 1 1',
'select user id 1 1 union select version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1 1',

```

```

'desc',

'select user id 1 1 1 1 1',
'anything x x',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'0x770061006900740066006f0072002000640065006c00',
'select user id 1 union select version version 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'isnull 1 0',
'insert user login password level value char 0x70 char 0x65 char 0x74 char 0x65 char 0x72 char 0x70
char 0x65 char 0x74 char 0x65 char 0x72 char 0x64',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1',
'admin 1 1',
'admin 1 1',
'user like',
'admin 1 1',
'sleep time',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'',
'select user id 1 union select version 1',
'admin 1 1',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 3',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'x 1 1 x',
'1 utl inaddr get host address select distinct table name select distinct table name rownum limit s
ys table limit 8',
'select user id 1 union select 1 version 1',
'x x',
'',
'select user id 1 1 union select version version 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select null version 1',
'select user id 1 union select version 1',
'select user id 1 1 union select null version 1',
'select user id 1 union select null version 1',
'select user id 1 1 union select null version 1',
'sleep time',
'1 1',
'',
'select user id 1 1 1 1 1 1',
'group userid 1 1',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'0 0',
'0 0',
'pg sleep time',
'1 1',
'select user id 1 1 1 union select version 1',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'1 utl inaddr get host address select count distinct username sys users',
'sqlvuln',
'select user id 1 1 1 1',
'select user id 1 1 union select version 1',
'0x730065006c0065006300740020004000400076006500',
'select user id 1 union select 1 1 version 1',
'select user id 1 union select version version 1',
'declare varchar 8000 select 0x73656c6563742040407665727369666e',
'select user id 1 1 1 1',
'',
'select user id 1 1 1 1',
'1 utl inaddr get host address select distinct column name select distinct column name rownum limit
tab column limit 6',
'select user id 1 union select 1 version 1',
'username like char 37',
'select user id 1 1 union select version version 1',
'something thing',
'name',
'1 select version',
'declare q nvarchar 4000 select q',
'select user id 1 union select 1 version 1',
'1 utl inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 2',
'select user id 1 1 1 1',
'select user id 1 union select null version 1',
'select user id 1 1 1 1 1',

```

```

'1 utl_inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 1',
'select user id 1 1 union select version version 1',
'select user id 1 union select la banner v version rownum 1 1',
'select user id 1 1 union select 1 version 1',
'admin 1 1',
'',
'select user id 1 1 1 1',
'1 1',
'declare q nvarchar 200 0x730065006c00650063007400200040004000760065007200730069006f006e00 exec q',
'admin',
'select user id 1 1 1 1 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 1 1',
'begin declare var varchar 8000 set var',
'select user id 1 1 1 1',
'pg sleep time',
'print variable',
'select user id 1 1 1 1 1',
'',
'x x',
'select user id 1 1 1 1 1',
'1 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'',
'2a 28 7c 28mail 3d 2a 29 29',
'select user id 1 1 1 1',
'1 1',
'x 1 select count tablename',
'1 utl_inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 3',
'select user id 1 union select 1 version 1',
'admin 1 1',
'',
'select user id 1 1 1 union select null version 1',
'select user id 1 union select 1 banner v version rownum 1 1',
'x x',
'union select',
'23 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'username like',
'elt 3 5 bin 15 ord 10 hex char 45',
'uname like',
'declare varchar 200 select 0x77616974666f722064656c61792027303a303a31302700 exec',
'1',
'select user id 1 union select null version 1',
'1 utl_inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 1',
'select user id 1 union select version 1',
'select user id 1 1 1 1',
'1 utl_inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 1',
'select user id 1 1 1 1',
'admin 1 1',
'1 sleep time',
'select user id 1 1 1 1',
'sqlattemp2',
'select user id 1 union select null version 1',
'x email null',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'1 utl_inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 8',
'select user id 1 1 union select version version 1',
'1234 1 0 union select admin 81dc9bdb52d04dc20036dbd8313ed055',
'select user id 1 union select 1 version 1',
'var select var var temp end',
'select user id 1 union select null version 1',
'',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 union select version 1',
'',
'waitfor delay 0 0 time',
'',
'select user id 1 1 1 1',
'select user id 1 1 1 1',
'1 utl_inaddr get host address select distinct granted role select distinct granted role rownum lim
it dba role privs grantee sys loginuser limit 7',

```

```

'select user id 1 union select 1 version 1',

'union select',
'select user id 1 1 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 1 1 1 1',
'select user id 1 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'exec master xp cmdshell nslookup www googl',
'benchmark 10000000 md5 1',
'union select',
'select user id 1 1 1 union select 1 version 1',
'select user id 1 1 1 1',
'',
'x member email null',
'select',
'1 1',
'',
'select user id 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'admin 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 union select 1 banner v version rownum 1 1',
'1 utl inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 5',
'1 utl inaddr get host address select count distinct granted role dba role privs grantee sys login
user',
'select user id 1 1 union select null version 1',
'1 utl inaddr get host address select distinct password select distinct password rownum limit sys u
ser limit 6',
'select user id 1 1 1 1',
'union select null null null select version',
'x x',
'select user id 1 union select version 1',
'1 1',
'select user id 1 1 1 1',
'3 3',
'1 benchmark 10000000 md5 1',
'1 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1 1',
'admin 1 1',
'text',
'',
'select user id 1 1 union select version 1',
'benchmark 10000000 md5 1',
'exec xp',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1 1',
'hi',
'',
'sqlattempt1',
'admin 1 1',
'select user id 1 1 1 1',
'1 1 1',
'select user id 1 1 union select 1 banner v version rownum 1 1',
'variable',
'2a 7c',
'select user id 1 union select 1 version 1',
'mail',
'not substring select version 25 1 8 waitfor delay 0 0 2',
'select user id 1 1 union select 1 version 1',
'select user id 1 1 1 union select version version 1',
'1 utl inaddr get host address select count distinct password sys user',
'select user id 1 1 union select 1 version 1',
'order',
'select user id 1 1 1 1',
'username not null username',
'admin 1 1',
'',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select version version 1',
'select user id 1 1 1 1 1',
'select user id 1 union select 1 version 1',
'',
'1 0 union',
'',
'0x770061006900740066006f0072002000640065006c00610079002000270030003a0030003a',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 union select 1 banner v version rownum 1 1',

```

```

'sqlvuln',
'text n text',
'1 1',
'0x730065006c00650063007400200040004000760065007200730069006f006e00 exec q',
'select user id 1 1 union select 1 version 1',
'select user id 1 union select version 1',
'select user id 1 1 1 1 1 1',
'select user id 1 1 1 1',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'1 exec sp exec xp',
'select user id 1 1 1 union select 1 banner v version rownum 1 1',
'3 3',
'admin',
'select top 1',
'select user id 1 1 1 1 1',
'',
'benchmark 10000000 md5 1',
'select user id 1 1 1 1',
'7659 7659',
'select user id 1 1 union select version version 1',
'select user id 1 1 1 1 1',
'1 1 1',
'',
'replace',
'select user id 1 union select 1 version 1',
'select user id 1 1 1 1',
'select user id 1 union select null version 1',
'select user id 1 union select la banner v version rownum 1 1',
'',
'objectclass',
'1 utl inaddr get host address select distinct granted role select distinct granted role rownum limit
it dba role privs grantee sys loginuser limit 4',
'limit',
'1 1',
'select user id 1 1 1 1',
'',
'1 utl inaddr get host address select distinct username select distinct username rownum limit sys u
sers limit 7',
'create user name identified pass123',
'select user id 1 union select null version 1',
'select user id 1 1 union select version 1',
'',
'select user id 1 1 1 1 1',
'',
'select user id 1 1 1 1 1',
'x full name like bob',
'select user id 1 union select version 1',
'select user id 1 1 1 1',
'select user id 1 union select version 1',
'benchmark 10000000 md5 1',
'waitfor delay 0 0 time',
'sleep time',
'wapiti',
'sleep time',
'sleep time',
'pg sleep time',
'sleep time',
'sleep time',
'pg sleep time',
'pg sleep time',
'sleep time',
'waitfor delay 0 0 time',
'1 8156 select count generate series 1 5000000',
'1 clye 7842 7842 char 109 char 79 char 70 char 90 regexp substring repeat right char 5012 0 500000
0000 null',
'4860 azyx 6901 6901 union select 6901 6901 6901 6901 6901',
'1 union select null null null null null',
'select benchmark 5000000 md5 0x4c4d6142 6866 6866',
'1 boed 6787 6787',
'1 vdbf 7969 7969 extractvalue 1297 concat 0x5c 0x7171706a71 select elt 1297 1297 1 0x717a767a71',
'1 4281 4281',
'1 8024 3560',
'1 5452 6050 6050 ciyc like ciyc',
'1 8148 like abcdefg upper hex randomblob 500000000 2',
'1 select case 9443 9443 sleep 5 else 9443 select 9443 information schema character set end',
'1 select rawn dual 4988 4988 char 68 char 69 char 97 char 85 regexp substring repeat right char 53
89 0 5000000000 null',
'4925 union select 5686 5686 5686 5686 5686 5686 5686',
'1 dnhd 2657 2657 4240 select 4240 pg sleep 5',
'9534 3038 3038',
'1 6941 6941 6537 dbms pipe receive message chr 76 chr 116 chr 117 chr 65 5',
'select case 7978 6009 7978 else 1 select 0 end',
'.....'

```



```
'1212 make set 7588 2306 2306',
'1 5466 5466 2388 benchmark 5000000 md5 0x6d457153',
'3794 union select 2485 2485 2485 2485 2485',
'1 5286 select count user t1 user t2 user t3 user t4 user t5 gnil gnil',
'1 8635 select count generate series 1 5000000',
'6400 union select 4650 4650 4650',
'1 select rttq dual 7368 7368 updatexml 1808 concat 0x2e 0x7171706a71 select elt 1808 1808 1 0x717a767a71 8666',
'1 ztkr 1532 1532',
'1 9842 9842 union select null null null null null null null null',
'1 exp select select concat 0x7171706a71 select elt 6270 6270 1 0x717a767a71 0x78 x fbsi like fbsi',
',
'1 char 68 char 69 char 97 char 85 regexp substring repeat right char 5389 0 5000000000 null uwep u
wep',
'4291 5023 ctxsys drithsx sn 5023 chr 113 chr 113 chr 112 chr 106 chr 113 select case 5023 5023 1 e
lse 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113 xyhi xyhi',
'1 boolean mode 3707 select count sysibm systables t1 sysibm systables t2 sysibm systables t3',
'1 1022 select count user t1 user t2 user t3 user t4 user t5',
'call regexp substring repeat left crypt key char 65 char 69 char 83 null 0 5000000000 null pawh paw
h',
'1589 1589 1',
'select count generate series 1 5000000 7240 7240',
'1 3715 char 113 char 113 char 112 char 106 char 113 select case 3715 3715 char 49 else char 48 end
char 113 char 122 chr 118 chr 122 chr 113 9548 9548',
'1 6240 qppjq select case 6240 6240 1 else 0 end rdb database qzvzq 6406 6406',
'5021 select yadq 4285 4285 order 1',
'1 4386 utl inaddr get host address chr 113 chr 113 chr 112 chr 106 chr 113 select case 4386 4386 1
else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113',
'1 6793 select 6793 pg sleep 5 baia baia',
'1 3824 benchmark 5000000 md5 0x76555642 jozh jozh',
'9508 union select 9950 9950 9950 9950 9950 9950 9950',
'1 8514 benchmark 5000000 md5 0x544d5a4c',
'1064 make set 6439 2937 2937 qojd qojd',
'1 rbpx 1264 1264',
'1 exp select select concat 0x7171706a71 select elt 6270 6270 1 0x717a767a71 0x78 x',
'8284 6171 select 8284 else drop function mbih',
'1 sleep 5',
'7868 9323 9323',
'1 ekjw 5477 5477 union select null null null null null',
'1 2462 2462 2716 select count sysusers sys1 sysusers sys2 sysusers sys3 sysusers sys4 sysusers sys
5 sysusers sys6 sysusers sys7',
'select like abcdefg upper hex randomblob 500000000 2 mfib mfib',
'1 4240 select 4240 pg sleep 5',
'1 4595 4595',
'1 oknw 8777 8777',
'1 6501 6501',
'1 char 111 char 77 char 121 char 88 regexp substring repeat left crypt key char 65 char 69 char 83
null 0 5000000000 null 8929 8929',
'7999 8422 1336',
'5742 1314 1314 5903 qppjq select case 5903 5903 1 else 0 end rdb database qzvzq',
'1 potk 5040 5040 elt 5873 5873 sleep 5',
'4605 union select 8542 8542 8542 8542 8542 8542 8542',
'select count sysibm systables t1 sysibm systables t2 sysibm systables t3 njnr njnr',
'select case 6558 4327 1 else null end',
'7535 2724 char 113 char 113 char 112 char 106 char 113 select case 2724 2724 char 49 else char 48
end char 113 char 122 chr 118 chr 122 chr 113',
'end rqay like rqay',
'2312 union select 5282 5282 5282 5282 5282 5282 5282',
'1 8594 select 8594 pg sleep 5',
'select case 8993 8846 8993 else 8993 select 8993 mysql db end',
'select count sysibm systables t1 sysibm systables t2 sysibm systables t3',
'select pg sleep 5',
'call regexp substring repeat right char 3702 0 5000000000 null 4142 4142',
'8153 qh1b 4948 4948 union select 4948 4948 4948 4948 4948 4948',
'7319 4493 utl inaddr get host address chr 113 chr 113 chr 112 chr 106 chr 113 select case 4493 449
3 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113',
'1 bedq 8781 8781 8466 benchmark 5000000 md5 0x694a4745',
'1 2006 2006',
'5139 union select 3373 3373 3373 3373 3373 3373 3373',
'1 4411 select count sysusers sys1 sysusers sys2 sysusers sys3 sysusers sys4 sysusers sys5 sysusers
sys6 sysusers sys7 zcyc',
'1 rvch 1863 1863 8384 like abcdefg upper hex randomblob 500000000 2',
'1 8514 select count domain domain t1 domain column t2 domain table t3 loao',
'9446 wmrq 3705 3705 union select 3705',
'iif 7889 5114 1 1 0',
'select dbms pipe receive message chr 66 chr 67 chr 79 chr 101 5 dual ztmd ztmd',
'1 union select null null null null null null null',
'1 char 109 char 79 char 70 char 90 regexp substring repeat right char 5012 0 5000000000 null',
'1 elt 4249 4249 7259',
'1 4867 4867 rlike select case 7689 7689 1 else 0x28 end',
'8858 5680 select 8858 else drop function pyuo',
'1 procedure analyse extractvalue 9255 concat 0x5c benchmark 5000000 md5 0x52515a50 1 1748 1748',
'1 klie 2840 2840 8514 benchmark 5000000 md5 0x544d5a4c',
```

```
'3859 3440 cast chr 113 chr 113 chr 112 chr 106 chr 113 select case 3440 3440 1 else 0 end text chr
113 chr 122 chr 118 chr 122 chr 113 numeric 5846 5846',
'end qkkn like qkkn',
'1 pzoo 8036 8036 6793 select 6793 pg sleep 5',
'2120 8734 8844',
'1 lomw 9257 9257 union select null null null null null null',
'1 6784 6784 elt 3114 3114 sleep 5',
'1 select syr3 7699 7699 union select null null null',
'1 4595 4595',
'4984 union select 6980 6980 6980 6980 6980 6980 6980 6980',
'5224 1962 1962 union select 1962 1962 1962 1962 1962 1962 1962 1962 1962',
'1 8514 select count domain domain t1 domain column t2 domain table t3',
'1 select gboi 4191 4191 8514 select count domain domain t1 domain column t2 domain table t3',
'1 union select null null null null null',
'1 2633 dbms pipe receive message chr 112 chr 65 chr 65 chr 103 5 xmnd xmnd',
'1 row 6237 7469 select count concat 0x7171706a71 select elt 6237 6237 1 0x717a767a71 floor rand 0
2 x select 5192 union select 3785 union select 3931 union select 7158 group x ejul ejul',
'3721 union select 9050 9050',
'1 paai 4089 4089 3707 select count sysibm systables t1 sysibm systables t2 sysibm systables t3',
'1 elt 3114 3114 sleep 5',
'1 order 1',
'9281 8363 8363 make set 8220 5127 5127',
'1 6671 6671 char 119 char 100 char 99 char 121 regexp substring repeat right char 1441 0 5000000000
0 null',
'end vwbx vwbx',
'1 boolean mode 8189 select count sysibm systables t1 sysibm systables t2 sysibm systables t3',
'1 rlike sleep 5 iwct iwct',
'9087 order 1',
'7562 8571 8571',
'3707 8571 8571',
'1 select twyt 3376 3376 7756 dbms utility sqlid sqlhash chr 113 chr 113 chr 112 chr 106 chr 113 se
lect case 7756 7756 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113',
'call regexp substring repeat right char 3702 0 5000000000 null eevk eevk',
...]
```

In []:

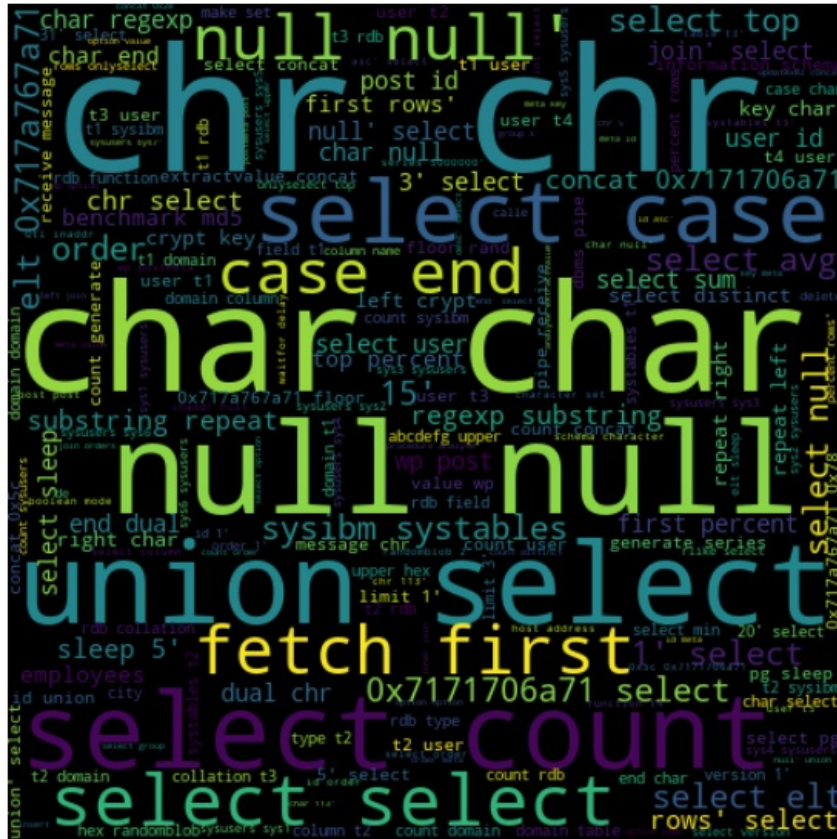
```
text = str(preprocessed_query)
```

Let's Do the Word Cloud Visualization

```
# Libraries
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Create the wordcloud object
wordcloud = WordCloud(width=480, height=480, margin=0).generate(text)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.margins(x=0, y=0)
plt.show()
```



Featurization

In [194]:

```
Some Feature Names ['00' '000001' '000003' '000004' '000023' '000045' '00033926r' '00041731d'
'00207473i' '00217e']
```

In []:

```
the type of count vectorizer <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer (30907, 2930)
```

TF-IDF Vectorizer

In [195]:

```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2),min_df=10)
tf_idf_vect.fit_transform(preprocessed_query)

print("some sample features(unique words in the corpus)",tf_idf_vect.get_feature_names_out()[0:10])
print('='*50)

final_tf_idf = tf_idf_vect.transform(preprocessed_query)
print("the type of count vectorizer ",type(final_tf_idf))
print("the shape of out text TFIDF vectorizer ",final_tf_idf.get_shape())

some sample features(unique words in the corpus) ['01' '01 1996' '06' '06 15' '07' '07 01' '07 31' '
09' '09 34' '0x28']
=====
the type of count vectorizer <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer (30907, 2930)
```

TF-IDF Weighted Word2Vec

In [186]:

```
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentence=[]
for sentence in preprocessed_query:
    list_of_sentence.append(sentence.split())
```

In [187]:

```
from gensim.models import Word2Vec
# min_count = 5 considers only words that occurred atleast 5 times
w2v_model=Word2Vec(list_of_sentence,min_count=5,size=50,workers=4)
```

In [188]:

```
w2v_words = list(w2v_model.wv.vocab)
print("number of words that occurred minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])
```

```
number of words that occurred minimum 5 times 3729
sample words ['pg', 'sleep', 'time', 'create', 'user', 'name', 'temp', '1', 'utl', 'inaddr', 'get',
'host', 'address', 'select', 'distinct', 'table', 'rownum', 'limit', 'sys', '5', 'id', 'union', 'ver
sion', 'load', 'file', 'char', '47', '101', '116', '99', '112', '97', '119', '100', 'admin', 'column
', 'tab', 'ascii', 'lower', 'substring', 'top', 'u', 'insert', 'password', '8', '27', 'null', 'banne
r', 'v', '2']
```

In [190]:

```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(list_of_sentence): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change this to 300 if you
use google's w2v
    cnt_words = 0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors.append(sent_vec)
print(len(sent_vectors))
print(len(sent_vectors[0]))
```

```
100%|██████████| 30907/30907 [00:07<00:00, 4039.76it/s]
```

```
30907
50
```

In [191]:

```
model = TfidfVectorizer()
model.fit(preprocessed_query)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(model.get_feature_names_out(), list(model.idf_)))
```

In [193]:

```
# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names_out() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf

tfidf_sent_vectors = []; # the tfidf-w2v for each sentence is stored in this list
row=0;
for sent in tqdm(list_of_sentence): # for each sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence
    for word in sent: # for each word in a sentence
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
            # tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole corpus
            # sent.count(word) = tf value of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors.append(sent_vec)
    row += 1
```

100%|██████████| 30907/30907 [02:23<00:00, 214.84it/s]