In [20]:
```
import pandas as pd
from bs4 import BeautifulSoup
import re
import string
```

In [12]:
```
#Importing all Query Results (4Files)
df1 = pd.read_csv("QueryResults.csv")
df2 = pd.read_csv("QueryResults (1).csv")
df3 = pd.read_csv("QueryResults (2).csv")
df4 = pd.read_csv("QueryResults (3).csv")
```

In [3]:
```
df1.head()
```

Out[3]:

	Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount
0	1402390	1	1402445.0	NaN	2009-09-09 22:06:26	NaN	128	127156
1	13707836	1	13707905.0	NaN	2012-12-04 16:50:39	NaN	35	127162
2	46540831	1	NaN	NaN	2017-10-03 08:58:54	NaN	35	127163
3	4344533	1	4344602.0	NaN	2010-12-03 10:19:46	NaN	97	127164
4	15751241	1	15751300.0	NaN	2013-04-01 20:26:23	NaN	34	127166

5 rows × 23 columns

In [7]:
```
def rmv_html_tags(raw_html):
    clean_text=BeautifulSoup(raw_html, "lxml").text
    return clean_text
```

In [10]:
```
#Removing HTML tags from body element
df1["Body"]=df1["Body"].apply(rmv_html_tags)
```

```
df2["Body"]=df2["Body"].apply(rmv_html_tags)
df3["Body"]=df3["Body"].apply(rmv_html_tags)
df4["Body"]=df4["Body"].apply(rmv_html_tags)
```

In [9]:

```
df1.head()
```

Out[9]:

	Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount
0	1402390	1	1402445.0	NaN	2009-09-09 22:06:26	NaN	128	127156
1	13707836	1	13707905.0	NaN	2012-12-04 16:50:39	NaN	35	127162
2	46540831	1	NaN	NaN	2017-10-03 08:58:54	NaN	35	127163
3	4344533	1	4344602.0	NaN	2010-12-03 10:19:46	NaN	97	127164
4	15751241	1	15751300.0	NaN	2013-04-01 20:26:23	NaN	34	127166

5 rows × 23 columns

In [14]:

```
#Removing punctuations
def rmv_punc(word):
    pattern = r'[' + string.punctuation + ']'
    return re.sub(pattern, '', word)
```

In [17]:

```
#Removing punctuations from Body coloumn
df1["Body"]=df1["Body"].apply(rmv_punc)
df2["Body"]=df2["Body"].apply(rmv_punc)
df3["Body"]=df3["Body"].apply(rmv_punc)
df4["Body"]=df4["Body"].apply(rmv_punc)
```

In [18]:

```
#Removing punctuations from Title coloumn
df1["Title"]=df1["Title"].apply(rmv_punc)
df2["Title"]=df2["Title"].apply(rmv_punc)
```

```
df3["Title"]=df3["Title"].apply(rmv_punc)
df4["Title"]=df4["Title"].apply(rmv_punc)
```

In [19]:

```
df1.head()
```

Out[19]:

	Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount
0	1402390	1	1402445.0	NaN	2009-09-09 22:06:26	NaN	128	127156
1	13707836	1	13707905.0	NaN	2012-12-04 16:50:39	NaN	35	127162
2	46540831	1	NaN	NaN	2017-10-03 08:58:54	NaN	35	127163
3	4344533	1	4344602.0	NaN	2010-12-03 10:19:46	NaN	97	127164
4	15751241	1	15751300.0	NaN	2013-04-01 20:26:23	NaN	34	127166

5 rows × 23 columns

In [23]:

```
#Replacing not required elements
df1 = df1.replace(r'\n',' ', regex=True)
df1 = df1.replace(r'\t',' ', regex=True)
df1 = df1.replace(r'\r',' ', regex=True)
df1 = df1.replace(r'\b',' ', regex=True)
df1 = df1.replace(r'\f',' ', regex=True)
```

In [22]:

```
df2 = df2.replace(r'\n',' ', regex=True)
df2 = df2.replace(r'\t',' ', regex=True)
df2 = df2.replace(r'\r',' ', regex=True)
df2 = df2.replace(r'\b',' ', regex=True)
df2 = df2.replace(r'\f',' ', regex=True)

df3 = df3.replace(r'\n',' ', regex=True)
df3 = df3.replace(r'\t',' ', regex=True)
```

```
df3 = df3.replace(r'\r',' ', regex=True)
df3 = df3.replace(r'\b',' ', regex=True)
df3 = df3.replace(r'\f',' ', regex=True)

df4 = df4.replace(r'\n',' ', regex=True)
df4 = df4.replace(r'\t',' ', regex=True)
df4 = df4.replace(r'\r',' ', regex=True)
df4 = df4.replace(r'\b',' ', regex=True)
df4 = df4.replace(r'\f',' ', regex=True)
```

In [31]:

```
#Converting df files to txt and csv files
df1.to_csv("cleaned_data1.txt")
df1.to_csv("cleaned_data1.csv")

df2.to_csv("cleaned_data2.txt")
df2.to_csv("cleaned_data2.csv")

df3.to_csv("cleaned_data3.txt")
df3.to_csv("cleaned_data3.csv")

df4.to_csv("cleaned_data4.txt")
df4.to_csv("cleaned_data4.csv")
```