General libraries being loaded

In [1]:
```
# Python ≥3.5 is required
import sys
assert sys.version_info >= (3, 5)

# Scikit-Learn ≥0.20 is required
import sklearn
assert sklearn.__version__ >= "0.20"

# Common imports
import numpy as np
import os, time
import pandas as pd

# Our new Deep Learning imports
import tensorflow as tf
from tensorflow import keras

# To plot nice figures
# %matplotlib widget
%matplotlib inline

import matplotlib as mpl
import matplotlib.pyplot as plt
mpl.rc('axes', labelsize=14)
mpl.rc('xtick', labelsize=12)
mpl.rc('ytick', labelsize=12)

# For plotting statistical figures
import seaborn as sns; sns.set()

# For speeding up numpy operations
import cupy as cp

# For faster numpy computation
from numba import jit, cuda

# For Progress Bar
from tqdm.auto import tqdm,trange
tqdm.pandas()

# Vaex Dataframe library
import vaex as vx

# For Pyspark activation
import os
os.environ["PYARROW_IGNORE_TIMEZONE"] = "1"

# Pyspark Dataframe
from pyspark import pandas as ps

import os
os.environ['KMP_DUPLICATE_LIB_OK']='True'
```

Loading Stored Data for Sentence Tokenization

In [2]:
```
import tensorflow as tf
```

In [3]:
```
Article_Data_Cord_File_DF_Abstract = pd.read_pickle(r"D:\UoA\Tri 2\Big Data Analysis and Projects\Week 8\archive\cord_19_embeddings\article_data_cord_19_file_df_abstract.pkl")
```

In [4]:
```
Article_Data_Cord_File_DF_Body_Text = pd.read_pickle(r"D:\UoA\Tri 2\Big Data Analysis and Projects\Week 8\archive\cord_19_embeddings\article_data_cord_19_file_df_body_text.pkl")
```

Breaking Sentences to Tokens (Final Step) (Abstract)

In [5]:
```
Article_Data_Cord_File_DF_Abstract['Abstract_Tokens'] = Article_Data_Cord_File_DF_Abstract['Abstract_Sentences'].progress_apply(tf.keras.preprocessing.text.text_to_word_sequence)
```

```
  0%|          | 0/122754 [00:00<?, ?it/s]
```

In [6]:
```
del Article_Data_Cord_File_DF_Abstract['Abstract_Sentences']
```

In [7]:
```
import gc
gc.collect()
print('',end='')
```

In [8]:
```
Article_Data_Cord_File_DF_Abstract.to_pickle(r"D:\UoA\Tri 2\Big Data Analysis and Projects\Week 8\archive\cord_19_embeddings\article_data_cord_19_file_df_abstract_tokens.pkl")
```

Breaking Sentences to Tokens (Final Step) (Body Text)

In [9]:
```
Article_Data_Cord_File_DF_Body_Text['Body_Text_Tokens'] = Article_Data_Cord_File_DF_Body_Text['Body_Text_Sentences'].progress_apply(tf.keras.preprocessing.text.text_to_word_sequence)
```

```
  0%|          | 0/855743 [00:00<?, ?it/s]
```

In [10]:
```
del Article_Data_Cord_File_DF_Body_Text['Body_Text_Sentences']
```

In [11]:
```
import gc
gc.collect()
print('',end='')
```

In [12]:
```
Article_Data_Cord_File_DF_Body_Text.to_pickle(r"D:\UoA\Tri 2\Big Data Analysis and Projects\Week 8\archive\cord_19_embeddings\article_data_cord_19_file_df_body_text_tokens.pkl")
```

In [ ]:
```
```