In [1]:

```python
import findspark
findspark.init()
import pyspark
```

In [2]:

```python
sc = pyspark.SparkContext(appName="averageurl")
```

In [44]:

```python
rdd = sc.textFile("C:/Users/EASYFRONT/Documents/BD/requetessql/data/urlaverage.txt")
rdd = rdd.map(lambda x : x.split('\t'))
rdd = rdd.groupByKey()
rdd = rdd.map (lambda x : (x[0], list(x[1]) ) )
rdd = rdd.map (lambda x : (x[0], list(map(int,x[1]))) )
rdd = rdd.map (lambda x : (x[0], (sum(x[1]), len(x[1]))))
rdd = rdd.map (lambda x : (x[0], x[1][0]/x[1][1]))
```

In [ ]:

```python
rdd.collect()
```