Demo Spark Context

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q http://archive.apache.org/dist/spark/spark-2.4.0/spark-2.4.0-bin-hadoop2.7.tgz
!tar -xvf spark-2.4.0-bin-hadoop2.7.tgz
!pip install -q findspark
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.0-bin-hadoop2.7"
import findspark
findspark.init()
```

```
from google.colab import drive
drive.mount('/content/gdrive', force_remount=True)
```

```
Go to this URL in a browser:...

Enter your authorization code:
..........
Mounted at /content/gdrive
```

```
path = '/content/gdrive/My Drive/LDS9/Practice/Chapter2/'
```

```
import pyspark
```

```
from pyspark import SparkContext
sc =SparkContext(master="local", appName="First Spark App")
```

```
sc
```

SparkContext

Spark UI

Version
v2.4.0
Master
local
AppName
First Spark App

```
# file from hdfs
# file_name = "hdfs://172.24.40.251:19000/t8.shakespeare.txt"
# file in Drive
file_name = path + "t8.shakespeare.txt"
```

```
data = sc.textFile(file_name).cache()
numAs = data.filter(lambda s: 'face' in s).count()
numBs = data.filter(lambda s: 'world' in s).count()
print("Lines with 'face': %i, lines with 'world': %i" % (numAs, numBs))
```

```
Lines with 'face': 489, lines with 'world': 688
```