

# Individual Assignment Solution Guide

All code also available in my Github with the following link:  
<https://github.com/legiando/SparkStreamingMBD2018-19>

This report will guide you with the output that I have in the exercise:

## 1. Mandatory Assignment

### 1.1. Part one: Connect to Cluster

I successfully solved this question. As you mentioned in class, I don't need to submit anything to this part, because it should be justified if I can solve the other questions.

### 1.2. Part two: Madrid instead of #

The output is the following:

```
----- 2019-07-01 18:04:10 -----
+-----+
| hashtag | hashtag_count |
+-----+
| #CertamenJC2019 | 8 |
| #UltimaHora | 6 |
| #SGIP | 6 |
| #TAMPM | 6 |
| #Madrid | 5 |
| #HalaMadrid | 5 |
| #MadridCentral | 3 |
| #ULTIMAHORA | 3 |
| #RealMadrid | 3 |
| #MadridCentralFun... | 3 |
+-----+

Time: 2019-07-01 18:04:10

('RealMadrid', 3)
('#LOTEBI', 1)
('#CryptoPlaza', 1)
('#SaltaConBretaña', 1)
('#adaptación', 1)
('#MadridCentral', 1)
('#NOTICIA', 1)
('#Bale?', 2)
('#sosprisiones', 1)
('#ULTIMAHORA', 3)
...

----- 2019-07-01 18:04:12 -----
+-----+
| hashtag | hashtag_count |
+-----+
| #CertamenJC2019 | 8 |
| #Madrid | 7 |
| #UltimaHora | 6 |
| #SGIP | 6 |
| #TAMPM | 6 |
| #HalaMadrid | 5 |
| #MadridCentral | 3 |
| #ULTIMAHORA | 3 |
| #RealMadrid | 3 |
| #MadridCentralFun... | 3 |
+-----+

Time: 2019-07-01 18:04:12

('RealMadrid', 3)
('#LOTEBI', 2)
('#CryptoPlaza', 1)
('#SaltaConBretaña', 1)
('#adaptación', 1)
('#MadridCentral', 1)
('#NOTICIA', 1)
('#Bale?', 2)
('#sosprisiones', 1)
('#ULTIMAHORA', 3)
...

----- 2019-07-01 18:04:14 -----
+-----+
| hashtag | hashtag_count |
+-----+
| #CertamenJC2019 | 8 |
| #UltimaHora | 7 |
| #Madrid | 7 |
| #SGIP | 7 |
+-----+
```

### 1.3.Part three: Spark SQL

The output is the following:

```
----- 2019-07-04 10:47:40 -----
+-----+
| words|words_count|
+-----+
|Madrid|          22|
|  que|          18|
|   y|          17|
|  la|          16|
| del|          11|
|  un|           9|
|  no|           7|
| los|           6|
|  El|           6|
|  ha|           6|
+-----+

-----
Time: 2019-07-04 10:47:40
-----
('@Roquerol', 1)
('debería', 2)
('Tesla', 1)
('cambia', 1)
('los', 6)
('datos', 1)
('y', 17)
('española', 1)
(' ', 4)
('Carlos', 1)
...

----- 2019-07-04 10:47:50 -----
+-----+
| words|words_count|
+-----+
|  la|          27|
|Madrid|          17|
|  que|          17|
|   y|          16|
| del|          10|
|  un|           9|
|  En|           9|
| con|           8|
| los|           7|
| por|           7|
+-----+

-----
Time: 2019-07-04 10:47:50
-----
('Fernández', 2)
('los', 7)
(' ', 3)
('con', 8)
('antelación', 1)
('y', 16)
('Central', 1)
('empezaron', 1)
('https://t.co/EPDjDLiEeM', 1)
('O', 1)
```

## 2. Bonus Assignment

### 2.1. Personal Twitter Account Credential

I use my own twitter account credential as you can see below:

Apps > [SparkStreamingPratama](#)

App details

**Keys and tokens**

Permissions

#### Keys and tokens

Keys, secret keys and access tokens management.

##### Consumer API keys

9sKTPHAp31aYBlijFTeFcbvrP (API key) **NEW**

kMUsnNdgsPoqdSbyZs3adz3iUw6MoRtmSc4go0xnUgkKZSWtDs (API secret key)

**NEW**

Regenerate

##### Access token & access token secret

3106812974-drNpGWMkS3wT5JPCoj3Vbk7ZWHo5fN4KHOFma6F (Access token)

tpXPfEd0aNaL6WVxeJkt7xfEEA6DgXxWsv5d2KSGfYsQh (Access token secret)

Read and write (Access level)

Revoke

Regenerate

## 2.2. Flajolet-Martin Algorithm

The output is the following:

```
----- 2019-07-04 11:02:50 -----
{'haber': 675, 'gobierno': 676, 'de': 676, 'izquierda': 676,
'si': 676, 'Cs': 676, 'vota': 676, 'con': 676, 'ellos.': 676,
'En': 676, 'Mad...': 676}

----- 2019-07-04 11:03:00 -----
{'(re)definición.': 716, 'Desde': 716, '@riid_llss': 716,
'estaremos': 717, 'siempre': 717, 'apoyando': 717, 'y': 717,
'colaborando': 717, 'con': 717, 'los': 717, 'i...': 717}

----- 2019-07-04 11:03:10 -----
{'que': 751, 'Madrid': 751, 'Central': 751, 'no': 751, 'bajó': 751, 'la': 751,
'contaminación': 751, 'https://t.co/dvrnXair3h': 751, 'Por': 751, '@tatianagrivas':
751, '#MadridCentral': 752}

----- 2019-07-04 11:03:20 -----
{'de': 803, '@AthleticClub': 800, 'Porque': 801, 'si': 801, 'nos': 801, 'ponemos':
801, 'así.': 801, 'mi': 801, 'hijo.': 801, 'padres...': 801, 'https://t.co/NmXcZJJlb9':
801}

----- 2019-07-04 11:03:30 -----
{'de': 852, 'haber': 850, 'gobierno': 850, 'izquierda': 850, 'si': 851, 'Cs': 851,
'vota': 851, 'con': 851, 'ellos.': 851, 'En': 851, 'Mad...': 851}
```