Marine Mercier
SD201

# TP Spark

## Question 1 :

Il suffit de modifier la ligne «  pairs = words.map(lambda s: (s, 2))  » en changeant le 2 en 1. En effet, chaque mot est ainsi attribué au poids 1, et reduceByKey somme tous les poids affectés à un même mot, donc ses occurrences.

Résultat Obtenu :

```
[('', 10), ('out', 3), ('more."[12]', 1), ('was', 33), ('taking,', 1),
('computer-animated', 1), ('we', 2), ('German', 1), ('business', 2),
('According', 1), ('illegal', 1), ('2003', 1), ('undergraduate', 1), ('co-
founded', 1), ('start', 1), ('eventually', 3), ('breakthrough', 1),
('Jonathan', 1), ('consider', 1), ('pregnancy,', 1), ('as', 13), ('used',
1), ('platform,', 1), ('several', 1), ('where', 1), ('adoption', 4),
('II,', 2), ('way', 1), ('looking', 1), ('three', 1), ('graphical', 1),
('2001,', 1), ('single', 1), ('Story—an', 1), ('designer', 1),
('different"', 1), ('Walt', 1), ('Inc.;', 1), ('candidate,', 1),
('closely', 1), ('enlightenment', 1), ('her', 5), ('visual', 1), ('case,',
1), ('officer', 1), ('new', 2), ('son', 2), ('sheltered', 1), ('iTunes',
2), ('co-founder', 2)]
```

## Question 2 :

Voici les 5 mots avec le plus d'occurrences dans le texte étudié :

```
the: 66

and: 53

a: 45

to: 42

in: 41
```

## Question 3 :

Voici les 5 mots de plus de 5 lettres apparaissant le plus dans le texte étudié :

```
Jandali: 8

Schieble: 8

Jobs's: 8

Francisco: 6

Joanne: 5
```

## Question 4 :

Voici les 10 pages wikipédia avec le plus grand in-degree :

(Identifiant page, Nom de la page, Nombre d'occurrences)

Marine Mercier
SD201

```
('60589', (['United', 'States'], 8145))

('30594', (['France'], 7799))

('24449', (['Communes', 'of', 'France'], 5740))

('26539', (['Departments', 'of', 'France'], 5299))

('51359', (['Regions', 'of', 'France'], 4064))

('23683', (['City'], 3832))

('52174', (['Romania'], 3527))

('20409', (['Category:Rivers', 'in', 'Romania'], 2978))

('59931', (['Tributary'], 2799))

 ('28563', (['England'], 2277))
```