TP2 – Lab on Spark

By: Júlia Togashi de Miranda

Pedro Germano Almeida Machado

1 – When creating the list of pairs we changed the parameters of the function “pairs = words.map(lambda s: (s, 2))” to “pairs = words.map(lambda s: (s, 1))”. By this way the sum a + b corresponds exactly to the number of occurrences of the words.

```
# Counting the number of occurences of each word, by using 'reduceByKey'
counts = pairs.reduceByKey(lambda a, b: a + b)
for (word, count) in counts.take(5):
  print (word, count)

Steven 1
Jobs 23
(/dʒɒbz/; 1
was 33
an 10
```

2 –

```
# Sorting the list of tuples by 'sortBy' and choosing the second term (Question 2)
ordered = counts.sortBy(lambda x: x[1], False)
for a in ordered.take(5):
  print (a)

('the', 66)
('and', 53)
('a', 45)
('to', 42)
('of', 41)
```

3 –

```
# Sorting (descending order) words with largest number of occurrences
atLeast5 = ordered.filter(lambda x: len(x[0]) > 5)
for a in atLeast5.take(5):
  print (a)

("Jobs's", 8)
('Jandali', 8)
('Schieble', 8)
('Francisco', 6)
('biological', 5)
```

4 –

```
# Print the name of links with most occurrences
for a, b in counts_edge.sortBy(lambda x: x[1], False).take(10):
  print ('%s : %i occurrences' %(dict_labels[a], b - 1)) # The l
```

```
United States : 8145 occurrences
France : 7799 occurrences
Communes of France : 5740 occurrences
Departments of France : 5299 occurrences
Regions of France : 4064 occurrences
City : 3832 occurrences
Romania : 3527 occurrences
Category:Rivers in Romania : 2978 occurrences
Tributary : 2799 occurrences
England : 2277 occurrences
```