Project: Part 2
Pig + Spark

Assignment

You should provide the following programs, in Pig and in Spark.

Exercise A Write a program for transposing the output of the word count program in **Pig and in Spark**.

> **Pig**. Instead of concatenating the strings, as done in Hadoop-MapReduce exercises, we can generate a bag of strings as second element of the tuple. Save the output in `wordcount/pig_wordtrans`.
> *Optional*. Transform the bag in a string.
>
> **Spark**. Input from file, output to file as for exercise 1.

For instance, given the input

```
car 3
the 6
house 3
phone 5
pen 3
glass 3
battery 5
```

the following output has to be generated:

```
3 {car pen house glass}
5 {battery phone}
6 {the}
```

Exercise B Write a program that counts the number of words with the same frequency in Pig and Spark.

> **Pig**. Save the output in `wordcount/pig_wordfreq`.
>
> **Spark**. Input from file, output to file as for exercise 1).

For instance, given the input

```
car 3
the 6
house 3
phone 5
pen 3
glass 3
battery 5
```

the following output has to be generated:

```
3 4
5 2
6 1
```

Exercise C Modify the **Spark** programs developed for exercise A and exercise B in order to use method `saveAsObjectFile` instead of `saveAsTextFile`.

> **Python**. You can use `saveAsPickleFile` or `saveAsSequenceFile` instead of `saveAsObjectFile`.

Then, run the modified transpose program (exercise A) passing as output directory `wordcount/spark_wordtrans` and look at the result. Run the modified frequency program (exercise B) passing as output directory `wordcount/spark_wordfreq` and look at the result.

Exercise D Write a program for computing maximum frequency and its associated words in **Pig and Spark**.

> **Spark**. Use `wordcount/spark_wordtrans` as input directory. Note that the `Comparator` must implements `Serializable`.

For instance, given the input

```
3 {car pen house glass}
5 {battery phone}
6 {the}
```

the following output has to be generated:

```
6 {the}
```

Exercise E Write a program for computing the average frequency of words in a text **Pig and Spark**.

> **Pig**. You can choose to use as input either `wordcount/pig_output` or `wordcount/pig_wordfreq`.

> **Spark**. Use `wordcount/spark_wordfreq` as input directory.

For instance the average of:

```
car 3
the 6
house 3
phone 5
pen 3
glass 3
battery 5
```

is

```
4.0
```

Exercise F Write a program in **Pig and Spark** for performing a simple join-like operation between the data stored respecitvely in `wordcount/pig_wordtrans` and `wordcount/pig_wordfreq` and in `wordcount/spark_wordtrans` and `wordcount/spark_wordfreq`.
More precisely, in exercise A we generated transposed data like:

```
3 {car pen house glass}
5 {battery phone}
6 {the}
7 {one}
```

In exercise B, we generated count frequency data like:

```
3 4
5 2
6 1
8 5
```

Now, you should join the rows sharing the same key, obtaining:

```
4 {car pen house glass}
2 {battery phone}
1 {the}
```

Rules for project development and delivery

- The project, Part 2, can be developed by groups of up to two persons.
- Each student should upload on AulaWeb, a single zip file of name CognomeN (where N is the initial name letter). The file should contain: (i) one folder of name ExercizeX containing the Pig program (when requested) and the Spark program file developed for Exercize X; (ii) a short document specifying the names of the group members and describing the proposed solution for each exercize.
- The zip file should be uploaded by December 9.
- To Project Part 2 will be assigned a rating among {A, B, C, D}, according to the following rules:
  - A+: all exercizes have been correctly solved

– A: all exercizes except D have been correctly solved
– B: exercizes A, B, C, E have been correctly solved
– C: exercizes A, B, C have been correctly solved
– D: exercizes A and B have been correctly solved
– In all other cases, no rating will be provided.