In [ ]:

Big Data Algorithms Techniques & Platforms

Spark and DataFrames

A. Analysis of the `Great Expectations` - Charles Dickens

Suppose you have a file containing the text of the `Great Expectation` , a novel written in English by Charles Dickens. You can see an excerpt below :

"I pointed to where our village lay, on the flat in-shore among the alder-trees and pollards, a mile or more from the church.

The man, after looking at me for a moment, turned me upside down, and emptied my pockets. There was nothing in them but a piece of bread. When the church came to itself,—for he was so sudden and strong that he made it go head over heels before me, and I saw the steeple under my feet,—when the church came to itself, I say, I was seated on a high tombstone, trembling while he ate the bread ravenously."

Exercise 0. Support functions

In [ ]:
```
### Write here all the import function and the support function you need for proces
```

Exercise 1 Word number in sentences

The length of a sentence is the number of words that compose the sentence.

Write and comment on the set of Spark operations that return how many sentences we have for each available length in the text. You must also show the five most common lengths.

Notice that in the available text, a sentence is a set of lines that ends with a strong punctuation mark (i.e., ".", "!", "?", etc.).

Notice that:

- The novel starts after the `*** START OF THE PROJECT GUTENBERG EBOOK GREAT EXPECTATIONS ***`
- `[Illustration]` is not a sentence
- `Chapter VIII.` is not a sentence

You can introduce your constraints for the parsing. Multiple solutions and points of view are correct. You must comment on your point of view (i.e., the definition of ``sentence'' in your analysis).

You can choose if you are considering the stop-words in the count: add your point of view in the comment.

In [ ]:
```
############## WRITE YOUR CODE HERE ##############
```

Exercise 2. The average length of sentences in the entire novel

Write and comment on the set of Spark operations that returns and shows the average length of the sentences in the novel.

In [ ]:
```
############## WRITE YOUR CODE HERE ##############
```

Exercise 3. Average length in chapters

Write and comment on the set of Spark operations that returns and shows the average length of sentences in each chapter.

In [ ]:
```
############## WRITE YOUR CODE HERE ##############
```

Exercise 4. The most repeated words in the sentences

Write and comment on the set of Spark operations that returns the ten most repeated words in a sentence and their repetition rate.

For example, the most repeated word in this set of sentences is `cat`, with an average repetition rate of 2. The word `table` is not considered as repeated.

- The cat is on the cat table.
- The table is red.
- The wooden table is broken.

In [ ]:

B. Coffee Dataset

For running this series of exercises, we are going to use a dataset coming from Kaggle.

As stated in the description of the dataset: "Dataset contains information about coffee production and consumption.

All data are available from the official ICO website: https://www.ico.org/new_historical.asp".

The dataset

The dataset is in a .csv file, and among the columns, you can find:

- `name` The name of the blend
- `roaster` The name of the roaster
- `roast` The type of roast
- `loc_country` The country of the roaster
- `...` ...

In [ ]:
```
! pip install kaggle

! mkdir ~/.kaggle

! cp kaggle.json ~/.kaggle/

! chmod 600 ~/.kaggle/kaggle.json
```

In [ ]:
```
! kaggle datasets download schmoyote/coffee-reviews-dataset
```

In [ ]:
```
!unzip coffee-reviews-dataset.zip
```

In [ ]:
```
# Add your imports
```

Spark and Pandas.

For this set of exercises, you must import data in Spark. After this first import, you can pass any dataset to Pandas for data analysis. At the end of each exercise (when the question is pertinent), you must return (reconvert) the dataframe in Spark.

Each time you do this conversion you must comment about this. Example:

```
# creating a Spark dataframe

df = ...

# using Pandas and creating a Pandas dataframe

dfp = df. ...

# back to Spark

dfs = ...
```

Exercise 5. First import and data type

Import the .csv file in Spark DataFrame and show the structure of the dataframe.

Check and comment about the data type of each column. As you know, a good data analysis always starts from understanding your dataset.

In [ ]:
```
# Write the command that creates (reads) a Spark DataFrame and stores the reference

#'''############## WRITE HERE YOUR CODE ##############'''
dfs =

# show the DataFrame schema
dfs

###### Write here your comment ##############
```

Exercise 6. Data modeling choices

Comment on the data-modeling choices and if you consider them correct from a general data-modeling point of view.

In [ ]:
```
###### Write here your comment
```

Exercise 7. Best rated coffees

We want to find the 5 best rated coffees.

In [ ]:
```
############## WRITE YOUR CODE HERE ##############
```

Exercise 8. The best 10 roasters

We want to find the ten best roasters. It is up to you to define and refine the ``best'' metric (best in the platform from the beginning of data collection, best in the last three years, best and with a minimum number of ratings, etc.).

Add the definition of your best metric in the comments.

In [ ]:
```
############## WRITE YOUR CODE HERE ##############
```

Exercise 9. Best country

If you were a roaster, in which three countries would you try to set your business?

Show them and refine your metric definition if necessary.

In [ ]:
```
############## WRITE YOUR CODE HERE AND THE DESCRIPTION OF YOUR "SOMEHOW" #########
```

Exercise 10. Less common origin

Show the ten less common origins of the coffees.

In [ ]:
```
############## WRITE YOUR CODE HERE AND THE DESCRIPTION OF YOUR "SOMEHOW" #########
```

Exercise 11. Propose your own analysis

Propose here an analysis on the dataframe.

In [ ]:
```

```