Développer des programmes avec Spark SQL

Exercice: moyenne de nombre d'amis (friends) par âge

- Le fichier friends-header contient des entêtes
- Utiliser: select ("titte-col1", "titre-col2") pour avoir les colonnes nécessaires
- Utiliser les fonctions de DataFrame:
  avg ("columnName"), groupBy("columnName"), show()

Exercice: fréquences de mots avec dataFrames

Fonctions SQL

- from pyspark.sql import functions as func
- func.explode() – comme flatmap, éclater les colonnes en lignes
- func.split()
- func.lower()
- Passer les colonnes en paramètres:
  - func.split(inputDF.value, \\w+)
  - filter(wordsDF.word, !="")
  - C'est possible d'utiliser func.col("culmnName") to refer to a column
- L'utilisation de dataframes avec ce texte non structuré n'est pas un ajustement parfait
- Le dataframe initial aura des lignes avec une colonne appellée "valeur" pour chaque ligne du texte
- Ceci est un cas où l'utilisation d'un RDD est plus simple que celle du dataframe

➢ Dataframe est mieux utilisée avec les données structurées

Exercice: MIN-MAX Temperature avec dataFrames