FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

DEI
DEPARTAMENTO ENGENHARIA INFORMÁTICA

Project #2

Assignment #1 – Data Processing with Hadoop MapReduce

16th November 2023

Advanced Infrastructures for Data Science

Pedro Neves, pedroneves@dei.uc.pt, 2023/2024

Master in Data Science and Engineering (MDSE) Course

Task #1 – Average Number of Friends by Age

Assignment #1

Context: Social Network dataset

Dataset

ID	Name	Age	Friends
0	Will	33	385
1	Jean-Luc	33	2
2	Hugh	55	221
3	Deanna	40	465
4	Quark	68	21

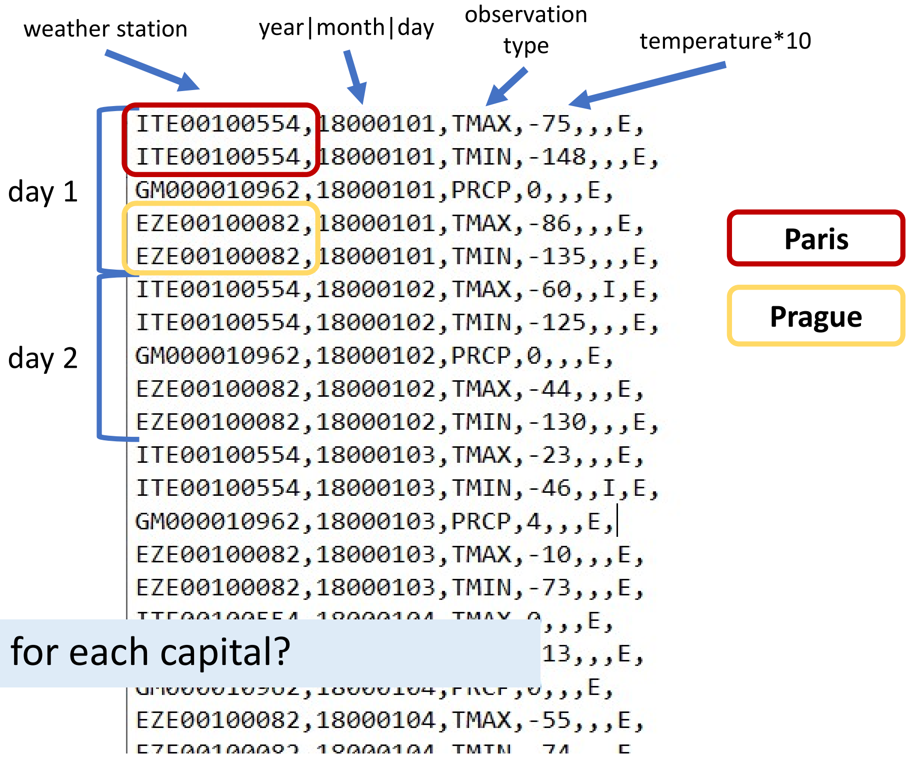
Objective: What is the average number of friends by age?

Task #2 – Minimum Temperature Per Capital

Assignment #1

Context: Daily weather data for year 1800 dataset

Dataset



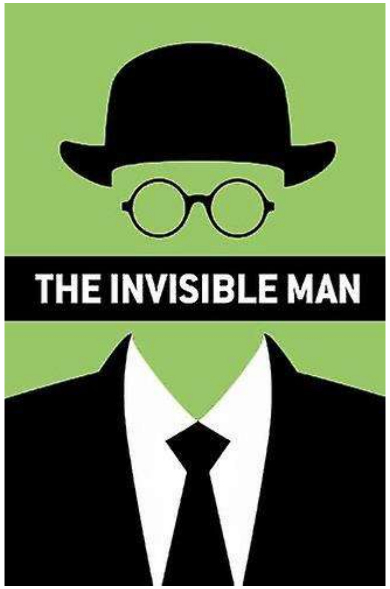
```
ITE00100554,18000101,TMAX,-75,,,E,
ITE00100554,18000101,TMIN,-148,,,E,
GM000010962,18000101,PRCP,0,,,E,
EZE00100082,18000101,TMAX,-86,,,E,
EZE00100082,18000101,TMIN,-135,,,E,
ITE00100554,18000102,TMAX,-60,,I,E,
ITE00100554,18000102,TMIN,-125,,,E,
GM000010962,18000102,PRCP,0,,,E,
EZE00100082,18000102,TMAX,-44,,,E,
EZE00100082,18000102,TMIN,-130,,,E,
ITE00100554,18000103,TMAX,-23,,,E,
ITE00100554,18000103,TMIN,-46,,I,E,
GM000010962,18000103,PRCP,4,,,E,
EZE00100082,18000103,TMAX,-10,,,E,
EZE00100082,18000103,TMIN,-73,,,E,
ITE00100554,18000104,TMAX,0,,,E,
13,,,E,
GM000010962,18000104,PRCP,0,,,E,
EZE00100082,18000104,TMAX,-55,,,E,
EZE00100082,18000104,TMIN,-74,,,E
```

Objective: What is the minimum temperature for each capital?

Task #3 – Sort the Word Frequency in a Book

Assignment #1

Context: Book dataset

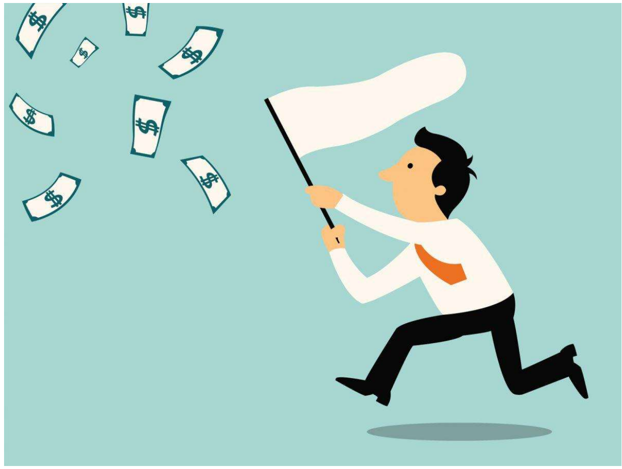


Objective: What is the (sorted) word frequency in the book?

Task #4 – Sort the Total Amount Spent by Customer

Assignment #1

Context: Shopping Store Dataset



Dataset ⇒

Customer_ID	Product_ID	Amount
344	983	45.1
99	574	7.08
344	24	102.2
43	241	37.08
99	230	61.89

Objective: What is the (sorted) total amount spent by each customer?

Good
Work