Project #2

Assignment #3 – Data Processing with Spark RDD

28th November 2024

Advanced Infrastructures for Data Science

Pedro Neves, pedroneves@dei.uc.pt, 2024/2025

Master in Data Science and Engineering (MDSE) Course

Task #1 – Minimum Temperature Per Capital
Assignment #3

Context: Daily weather dataset

Dataset →

weather station → year|month|day → observation type → temperature*10

```
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,
```
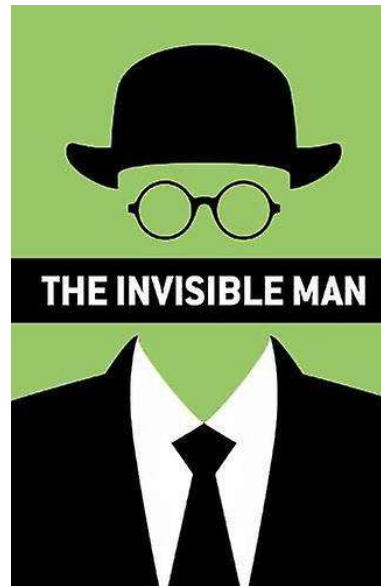
day 1

day 2

Paris

Prague

Objective: What is the minimum temperature for each capital?

Task #2 – Obtain the Word Frequency in a Book
Assignment #3
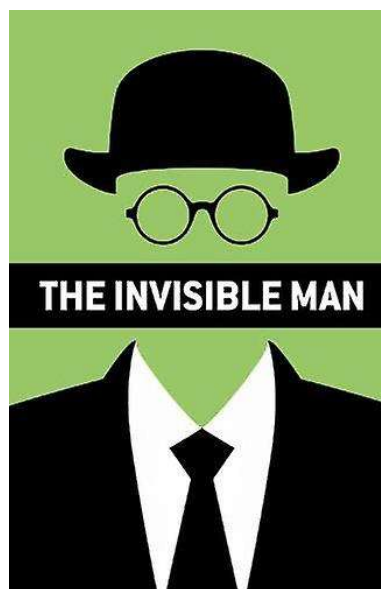
Context: Book dataset

Objective: What is the word frequency in the book?

Task #3 – Sort the Word Frequency in a Book
Assignment #3

Context: Book dataset

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

Task #4 – Obtain the Total Amount Spent by Customer
Assignment #3

Context: Shopping Store Dataset

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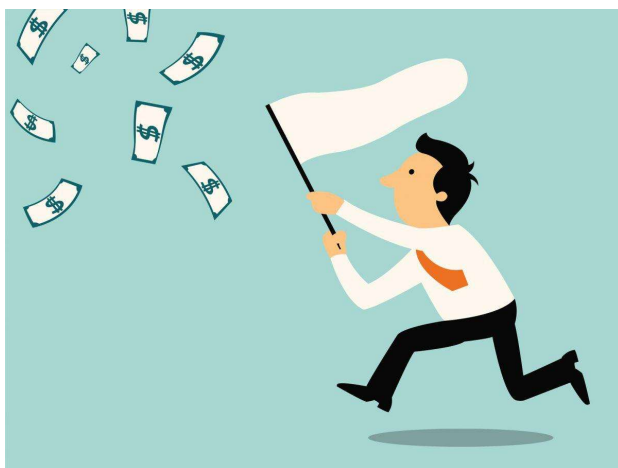
Customer_ID	Product_ID	Amount
344	983	45.1
99	574	7.08
344	24	102.2
43	241	37.08
99	230	61.89

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

Task #5 – Sort the Total Amount Spent by Customer
Assignment #3

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?

Task #6 – Most Popular Superhero
Assignment #3

Context: Superheroes dataset

Dataset →

```
C: > Users > 10050509 > Downloads >  $  Marvel+Names (1)
   1     1  "24-HOUR MAN/EMMANUEL"
   2     2  "3-D MAN/CHARLES CHAN"
   3     3  "4-D MAN/MERCURIO"
   4     4  "8-BALL/"
   5     5  "A"
   6     6  "A'YIN"
   7     7  "ABBOTT, JACK"
   8     8  "ABCISSA"
   9     9  "ABEL"
  10    10  "ABOMINATION/EMIL BLO"
```

```
C: > Users > 10050509 > Downloads >  ◐  Marvel+Graph (1)
   1    5988 748 1722 3752 4655 5743 1872 3413 5527 6368
   2    5989 4080 4264 4446 3779 2430 2297 6169 3530 327
   3    5982 217 595 1194 3308 2940 1815 794 1503 5197 8
   4    5983 1165 3836 4361 1282 716 4289 4646 6300 5084
   5    5980 2731 3712 1587 6084 2472 2546 6313 875 859
   6    5981 3569 5353 4087 2653 2058 2218 5354 5306 313
   7    5986 2658 3712 2650 1265 133 4024 6313 3120 6066
   8    5987 2614 5716 1765 1818 2909 6436 1587 6451 566
```

Objective: Find the most popular superhero?

Task #7 – Least Popular Superhero
Assignment #3

Context: Superheroes dataset

Dataset →

```
C: > Users > 10050509 > Downloads > $ Marvel+Names (1)
   1    1  "24-HOUR MAN/EMMANUEL"
   2    2  "3-D MAN/CHARLES CHAN"
   3    3  "4-D MAN/MERCURIO"
   4    4  "8-BALL/"
   5    5  "A"
   6    6  "A'YIN"
   7    7  "ABBOTT, JACK"
   8    8  "ABCISSA"
   9    9  "ABEL"
  10   10  "ABOMINATION/EMIL BLO"
```

```
C: > Users > 10050509 > Downloads > ● Marvel+Graph (1)
   1    5988 748 1722 3752 4655 5743 1872 3413 5527 6368
   2    5989 4080 4264 4446 3779 2430 2297 6169 3530 327
   3    5982 217 595 1194 3308 2940 1815 794 1503 5197 8
   4    5983 1165 3836 4361 1282 716 4289 4646 6300 5084
   5    5980 2731 3712 1587 6084 2472 2546 6313 875 859
   6    5981 3569 5353 4087 2653 2058 2218 5354 5306 313
   7    5986 2658 3712 2650 1265 133 4024 6313 3120 6066
   8    5987 2614 5716 1765 1818 2909 6436 1587 6451 566
```

Objective: Find the least popular superhero?

👍 Good Work