Big Data algorithms, techniques and plateforms

Madriss Seksaoui
Msc in Data Sciences and Business Analytics

LAB 1:

Questions 2.7: Displaying the content of a CSV File

Here is the output of my program (CompterLigneFile.java): (5 last lines of output)

```
1852 30.0

1863 12.0

1896 16.0

1918 32.0

1860 22.0

1870 15.0

The number of lines is 98
```

Questions 2.8: Displaying a compact file

Here is the output of my program (Display28.java):

```
station : BRISTOL/LULSGATE          FIPS : UK Altitude : +0189.0

station : BRISTOL                   FIPS : UK Altitude : +0189.6

station : BRISTOL WEA CENTER          FIPS : UK Altitude : +0011.0

station : LYNEHAM                   FIPS : UK Altitude : +0156.4

station : LARKHILL                  FIPS : UK Altitude : +0133.0

station : GREENHAM COMMON RAF          FIPS : UK Altitude : +0122.0

station : UPAVON                    FIPS : UK Altitude : +0175.0

station : NETHERAVON(RA)            FIPS : UK Altitude : +0139.0

station : BOSCOMBE DOWN             FIPS : UK Altitude : +0124.1

station : WINCHESTER                FIPS : UK Altitude : +0083.0

station : MIDDLE WALLOP             FIPS : UK Altitude : +0091.0

The number of lines is 1727
```

In both 2.7 and 2.8 we have missing values sometimes. Note that I choose to display the number of line at the end of the output.

Questions 4.1: TF-IDF

As I was reading the pdf file with the detailed diagram of a tf-idf calculation, I choose to write 3 separate java files, one for each round.

My tf-idf program runs like this:

round1(input text) -> output_round1
round2(output_round1) -> output_round2
round3(output_round2) -> output_tfidf

Here are the top 20 words with the highest tf-idf values:

Word	TF-IDF
Buck	0,005477549
Dogs	0,001315474
thornton	0,001293909
myself	0,001234037
buck's	9,70E-04
spitz	9,27E-04
francois	9,06E-04
john	8,41E-04
sled	8,19E-04
buck,	7,12E-04
dogs,	6,69E-04
friday	6,39E-04
shore,	5,72E-04
perrault	5,39E-04
hal	5,18E-04
god	5,05E-04
trail	4,96E-04