Exercise #1

- Word count problem
  - Input: (unstructured) textual file
  - Output: number of occurrences of each word appearing in the input file

Exercise #1 - Example

- Input file
  - Toy example file for Hadoop. Hadoop running example.

- Output pairs
  - (toy, 1)
  - (example, 2)
  - (file, 1)
  - (for, 2)
  - (hadoop, 2)
  - (running, 1)

Exercise #2

- Word count problem
  - Input: a HDFS folder containing textual files
  - Output: number of occurrences of each word appearing in at least one file of the collection (i.e., files of the input directory)
  - The only difference with respect to exercise #1 is given by the input
  - Now the input is a collection of textual files

Exercise #2 - Example

- Input files
  - Toy example file for Hadoop. Hadoop running example.
  - Another file for Hadoop.

- Output pairs
  - (another, 1)
  - (example, 2)
  - (file, 2)
  - (for, 2)
  - (hadoop, 3)
  - (running, 1)
  - (toy, 1)
**Exercise #3**

- **PM_{10}** pollution analysis
  - Input: a (structured) textual file containing the daily value of PM_{10} for a set of sensors
    - Each line of the file has the following format:
      
      \[ \text{sensorId, date, PM}_{10} \text{ value (\mu g/m}^3) \]n
  - Output: report for each sensor the number of days with PM_{10} above a specific threshold
    - Suppose to set threshold = 50 \mu g/m^3

**Exercise #3 - Example**

- **Input file**
  
  \[
  \begin{align*}
  \text{s1,2015-01-01} & \quad 20.5 \\
  \text{s2,2015-01-01} & \quad 30.2 \\
  \text{s1,2015-01-02} & \quad 60.2 \\
  \text{s2,2015-01-02} & \quad 20.4 \\
  \text{s1,2015-01-03} & \quad 55.5 \\
  \text{s2,2015-01-03} & \quad 52.5 \\
  \end{align*}
  \]

  - **Output pairs**
    \( (s1, 2) \)
    \( (s2, 3) \)

**Exercise #4**

- **PM_{10}** pollution analysis per city zone
  - Input: a (structured) textual file containing the daily value of PM_{10} for a set of city zones
    - Each line of the file has the following format:
      
      \[ \text{zoneId, date, PM}_{10} \text{ value (\mu g/m}^3) \]n
  - Output: report for each zone the list of dates associated with a PM_{10} value above a specific threshold
    - Suppose to set threshold = 50 \mu g/m^3

**Exercise #4 - Example**

- **Input file**
  
  \[
  \begin{align*}
  \text{zone1,2016-01-01} & \quad 20.5 \\
  \text{zone2,2016-01-01} & \quad 30.1 \\
  \text{zone1,2016-01-02} & \quad 60.2 \\
  \text{zone1,2016-01-02} & \quad 20.4 \\
  \text{zone1,2016-01-03} & \quad 55.5 \\
  \text{zone2,2016-01-03} & \quad 52.5 \\
  \end{align*}
  \]

  - **Output pairs**
    \( \{\text{zone1, 2016-01-03, 2016-01-02}\} \)
    \( \{\text{zone2, 2016-01-01}\} \)