Exercise #44

- Misleading profile selection
- Input:
  - A textual file containing the list of movies watched by the users of a video on demand service
    - Each line of the file contains the information about one visualization
      userid,movieid,start-timestamp,end-timestamp
    - The user with id userid watched the movie with id movieid from start-timestamp to end-timestamp

Exercise #44

- Input:
  - A second textual file containing the list of preferences for each user
    - Each line of the file contains the information about one preference
      userid,movie-genre
    - The user with id userid liked the movie of type movie-genre

Exercise #44

- Input:
  - A third textual file containing the list of movies with the associated information
    - Each line of the file contains the information about one movie
      movieid,title,movie-genre
    - There is only one line for each movie
      - i.e., each movie has one single genre

Exercise #44

- Output:
  - Select the userid of the list of users with a misleading profile
    - A user has a misleading profile if more than threshold% of the movies he/she watched are not associated with a movie genre he/she likes
      - threshold is an argument/parameter of the application and it is specified by the user
    - Store the result in an HDFS file
Exercise #45

- Profile update
- Input:
  - A textual file containing the list of movies watched by the users of a video on demand service
    - Each line of the file contains the information about one visualisation
      - userid,movieid,start-timestamp,end-timestamp
    - The user with id userid watched the movie with id movieid from start-timestamp to end-timestamp

Exercise #45

- Input:
  - A second textual file containing the list of preferences for each user
    - Each line of the file contains the information about one preference
      - userid,movie-genre
    - The user with id userid liked the movie of type movie-genre

Exercise #45

- Output:
  - Select for each user with a misleading profile (according to the same definition of Exercise #44) the list of movie genres that are not in his/her preferred genres and are associated with at least 5 movies watched by the user
  - Store the result in an HDFS file
    - Each line of the output file is associated with one pair (user, selected misleading genre) associated with him/her
    - The format is
      - userid, selected (misleading) genre
    - Users associated with a list of selected genres are associated with multiple lines of the output file

Exercise #46

- Time series analysis
- Input:
  - A textual file containing a set of temperature readings
  - Each line of the file contains one timestamp and the associated temperature reading
    - The format of the timestamp is the Unix timestamp that is defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time (UTC), Thursday, 1 January 1970
  - The sample rate is 1 minute
    - i.e., the difference between the timestamps of the two consecutive readings is 60
Exercise #46 - Example

- Input file
  1451606400,12.1
  1451606500,12.2
  1451606600,13.5
  1451606800,14.0
  1451606900,14.0
  1451606800,15.5
  1451606900,15.0

- Output file
  1451606400,12.2,1451606400,12.2,1451606500,13.5
  1451606600,12.2,1451606700,13.5,1451606800,14.0