Report SQUAD ETL

Hernani Chavez, Marco Antonio
Navarro, Helen
Vilaró Antunes Dos Santo, Francesc
Chin Jiménez, Lien

Structure Database of data.txt

In the data.txt file content we found 3 kinds of lines:

- Data lines.
- Comments.
- Empty lines.

Data lines: Each one of these lines contained three elements divided by commas. The first one was a name, the second one was an url, and the last one was a description. Some lines had an empty space with no element in it . However, these lines always have the following format:

name, url, description

Each of the elements sometimes also contained non-wanted punctuation characters and whitespaces.

Comments: These were useful text lines with information about the data for the user. But we didn't need them as they do not contain data to work. They began with any of the next characters *,#, %.

Empty lines: Lines without characters. Despite having no printable characters, they had the unprintable character "\n", which means end of line.

After the execution of the function `read_and_clean()` the data.txt content was changed. In this new version, comments, lines and also all non desirable characters (punctuation characters and non desired whitespace) were deleted. So after that, data.txt contained just the data in the same format than data.txt.:

name, url, description