Exercise:

Raw Data - rt-feed-record.rar

This file contains a piece of a recording gathered by subscribing to a real-time feed stream. Every line contains a JSON line analytic. The analytics can belong to the same document and they are related by the **RP\_DOCUMENT\_ID** hash.

The field **DOCUMENT\_RECORD\_COUNT** indicates the total number of analytics of a given story.

The field **DOCUMENT\_RECORD\_INDEX** indicates the number/place of the analytic within the total number of analytics belonging to a given story.

Could you provide a script in Python to report:

* The total number of distinct stories that appear in the record file
* Sometimes, the system is dropping some analytics for some stories. This is an error and we are interested in building a mechanism that, given a piece of real time record, will be able to spot missing analytics. Could you implement a method to spot these issues related to missing analytics based on the information of the fields provided above?
* The **RP\_ENTITY\_ID** field follows a clear format pattern. How would you implement a method to validate that the data received on this field conforms to the expected format.

Please also provide a brief description of your solution and clearly document the validation process it performs.

Bonus:

It is advantageous if your solution is containerized using Docker.