ETL-Helper

a little helper to help find the probable title

Overview

An analyst "John Doe" in a [reserved name] company receive monthly reports of the results of some machines, that reports majorly are in excel; all the reports are hand filled. In the process to most common mailformed data and need to understand is the "TitleOfGame", in an internal process all the data need to be formated to upload automatically to a server and evidently the malformed title need to be linked to their correct title in a dictionary table.

Objective

The last described process is worked by hand and is time consuming as it is a sensitive process and needs to be monitored, it cannot be completed automatically. So the proposed solution is a little helper to help find the probable title by Similarity of Characters.

Data

Due to data sensitivity and privacy policies, I only have the necessary part of the database structure and some examples. In this case, to show the work I change the column names to comply with privacy policies.

- 221107Daniel_ETL.ipynb:
  - The first observation, how this work? The most common procedure: Jaccard.
- 221108Daniel_ETL.ipynb:
  - How this work? Perform options evaluation.
    - Hamming distance
    - Jaccard distance
    - Votting of the distance in the last part of title vs all title
- 221114Daniel_ETL.ipynb
  - The project have a little pause, but the first draft is ready and only need this for now.
- 230103Daniel_ETL.ipynb
  - The little helper has a bug: sometimes the report has duplicates and I was working with dictionaries, I changed it to lists so I can work with duplicates and save it to a new XLS file too, so as not to overwrite the source file. The above was overwritten because it was supposed to be a draft but it was used at work.

machine learning tools are too much for such task

Disclaimer In this repository evidently des not exist the production sys.

```
:                      ლ(ò_óˇ)                        :
:                         /  !                         :
```