Report

June 7, 2020

1 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ (Project)

1.1 In this Project we read csv file, remove duplicates, and make the proper changes on csv files so that we can import them later on our database

First of all we import the libraries that we are going to use

```
[35]: import pandas as pd
```

1.1.1 Editing csv file text errors, and removing rows with null values from 'keywords.csv' file

```
[36]: data = pd.read_csv("keywords.csv").replace('\':', '":', regex = True)
      data = data.replace("{\'", '{"', regex = True)
      data = data.replace(", \'", ', "', regex = True)
      data = data.replace(": \'", ': "', regex = True)
      data = data.replace("\'}", '"}', regex = True)
      data = data.replace("\\\\", ' ', regex = True)
      data = data.replace('\"trudy jackson"', "trudy jackson", regex = True);
      data = data[data.keywords != '[]']
```

1.1.2 Removing duplicates using pandas library based on id of each row

```
[37]: data = data.drop_duplicates(subset = 'id')
```

1.1.3 Creating new csv file that will be imported on our database

```
[38]: data.to_csv('keywords_wo_duplicates.csv', encoding='utf-8', index = False)
```

1.1.4 Editing csv file text errors, and removing rows with null values from 'credits.csv' file

```
[39]: data = pd.read_csv("credits.csv").replace("\"", "\'", regex = True).replace('\':
       ↪', '":', regex = True)
      data = data.replace("{\'", '{"', regex = True)
      data = data.replace(": \'", ': "', regex = True)
      data = data.replace("\'}", '"}', regex = True)
data = data.replace("\\\\", ' ', regex = True)
data = data.replace("\', \W", "\", \"", regex = True)
data = data.replace("\", \W", "\", \"", regex = True)
data = data.replace("None", "\"None\"", regex = True)
data = data.replace("\'cast_id\'", '"cast_id"', regex = True)
data = data.replace("\'character\"", '"character"', regex = True)
data = data.replace("\'credit_id\"", '"credit_id"', regex = True)
data = data.replace("\'gender\"", '"gender"', regex = True)
data = data.replace("\'id\"", '"id"', regex = True)
data = data.replace("\'name\"", '"name"', regex = True)
data = data.replace("\'profile_path\"", '"profile_path"', regex = True)
data = data.replace("\'department\"", '"department"', regex = True)
data = data.replace("\'job\"", '"job"', regex = True)
data = data.replace("\"Jack Jones\"", 'Jack Jones', regex = True)
data = data.replace("Art\", a", 'Art\', a', regex = True)
data = data.replace("\(\"None\"", '(\'None\'', regex = True)
data = data.replace("Book Person: \"The", 'Book Person: \'The', regex = True)
data = data.replace("\"order", '\'order', regex = True)
data = data.replace("\'order\":", '"order":', regex = True)
data = data.replace("Clips from 'For Me and My Gal\", \"Easter Parade\"",␣
↪"Clips from 'For Me and My Gal\', \'Easter Parade\'", regex = True)
data = data.replace("\" \'Girl Crazy\'", '\'Girl Crazy\'', regex = True)
data = data.replace("Masa\'s Uncle \"None\"", 'Masa\'s Uncle \'None\'', regex =␣
↪True)
data = data.replace("\"character\": Jack Jones", '\"character\": "Jack Jones"',␣
↪regex = True)
data = data.replace("\"name\": Jack Jones", '\"name\": "Jack Jones"', regex =␣
↪True)
data = data.replace("Duke of \"None\"", 'Duke of \'None\'', regex = True)
data = data.replace("\"\"None\"""", '\"None', regex = True)
data = data.replace("story \'Imomushi\", \"Kagami jigoku\", \"Kasei", 'story␣
↪\'Imomushi\', \'Kagami jigoku\', \'Kasei', regex = True)
data = data.replace("Mom \(segment: \"The", 'Mom (segment: \'The', regex = True)
data = data.replace("Demon \(segment: \"The", 'Demon (segment: \'The', regex =␣
↪True)
data = data.replace("Demon Voice \(segment: \"The", 'Demon Voice (segment:␣
↪\'The', regex = True)
data = data.replace("The Man \(segment: \"The", 'The Man (segment: \'The',␣
↪regex = True)
data = data.replace("Maggie \(segment: \"The", 'Maggie (segment: \'The', regex␣
↪= True)
data = data.replace("Alan \(segment: \"The", 'Alan (segment: \'The', regex =␣
↪True)
data = data.replace("Man \(segment: \"Undying", 'Man (segment: \'Undying',␣
↪regex = True)
```

```
data = data.replace("Woman \(segment: \"Undying", 'Woman (segment: \'Undying', regex = True)
data = data.replace("Girlfriend \(segment: \"Undying", 'Woman (segment: \'Undying', regex = True)
data = data.replace("Alley Zombie \(segment: \"Undying", 'Alley Zombie (segment: \'Undying', regex = True)
data = data.replace("Zombie \(segment: \"Undying", 'Zombie (segment: \'Undying', regex = True)
data = data.replace("\"Undying", '\'Undying', regex = True)
data = data.replace("\"Death Scenes", '\'Death Scenes', regex = True)
data = data.replace("\"Evaded", '\'Evaded', regex = True)
data = data.replace("\"Banishing", '\'Banishing', regex = True)
data = data.replace("segment: \"The Sleeping Plot", 'segment: \'The Sleeping Plot', regex = True)
data = data.replace("Herself - Author: \"The", 'Herself - Author: \'The', regex = True)
data = data.replace("Dynamiitti-Lahti\", \"Tyny", 'Dynamiitti-Lahti\', \'Tyny', regex = True)
data = data.replace("Himself: \"Mr", 'Himself: \'Mr', regex = True)
```

1.1.5 Removing duplicates using pandas library based on id of each row

[40]:

```
data = data.drop_duplicates(subset = 'id');
```

1.1.6 Creating new csv file that will be imported on our database

[41]:

```
data.to_csv('credits_wo_duplicates.csv', encoding='utf-8', index = False)
```

1.1.7 Editing csv file text errors, and removing rows with null values from 'links.csv' file

[42]:

```
data = pd.read_csv("links.csv").replace('\'', '"', regex = True)
data = data.fillna(-1)
data = data.astype('int64')
```

1.1.8 Removing duplicates using pandas library based on movieId, imdbId, tmdbId of each row

[43]:

```
data = data.drop_duplicates(subset = ['movieId', 'imdbId', 'tmdbId'])
```

1.1.9 Creating new csv file that will be imported on our database

[44]:

```
data.to_csv('links_wo_duplicates.csv', encoding='utf-8', index = False)
```

1.1.10 Editing csv file text errors, and removing rows with null values from ‘links.csv’ file

[45]:
```
data = pd.read_csv("movies_metadata.csv").replace("\"", "\'", regex = True).
 ↪replace('\':', '":', regex = True)
data = data.replace(", \'", ", \"", regex = True)
data = data.replace("{\'", '{"', regex = True)
data = data.replace(": \'", ': "', regex = True)
data = data.replace("\'}", '"}', regex = True)
data = data.replace("\\\\", ' ', regex = True)
data = data.replace("\', \W", "\", \"", regex = True)
data = data.replace("\", \W", "\", \"", regex = True)
data = data.replace("None", "\"None\"", regex = True)
data = data.replace("tt", " ", regex = True);
data = data.replace("\"\"None\"theless Productions", "\"\'None\'theless␣
 ↪Productions", regex = True)
```

1.1.11 Removing duplicates using pandas library based on movieId, imdbId, tmdbId of each row

[46]:
```
data = data.drop_duplicates(subset = ['id'])
```

1.1.12 Creating new csv file that will be imported on our database

[47]:
```
data.to_csv('movies_metadata_wo_duplicates.csv', encoding='utf-8', index =␣
 ↪False)
```