Wikipedia Movie Plots

This paper was done within the course **"Deep Learning"** at the University of Thessaly, Department of Computer Science and Biomedical Informatics. Prepared by: *Christina Panagiota Kommata, 4th year. AM: 01637*

Summary of the paper

In this paper, we will be reviewing an existing work on Wikipedia Movie Plots and the way it works. We will be seeing the bnefits of a system like this, along with the coding that has been used to create this project. This assignment will contain: A summary of the problem and how to solve it. (Assumptions made)

- Brief bibliographic review of the problem.
- Brief description of the methods used.
- Presentation of executable code (Python) and results following the flow execution.
- Commentary and interpretation of results.

> This dataset contains features from 34,886 movies from around the world. One of the goals is to predict the type of film based on its description plot of the film.

Introduction to Wikipedia

Wikipedia is a multilingual free online encyclopedia written and maintained by a community of volunteers through open collaboration and a wiki-based editing system. Individual contributors, also called editors, are known as Wikipedians. Wikipedia is the largest and most-read reference work in history. It is consistently one of the 10 most popular websites ranked by the Similarweb and former Alexa; as of 2022, Wikipedia was ranked the 7th most popular site.It is hosted by the Wikimedia Foundation, an American non-profit organization funded mainly through donations.

On January 15, 2001, Jimmy Wales and Larry Sanger launched Wikipedia. Sanger coined its name as a blend of "wiki" and "encyclopedia." Wales was influenced by the "spontaneous order" ideas associated with Friedrich Hayek and the Austrian School of economics, after being exposed to these ideas by Austrian economist and Mises Institute Senior Fellow Mark Thornton.Initially available only in English, versions in other languages were quickly developed. Its combined editions comprise more than 58 million articles, attracting around 2 billion unique device visits per month and more than 17 million edits per month (1.9 edits per second) as of November 2020.In 2006, Time magazine stated that the policy of allowing anyone to edit had made Wikipedia the "biggest (and perhaps best) encyclopedia in the world."

Wikipedia has received praise for its enablement of the democratization of knowledge, extent of coverage, unique structure, culture, and reduced degree of commercial bias; but criticism for exhibiting systemic bias, particularly gender bias against women and alleged ideological bias. Its reliability was frequently criticized in the 2000s but has improved over time, as Wikipedia has been generally praised

in the late 2010s and early 2020s. The website's coverage of controversial topics such as American politics and major events like the COVID-19 pandemic has received substantial media attention. It has been censored by world governments, ranging from specific pages to the entire site. Nevertheless, Wikipedia has become an element of popular culture, with references in books, films, and academic studies. In April 2018, Facebook and YouTube announced that they would help users detect fake news by suggesting fact-checking links to related Wikipedia articles. Articles on breaking news are often accessed as a source of frequently updated information about those events.

Movie plots

In a literary work, film, or other narrative, the plot is the sequence of events where each affects the next one through the principle of cause-and-effect. The causal events of a plot can be thought of as a series of events linked by the connector "and so". Plots can vary from the simple—such as in a traditional ballad—to forming complex interwoven structures, with each part sometimes referred to as a subplot or imbroglio. Plot is similar in meaning to the term storyline.In the narrative sense, the term highlights important points which have consequences within the story, according to American science fiction writer Ansen Dibell.The term plot can also serve as a verb, referring to either the writer's crafting of a plot (devising and ordering story events), or else to a character's planning of future actions in the story. The term plot, however, in common usage (for example, a "movie plot") can mean a narrative summary or story synopsis, rather than a specific cause-and-effect sequence.

The definition of a plot is: Early 20th-century English novelist E. M. Forster described plot as the cause-and-effect relationship between events in a story. According to Forster, "The king died, and then the queen died, is a story, while The king died, and then the queen died of grief, is a plot." Teri Shaffer Yamada, Ph.D., of CSULB agrees that a plot does not include memorable scenes within a story which do not relate directly to other events but only "major events that move the action in a narrative." For example, in the 1997 film Titanic, when Rose climbs on the railing at the front of the ship and spreads her hands as if she's flying, this scene is memorable but does not directly influence other events, so it may not be considered as part of the plot. Another example of a memorable scene which is not part of the plot occurs in the 1980 película The Empire Strikes Back, when Han Solo is frozen in carbonite.

About the Dataset

The dataset contains descriptions of 34,886 movies from around the world. Column descriptions are listed below:

- Release Year - Year in which the movie was released
- Title - Movie title
- Origin/Ethnicity - Origin of movie (i.e. American, Bollywood, Tamil, etc.)
- Director - Director(s)
- Plot - Main actor and actresses
- Genre - Movie Genre(s)
- Wiki Page - URL of the Wikipedia page from which the plot description was scraped

- Plot - Long form description of movie plot **WARNING: May contain spoilers!!!**

Inspiration for this project was found in:

Content-Based Movie Recommender: Recommend movies with plots similar to those that a user has rated highly.

Movie Plot Generator: Generate a movie plot description based on seed input, such as director and genre

Information Retrieval: Return a movie title based on an input plot description

Text Classification: Predict movie genre based on plot description

Acknowledgements This data was scraped from Wikipedia

Code that was used for the classification of the dataset (in Python)

In this part, we will be seeing the code that was used combined with the file "/kaggle/input/wikipedia-movie-plots/wiki_movie_plots_deduped.csv" that contains the dataset of the movie plots.

Exploratory Analysis To begin this exploratory analysis, first use `matplotlib` to import libraries and define functions for plotting the data.

First we will import the libraries that we need to use for our program.

Those libraries are for plotting,accessing directory structure, data processing

In [1]:

```
from mpl_toolkits.mplot3d import Axes3D
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt # plotting
import numpy as np # linear algebra
import os # accessing directory structure
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

There is 1 csv file in the current version of the dataset:

In [2]:

```
print(os.listdir('../input'))
```

```
['wiki_movie_plots_deduped.csv']
```

Following the previous step, we will be handling the distribution of the graphs by creating the rows and columns

Show hidden code

On the next step, we can see the correlation matrix for this file

Show hidden code

Here, the code helps with the scatter and density plots

Show hidden code

Now it is ready to read the data that we give to the program, which is the file that contains approximately 35.000 movies

In [6]:

```
nRowsRead = 1000 # specify 'None' if want to read whole file
# wiki_movie_plots_deduped.csv has 34886 rows in reality, but we are only loading/previewing the first 1000 rows
df1 = pd.read_csv('../input/wiki_movie_plots_deduped.csv', delimiter=',', nrows = nRowsRead)
df1.dataframeName = 'wiki_movie_plots_deduped.csv'
nRow, nCol = df1.shape
print(f'There are {nRow} rows and {nCol} columns')
```

```
There are 1000 rows and 8 columns
```

Here we can see what the data looks like!

In [7]:

```
df1.head(5)
```

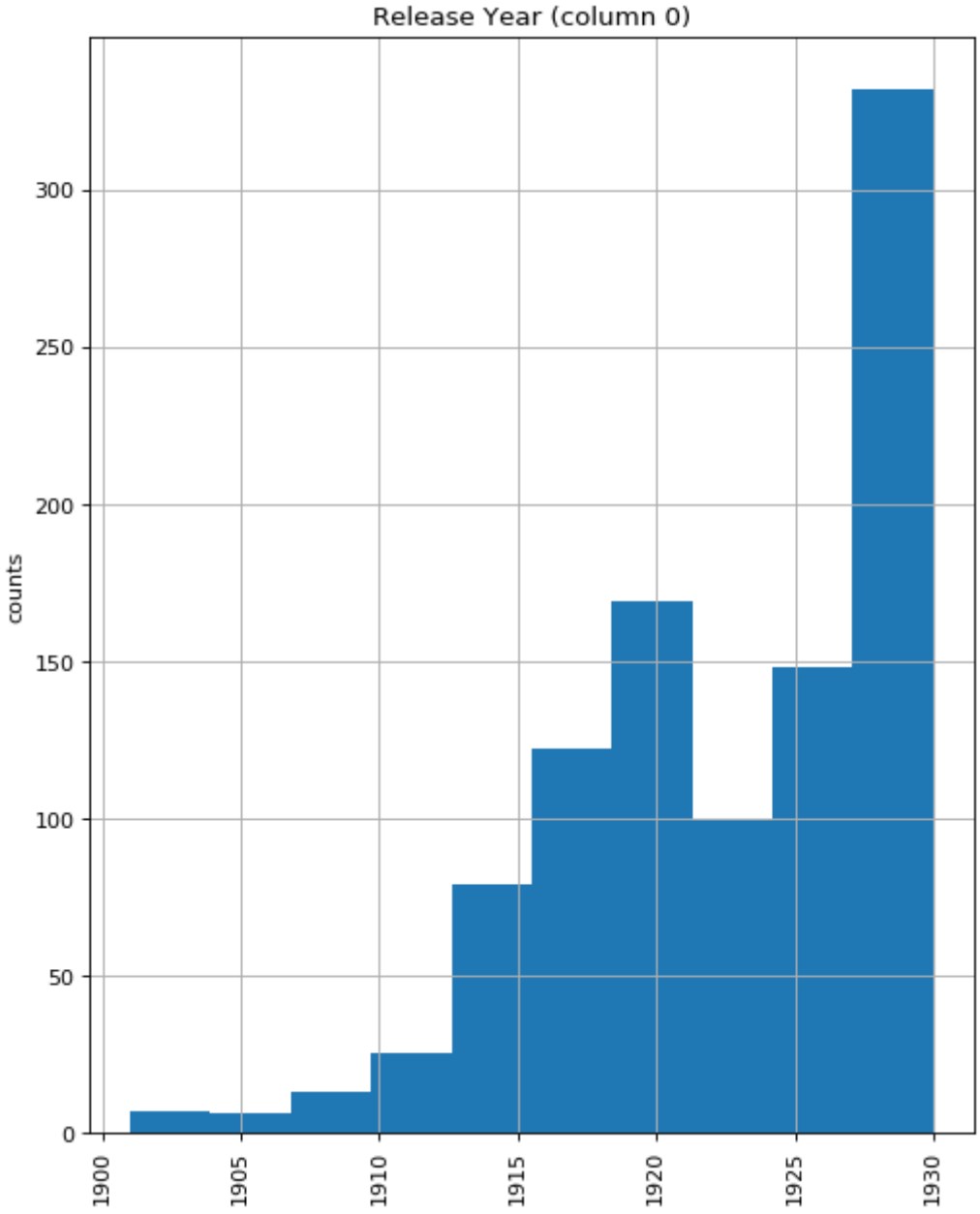
Out[7]:

	Release Year	Title	Origin/Ethnicity	Director	Cast	Genre	Wiki Page
0	1901	Kansas Saloon Smashers	American	Unknown	NaN	unknown	https://en.wikipedia.org/
1	1901	Love by the Light of the Moon	American	Unknown	NaN	unknown	https://en.wikipedia.org/
2	1901	The Martyred Presidents	American	Unknown	NaN	unknown	https://en.wikipedia.org/
3	1901	Terrible Teddy, the Grizzly King	American	Unknown	NaN	unknown	https://en.wikipedia.org/
4	1902	Jack and the Beanstalk	American	George S. Fleming, Edwin S. Porter	NaN	unknown	https://en.wikipedia.org/

And also we can see the distribution graphs regarding the movies and release year

In [8]:

```
plotPerColumnDistribution(df1, 10, 5)
```



Bibliography

- https://web.archive.org/web/20100313003400/http://www.linuxlibertine.org/index.php?id=2&L=1 (https://web.archive.org/web/20100313003400/http://www.linuxlibertine.org/index.php?id=2&L=1)
- http://www.npr.org/blogs/alltechconsidered/2010/05/12/126789933/new-globe-new-user-interface-for-wikipedia\ (http://www.npr.org/blogs/alltechconsidered/2010/05/12/126789933/new-globe-new-user-interface-for-wikipedia\)
- https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots (https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots)
- https://web.archive.org/web/20150821080004/http://www.ipl.org/div/farq/plotFARQ.html (https://web.archive.org/web/20150821080004/http://www.ipl.org/div/farq/plotFARQ.html)
- http://shakespearequotesandplays.com/2016/07/18/plot-definition-meaning-examples-whatisplot/ (http://shakespearequotesandplays.com/2016/07/18/plot-definition-meaning-examples-whatisplot/)
- https://archive.today/20140610141519/http://www.sadovaya6.ru/multimedia/y-slavutin-v-pimonov-the-minimal-plot/ (https://archive.today/20140610141519/http://www.sadovaya6.ru/multimedia/y-slavutin-v-pimonov-the-minimal-plot/)