

Digital humanities

Cvičení: Vizualizace dat

Jindřich Marek

Přehled

- toto cvičení se zaměřuje na vytváření efektivních datových vizualizací pomocí knihovny Plotly Express
- pracujeme s datasetem českých filmů z IMDB a vytváříme interaktivní grafy pro jejich analýzu

Cíle cvičení

- po dokončení tohoto cvičení budete schopni:
 - pochopit principy efektivní datové vizualizace
 - používat Plotly Express pro vytváření interaktivních grafů
 - vytvářet různé typy grafů (pruhový, rozptylový, krabicový, houslový apod.)
 - filtrovat a transformovat data pro lepší prezentaci

Požadované nástroje

- stejné jako v předchozím cvičení

Dataset: české filmy z IMDB

- zdroj: Kaggle z Internet Movie Database (IMDB)
- vyfiltrováno v rámci cvičení č. 2
- formát: CSV
- počet záznamů: 200-300 filmů
- klíčové sloupce
 - stejné jako ve cvičení č. 2

Postup

- import knihoven a načtení dat
- příprava dat - filtrování, transformace
- analýza základních statistik datasetu
- vytváření jednotlivých vizualizací

Explorativní vizualizace dat v Pythonu

- notebook:
 - imdb_ceske_filmy.ipynb
- dataset:
 - final-dataset_filtered.csv
 - vznikl při cvičení k explorativní analýze dat v OpenRefine
- cílem je poznat hlouběji data k českým filmům z většího datasetu
- postupujeme stejně jako v minulém cvičení
 - viz instrukce tam

České filmy od roku 1920 z datasetu s finančně úspěšnými filmy v IMDB

1. Import knihoven a načtení dat

```
[1]: # načte knihovnu pandas a pojmenuje ji jako pd (pro práci s datovými rámci)
import pandas as pd

# načte Plotly Express - jednoduché rozhraní pro tvorbu interaktivních grafů
import plotly.express as px

# načte Plotly Graph Objects - detailnější kontrolu nad grafickými objekty
import plotly.graph_objects as go

# funkce pro tvorbu figure s více podgrafy (subplots)
from plotly.subplots import make_subplots

# načte knihovnu NumPy a pojmenuje ji jako np (pro numerické výpočty a pole)
import numpy as np
```

```
[2]: # načtení dat z CSV souboru do pandas DataFrame
df = pd.read_csv("final-dataset_filtered.csv")

# výpis základní informace o rozměrech datové sady (počet řádků, počet sloupců)
print(f"Rozměry datasetu: {df.shape}")

# oznámení, že následuje ukázka prvních řádků
print("\nPrvní řádky:")

# vrátí DataFrame s prvních 5 řádky; v notebooku se automaticky vykreslí
df.head()
```

Rozměry datasetu: (247, 23)

První řádky:

[2]:		id	title	duration	mpa	rating	votes	méta_score	description	movie_link	writers	...	opening_weekend_gross	gross_worldwide	gross_us_canada	release_date	countries_origin
0	tt0176345	Zlaté kapradí	115	NaN	7.0	166.0	NaN	Czech fairy tale about a shepherd who finds a ...	https://www.imdb.com/title/tt0176345/	['Jan Drda', 'Jirí Weiss']	...	NaN	NaN	NaN	1963-11-22T00:00:00Z	['Czechoslovakia']	
		Snail on a Rock							https://www.imdb.com/title/tt0176345/	['Jaroslav Hrubý']							

Volitelný úkol

Vytvořte vlastní vizualizaci, která zodpoví jednu z následujících otázek:

1. Jak se změnila průměrná délka filmu v průběhu desetiletí?
2. Kteří režiséři mají filmy s nejvyšším hodnocením?
3. Existuje korelace mezi délkou filmu a jeho hodnocením?