

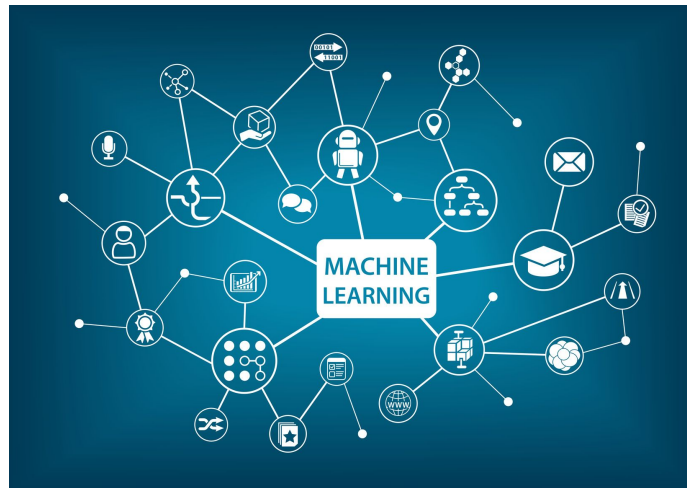
DW Poznań - projekt filmweb-rekomendacje #5

2020-02-07

Co udało nam się zrobić podczas projektu

Agenda

01. Pokaz projektu,
02. Motywacja
03. Historia projektu
04. Struktura projektu
05. Wyzwania
06. Kodowanie produktu
07. Zakończenie i pytania



<https://github.com/alexiei/filmweb-rekomendacje>

Motywacja

- standardowe początki z uczeniem maszynowym:

```
df = pd.read_csv('assets/train.csv')  
df.head()
```

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|-------------|----------|--------|---|--------|------|-------|-------|-----------|---------|-------|----------|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | | | | | | | |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | | | | | | | |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | | | | | | | |

iris setosa



petal sepal

iris versicolor



petal sepal

iris virginica



petal sepal

- własne dane - <https://www.filmweb.pl/user/Kapela86>

Motywacja

- jakość pozyskanych danych

```
In [3]: data = pd.read_csv('oceny.csv', parse_dates=['Data'])
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1192 entries, 0 to 1191
Data columns (total 10 columns):
ID                1192 non-null int64
Tytuł polski      1192 non-null object
Tytuł oryginalny  904 non-null object
Rok produkcji    1192 non-null int64
Ulubione          3 non-null object
Ocena             1192 non-null object
Komentarz         0 non-null float64
Kraj produkcji   1192 non-null object
Gatunek           1192 non-null object
Data              1192 non-null datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(2), object(6)
memory usage: 93.2+ KB
```

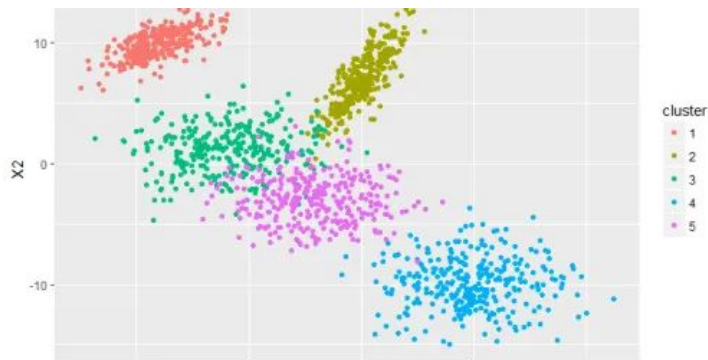
- dodatkowe dane - <https://github.com/ajbrzoz/FWapi>

Motywacja

- przygotowanie danych
 - puste kolumny (ulubione, komentarz)
 - one hot encoding (kraj produkcji, gatunek):

| kaj produkcji | | | |
|---------------|-------|---------|---|
| usa | chiny | francja | |
| USA, Chiny | 1 | 1 | 0 |
| Francja | 0 | 0 | 1 |
| USA | 1 | 0 | 0 |

- K-means (budżet, boxoffice):



Motywacja



Klasyfikacja ocen (1-10)

- DecisionTreeClassifier
- RandomForestClassifier
- KNeighborsClassifier
- SVC
- ...

Accuracy Score : 0.2073732718894009

Precision: 0.26
Accuracy: 0.26
F1: 0.26



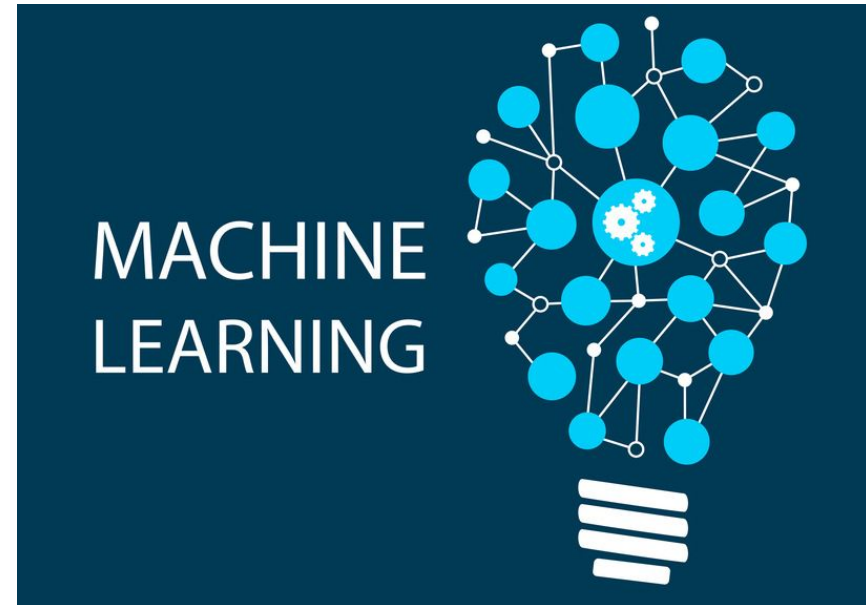
Accuracy Score : 0.2350230414746544

Report :

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 2.0 | 0.00 | 0.00 | 0.00 | 6 |
| 3.0 | 0.00 | 0.00 | 0.00 | 28 |
| 4.0 | 0.00 | 0.00 | 0.00 | 21 |
| 5.0 | 0.00 | 0.00 | 0.00 | 34 |
| 6.0 | 0.24 | 1.00 | 0.38 | 51 |
| 7.0 | 0.00 | 0.00 | 0.00 | 54 |
| 8.0 | 0.00 | 0.00 | 0.00 | 20 |
| 9.0 | 0.00 | 0.00 | 0.00 | 1 |
| 10.0 | 0.00 | 0.00 | 0.00 | 2 |
| accuracy | | | 0.24 | 217 |
| macro avg | 0.03 | 0.11 | 0.04 | 217 |
| weighted avg | 0.06 | 0.24 | 0.09 | 217 |

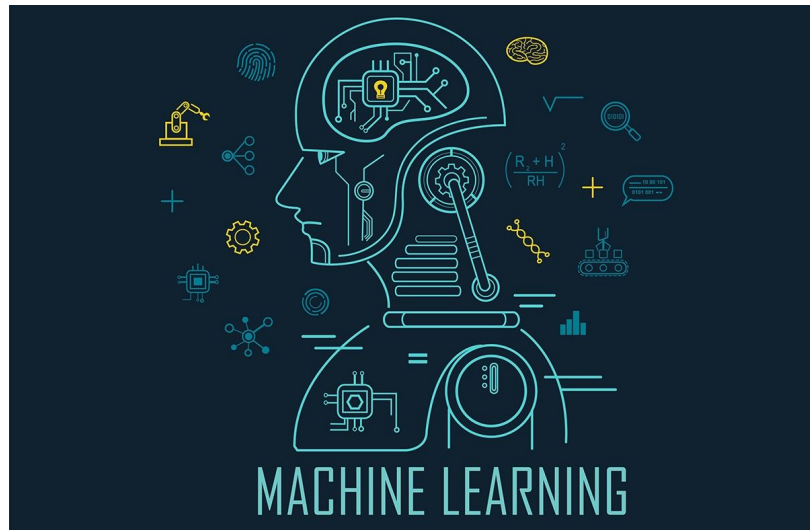
Historia projektu

- Pierwsze spotkanie - wymiana doświadczeń.
- Wymiana pomysłów, ale wygrała idea ...
- Utworzenie repo na github
- Potem było z już łatwiej ...



Historia projektu

- ... nie do końca łatwiej
- Dwie nie związane z sobą bazy danych: Filmweb i IMDB
- Google extension - mała wygrana
- Spotkanie odnośnie modeli rekomendacji
- Kodzenie
- Finał

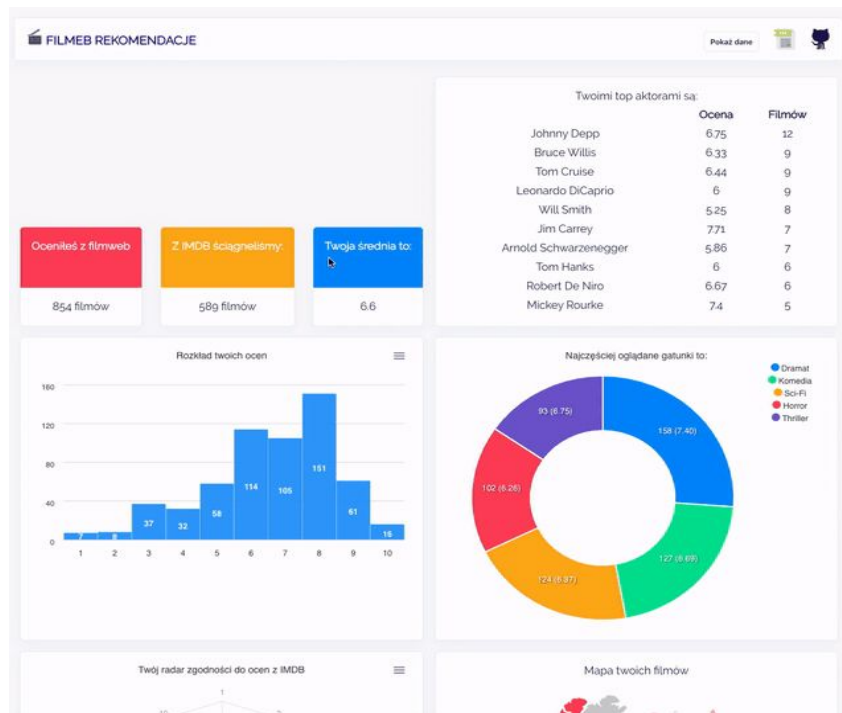


Historia projektu

- Zaczęliśmy 22-10-2019
- Nigdy projekty nie kończą się o czasie
- Gotowy produkt na stronie pod koniec stycznia



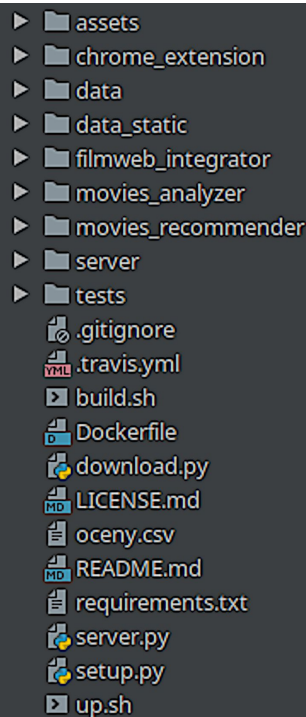
Struktura Projektu



Struktura Projektu - biblioteki



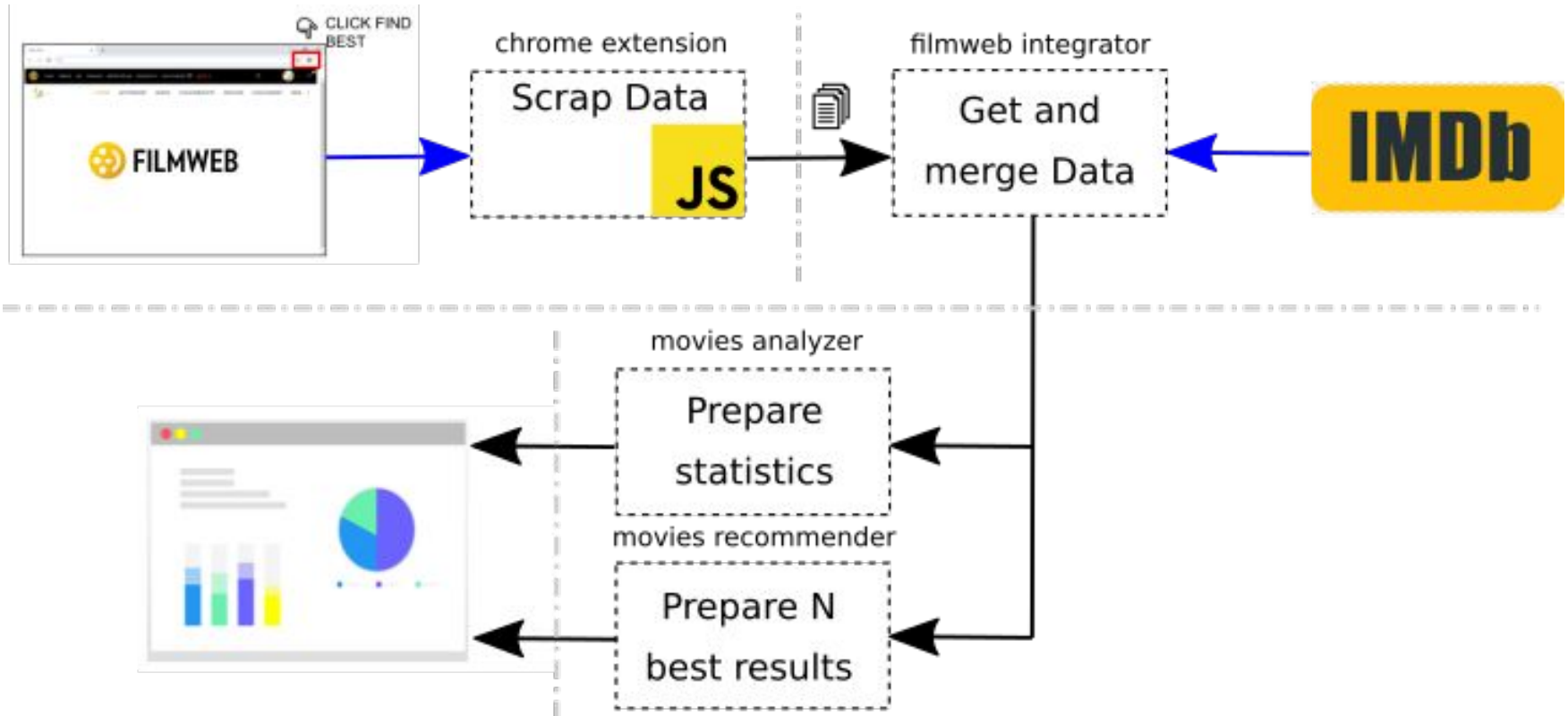
Struktura Projektu - fizyczna



A screenshot of a file explorer window with a dark background, showing the physical structure of a project. The left sidebar shows a tree view with folders expanded. The main area lists the files and folders. The folders are: assets, chrome_extension, data, data_static, filmweb_integrator, movies_analyzer, movies_recommender, server, and tests. The files are: .gitignore, .travis.yml, build.sh, Dockerfile, download.py, LICENSE.md, oceny.csv, README.md, requirements.txt, server.py, setup.py, and up.sh.

- ▶ assets
- ▶ chrome_extension
- ▶ data
- ▶ data_static
- ▶ filmweb_integrator
- ▶ movies_analyzer
- ▶ movies_recommender
- ▶ server
- ▶ tests
- .gitignore
- .travis.yml
- build.sh
- Dockerfile
- download.py
- LICENSE.md
- oceny.csv
- README.md
- requirements.txt
- server.py
- setup.py
- up.sh

Struktura Projektu - logiczna



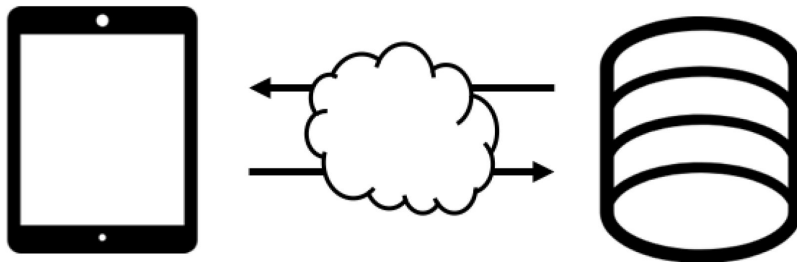
Wyzwania

- Wielkość plików
 - baza IMDB ok. 500MB
- Filtrowanie danych IMDB
 - typ “movie”
- Ładowanie plików
 - pickle 50MB -> parquet 15MB
- Połączenie Filmweb i IMDB: brak wspólnego klucza
 - łączenie po nazwie i roku produkcji
 - duplikaty!
 - różne gatunki filmów pomiędzy dwoma bazami - słownik gatunków
 - Problemy z tłumaczeniem filmu i gatunków: *anime/animation/cartoon* -> *animacja*
 - podobieństwo na podstawie gatunków
- Szybkość API
 - scrapping bezpośrednio ze strony filmweb.pl :(

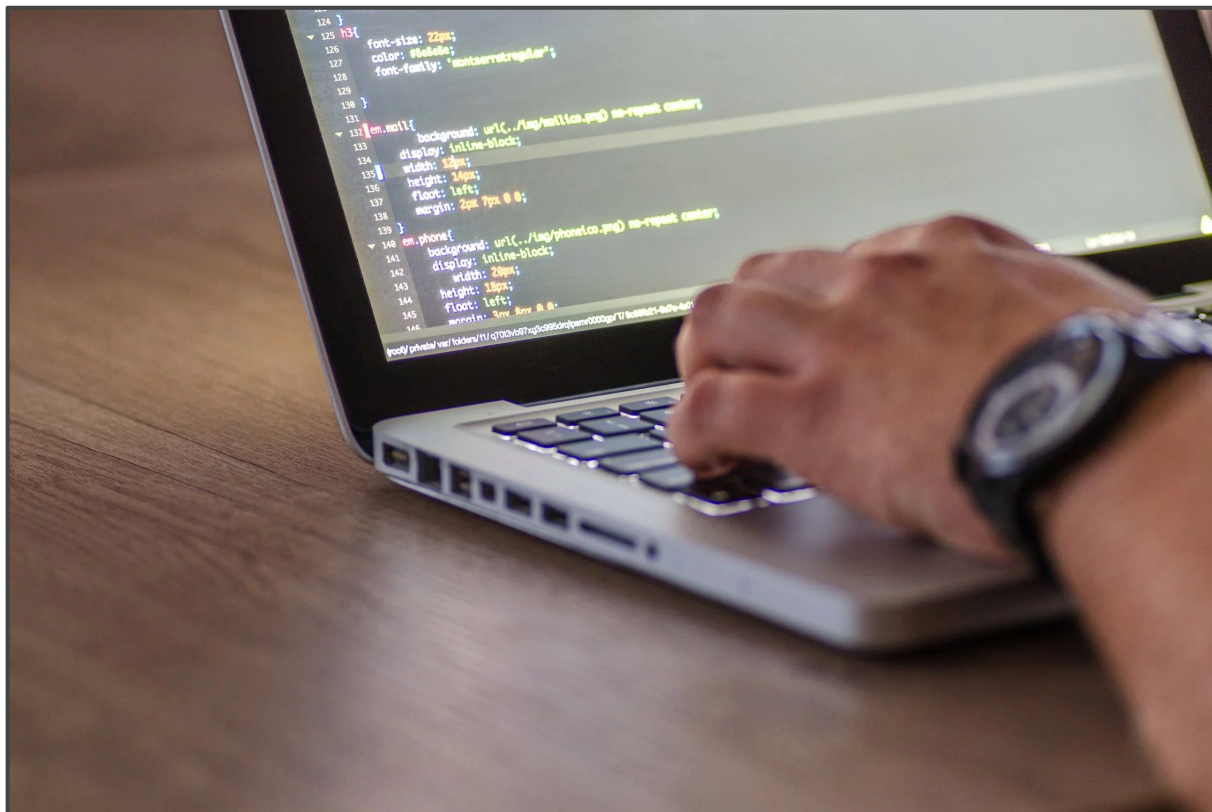


Wyzwania

- Wydajność algorytmów rekomendacyjnych:
 - ItemBased -> 20 sekund liczenie, 0.6 sekundy odpowiedź
 - SVD -> 6 sekund liczenie, 6 sekund odpowiedź
 - SVDpp -> 872 sekundy liczenie, 850 sekund odpowiedź
 - UserBased -> 0.4 sekundy liczenie, 0.5 sekundy odpowiedź
 - SVDSimilarUser -> 6 sekund liczenie, 0.7 sekundy odpowiedź
 - SVDppSimilarUser -> 800 sekund liczenie, 2 sekundy odpowiedź

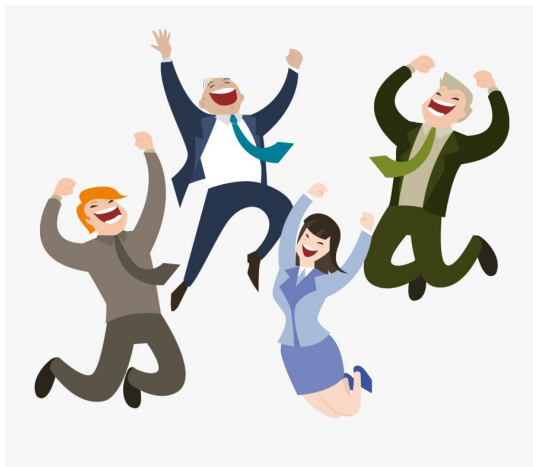


Kodowanie produktu



Podsumowanie

- Było to spore wyzwanie
- Dużo do nauki
- Integracja wielu systemów nigdy nie jest łatwa
- Sukces!



Kolejne kroki

- Spotkanie przy piwie (środa: 12-go lutego, 2020)
- Proszę o wypełnienie formularza ze spotkania :)

<https://forms.gle/LDKuSTJeqnzSQLbe9>

<https://github.com/dataworkshop/dw-poznan-project>



Dziękuję