

# System rekomendacji filmów bazujący na modelach językowych

---

Szymon Fica


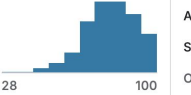
# Cel projektu

Stworzenie modelu rekomendującego filmy, gdzie wejściem jest tytuł filmu, a wyjściem tytuł innego, podobnego filmu.

Generowanie krótkich opisów filmów na podstawie tytułu, gatunków i reżysera.



# Zbiór danych

Poster_Link	Series_Title	Released_Year		Certificate		Runtime		Genre		# IMDB_Rating	Overview	# Meta_score	Director	
Poster Link of Movie	Name of the Movie	Released Year of the Movie		Certificate of the Movie		Total Runtime of the Movie		Genre of the Movie		IMBD Rating of the Movie	Overview of the Movie	Metascore earned by the Movie	Name of the Director	
1000 unique values	999 unique values	2014	3%	U	23%	130 min	2%	Drama	9%		1000 unique values		Alfred Hitchcock	1%
		2004	3%	A	20%	100 min	2%	Drama, Romance	4%				Steven Spielberg	1%
		Other (937)	94%	Other (569)	57%	Other (954)	95%	Other (878)	88%				Other (973)	97%
https://m.media-amazon.com/images/M/MV5BMDFkYTc0MGEtZmNhMC00ZDIzLWFmNTc0ODM1ZmRlYWVmMWFmXkEyXkFqcGde...	The Shawshank Redemption	1994		A		142 min		Drama		9.3	Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts ...	80	Frank Darabont	
https://m.media-amazon.com/images/M/MV5BM2MyNjYxNmUtYTawNi00MTYxLWJmNWYtYzZlODY3ZTk3OTFlXkEyXkFqcGde...	The Godfather	1972		A		175 min		Crime, Drama		9.2	An organized crime dynasty's aging patriarch transfers control of his clandestine empire to his relu...	100	Francis Ford Coppola	
https://m.media-amazon.com/images/M/MV5BMTgxNTMwODM0NF5BM15BanBnXkFtZTcwODAyMTk2Mw@@._V1_UX67_CR0,0,...	The Dark Knight	2008		UA		152 min		Action, Crime, Drama		9	When the menace known as the Joker wreaks havoc and chaos on the people of Gotham, Batman must accep...	84	Christopher Nolan	
https://m.media-amazon.com/images/M/MV5BMWMwMGQzZTItYzJlNC00OWZiLWIyMDctNDk2ZDQ2YjRjMw00XkEyXkFqcGde...	The Godfather: Part II	1974		A		202 min		Crime, Drama		9	The early life and career of Vito Corleone in 1920s New York City is portrayed, while his son, Micha...	90	Francis Ford Coppola	

<https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>

# Word2Vec

Użycie **Word2Vec** do nauki relacji między filmami na podstawie ich cech (gatunek, czas trwania, reżyser, oceny).

Przetwarzanie danych: Ekstrakcja i tokenizacja cech.

```
movie_info = ["genre", "director", "runtime",  
"IMDB_rating"]
```

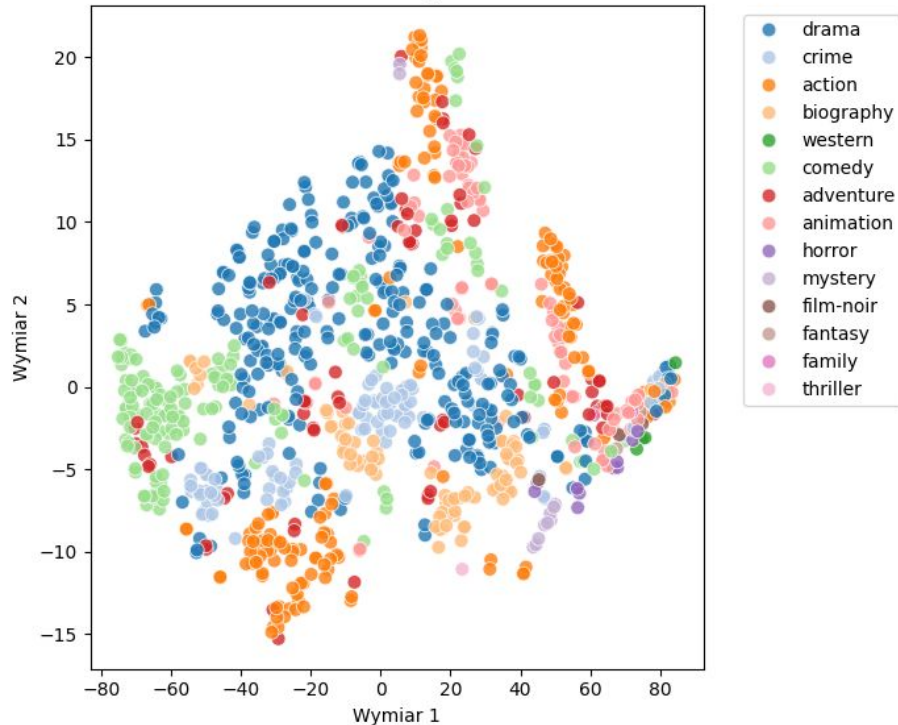
# Bert

Wykorzystanie pretrenowanego **BERT-a** do analizy opisów fabuły i porównywania ich reprezentacji wektorowych.

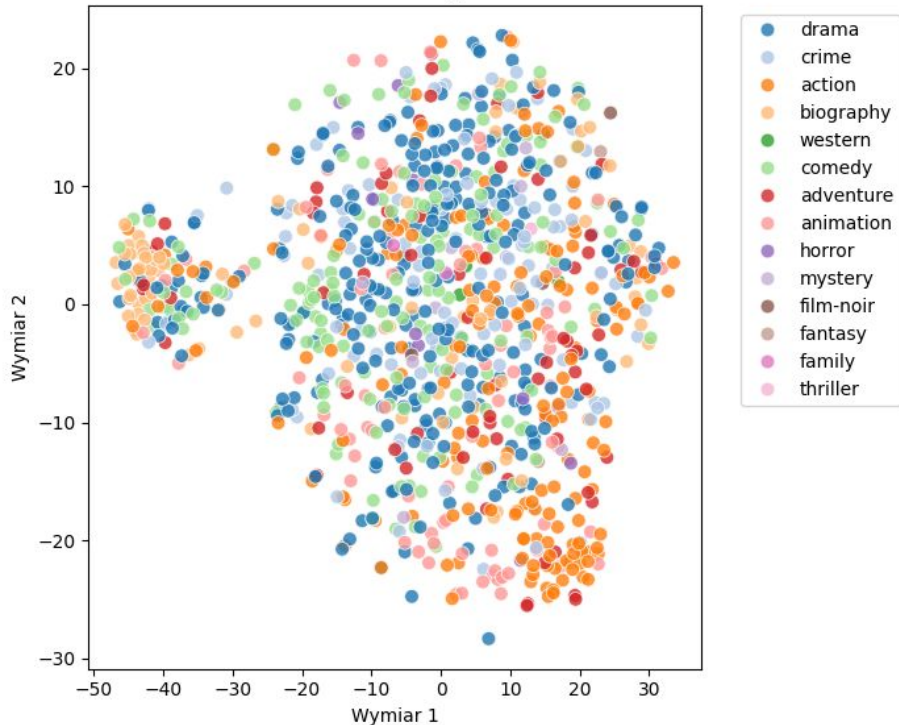
Przetwarzanie danych: Tokenizacja opisów fabuły i obliczenie reprezentacji wektorowej tokenu [CLS].

```
movie_info = [ "title", "overview"]
```

t-SNE dla embeddingów Word2Vec



t-SNE dla embeddingów BERT



# The Godfather

## Rekomendacje oparte na Word2Vec:

The Godfather: Part II (podobieństwo: 0.999)  
The Godfather: Part III (podobieństwo: 0.999)  
After Hours (podobieństwo: 0.999)  
Goodfellas (podobieństwo: 0.999)  
Wind River (podobieństwo: 0.999)  
12 Angry Men (podobieństwo: 0.999)  
Taxi Driver (podobieństwo: 0.999)  
The King of Comedy (podobieństwo: 0.999)  
The Sting (podobieństwo: 0.999)  
Adams æbler (podobieństwo: 0.999)

## Rekomendacje oparte na BERT:

Sunset Blvd. (podobieństwo: 0.922)  
Gladiator (podobieństwo: 0.922)  
Chinatown (podobieństwo: 0.921)  
Dead Man's Shoes (podobieństwo: 0.915)  
Kingsman: The Secret Service (podobieństwo: 0.915)  
Inside Man (podobieństwo: 0.914)  
The Pursuit of Happyness (podobieństwo: 0.913)  
Hell or High Water (podobieństwo: 0.913)  
Dip huet seung hung (podobieństwo: 0.911)  
Barry Lyndon (podobieństwo: 0.909)

# The Dark Knight

## Rekomendacje oparte na Word2Vec:

Lord of War (podobieństwo: 0.999)  
Haider (podobieństwo: 0.999)  
Hell or High Water (podobieństwo: 0.999)  
Män som hatar kvinnor (podobieństwo: 0.999)  
Enter the Dragon (podobieństwo: 0.999)  
3:10 to Yuma (podobieństwo: 0.999)  
Sicario (podobieństwo: 0.999)  
End of Watch (podobieństwo: 0.999)  
Baby Driver (podobieństwo: 0.999)  
The Fugitive (podobieństwo: 0.999)

## Rekomendacje oparte na BERT:

Joker (podobieństwo: 0.891)  
Spider-Man: Into the Spider-Verse (podobieństwo: 0.886)  
Batman Begins (podobieństwo: 0.881)  
Avengers: Infinity War (podobieństwo: 0.877)  
Star Wars: Episode VII – The Force Awakens (podobieństwo: 0.874)  
Thor: Ragnarok (podobieństwo: 0.873)  
The Avengers (podobieństwo: 0.873)  
Harry Potter and the Deathly Hallows: Part 1 (podobieństwo: 0.869)  
Terminator 2: Judgment Day (podobieństwo: 0.867)  
Indiana Jones and the Last Crusade (podobieństwo: 0.866)

---

# Klasyfikacja gatunku - TF-IDF

- Przygotowanie danych: ["Genre", "Overview" i "Director"].  
(Ekstrakcja pierwszego gatunku z listy, który traktujemy jako główny gatunek filmu).
- Przekształcenie tekstu na wektor cech przy użyciu TfidfVectorizer.
- Następnie zastosowany zostaje klasyfikator logistyczny (LogisticRegression), który uczy się przyporządkowywać wektor cech do konkretnego gatunku.
- test\_size=0.2

	precision	recall	f1-score	support
Action	0.57	0.45	0.50	29
Adventure	0.00	0.00	0.00	14
Animation	0.00	0.00	0.00	20
Biography	0.00	0.00	0.00	21
Comedy	0.29	0.07	0.11	30
Crime	0.00	0.00	0.00	16
Drama	0.33	0.87	0.47	63
Fantasy	0.00	0.00	0.00	1
Film-Noir	0.00	0.00	0.00	1
Horror	0.00	0.00	0.00	1
Mystery	0.00	0.00	0.00	4
accuracy			0.35	200
macro avg	0.11	0.13	0.10	200
weighted avg	0.23	0.35	0.24	200

# Generowanie opisów filmów

## Few-shot learning

Wykorzystujemy już wytrenowany model GPT-2 i przekazujemy mu przykłady w promptach. Przygotowaliśmy przykładowe dane wejściowe dla filmów takich jak *Incepcja* i *La La Land*, a następnie generujemy opis na podstawie nowego promptu zawierającego tytuł, gatunki i reżysera. Podejście to jest szybkie i nie wymaga dużych zasobów obliczeniowych, generuje zazwyczaj spójne wyniki w odniesieniu do specyfiki naszych danych, jednak nie zawsze prawdziwe i bardziej ogólne.

## Fine-tuning

Aby uzyskać lepsze dopasowanie do naszego zadania, przeprowadziliśmy fine-tuning modelu GPT-2 na przygotowanych danych z bazy Kaggle. Wykorzystaliśmy bibliotekę Hugging Face Transformers, definiując dataset oraz parametry treningu. Dane treningowe:

Title: <Series\_Title>  
Genres: <Genre>  
Director: <Director>  
Overview: <Overview>



# Wyniki - few-shot

```
title = "Matrix"
genres = "Action, Sci-Fi"
director = "Lana Wachowski, Lilly Wachowski"
description = generate_description_few_shot(title, genres, director)
print("Generated description (few-shot):")
print(description)
```

Generated description (few-shot):  
Matrix is about dreams, love, and music that touches the hearts of viewers.

```
title = "The Joker"
genres = "Crime, Drama, Thriller"
director = "Todd Phillips"

description = generate_description_few_shot(title, genres, director)
print("Generated description (few-shot):")
print(description)
```

Generated description (few-shot):  
In this classic DC film, a man must decide if he wants to be a vigilante or a thief.

```
title = "The Godfather"
genres = "Crime, Drama"
director = "Francis Ford Coppola"
description = generate_description_few_shot(title, genres, director)
print("Generated description (few-shot):")
print(description)
```

Generated description (few-shot):  
A young man is forced into a terrifying fight with a madman. This story will take you on an adventure that will change your life forever.

```
title = "The Dark Knight"
genres = "Action, Crime, Drama"
director = "Christopher Nolan"

description = generate_description_few_shot(title, genres, director)
print("Generated description (few-shot):")
print(description)
```

Generated description (few-shot):  
Dark Knight is an action and suspense-packed movie about the rise of the Dark Knight and how he can take his place in the history of crime.

# Wyniki - fine-tuning

```
title = "Harry Potter and the Philosopher's Stone"
genres = "Family, Fantasy "
director = "Chris Columbus"

description = generate_movie_description(model_tuned, tokenizer_tuned, title, genres, director)
print("Generated description:")
print(description)
```

Generated description:  
Two friends are asked to help Harry and Ron get through the summer holidays and prevent a child from

```
title = "Goodfellas"
genres = "Biography, Crime, Drama"
director = "Martin Scorsese"

description = generate_movie_description(model_tuned, tokenizer_tuned, title, genres, director)
print("Generated description:")
print(description)
```

Generated description:  
A man learns about life in a small town and becomes obsessed with finding out what's going on.

```
title = "The Lord of the Rings: The Return of the King"
genres = "Action, Adventure, Drama"
director = "Peter Jackson"

description = generate_movie_description(model_tuned, tokenizer_tuned, title, genres, director)
print("Generated description:")
print(description)
```

Generated description:  
During the Battle of the Great Hall of the East Room, Gandalf and the other hobbits are surrounded by the armies of Sauron, and as they fight, Bilbo

# Thank You for Your Attention!

Code: [https://github.com/szymonfica/language\\_models/tree/main/project](https://github.com/szymonfica/language_models/tree/main/project)

Dataset: <https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>