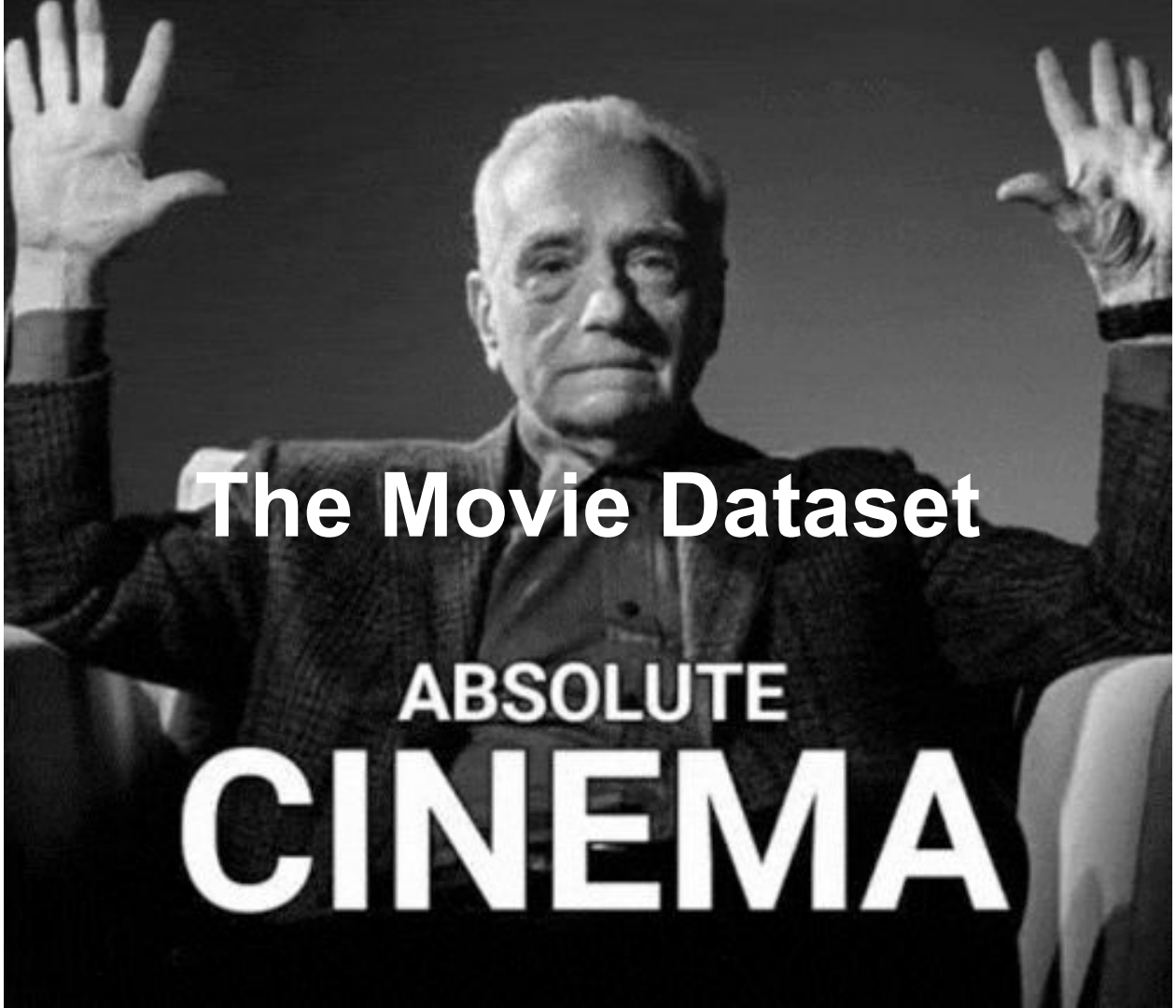
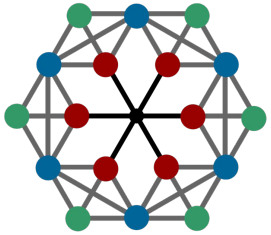
Data Literacy Group 3


The Movie Dataset
ABSOLUTE
CINEMA

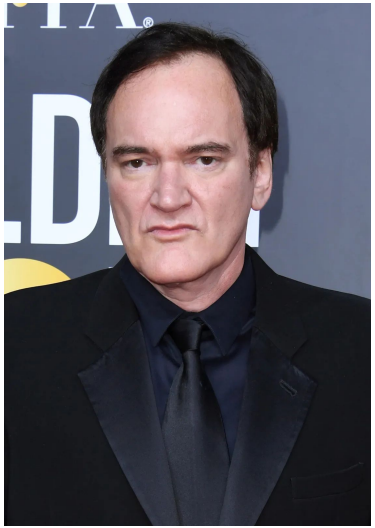
The Movies Dataset



- API access
- Name / year / directors / actors
- Movie reviews

Plot / summary section scraping

The Movies Dataset



Questions to explore:

1. “Quentin Tarantino: The Most Influential Movie Director”
    a. Some clickbait title
    b. Main idea is exploring semantic shifts of movie plot embeddings
    c. Do movies converge toward a particular centroid (genre/director)?
        i. E.g. Fast and furious -> racing becomes action
        ii. When we say converge, it could mean many things
            1. Emotions
            2. Plot
            3. Cycles ?
2. “Hollywood has no new ideas”
    a. Novelty premium or penalty
    b. Do movies who try novel plots and ideas do well or do people actually enjoy the comfort of mainstream plots?
        i. e.g. Movies with bad endings
        ii. Extremely novel plots, e.g. Squid Game

Previous research

1. Screenplay Summarization Using Latent Narrative Structure

   The paper argues that existing extractive summarization models, mostly trained on short news articles, rely too heavily on sentence position and thus fail on long, complex narratives. To address this, the authors propose incorporating narrative structure (specifically, key events or turning points) into both supervised and unsupervised summarization models. They apply this approach to TV screenplays (the CSI corpus), treating turning points as latent variables to extract the most relevant scenes. Experiments show that these latent narrative elements align well with important story moments and produce more complete and diverse summaries than standard extractive methods.

2. Beyond Labels: Leveraging Deep Learning and LLMs for Content Metadata

   The abstract discusses the importance of content metadata in movie recommendation systems, especially focusing on genre labels that categorize titles and shape audience expectations. It highlights the challenges of using traditional genre information and introduces a new concept called the Genre Spectrum, which captures more nuanced genre representations for each title. Experiments show that this approach improves recommendation quality. The talk also explores how large language models (LLMs) can enhance content metadata, enabling more effective and personalized organization of recommendations in a user's home interface.

3. TrUMAn: Trope Understanding in Movies and Animations

   The paper introduces a new task and dataset called TrUMAn (Trope Understanding in Movies and Animations), designed to push video understanding beyond surface-level visual cues toward deeper reasoning about intentions, motivations, and causality. The dataset contains 2,423 videos labeled with 132 storytelling tropes, challenging models to recognize abstract narrative patterns. To address this, the authors propose TrUSt (Trope Understanding and Storytelling), which includes a Conceptual Storyteller module that guides video encoding through latent-space storytelling. Experiments show that existing models perform poorly on this task (around 12% accuracy), and even with human annotations, results remain limited (28%). TrUSt improves performance modestly to 13.94%, and the authors provide analyses to support future research in deep narrative reasoning.

4. Is Cinema Becoming Less and Less Innovative With Time? Using neural network text embedding model to measure cultural innovation

Remarks

Start getting data asap

Careful when handling shows vs movies (we just start with movies first)

Document each finding along the way while processing the data

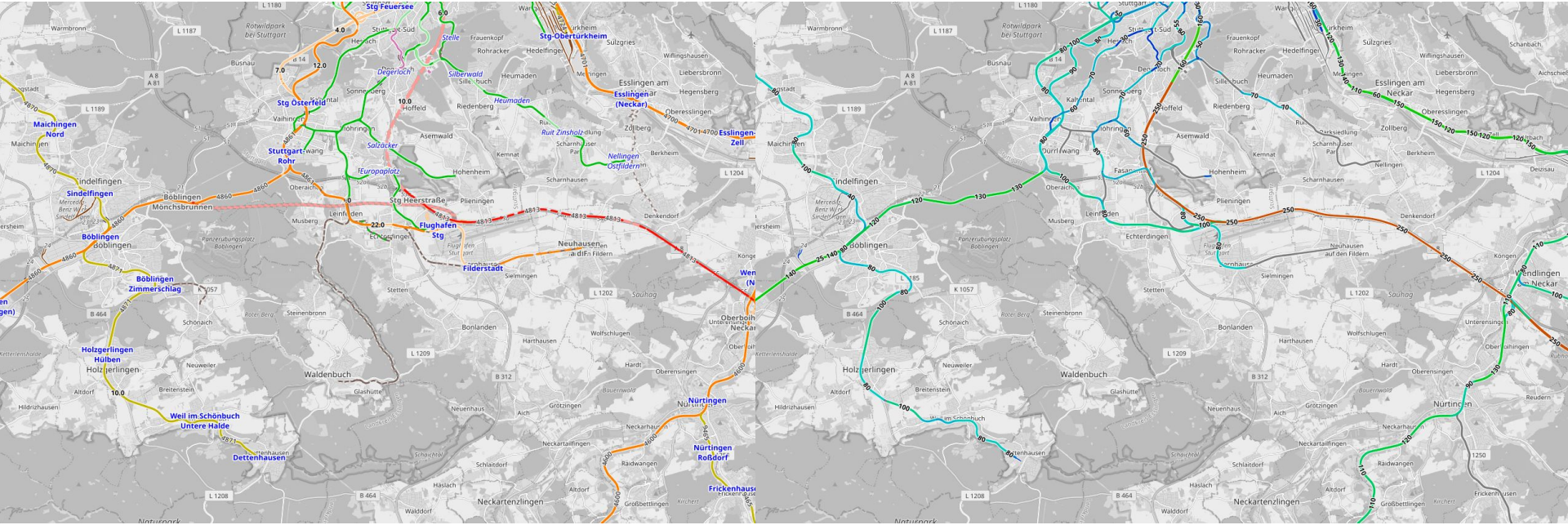
- best not to just “chase a lead”, but rather, follow a main research question
- shouldn't be an issue for us since we have questions we can ask

Why is your Deutsche Bahn late?

Route Features and Train Delay Relationships

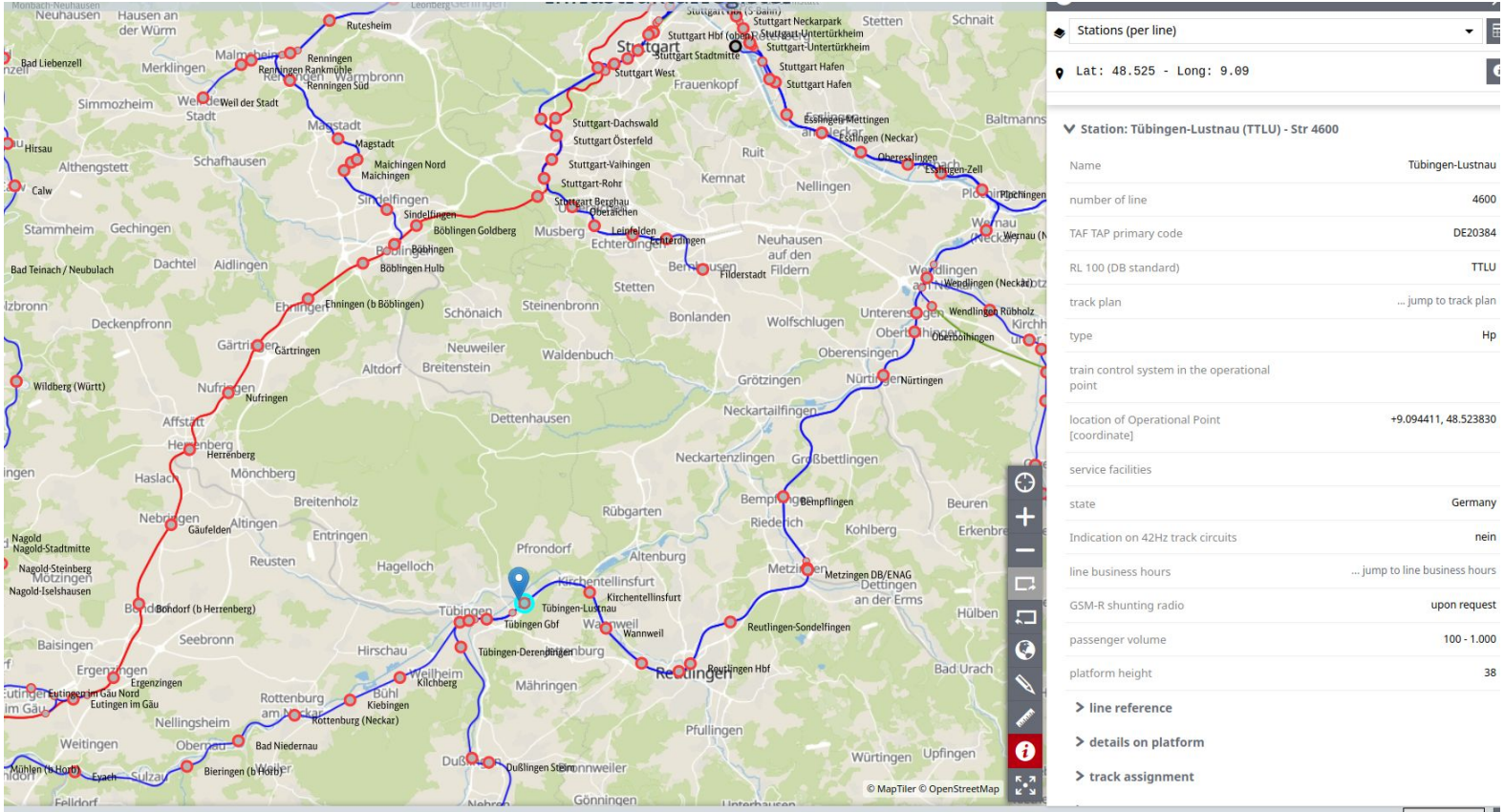
- Analysis of how following variables are related to train delays and the variability of train delays
  - Cumulative ascent/descent
  - Maximal/average gradient
  - Absolute elevation
  - Number of tunnels and bridges
  - Track type
  - Electrified vs non-electrified sections
  - Time features (e.g. time of day, day of week, month)
  - Number of days before a holiday
  - Weather at the time of the trip
  - Train type (but mostly stick to regional trains?)
  - Vegetation near the railway track
- Do train delays vary more on routes with big elevation difference?

Maps and Open Access



Infrastructure, Max speed, Electrification, Track gauge, …

Infrastrukturregister by DB



Dataset

- Deutsche Bahn data
  - github repository with historical delays since 1st July 2024
  - alternatively directly collect data from API for a certain month
- DB Infrastructure API
  - contains detailed informations about train stations and railway tracks including coordinates of the stations and railway segments
- DWD - German Weather Service
  - historical weather data
- Copernicus Digital Elevation Model

Remarks

- Be aware of the granularity of the copernicus data
  - e.g. If trains are per minute, then we need to scrape the minutes geo data
- Also be aware that some of the datasets might be really large data

Group 4

Alessio Villa, Ansel Cheung, Bartol Markovinović, Martin Lopez, Niklas Abraham

Starting point from last week …

First things first: IS IT LEGAL?

Creative Commons Attribution-ShareAlike 4.0 International License

You are free:

- to **Share**: copy and redistribute the material in any medium or format
- to **Adapt**: remix, transform, and build upon the material
- for any purpose, **even commercially.**

The licensor cannot revoke these freedoms as long as you follow the license terms.

- Under the following terms:
- Attribution: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- Share Alike: If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- No additional restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits

Ref: https://en.wikipedia.org/wiki/Wikipedia:Copyrights

Pipeline of the data acquisition

Step 1: WikiData

For each year:

```
    query = f"""
SELECT ?movie ?movieLabel ?sitelinks
       (SAMPLE(?article_) AS ?article)
       (SAMPLE(?plotSummary_) AS ?plotSummary)
       (MIN(?releaseDate_) AS ?releaseDate)
       (SAMPLE(?imdbID_) AS ?imdbID)
       (SAMPLE(?duration_) AS ?duration)
       (MIN(?budget_) AS ?budget)
       (MIN(?boxOffice_) AS ?boxOffice)
       (MIN(?countryLabel_) AS ?countryLabel)
       ?genres ?directors ?actors ?awards ?setPeriods
WHERE {{
```

Step 1: WikiData

For each year:

```
    query = f"""
SELECT ?movie ?movieLabel ?sitelinks
       (SAMPLE(?article_) AS ?article)
       (SAMPLE(?plotSummary_) AS ?plotSummary)
       (MIN(?releaseDate_) AS ?releaseDate)
       (SAMPLE(?imdbID_) AS ?imdbID)
       (SAMPLE(?duration_) AS ?duration)
       (MIN(?budget_) AS ?budget)
       (MIN(?boxOffice_) AS ?boxOffice)
       (MIN(?countryLabel_) AS ?countryLabel)
       ?genres ?directors ?actors ?awards ?setPeriods
```

Todd Phillips (male, United States)	Ben Warheit (male,), Bill Camp (male, United States), Brett

Step 1: WikiData

Todd Phillips (male, United States)	Ben Warheit (male,), Bill Camp (male, United States), Brett

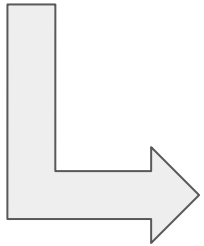
Step 2: MovieDB

- https://github.com/celiao/tmdbsimple/ - python wrapper for MovieDB's api
- Extract movies with Wiki_ID
-
    - ➔ Basic Info
    - ➔ Ratings
    - ➔ Financial Info
    - ➔ Cast
    - ➔ Production

Step 3: Wikipedia

- wikipedia-api - python wrapper for Wikipedia's api

https://en.wikipedia.org/wiki/Inception



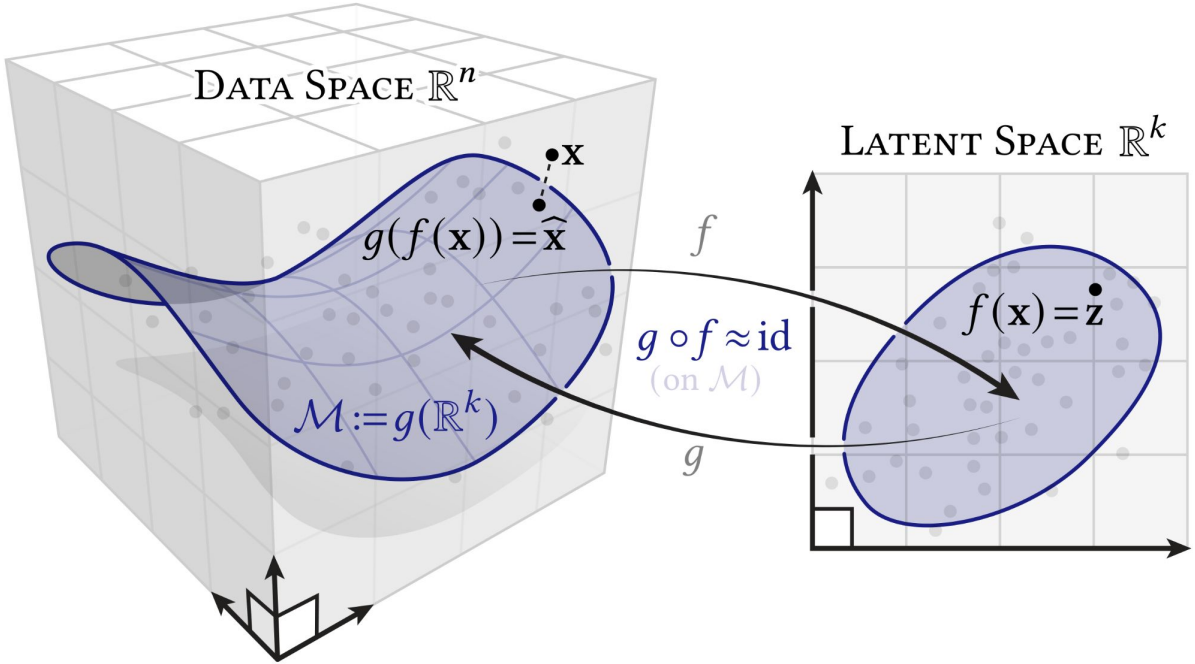
more (Best Picture, Best Original Screenplay, Best Art Direction, Best Original Score) at the 83rd Academy Awards.

Plot [edit]

Dom Cobb and Arthur are "extractors" who perform corporate espionage using experimental dream-sharing technology to infiltrate their targets' subconscious and extract information. Their latest target, Saito, is impressed with Cobb's ability to layer multiple dreams within each other. He offers to hire Cobb for the ostensibly impossible job of implanting an idea into a person's subconscious; performing "inception" on Robert Fischer, the son of Saito's competitor Maurice Fischer, with the idea to dissolve his father's company. In return, Saito promises to clear Cobb's criminal status, allowing him to return home to his children.

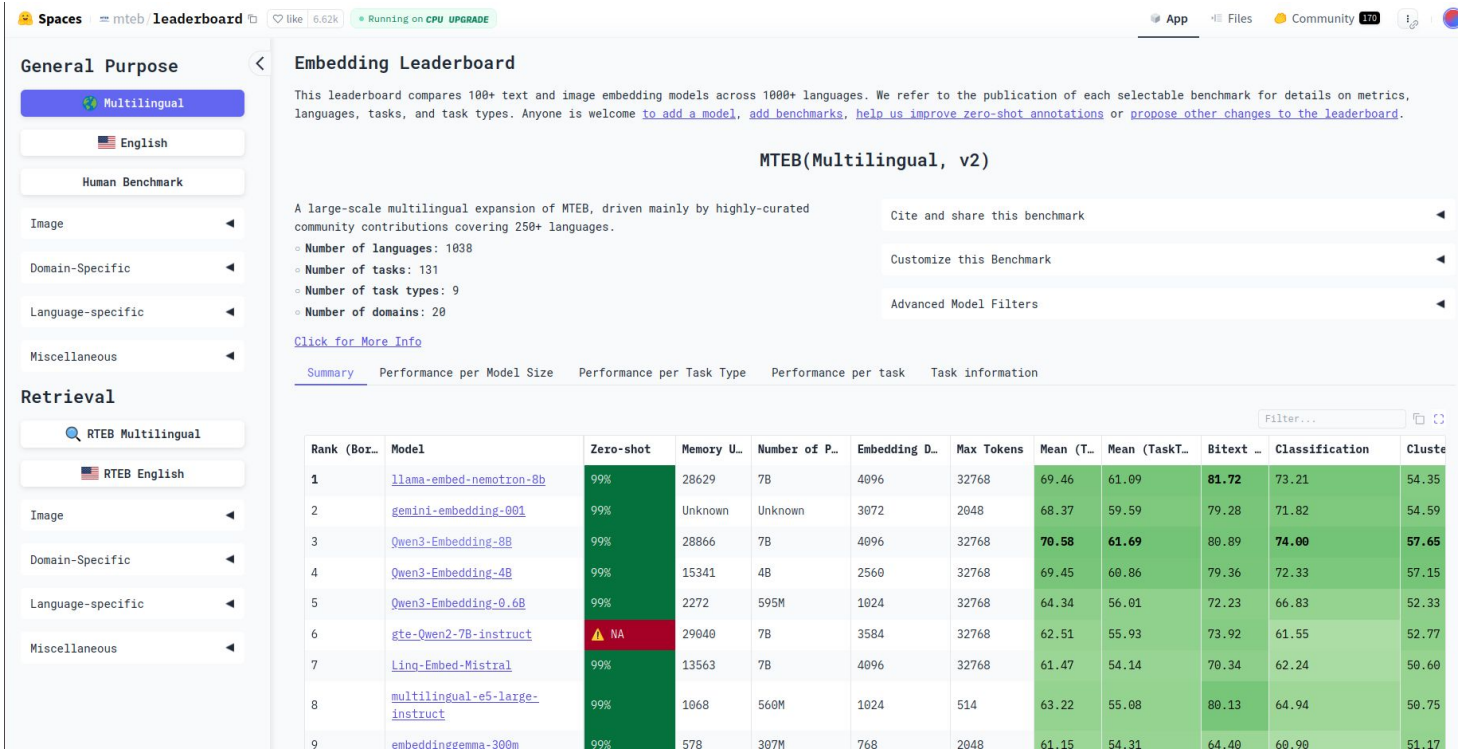
Cobb accepts the offer and assembles his team: a forger named Eames, a chemist named Yusuf, and a college student named Ariadne. Ariadne is tasked with designing the dream's architecture, something Cobb himself cannot do for fear of being sabotaged by his mind's projection of his late wife, Mal. Maurice Fischer dies, and the team sedates Robert Fischer into a three-layer shared dream on an

Step 4: Embedding 1/3



https://www.threads.com/geometry.collective/post/DORqqZpkst2/with-autoencoders-the-first-and-last-picture-we-see-often-looks-like-this-one-a-

Step 4: Embedding 2/3



https://huggingface.co/spaces/mteb/leaderboard

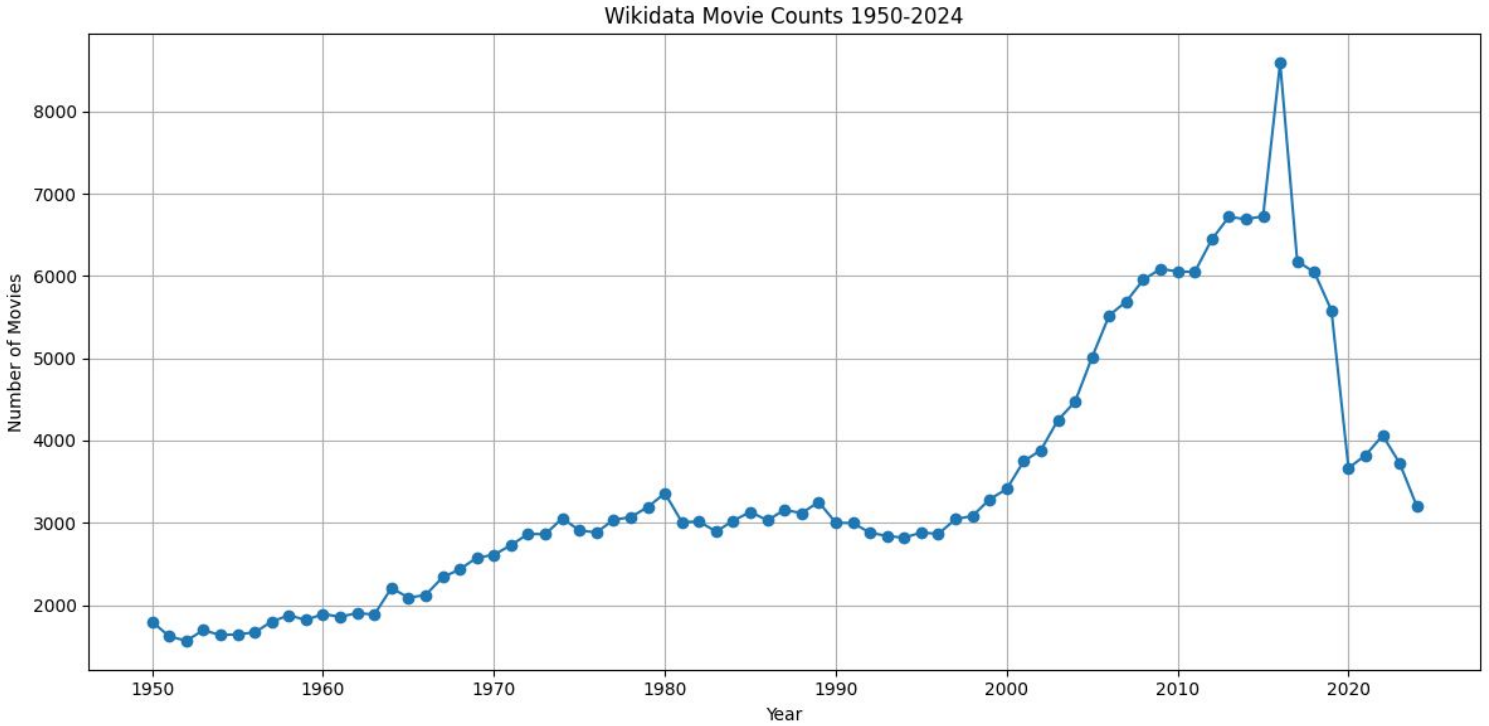
Step 4: Embedding 3/3

25	jina-embeddings-v3	99%	1092	572M	1024	8194	58.37	50.66	65.25	58.77	45.65
26	bge-m3	98%	2167	568M	1024	8194	59.56	52.18	79.11	60.35	40.88
27	KaLM-embedding-multilingual-mini-v1	92%	1885	494M	896	512	57.05	50.05	64.77	57.57	45.61

- in a first test we went with bge-m3
- Pros: quite small 500M, easy to use, fast to run, and very well documented
- Cons: Not benchmarking as number one
- is a hybrid retrieval model, meaning: dense retrieval, sparse retrieval, and multi vector retrieval - *TLDR: many options which we can at some point use*

https://huggingface.co/BAAI/bge-m3

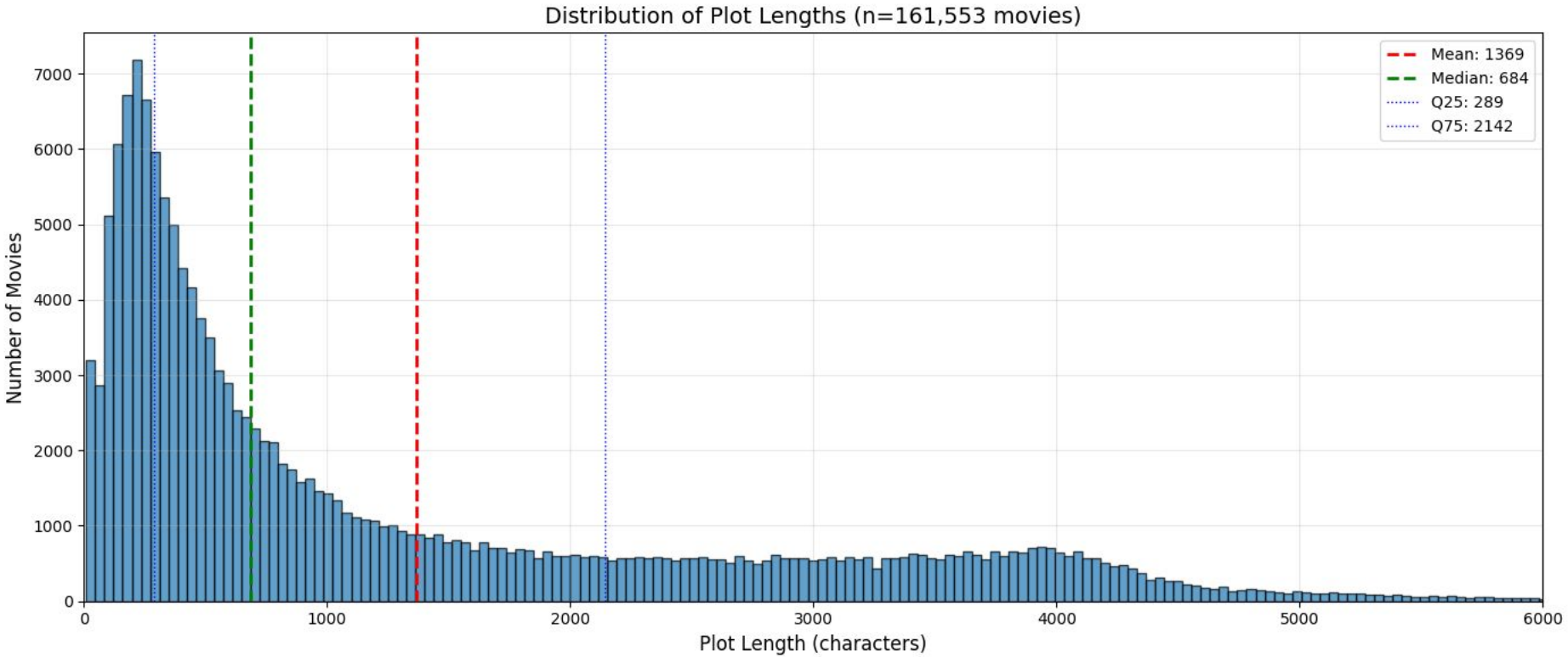
Simple Data Stats - Number of movies



Simple Data Stats - Features

field year	actors	director	duration	genre	imdb_id	plot	popularity	title	tmdb_id	vote_average	vote_count
1950	77.540000	96.6500	68.9400	86.540000	97.9600	91.4600	92.450000	100.0	92.450000	92.450000	92.450000
1951	76.410000	95.5700	62.4300	84.300000	98.0500	91.6600	90.460000	100.0	90.460000	90.460000	90.460000
1952	78.780000	96.5400	64.1900	83.570000	96.6800	87.9900	89.610000	100.0	89.610000	89.610000	89.610000
1953	80.270000	94.9800	58.5100	85.290000	98.5100	92.0700	90.030000	100.0	90.030000	90.030000	90.030000
1954	77.950000	95.7100	62.5300	82.320000	97.5900	87.4300	89.990000	100.0	89.990000	89.990000	89.990000
...	...	...	...	...	...	...	...	...	...	...	...
2021	54.540000	81.0600	65.1000	78.440000	95.4100	70.2600	88.030000	100.0	88.030000	88.030000	88.030000
2022	61.090000	82.0700	68.8400	77.130000	95.1400	69.8500	90.860000	100.0	90.860000	90.860000	90.860000
2023	67.360000	79.8500	65.9600	73.530000	95.5900	76.9600	89.900000	100.0	89.900000	89.900000	89.900000
2024	62.430000	80.3900	58.4400	74.630000	96.5800	77.1500	90.500000	100.0	90.500000	90.500000	90.500000
Average	71.237067	90.5632	61.1184	81.283867	94.5296	76.6608	86.144267	100.0	86.145067	86.144267	86.144267

Simple Data Stats



Data is ready, now the real work …

Pitch: *Movie Analysis based on Embeddings*

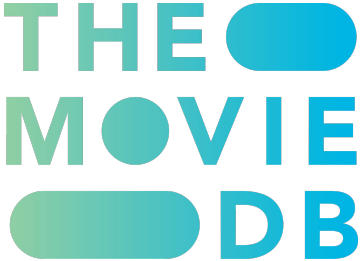
Ansel Cheung, Alessio Villa, Bartol Lastname, Niklas Abraham, Martin Lopez de Ipina Lastname2 Lastname3 SpanishLastname 4

Outline

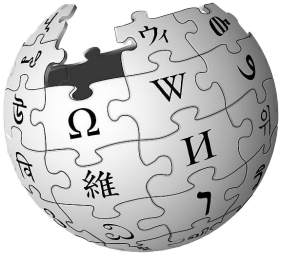
1. Data
    a. Sources + joins/merges
    b. Features
2. Related works
3. Research Question
4. Methods
    a. Embeddings
    b. Comparing embeddings
    c. Clustering
5. First works

Data - Quick Recap

- Title
- Release date
- **Genre, Director,** Actors
- Duration
- Budget (limited), Box office (limited), Awards (limited)
- **Year of release**



- Popularity
- Vote average
- Vote count



- **Plot** (for embeddings)

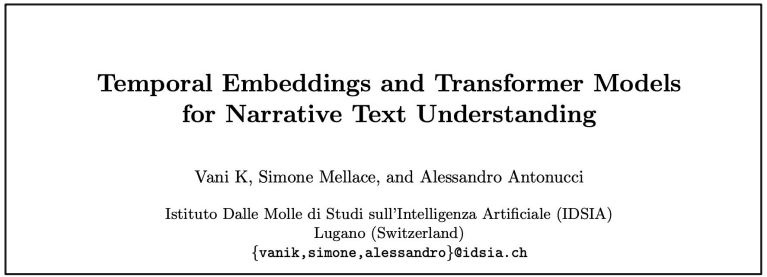
Related works

Many papers on:

- Classification
- Recommendation
- Multimodal (plot + movie/images)

Related works

More relevant: Temporal, embeddings, genres



Temporal Embeddings and Transformer Models for Narrative Text Understanding

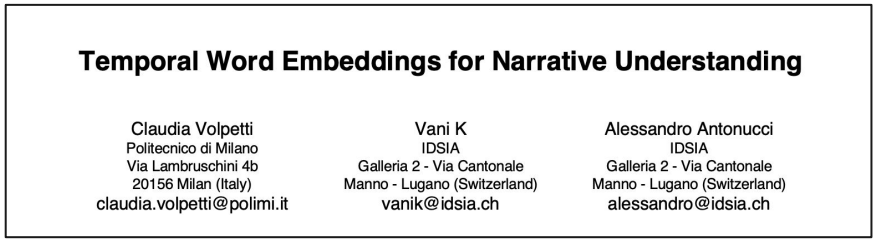
Vani K, Simone Mellace, and Alessandro Antonucci

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Lugano (Switzerland)
{vanik,simone,alessandro}@idsia.ch

Representing Complex Relative Chronology Across Narrative Levels in Movie Plots

Pablo Gervás[1], José Luis López Calle[1]

Facultad de Informática, Universidad Complutense de Madrid, Madrid, 28040 Spain



Temporal Word Embeddings for Narrative Understanding

Claudia Volpetti
Politecnico di Milano
Via Lambruschini 4b
20156 Milan (Italy)
claudia.volpetti@polimi.it

Vani K
IDSIA
Galleria 2 - Via Cantonale
Manno - Lugano (Switzerland)
vanik@idsia.ch

Alessandro Antonucci
IDSIA
Galleria 2 - Via Cantonale
Manno - Lugano (Switzerland)
alessandro@idsia.ch

Related works

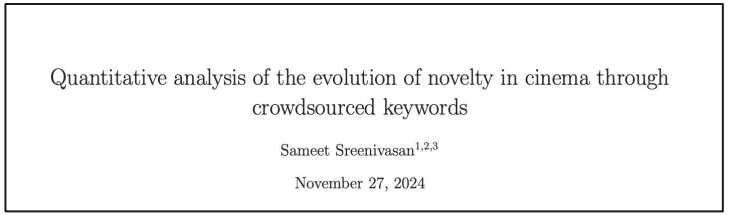
More relevant: Temporal, embeddings, genres

Mapping movie genre evolution (1994 – 2019) using the role of cultural and temporal shifts: a thematic analysis

[version 1; peer review: 2 approved with reservations]

Anshuman Mohanty[1], Aditi Mudgal[2], Shirshendu Ganguli[3]

[1]Manipal Academy of Higher Education, T A Pai Management Institute, Manipal, Karnataka, 576104, India
[2]Jagdish Sheth School of Management, Bengaluru, Karnataka, India
[3]International Management Institute, Bhubaneshwar, Gothapatna, Bhubaneshwar, Odisha, India



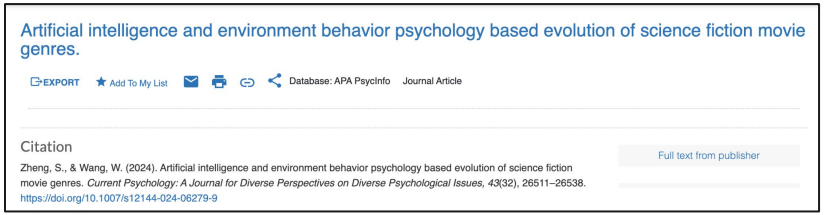
Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords

Sameet Sreenivasan[1,2,3]

November 27, 2024

Genre analysis of movies using a topic model of plot summaries

Paul Matthews | **Kathrina Glitre**



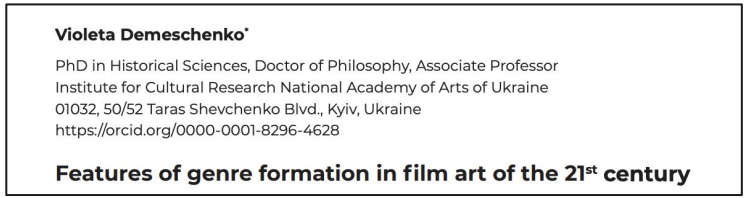
Artificial intelligence and environment behavior psychology based evolution of science fiction movie genres.

EXPORT ★ Add To My List Database: APA PsycInfo Journal Article

Citation

Zheng, S., & Wang, W. (2024). Artificial intelligence and environment behavior psychology based evolution of science fiction movie genres. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, *43*(32), 26511–26538. https://doi.org/10.1007/s12144-024-06279-9

Full text from publisher



Violeta Demeschenko[*]

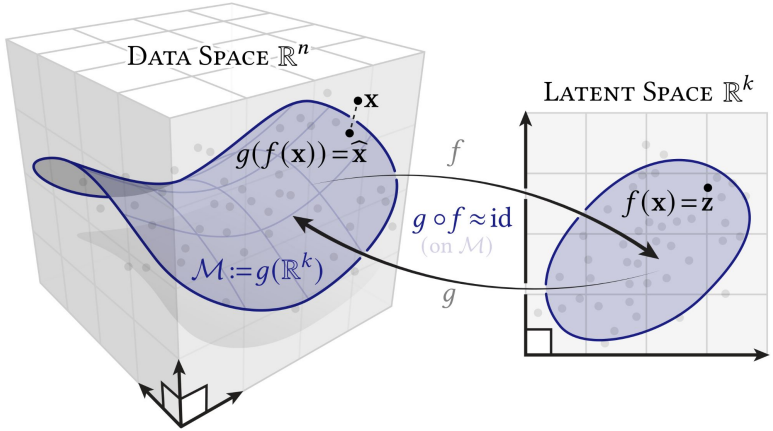
PhD in Historical Sciences, Doctor of Philosophy, Associate Professor
Institute for Cultural Research National Academy of Arts of Ukraine
01032, 50/52 Taras Shevchenko Blvd., Kyiv, Ukraine
https://orcid.org/0000-0001-8296-4628

Features of genre formation in film art of the 21st century

Research Questions

- From Flying Cars to Digital Dystopias: How Science Fiction Rewrites the Future Across Generations
- Novelty or Popularity? Examining Genre Convergence in Modern Film Production
- Genre shifts in the last decades how have genres merged and gotten further apart?
- Can we in movie plots see social shifts, e.g. suddenly people might wear masks, use a new app, a new trend, how long does it take to catch on?

Methods



Embeddings

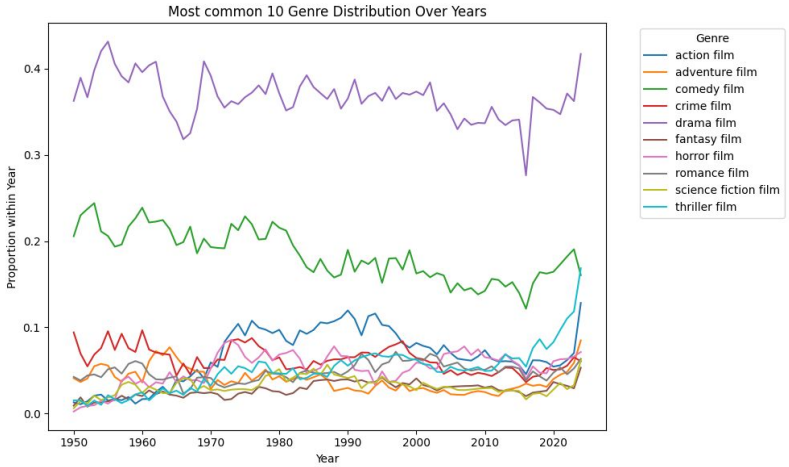
- bge-m3 (1024 dim)
- CLS token embeddings for each movie plot → dense embedding
- word tokens, each dimension on token (100 000 dims) → sparse embedding

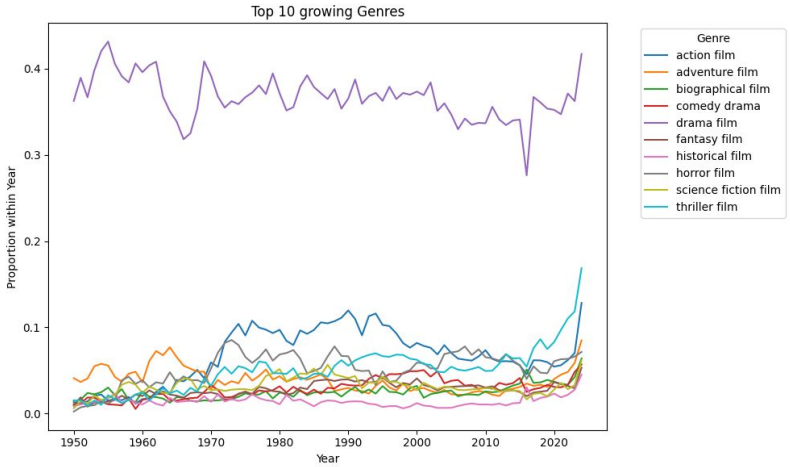
Methods

Comparing embeddings

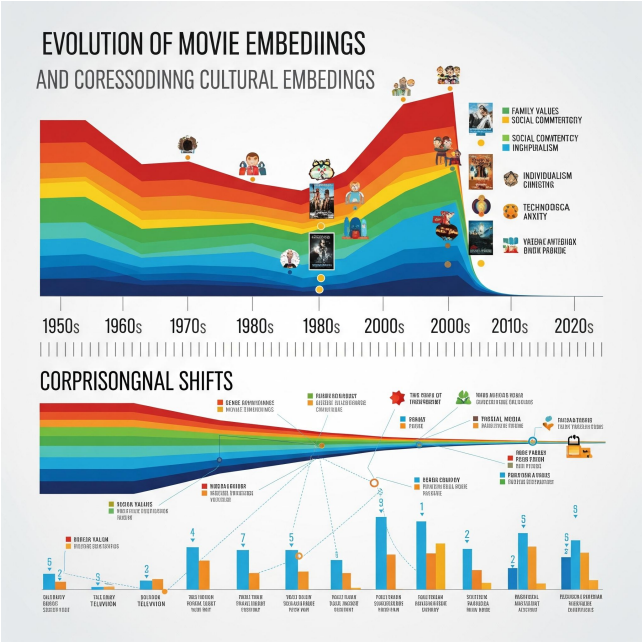
- Cosine similarity
- Euclidean distance (pretty bad for high dim)
- Dot product similarity
- Wasserstein Distances
- Velocity
- Acceleration
- Curvature

First works


Most common 10 Genre Distribution Over Years
Proportion within Year
Year
0.0
0.1
0.2
0.3
0.4
1950
1960
1970
1980
1990
2000
2010
2020
Genre
action film
adventure film
comedy film
crime film
drama film
fantasy film
horror film
romance film
science fiction film
thriller film


Top 10 growing Genres
Proportion within Year
Year
0.0
0.1
0.2
0.3
0.4
1950
1960
1970
1980
1990
2000
2010
2020
Genre
action film
adventure film
biographical film
comedy drama
drama film
fantasy film
historical film
horror film
science fiction film
thriller film

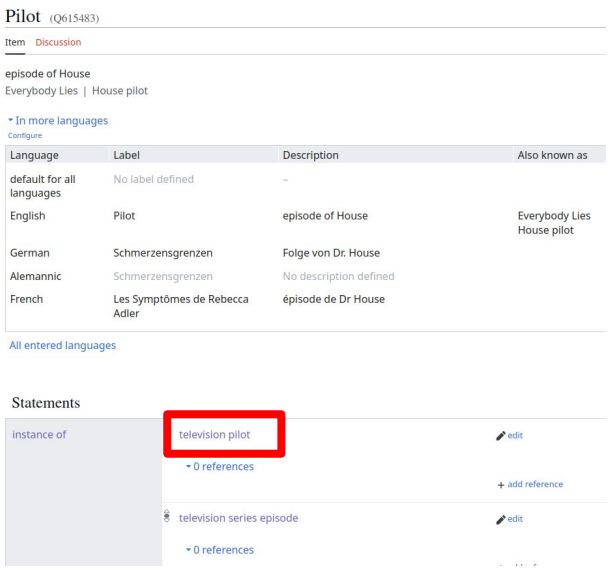
Pitch: *Movie Analysis based on Embeddings*

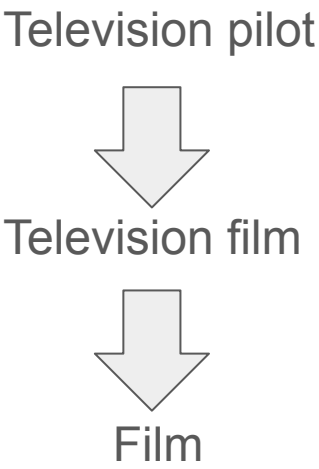


Outline - Target Today Data Preprocessing/Cleaning

- ➔ Data cleaning
- ➔ Genre fixes
- ➔ Some data analysis
- ➔ Chunking and metrics
- ➔ What to do with the very long movies?

Filtering TV show episodes from data

- Problem noticed: some TV show episodes are part of the collected data
  - But the only episodes that seem to appear are TV show pilot episodes?!
- old query: *“get all instances of some (indirect) subclass of film”*
- Explanation of the problem:
- 



Filtering TV show episodes from data - solution

- new query: *“get all (indirect) subclasses of film, exclude subclasses of television series episode and short film, then find all instances of these subclasses”*
- Added feature that is equal to the wikidata class of the film
- Excluded movies that have no wikipedia entry immediately

Removing duplicates

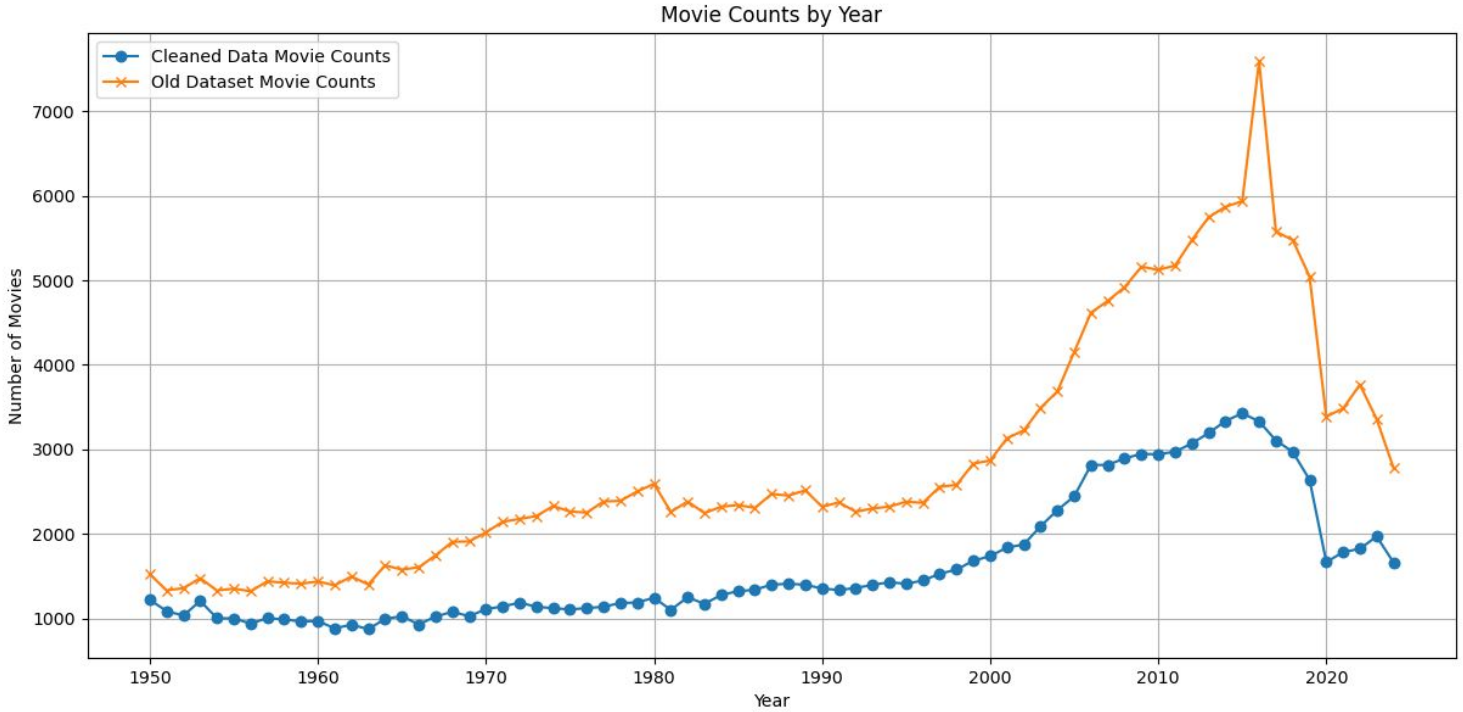
- Around 15000 duplicates in the collected dataset
- Root cause of the problem are publication dates

2005 data

2006 data

Movie counts

- Number of movies in the cleaned dataset: 122 626
- Number of movies in the old dataset: 214 682



Genre fix

- Genres raw string “genre1, genre2, genre3, …”

```
action film, war film, science fiction film, thriller film, speculative fiction film, apocalyptic film
horror film, erotic film, zombie film
fantasy film, action film, horror film, adventure film, science fiction film, superhero film
```

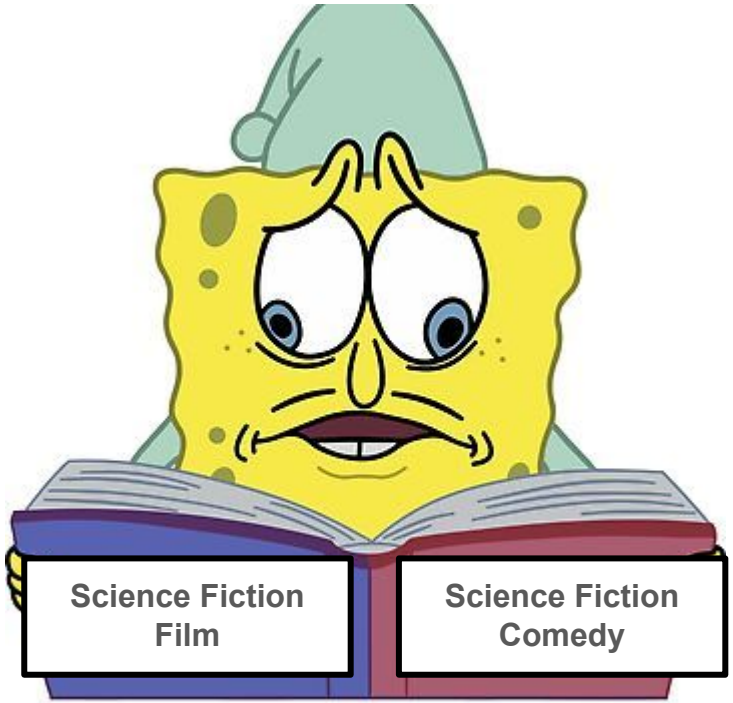
- Split by “,”
- Strip “film” and \s

Genre fix

- 975 unique genres
- Too granular

Options:

- Drop non-common genres
- Cluster/Group similar genres



Genre fix

Option 1: Drop movies with uncommon movies

- Movies with non-na genres: 176705
- Movies with no top **10** genres: 40754

NO

- Movies with no top **20** genres: 33459

```
[('drama', 4274),
 ('action', 2416),
 ('comedy', 2388),
 ('thriller', 1783),
 ('science fiction', 1655),
 ('horror', 1632),
 ('fantasy', 1623),
 ('adventure', 1619),
 ('crime', 1394),
 ('romance', 1073)]
```

Genre fix

Option 2: Quick clustering of genres

- Sentence transformers: all-MiniLM-L6-v2
- 10 clusters
- K-means
- Embed cleaned genres (could be improved but quick and dirty)

Genre fix

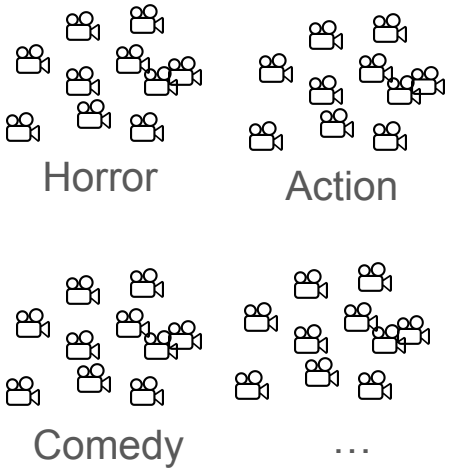
1. Others
2. Comedy
3. Romance
4. Drama
5. Fantasy
6. Action
7. Science Fiction
8. Family
9. Mystery
10. Thriller

```
[('drama', 4274),
 ('action', 2416),
 ('comedy', 2388),
 ('thriller', 1783),
 ('science fiction', 1655),
 ('horror', 1632),
 ('fantasy', 1623),
 ('adventure', 1619),
 ('crime', 1394),
 ('romance', 1073)]
```

Data analysis - What is the genre shift velocity?

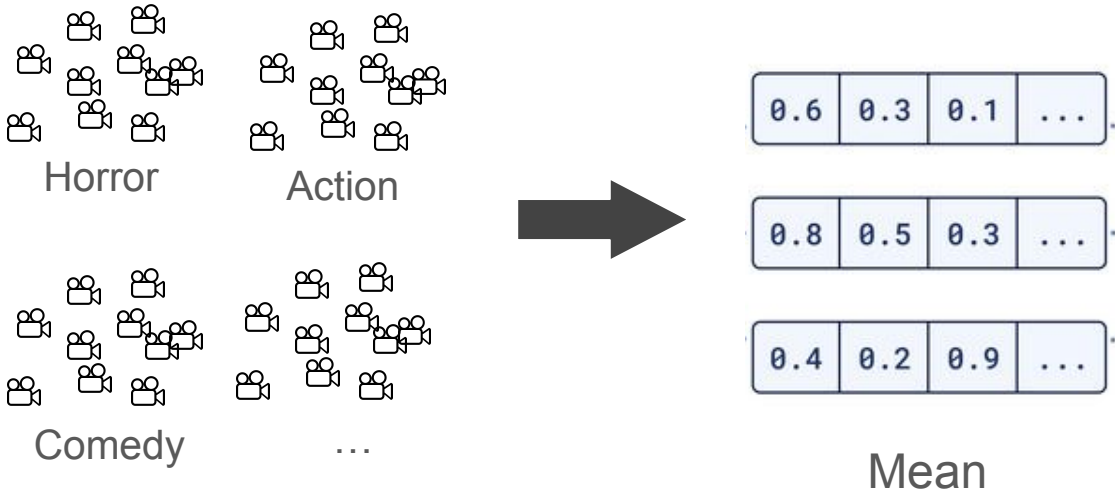
Data analysis - What is the genre shift velocity?

Year X:



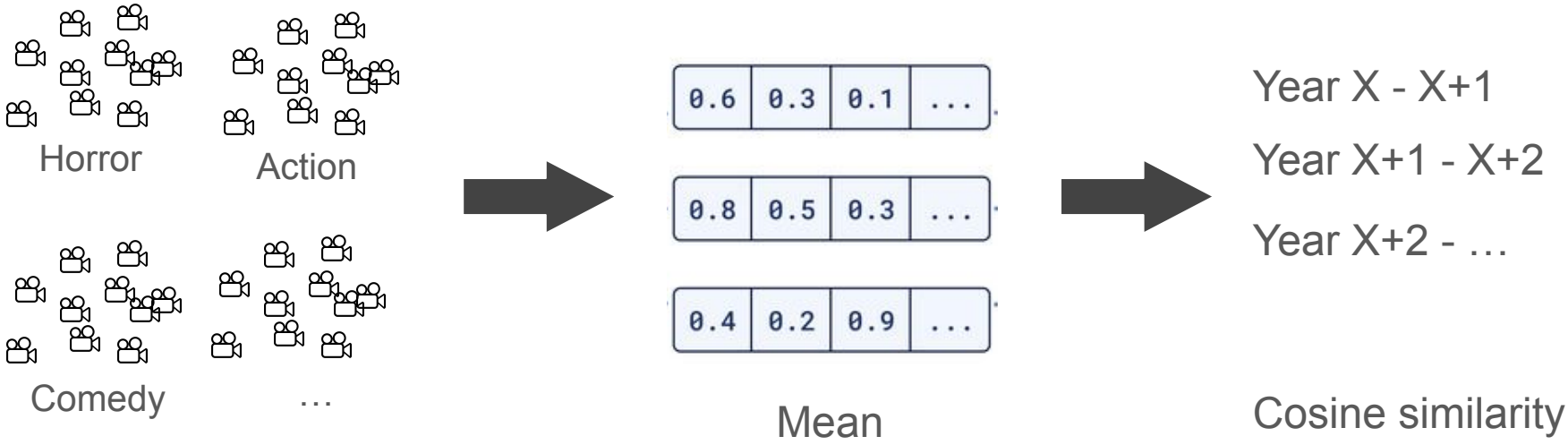
Data analysis - What is the genre shift velocity?

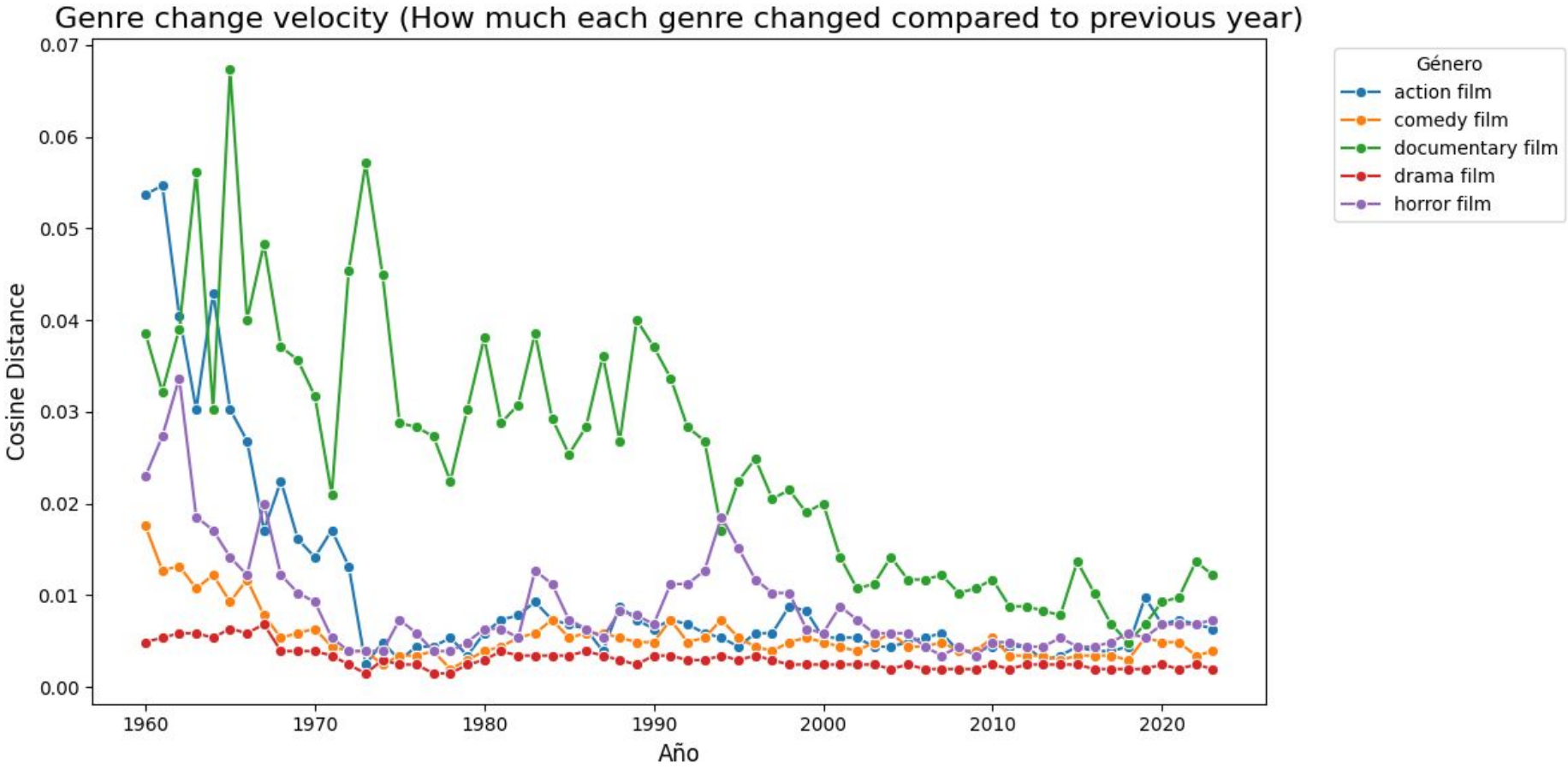
Year X:

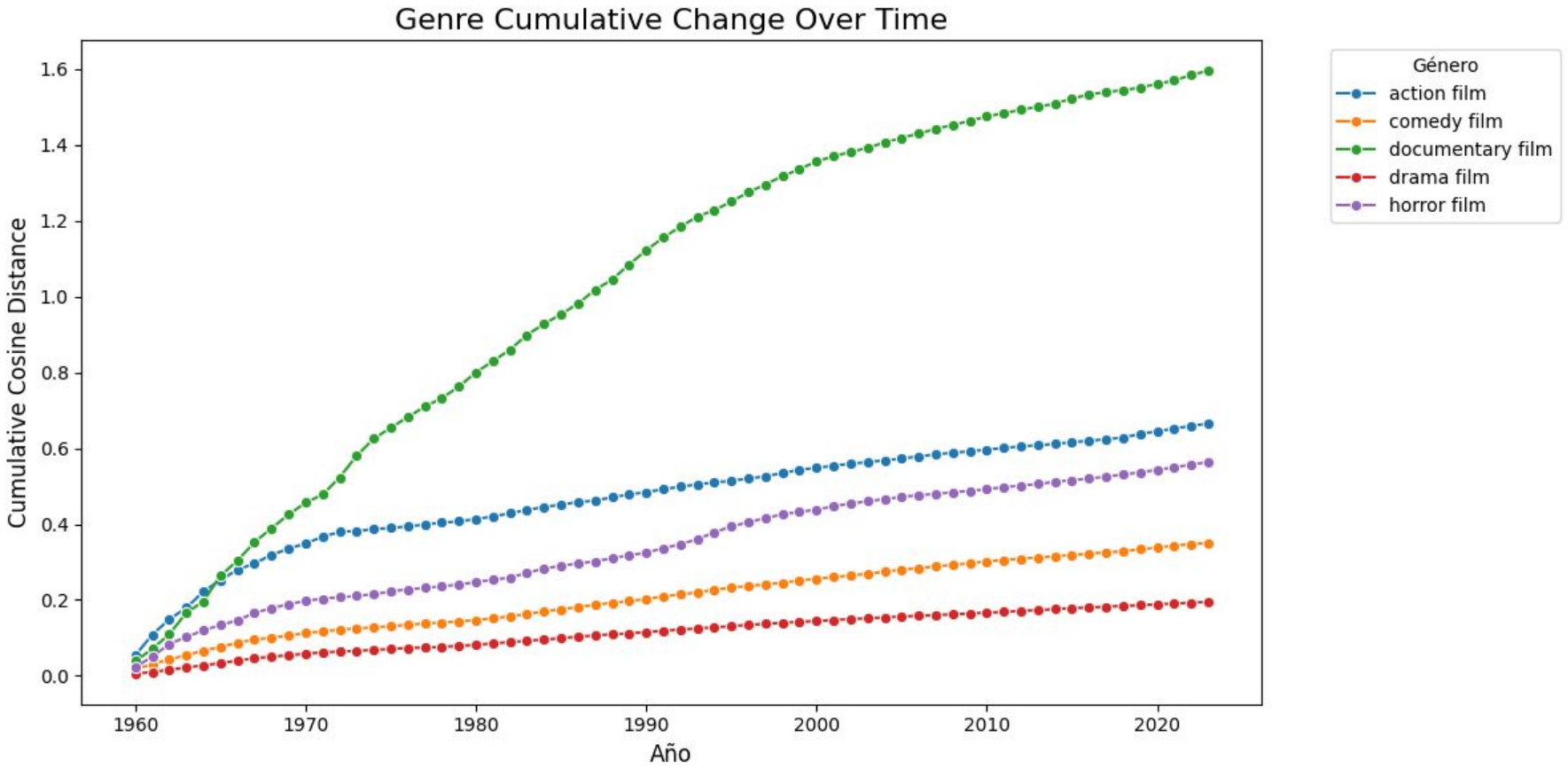


Data analysis - What is the genre shift velocity?

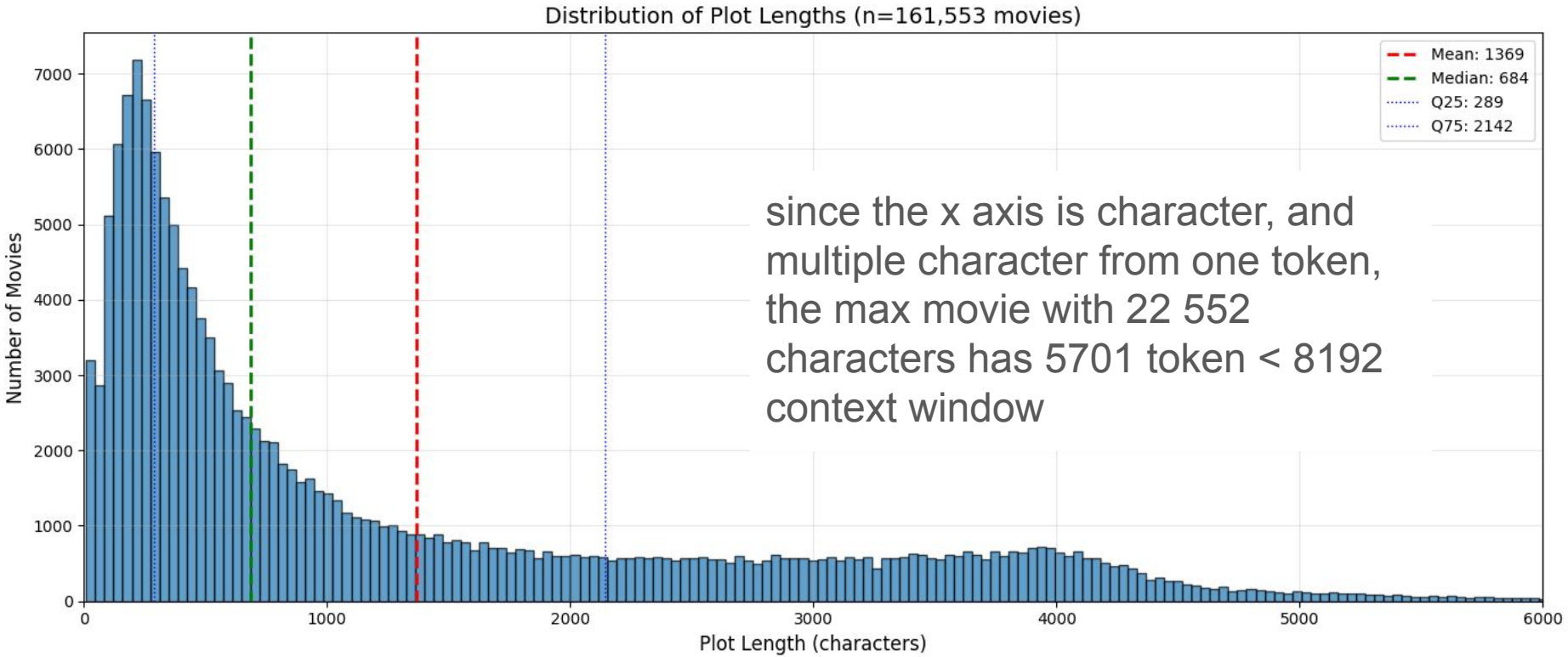
Year X:




Genre change velocity (How much each genre changed compared to previous year)
Cosine Distance
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1960
1970
1980
1990
2000
2010
2020
Año
Género
action film
comedy film
documentary film
drama film
horror film


Genre Cumulative Change Over Time
Cumulative Cosine Distance
Año
Género
action film
comedy film
documentary film
drama film
horror film
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1960
1970
1980
1990
2000
2010
2020

Context Window Size of bge-m3 → 8192 tokens

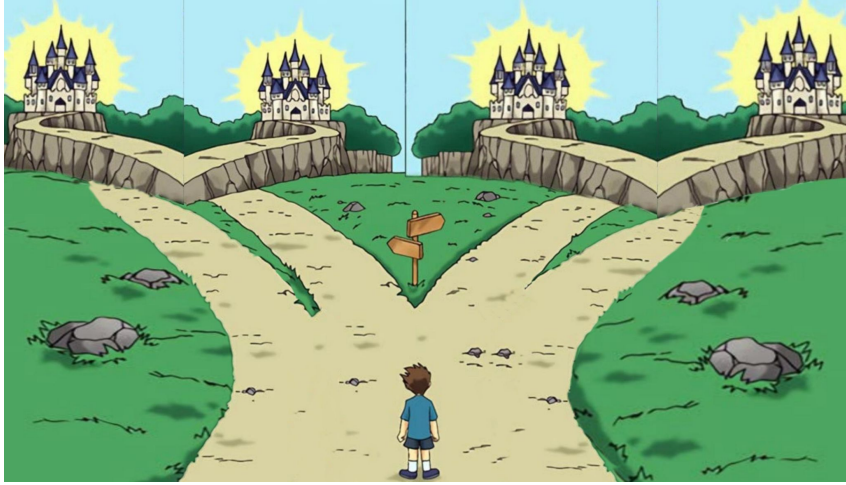


Chunking - Variable Text Length in Plots

- Problem definition: *Corpus contains **variable-length** documents (a few sentences to several thousand words). Transformer models have **fixed token limits** and a **CLS token bias toward input length** and structure. Therefore, one has to choose a pooling and chunking strategy that **minimizes length bias** while preserving semantic fidelity.*

	Core Idea	References
Mean Pooling (Global Average)	Take the mean of all token embeddings	Classical sentence representation
Chunk-then-Embed (Early Chunking)	Split long text into chunks, embed each separately, aggregate	Hierarchical Attention Networks (Yang et al. 2016)
Embed-then-Chunk (Late Chunking)	Embed full text once, then pool token embeddings over fixed-size windows	Late Chunking: Contextual Chunk Embeddings for RAG (2024)

Back to the core problem



- We have four methods and no idea which one to use:
  - Mean pooling → low variance but high bias;
  - Chunking → higher variance, less bias;
  - Late chunking → controlled variance, minimal bias.
  - just take the CLS Token
- but wait shouldn't the CLS token already handle all this?

→ well kinda, but the CLS token is not a semantic summary token, in training it serves as the token which is supposed to predict the next sentence

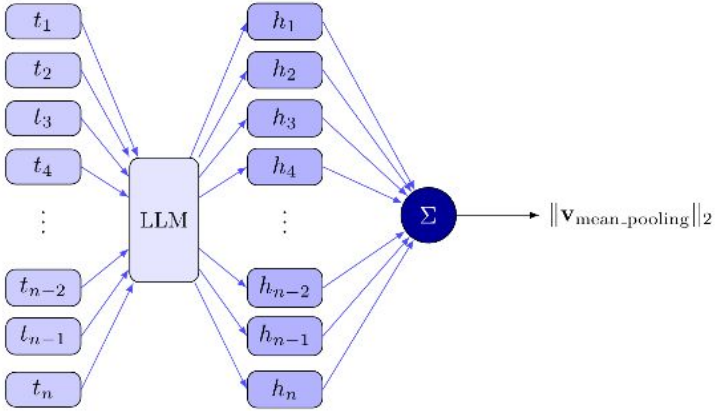
CLS Token problems (Devlin et al. (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.)

- They show attention entropy at [CLS] drops sharply as sequence length increases; the token focuses on the first ~128 tokens *(Gong et al. (2021), Efficient Training of BERT on Long Sequences (ICLR)*
- Report that learned summary tokens underperform mean pooling for long documents; decoder attention favors earlier positions *(Li et al. (2020), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5))*
- Quantify that BERT's effective context window saturates around 512–1024 tokens; beyond that, semantic retention in [CLS] collapses. *(Wu et al. (2020), On the Limitations of Pre-trained Transformers for Long-Range Context (arXiv 2009.11893)*

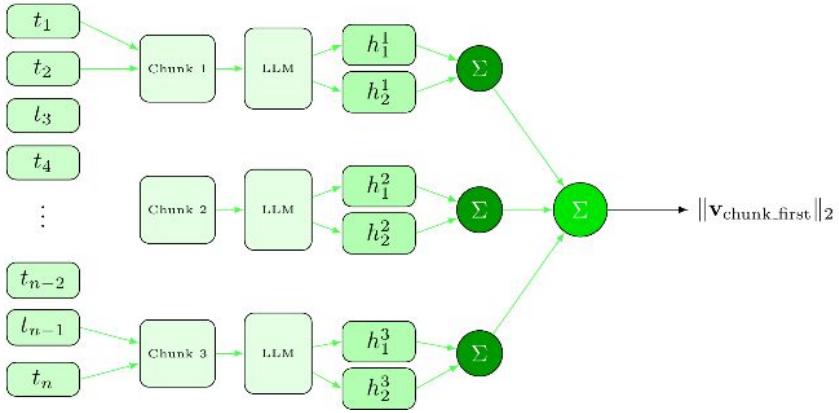
→since a significant amount of over text more than 75% is longer than 512 we have to consider this

Overview of methods

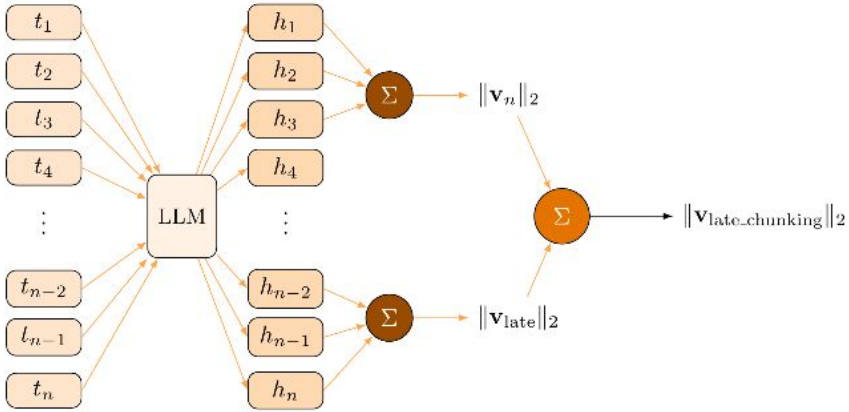
(a) Mean Pooling



(b) First Chunk then Embed



(c) Late Chunking



Results of Chunking Methods

Table 1: Summary of results for each chunking method and embedding configuration. The best value in each column is shown in bold.

Method	Length normcorr	Isotropy 1st PC	Isotropy abtt2	Mean Between Dist.	Silhouette Score	Genre Accuracy	Genre F1
MeanPooling	0.643	12.23	**3.43**	**0.59**	-0.18	0.46	0.35
CLS Token	-0.01	5.14	2.22	0.57	-0.11	**0.47**	**0.36**
ChunkFirst E. 512/256	-0.38	5.83	2.26	0.56	-0.12	**0.47**	**0.36**
ChunkFirst E. 1024/512	-0.32	5.28	2.22	0.57	-0.11	**0.47**	**0.36**
ChunkFirst E. 2048/1024	-0.09	5.14	2.22	0.57	-0.11	**0.47**	**0.36**
LateChunking 512/256	**0.84**	**12.58**	**3.43**	**0.59**	-0.18	0.46	0.35
LateChunking 1024/512	0.75	12.32	**3.43**	**0.59**	-0.18	0.46	0.35
LateChunking 2048/1024	0.66	12.23	**3.43**	**0.59**	-0.18	0.46	0.35
LateChunking 2048/512	0.66	12.23	**3.43**	**0.59**	-0.18	0.46	0.35
LateChunking 512/0	**0.84**	12.24	3.42	**0.59**	-0.18	0.46	0.35
LateChunking 1024/0	0.75	12.16	3.42	**0.59**	-0.18	0.46	0.35
LateChunking 2048/0	0.66	12.23	**3.43**	**0.59**	-0.18	0.46	0.35

Related works

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

William L. Hamilton, Jure Leskovec, Dan Jurafsky

Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords

Sameet Sreenivasan[1,2,3]

Decoding Musical Evolution Through Network Science

Niccolò Di Marco[1,*], Edoardo Loru[2], Alessandro Galeazzi[3], Matteo Cinelli[1], Walter Quattrociocchi[1]

Is Cinema Becoming Less and Less Innovative With Time? Using neural network text embedding model to measure cultural innovation ⋆

Edgar Dubourg*, Andrei Mogoutov and Nicolas Baumard

Super long movies …

Strange Frequencies: Taiwan Killer Hospital

Add languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Strange Frequencies: Taiwan Killer Hospital is a 2024 Philippine found footage supernatural horror film based on the 2018 South Korean film *Gonjiam: Haunted Asylum*. The film stars Enrique Gil, Jane de Leon, Alexa Miro, MJ Lastimosa, Raf Pineda and Ryan "Zarckaroo" Azurin. Produced by Reality MM Studios and Creative Leaders Group 8, it serves as an official entry to the 2024 Metro Manila Film Festival.[2][3]

The film was later released in Netflix on September 5, 2025, and it debuted at the number 1 spot on Netflix's top 10 movies in the Philippines.

Plot [edit]

> This article's plot summary **may be too long or excessively detailed.** Please help improve it by removing unnecessary details and making it more concise. *(September 2025) (Learn how and when to remove this message)*

The film begins with a grainy, glitchy livestream from two foreign ghost hunting vloggers joking around as they explore the ruins of Xinglin General Hospital in Taiwan. After finding a pile of bones, the pair become scared and try to find a way out, but somehow are lost. Ethan claims that someone is calling for help, so he

Strange Frequencies: Taiwan Killer Hospital

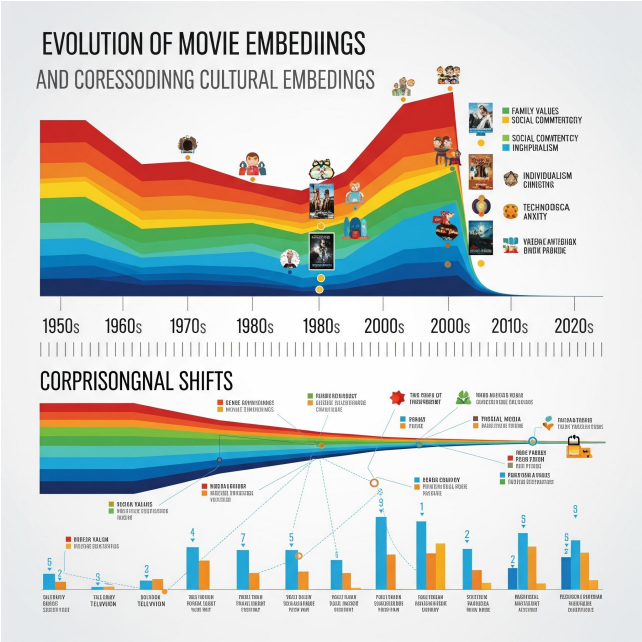
Theatrical release poster

Directed by	Kerwin Go
Written by	Kerwin Go Dustin Celestino

we are not the first taking issue with the super long plots …

wikipedia definition for to long anything > 14 000 characters, in our case that's 20 movies …

Pitch: *Movie Analysis based on Embeddings*



Outline - Target Today Data Preprocessing/Cleaning

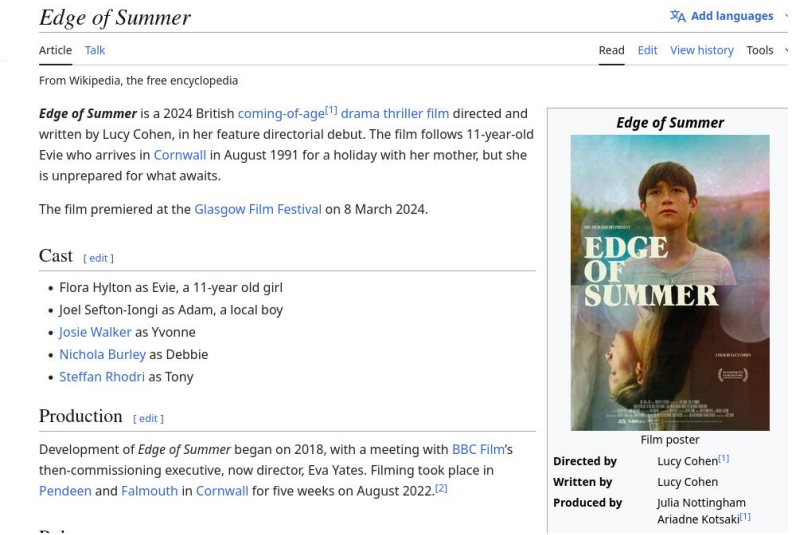
- ➔ Data cleaning
- ➔ Some data analysis

Recollecting the data

- No more duplicates
- Filter out tv show episodes and short films
- Fetch wikidata class - to filter out other non movie classes
- Assure that all durations are in minutes
- Fetch currencies for box office and budget
- Fetch only worldwide budget
- Fetch all estimates of duration/budget/box office if multiple exist
- Fetch wikidata QIDs of actors/
- Fetch all movies from 1930 onwards
- Using CLS token in the embedding

Recollecting the data - movie plots

- Previously: if plot does not exist -> use summary (wikipedia header paragraph)
- Now: fetch plot subsection or some synonym subsection
- plot, plot summary, plot synopsis, synopsis, summary, story, storyline, premise



Edge of Summer

Add languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Edge of Summer is a 2024 British coming-of-age[1] drama thriller film directed and written by Lucy Cohen, in her feature directorial debut. The film follows 11-year-old Evie who arrives in Cornwall in August 1991 for a holiday with her mother, but she is unprepared for what awaits.

The film premiered at the Glasgow Film Festival on 8 March 2024.

Cast [edit]

- Flora Hylton as Evie, a 11-year old girl
- Joel Sefton-Iongi as Adam, a local boy
- Josie Walker as Yvonne
- Nichola Burley as Debbie
- Steffan Rhodri as Tony

Production [edit]

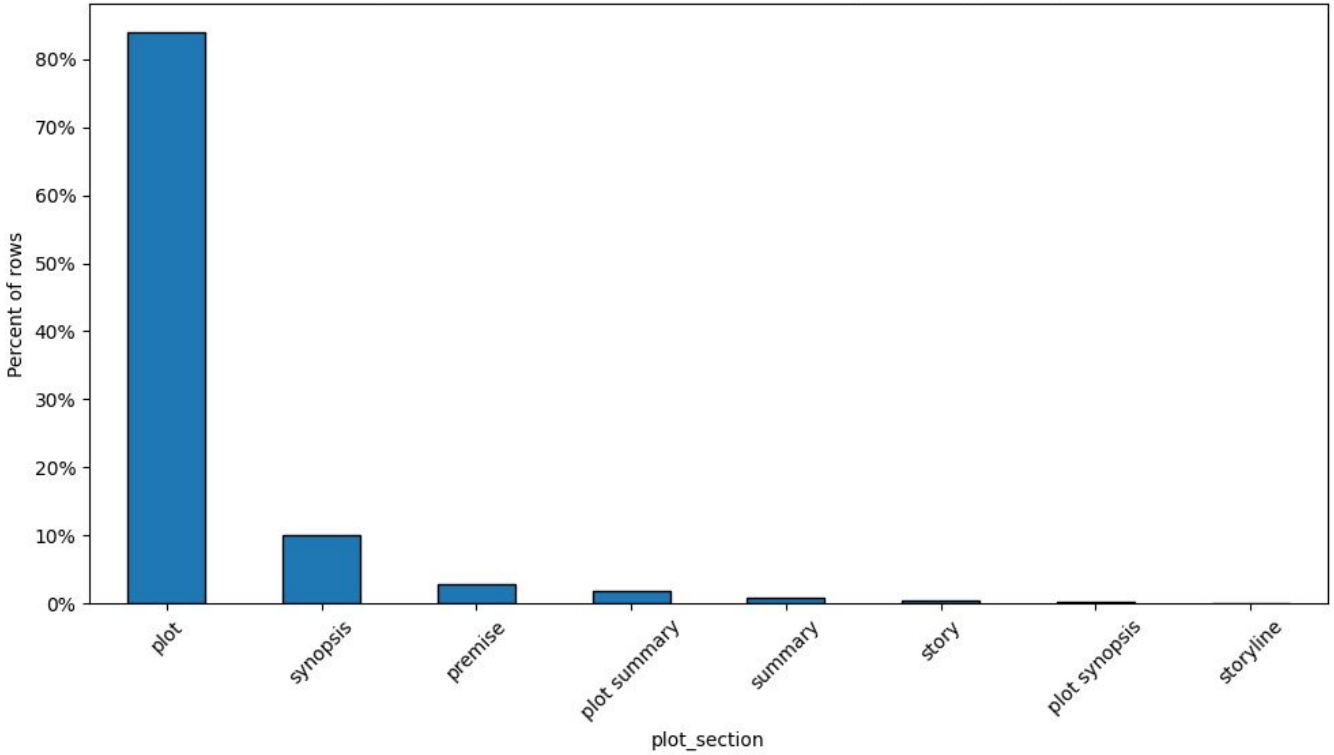
Development of *Edge of Summer* began on 2018, with a meeting with BBC Film's then-commissioning executive, now director, Eva Yates. Filming took place in Pendeen and Falmouth in Cornwall for five weeks on August 2022.[2]

Edge of Summer

Film poster

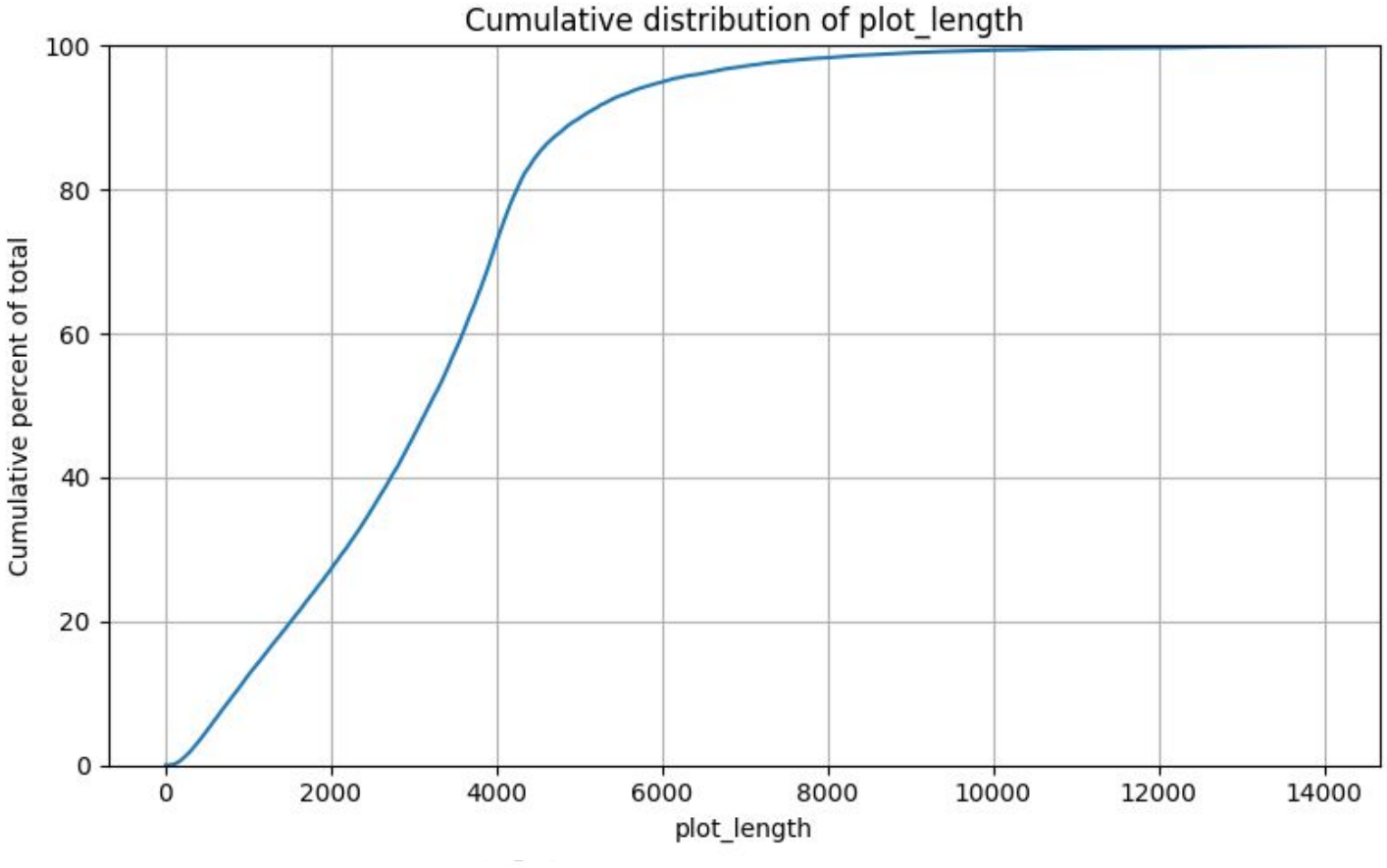
Directed by	Lucy Cohen[1]
Written by	Lucy Cohen
Produced by	Julia Nottingham Ariadne Kotsaki[1]

Recollecting the data - movie plots

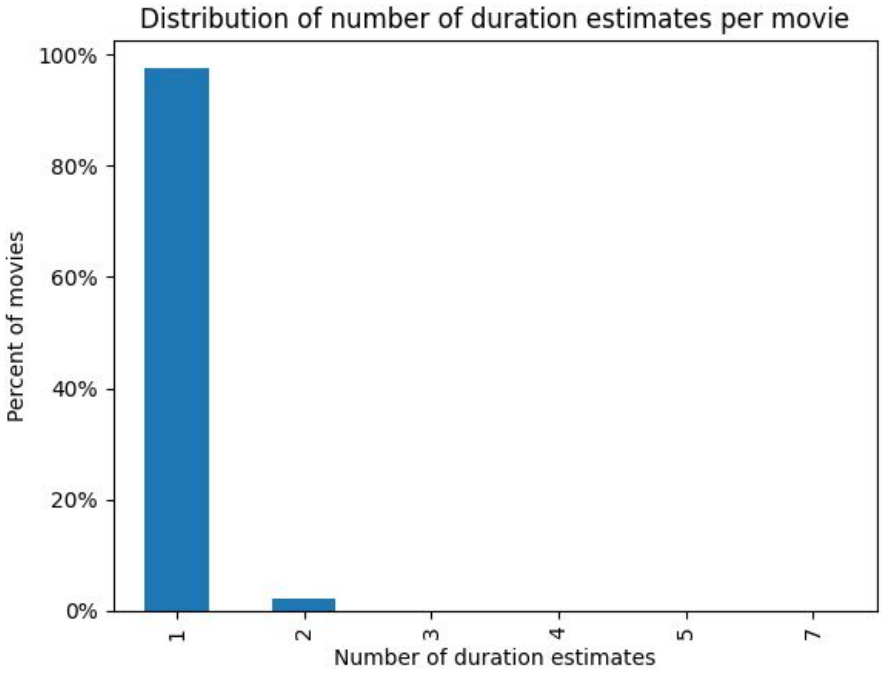


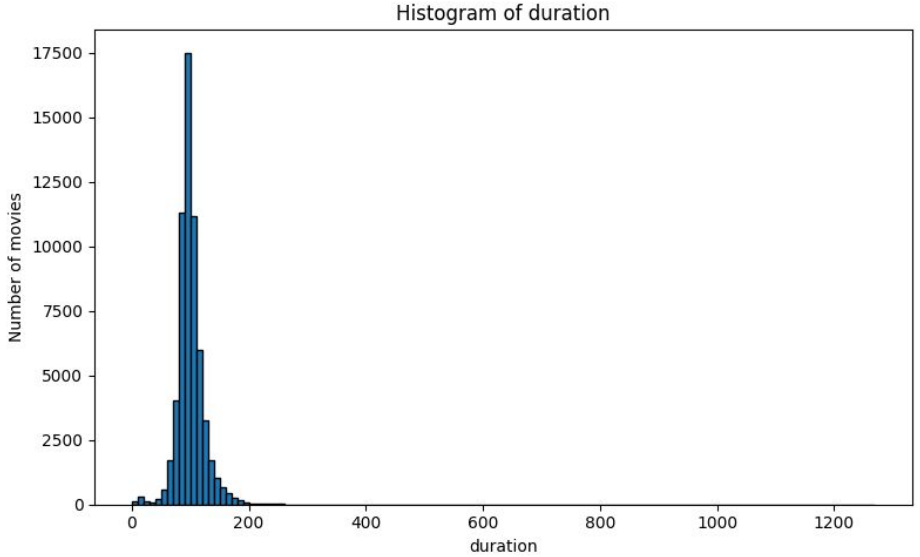
plot_section	
plot	82617
synopsis	9971
premise	2740
plot summary	1758
summary	767
story	376
plot synopsis	208
storyline	76

Data cleaning

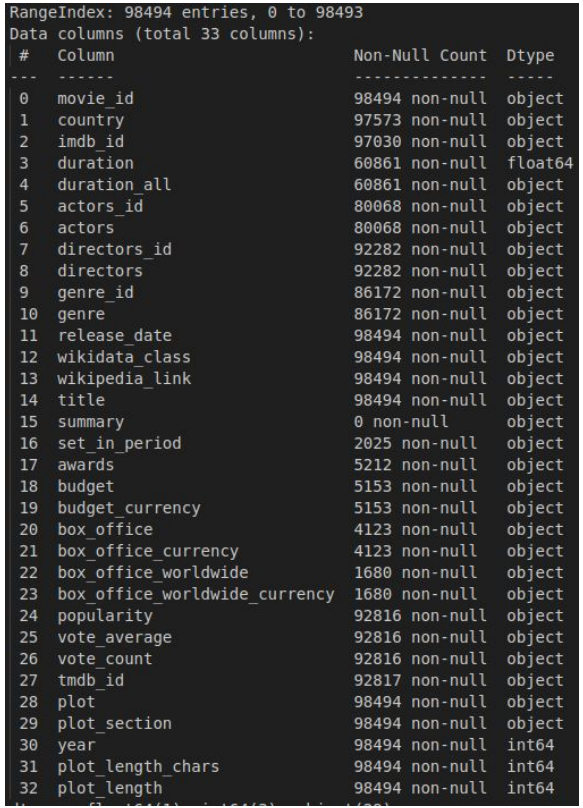


Movie durations





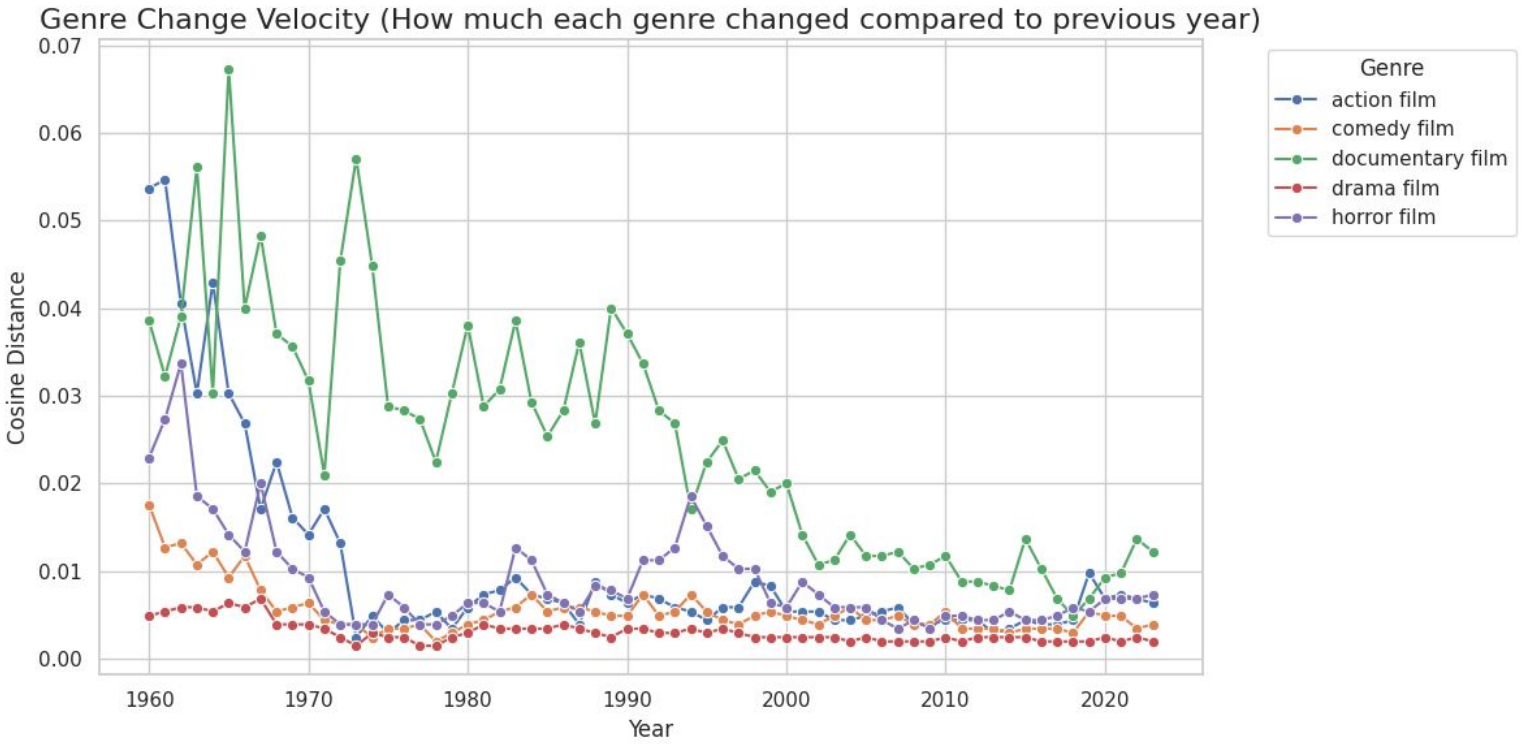
Features



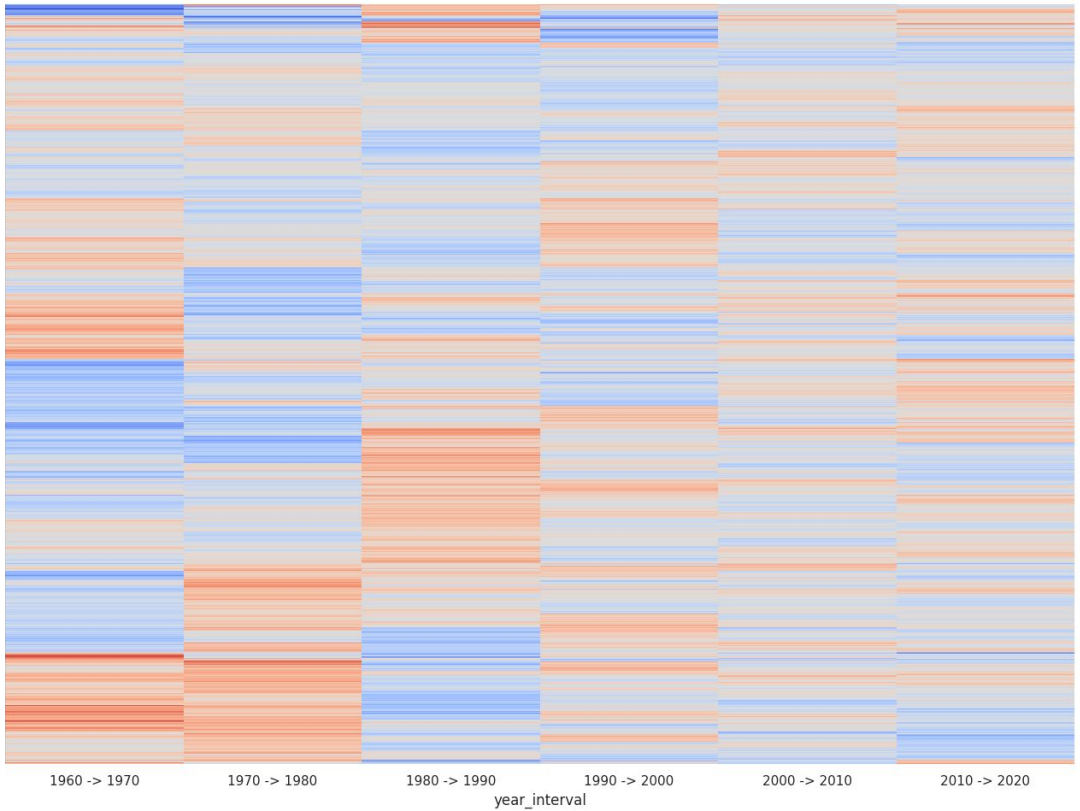
```
RangeIndex: 98494 entries, 0 to 98493
Data columns (total 33 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   movie_id                      98494 non-null  object
 1   country                       97573 non-null  object
 2   imdb_id                       97030 non-null  object
 3   duration                      60861 non-null  float64
 4   duration_all                  60861 non-null  object
 5   actors_id                     80068 non-null  object
 6   actors                        80068 non-null  object
 7   directors_id                  92282 non-null  object
 8   directors                     92282 non-null  object
 9   genre_id                      86172 non-null  object
 10  genre                         86172 non-null  object
 11  release_date                  98494 non-null  object
 12  wikidata_class                98494 non-null  object
 13  wikipedia_link                98494 non-null  object
 14  title                         98494 non-null  object
 15  summary                       0 non-null      object
 16  set_in_period                 2025 non-null   object
 17  awards                        5212 non-null   object
 18  budget                        5153 non-null   object
 19  budget_currency               5153 non-null   object
 20  box_office                    4123 non-null   object
 21  box_office_currency           4123 non-null   object
 22  box_office_worldwide          1680 non-null   object
 23  box_office_worldwide_currency 1680 non-null   object
 24  popularity                    92816 non-null  object
 25  vote_average                  92816 non-null  object
 26  vote_count                    92816 non-null  object
 27  tmdb_id                       92817 non-null  object
 28  plot                          98494 non-null  object
 29  plot_section                  98494 non-null  object
 30  year                          98494 non-null  int64
 31  plot_length_chars             98494 non-null  int64
 32  plot_length                   98494 non-null  int64
```

Too little data

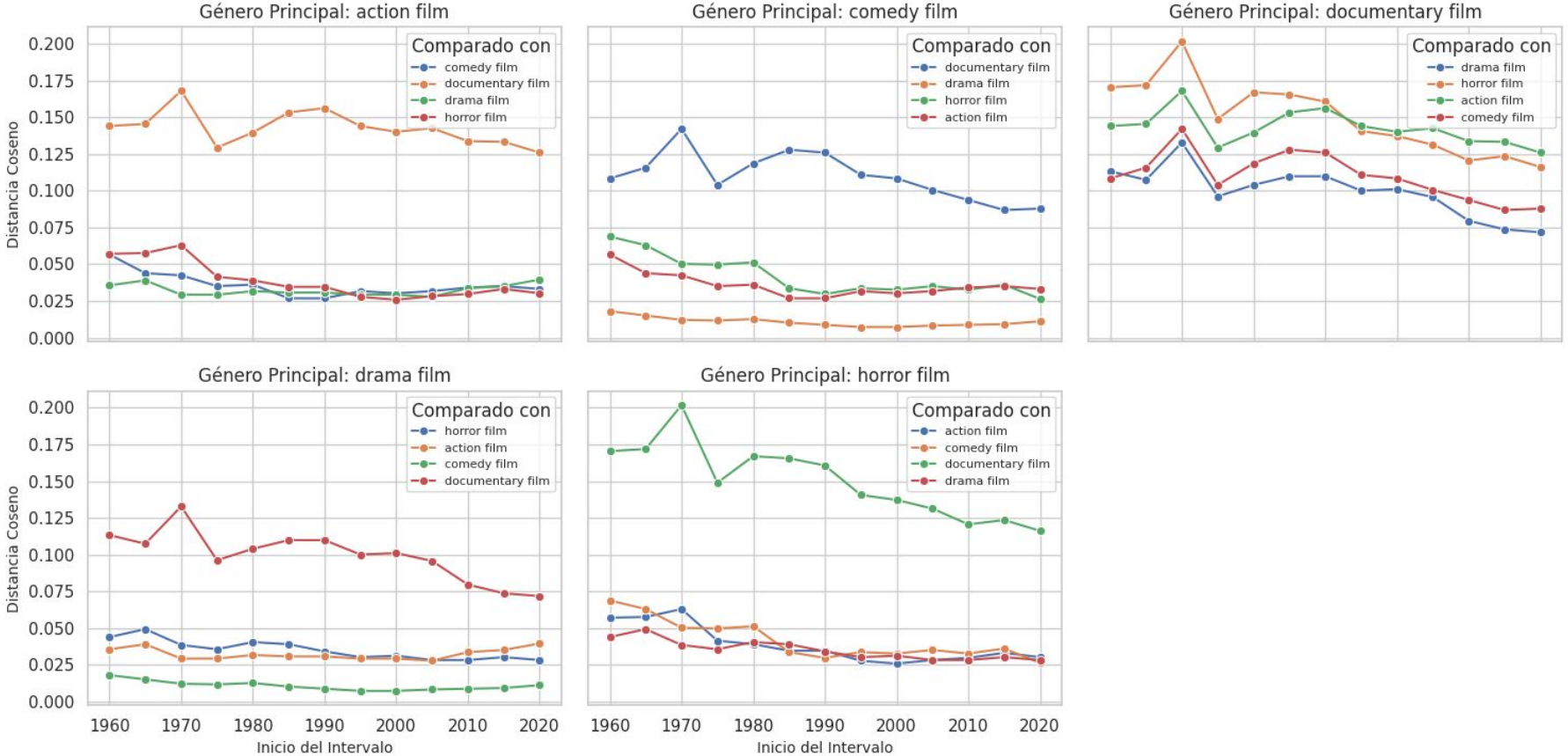
Previously on “some data analysis”:



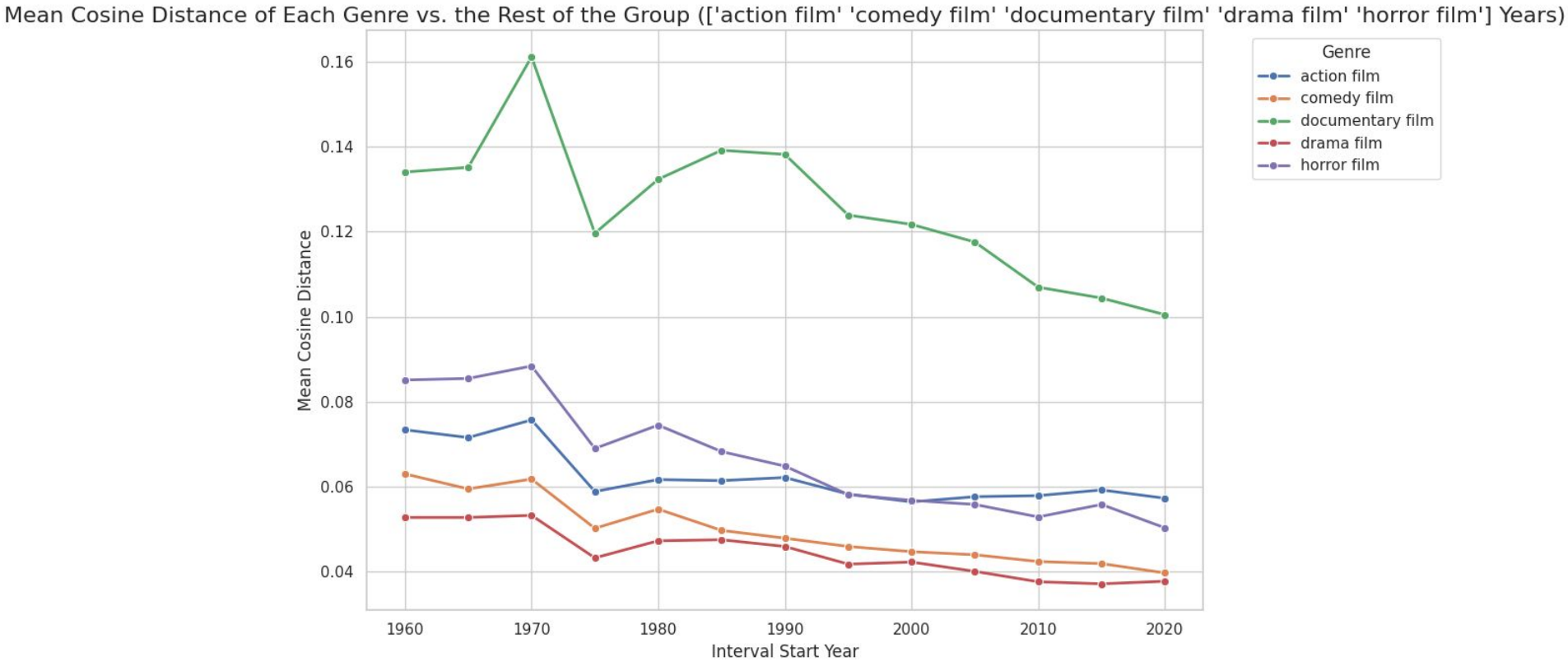
Does the genre drift converge? -> horror movies



Pairwise genre cosine distance per decade



Cosine distance of each genre w.r.t all the others



Outline

- Recap data pipeline
- Cut off for text length
- Data cleaning
- Genre classification
- Sanity Check with the data
- Literature review
  - Dimension reduction methods
  - Latent space comparison methods
  - Average embeddings
  - How to use tabular data (not just embeddings)

From Plot to Vector: The Movie Data Pipeline

A multi-stage process of acquiring movie data and transforming it into a machine-learning-ready format.

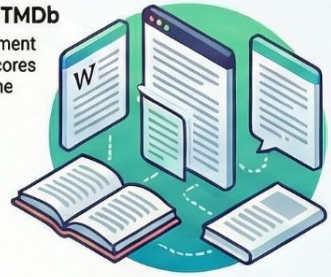
Phase 1: Data Acquisition

Step 1: Query Wikidata

Acquired foundational movie data, including titles, release years, and linke to Wikipedia pages.

Step 2: Enrich with TMDb

Added audience engagement metrics like popularity scores and vote counts using The Movie Database.



Step 3: Extract from Wikipedia

Collected rich, full-text plot summaries from corresponding English Wikipedia articles.

Phase 2: Data Transformation

Step 4: Input Raw Text

Used the variable-length plot summaries as the primary text input for processing.

Step 5: Select Embedding Model

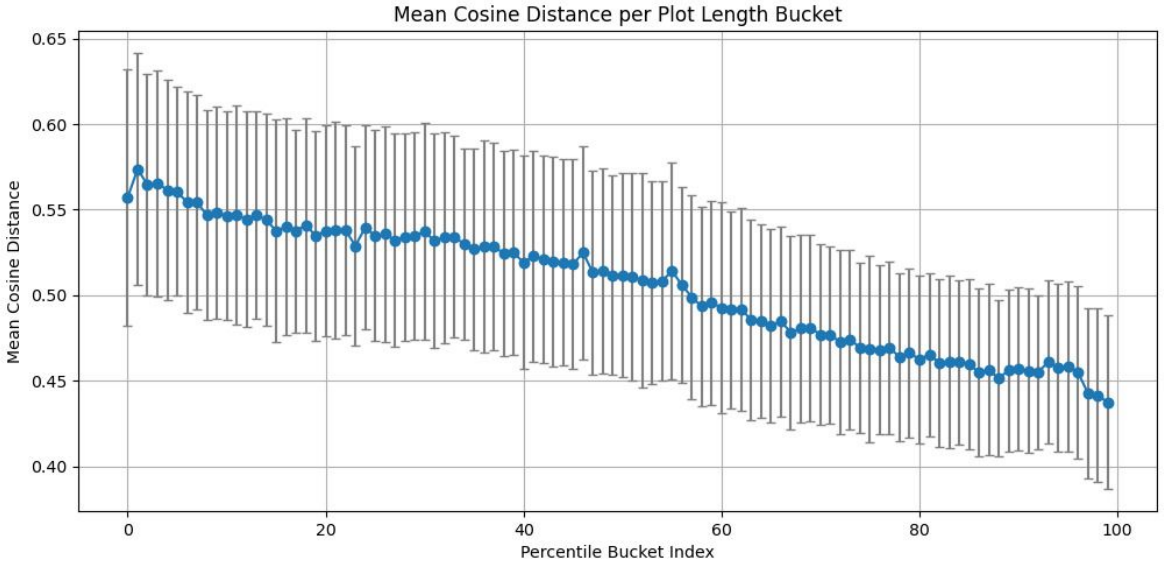
Chose the BGE-M3 model to convert text into meaningful numerical representations (vectors).

Step 6: Generate Final Vectors

Applied chunking strategies to handle long plots and create a single, unified vector for each movie.

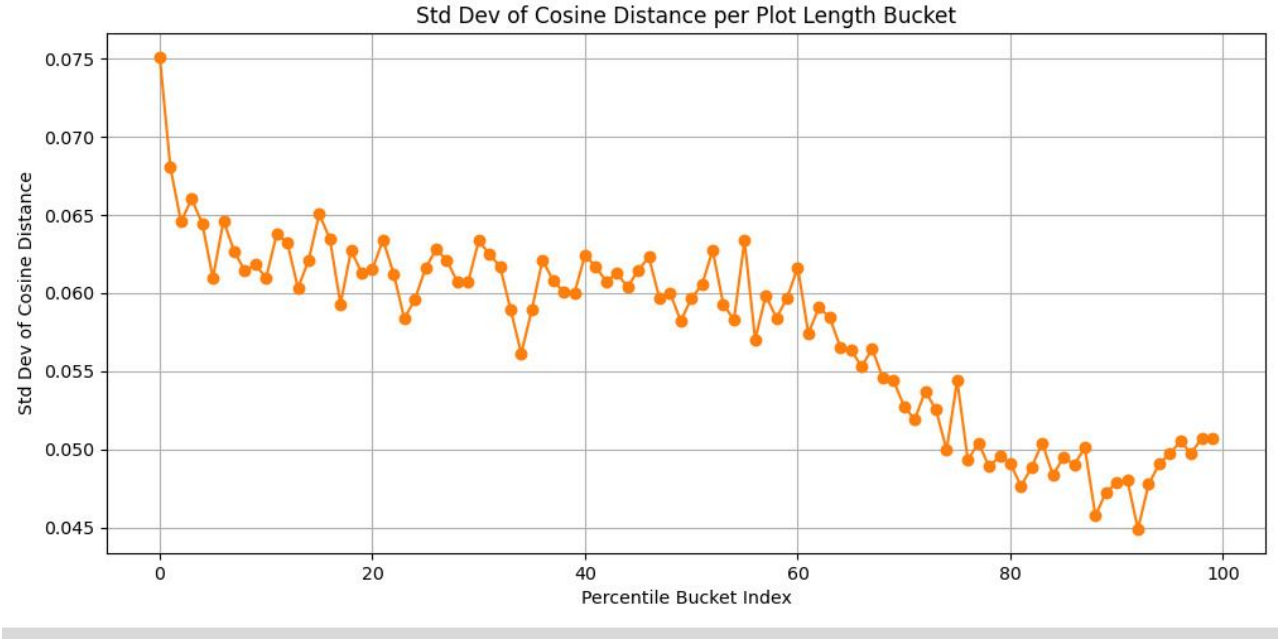
Plot length cut-off – 1st try

- Our thinking: simple/short plots will be very similar to each other -> the variance of the distance between their embeddings will be significantly smaller
- Reality: completely opposite



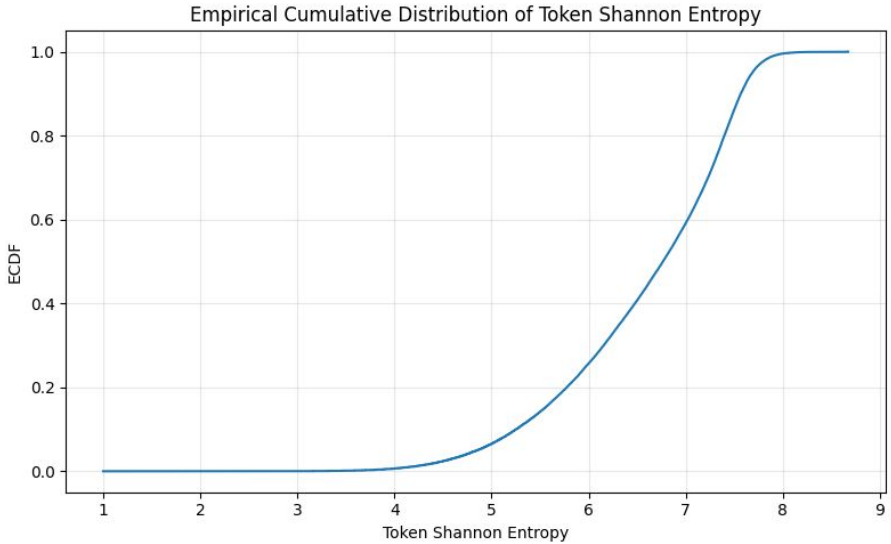
Plot length cut-off – 1st try

- Possible cause: Very short plots have a very small amount of context and the weight is likely distributed among a smaller number of embedding coordinates

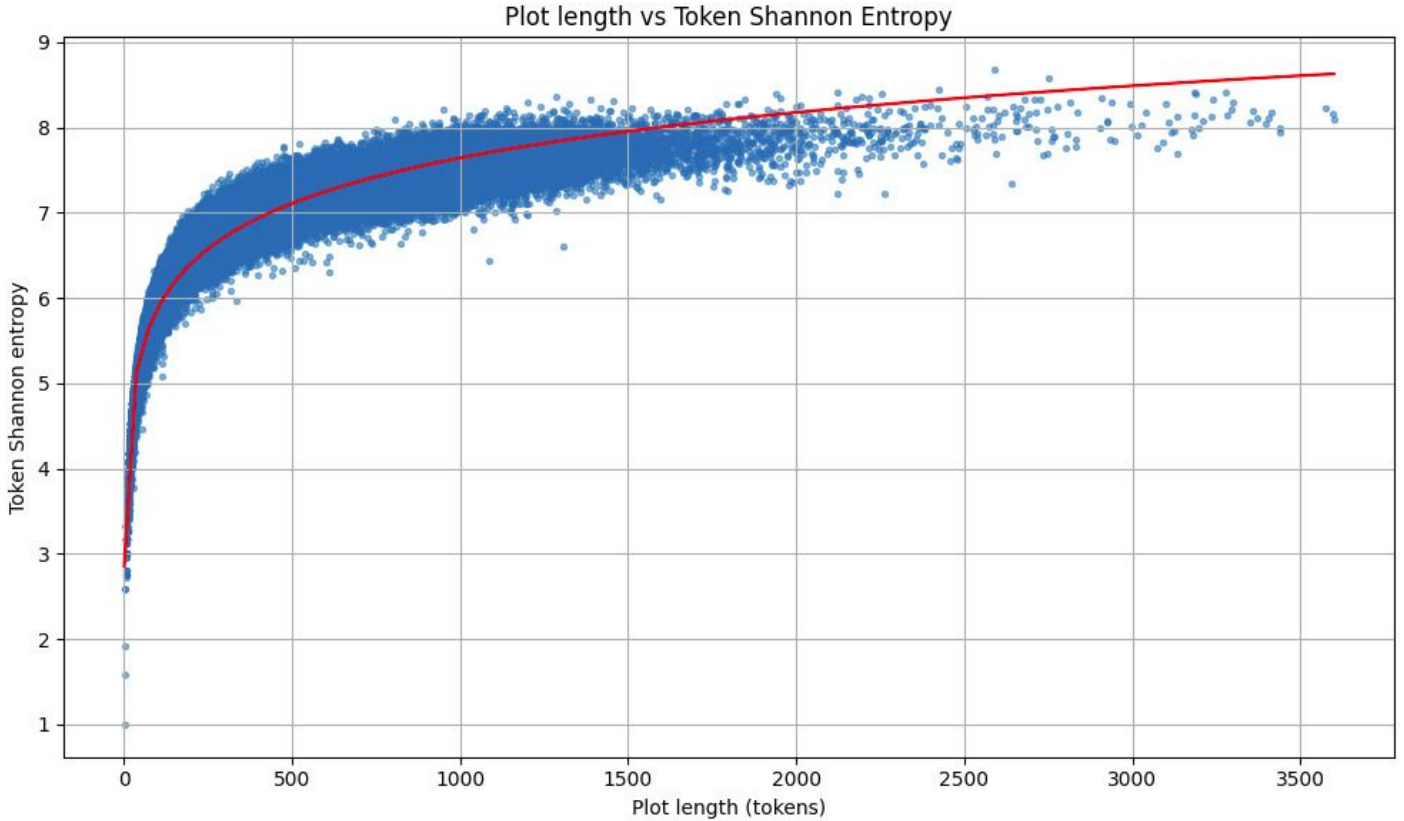


Plot length cut-off – token distribution entropy

- (Wenzek et al., 2019.) used perplexity-based filtering to remove low-quality text from large-scale monolingual datasets
- We tokenized the plots and computed Shannon entropy of the empirical distribution (relative frequencies) of the tokens



Length of plot in tokens vs Shannon entropy



Exploration

Entropy = 3.0 ❌

1982 (2013 film)

Article Talk

From Wikipedia, the free encyclopedia

1982 is a 2013 drama film written and directed by Tommy Oliver and starring Hill Harper.[1] It is Oliver's directorial debut.[2] The film is also semi-autobiographical.[3] The film marked the final appearance of actress and activist Ruby Dee before her death in 2014.

Plot [edit]

This article's plot summary **needs to be improved**. Please help do so. Relevant discussion may be found on the talk page. *(August 2024)* *(Learn how and when to remove this message)*

The film is set in Philadelphia 1982.

Entropy = 6.0 ✅

The True Story of Tamara de Lempicka and The Art of Survival

Add languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The True Story of Tamara de Lempicka & The Art of Survival is a documentary film directed, written, and produced by Julie Rubio. The film explores the life and artistic legacy of Polish Art Deco painter Tamara de Lempicka, utilizing previously unseen 8 mm home movies, paintings, and newly discovered birth and baptism certificates. It reveals her true name, heritage, and identity for the first time.[1][2]

Synopsis [edit]

The documentary examines Lempicka's personal life, including her relationships with male and female models, her bisexuality, and her Jewish heritage. It features interviews with art historians, family members including her granddaughter and great-granddaughters, Broadway performer Eden Espinosa, collectors, and narration by Anjelica Huston. These elements are combined with archival materials to highlight themes of artistic survival, resilience, identity, and creative expression.[3][4]

What to do in the middle?

- Sample 150 movies from entropy range [4.0, 5.5] and manually annotate them as “good” or not
- “Not good” plot examples from the sample:

```
Q5055548,"Out of ambition, a woman sacrifices her true love and marries a man of fortune.",0
```

```
Q6450711,The film follows a story of King Kyansittha of Bagan Dynasty.,0
```

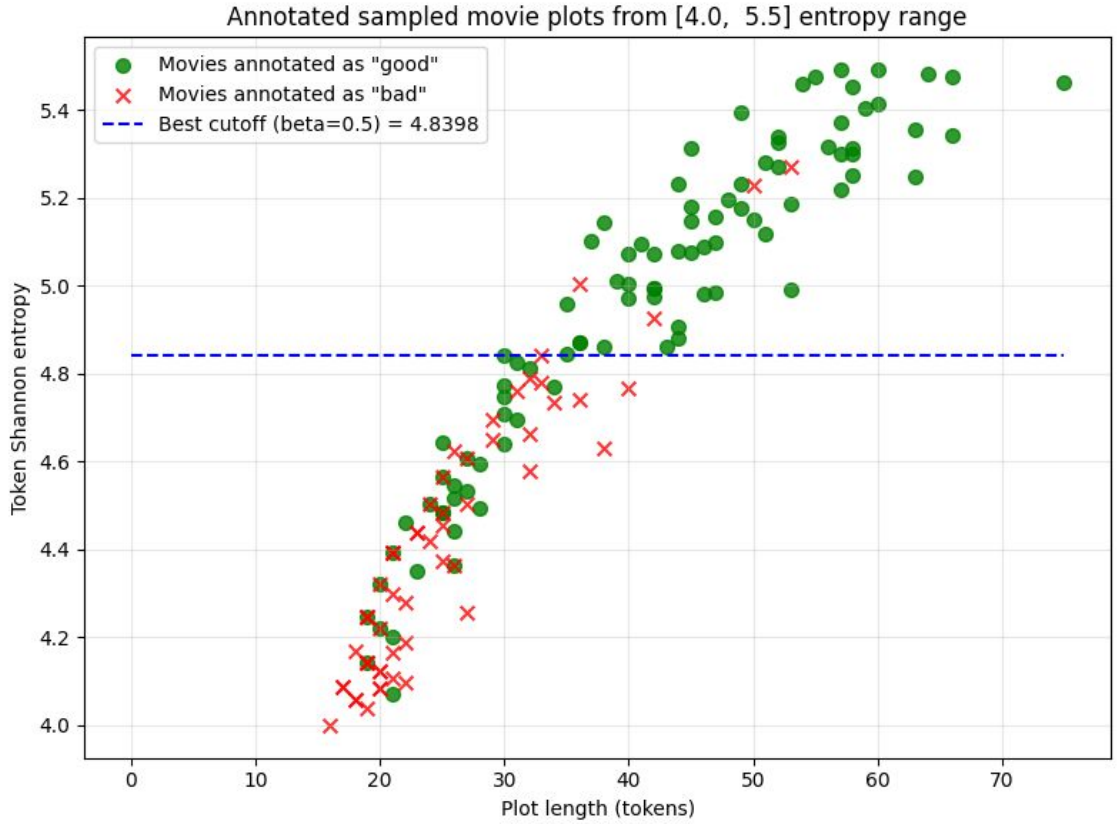
```
Q17026229,"The original TV listing in The Age newspaper described the plot as ""a woman's reactions to her husband's suspected affair with another woman""."",0
```

- “Good” plot examples from the sample:

```
Q7750102,A brilliant scientist suffering from amnesia is hunted by Communist agents in search of a secret formula.,1
```

```
Q12684212,"Captain Robert Rogers, a British Army officer, publishes a book about his father's exploits. After it is ridiculed as a hoax, Rogers leaves for the Malay Peninsula to prove the existence of Booloo, the legendary tiger that killed his father.",1
```

Sample results



Post cut-off extremes

Plot with the lowest entropy

Tiger Shark (film)

Article Talk

From Wikipedia, the free encyclopedia

Not to be confused with Tigershark (film).

Tiger Shark is a 1932 American pre-Code melodrama romantic film directed by Howard Hawks and starring Edward G. Robinson, Richard Arlen and Zita Johann.[2]

Plot [edit]

The wife of one-handed tuna fisherman Mike Mascarenhas falls for the man whose life Mike had saved while at sea.

Shortest plot (by number of tokens)

The Singer of Naples

Article Talk

From Wikipedia, the free encyclopedia

The Singer of Naples (Spanish: ***El cantante de Napoles***) is a 1935 American musical film directed by Howard Bretherton and Moreno Cuyar and starring Enrico Caruso Jr., Mona Maris and Carmen Río. It was made in Spanish by the Hollywood studio Warner Brothers. Unlike many other American Spanish language films of the era it was not a remake of an English language film.

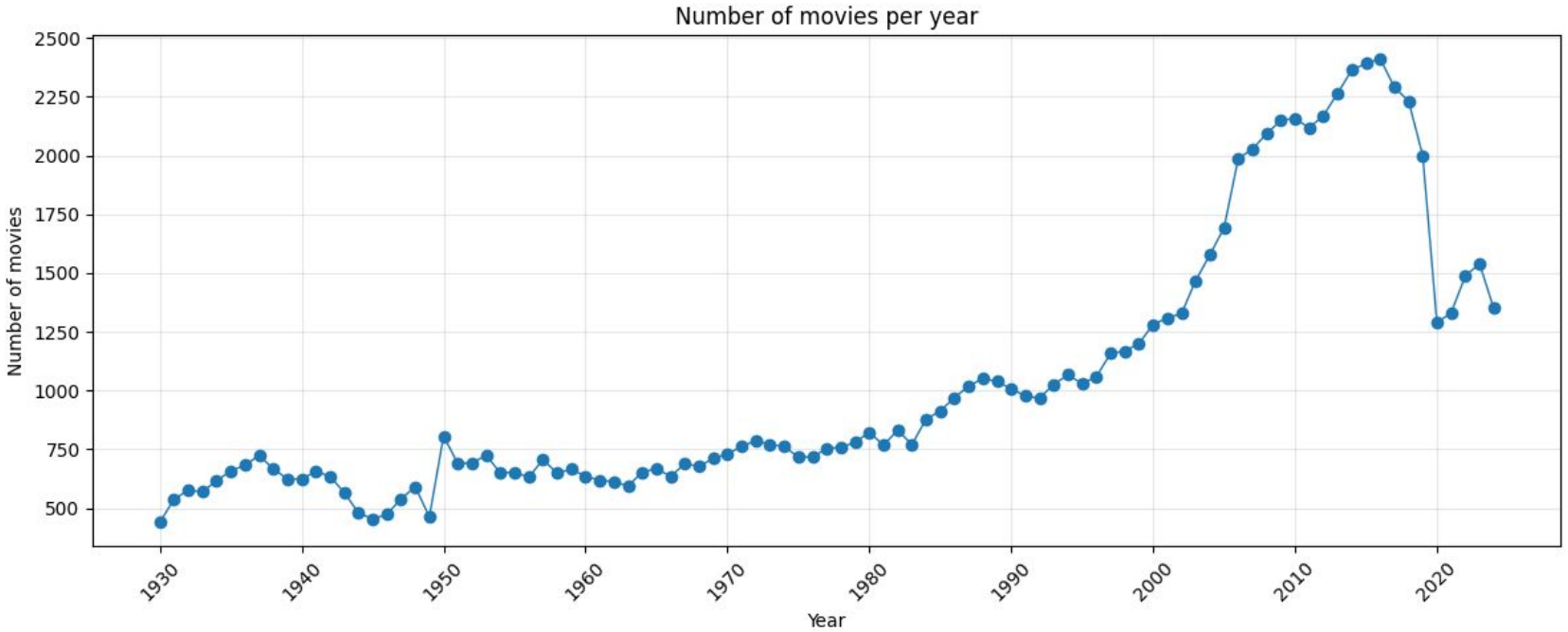
It was the last of Warner Brothers's Hollywood-made Spanish films.[1] The increasing success of dubbing meant that it was less viable to make separate Spanish films, and in future it became more common for films to be made in a single English version and then dubbed into a variety of other languages for global release.

Plot [edit]

A blacksmith's son from Naples rises to become a celebrated opera singer, performing at La Scala in Milan.

Final data

- 93696 movies in the dataset (4798 movies removed by cut off)



From Raw Data to Rich Insights: The Movie Data Pipeline

A 4-step process transforming raw movie information from multiple sources into an analysis-ready dataset.

STEP 1: Collect from Wikidata

Gathers foundational movie data like titles, genres, directors, and release dates.

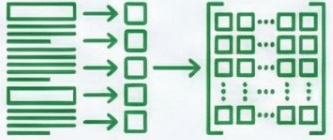
STEP 2: Enrich with TMDb

Adds valuable context by fetching user ratings and popularity scores.

STEP 3: Retrieve from Wikipedia

Extracts and cleans detailed plot summaries from movie Wikipedia pages.

STEP 4: Generate AI-Ready Data



Converts plot text into numerical vectors (embeddings) for advanced analysis and machine learning tasks.

Die Film-Datenpipeline: Von Rohdaten zu KI-Erkenntnissen

Datenquellen & Ziel

Ziel: Ein strukturierter Datensatz für ML-Analysen

Umwandlung von Rohdaten in einen Datensatz mit semantischen Embeddings für Aufgaben wie Ähnlichkeitssuche, Clustering und Klassifizierung.

Drei primäre Datenquellen

Wikidata:
Für grundlegende Metadaten (Titel, Genre, Regisseure, IMDB-ID).

TMDb:
Für Anreicherungsdaten (Popularität, Bewertungen, Stimmenanzahl).

Wikipedia:
Für textliche Inhalte (Handlungszusammenfassungen).

Inkrementelle Verarbeitung

Verarbeitet nur fehlende Daten, was die Pipeline wiederaufnehmbar (resumable) macht.

Skalierbarkeit

Die Organisation nach Jahren und die parallele Verarbeitung ermöglichen die Handhabung großer Datenmengen.

Der 4-stufige Verarbeitungsprozess

Schritt 1: Metadaten-Sammlung aus Wikidata

Sammelt grundlegende Filmdaten für die Jahre 1888-2024 mittels SPARQL-Abfragen und speichert sie als jährliche CSV-Dateien.

Inkrementelle Abfrage
Überspringt Jahre, für die bereits eine CSV-Datei (`wikidata_movies_{jahr}.csv`) existiert, um doppelte Arbeit zu vermeiden.

Wichtige extrahierte Felder
Titel, Genre, Regisseure, Schauspieler, Erscheinungsdatum, IMDB-ID, Wikipedia-Link, Budget & Einspielergebnis.

Schritt 2: Anreicherung mit TMDb-Daten

Fügt den bestehenden Daten Popularitätsmetriken und Nutzerbewertungen aus der TMDb-API hinzu.

Gezielte Anreicherung
Verarbeitet nur Filme, denen TMDb-Daten fehlen, und schont so API-Limits und Verarbeitungszeit.

Hinzugefügte Daten
Popularitäts-Score, durchschnittliche Bewertung (`vote_average`) und Anzahl der Stimmen (`vote_count`).

Schritt 3: Abruf der Handlung aus Wikipedia

Extrahiert Handlungszusammenfassungen von den englischen Wikipedia-Seiten der Filme.

Parallele Verarbeitung
Nutzt standardmäßig 8 Worker-Threads, um mehrere Filmhandlungen gleichzeitig und damit deutlich schneller herunterzuladen.

Intelligente Text-Extraktion
Sucht nach Sektionen wie "Plot", "Synopsis" oder "Story" und bereinigt den extrahierten Text automatisch.

Schritt 4: Erzeugung semantischer Embeddings

Wandelt die bereinigten Handlungstexte mithilfe von Transformer-Modellen (z.B. BAAI/bge-m3) in numerische Vektoren (Embeddings) um.

Multi-GPU-Unterstützung
Nutzt verfügbare GPUs parallel, um den rechenintensiven Embedding-Prozess massiv zu beschleunigen. Fällt bei Nichtverfügbarkeit auf die CPU zurück.

Flexible Chunking-Strategien
Unterstützt verschiedene Methoden, um lange Texte für die Verarbeitung durch die Modelle aufzuteilen (z.B. `cls_token`, `mean_pooling`).

Ergebnis: Analysebereite Datensätze

Strukturierte Ausgabedateien pro Jahr

Die Pipeline erzeugt für jedes verarbeitete Jahr ein Set von aufeinander abgestimmten Dateien, die eine einfache Verknüpfung von Metadaten und Embeddings ermöglichen.

Data Table

Metadaten & Handlung
`wikidata_movies_{jahr}.csv` | Alle gesammelten und angereicherten Informationen als Tabelle.

Embeddings
`movie_embeddings_{jahr}.npy` | NumPy-Array mit den Vektoren der Handlungen (n_filme × embedding_dim).

Film-IDs
`movie_ids_{jahr}.npy` | NumPy-Array mit den Wikidata-IDs, das die Reihenfolge der Embeddings definiert.

Lexikalische Gewichte (Optional)
`movie_lexical_weights_{jahr}.npz` | Sparse-Gewichte zur Identifikation wichtiger Begriffe im Text.

Fehlertoleranz

Einzelfehler (z.B. bei einem Film) führen nicht zum Abbruch der gesamten Pipeline.

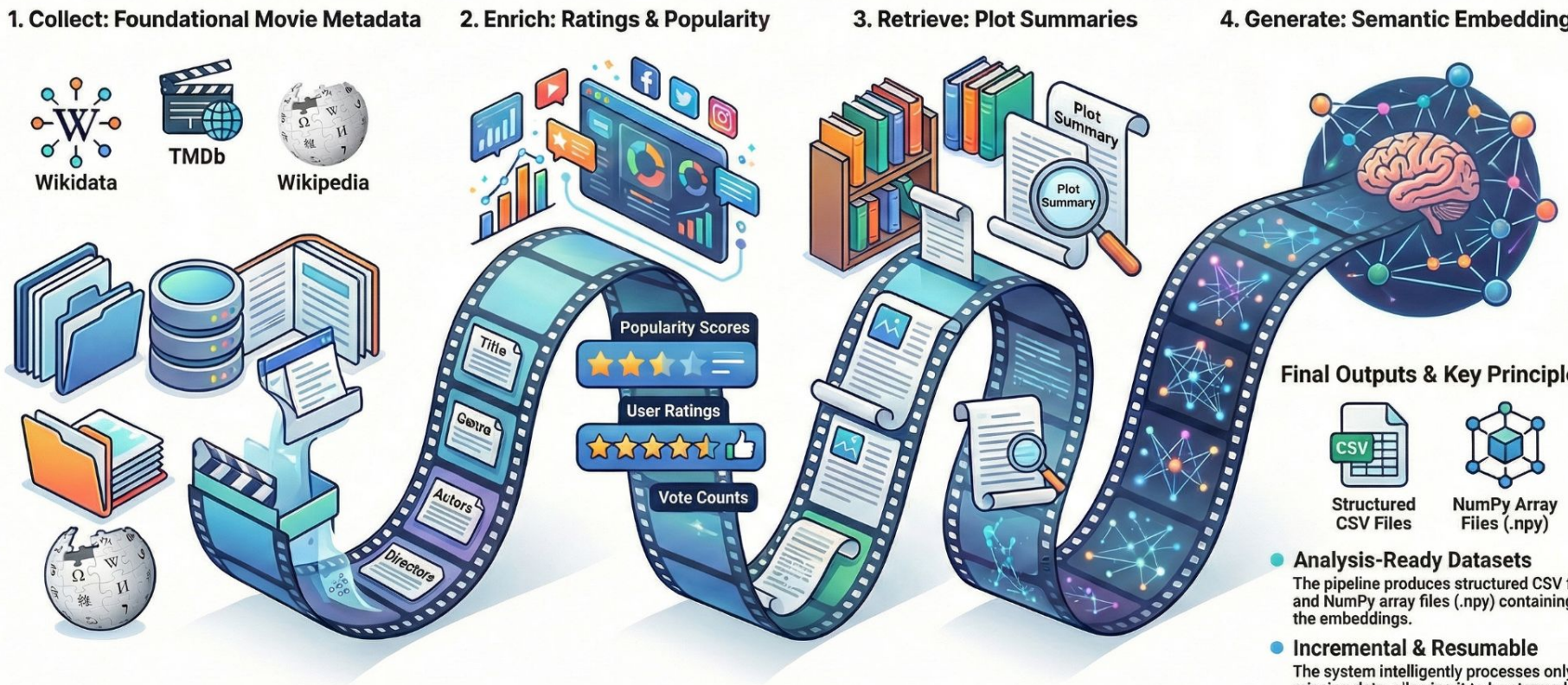
Flexibilität & Konfigurierbarkeit

Parameter wie Jahresbereich, Embedding-Modell und Chunking-Methode sind anpassbar.

Datenvalidierung

In jedem Schritt wird die Integrität der Daten überprüft, um eine hohe Qualität des Endprodukts sicherzustellen.

The Movie Data Pipeline: From Raw Data to AI-Ready Insights



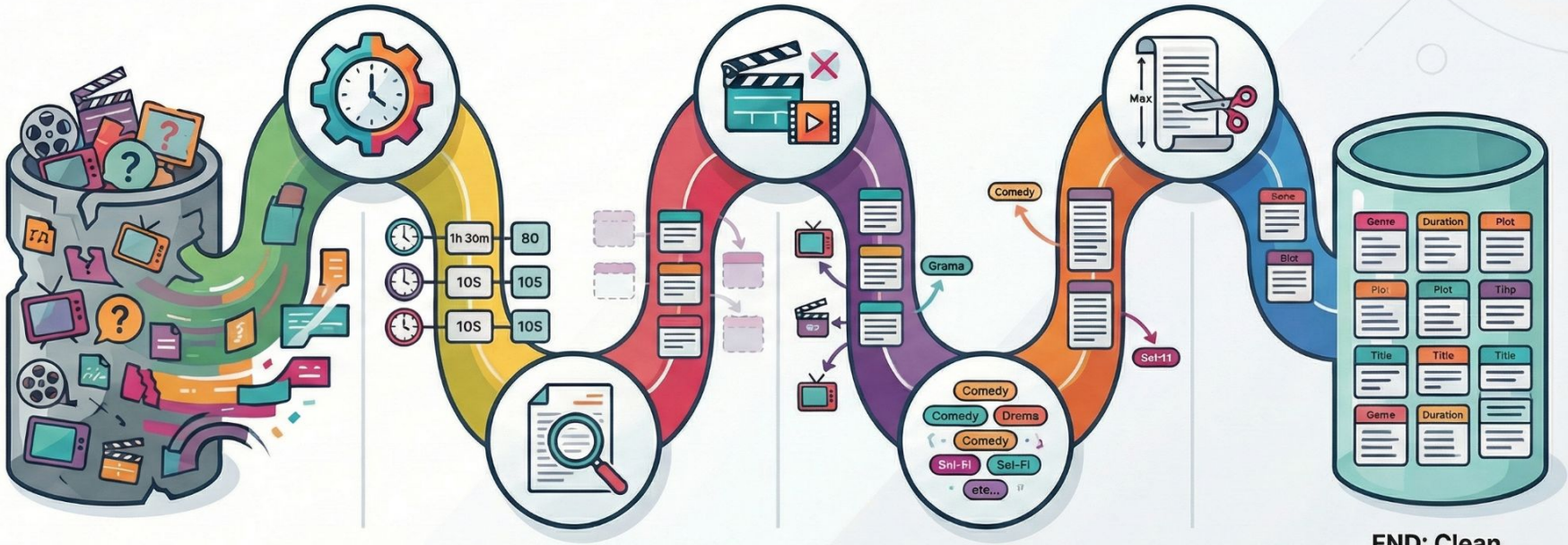
1. Collect
Gathers core movie info (title, genre, actors, directors) from the Wikidata database.

2. Enrich
Adds popularity scores, user ratings, and vote counts by matching movies with The Movie Database (TMDb).

3. Fetches and cleans the full text of plot summaries from each movie's English Wikipedia page.

4. Generate
Uses AI models to convert plot text into numerical vectors (embeddings) that capture semantic meaning.

The Movie Data Cleaning Pipeline



START: Raw Movie Dataset

The initial, unfiltered dataset containing movies, TV shows, and incomplete entries.

1. Convert Duration

Converts the 'duration' column to a numeric format for consistency.

2. Filter Movies Without Plots

Removes entries missing plot summaries, which are essential for semantic analysis.

3. Filter Non-Movies

Removes items like TV shows, short films, and trailers using a cached class list.

4. Normalize Genres (Optional)

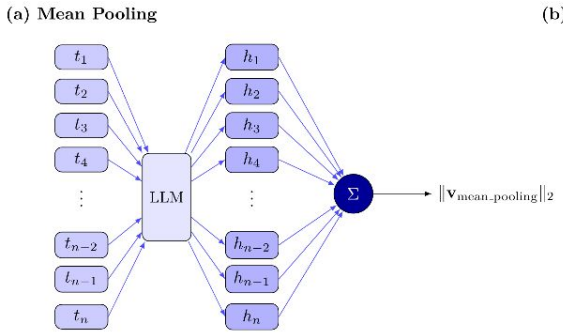
Removes genres that only appear once in the dataset to reduce noise.

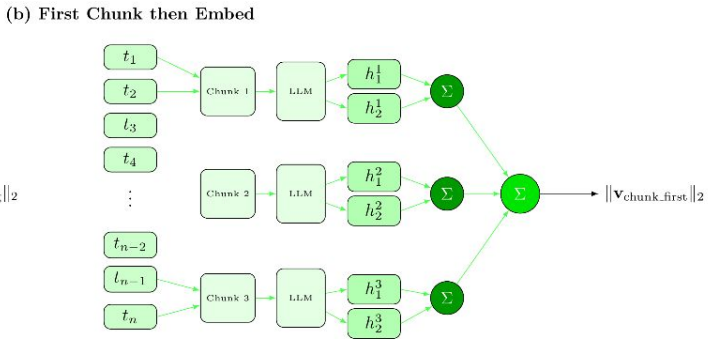
5. Filter by Plot Length

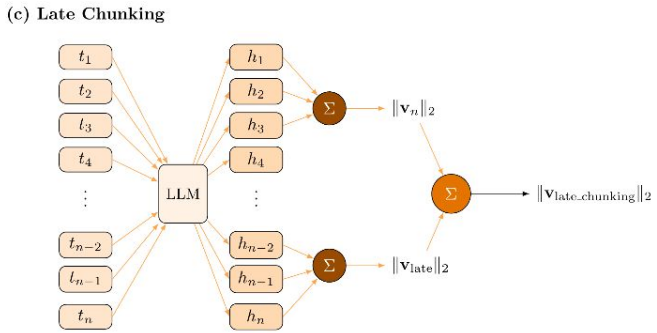
Removes movies with excessively long plots (over 14,900 characters) to avoid data errors.

END: Clean, Analysis-Ready Dataset

The final dataset is ready for genre classification, visualization, and analysis.


(a) Mean Pooling
LLM
$\|\mathbf{v}_{\text{mean\_pooling}}\|_2$


(b) First Chunk then Embed
Chunk 1
Chunk 2
Chunk 3
LLM
$\|\mathbf{v}_{\text{chunk\_first}}\|_2$


(c) Late Chunking
LLM
$\|\mathbf{v}_n\|_2$
$\|\mathbf{v}_{\text{late}}\|_2$
$\|\mathbf{v}_{\text{late\_chunking}}\|_2$

Genres - Classification

- there were 463 unique genres after cleaning, and dropping genres with only one movie
- based on the wikidata QID → 359 were left with description of the genres
- embedding these descriptions with the bge-m3 model
- with k means algorithmen, 20 cluster were fitted →Lloyd
- these 20 cluster were named based on their appearance by us

→ all movies with 20 genres classified (multi labeling allowed)

The Movie Genre Classification Pipeline

1. Extract Unique Genres

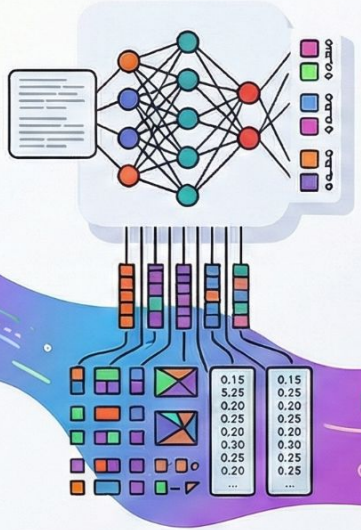
Identify and list all unique genres from a movie dataset.

463 unique genres

2. Fetch Genre Descriptions

Automatically fetch summary description from Wikipedia, caching results.

359 genres left

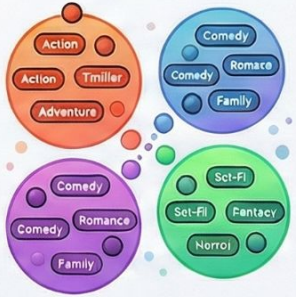


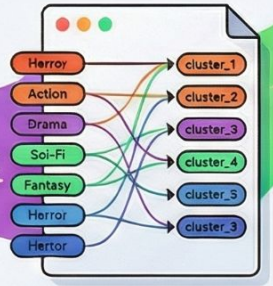
3. Convert Text to Vectors

A Sentence Transformer model reads text and converts to numerical vectors (embeddings).

4. Group Similar Genres

KMeans algorithm groups genres with similar meanings into a set number of clusters.





5. Generate a Genre Map

A JSON mapping file is created, assigning each genre to its final cluster label.

6. Analyze and Visualize Results

Display which genres belong to each cluster and plot charts showing movie counts per cluster.

Genres - Classification

Cluster	Cluster Name	Genres	QIDs	Genre Count	Total Movies
0	Action Crime & Exploitation	Satanic film, Z movie, anti-war film, arthouse science fiction film, boxing film, buddy cop film, gangster film, ghost film, girls with guns, mafia comedy film, mafia film, ninja film, nunsploitation, outlaw biker film, poliziotteschi, prison film, psychedelic film, rape and revenge film, stoner film, superhero film, supernatural horror film, vigilante film, zombie film	Q18355406, Q1117103, Q4774498, Q108084492,Q2484376, Q895583, Q4984974, Q17013749,Q7444356, Q16702172,Q31888058, Q1776156, Q6729489, Q33130924, Q3745429, Q1377546, Q4075563, Q1740789, Q586250, Q21858363, Q1067324,Q2254193, Q457832, Q1535153, Q3634883,Q43911809, Q2642760, Q2724311,Q3072049	23	2807
1	Satire & Dark Comedy	absurdist fiction, black comedy, black comedy film, burlesque, comedy of manners, humor, parable, political satire, religious satire, satire, surreal humour, tragedy	Q1770859, Q53094, Q5778924, Q217117, Q222926, Q60548032, Q208555, Q7225114, Q7311396, Q128758, Q1092460, Q80930	12	366
2	Action Western & International	B movie, B western, Bollywood, Bruceploitation, Christmas film, Filmfarsi, Masala film, Ostern, Revisionist Western, Rumberas film, Spaghetti Western, Tamil cinema, Western, Western film, acid Western, action film, action thriller, action-adventure film, art film, auteur film, blaxploitation film, disaster film, epic film, heist film, heroic bloodshed, hood film, independent film, kung fu film, martial arts film, pirate film, race film, road movie, science fiction action film, splatter film	Q223770, Q4836991, Q93196, Q580013, Q17175676,Q28026639, Q5712461, Q3028680, Q26268098, Q491158, Q7379160, Q212781, Q926324, Q21590660, Q172980, Q4674071,Q638364, Q188473,Q200092,Q2143665,Q542475, Q3990883, Q78461348, Q1135802, Q13209138, Q883179, Q814158,Q846544, Q576131,Q652256, Q496523, Q622310, Q11304653,Q5897543, Q157394,Q240911,Q459290, Q3072042, Q1033891, Q2096633, Q2125170, Q628165, Q20656232, Q644437,Q909586	34	13383
3	Horror & Sci-Fi	Weird West, alien invasion, biopunk, body horror, body horror film, cyberpunk, gothic film, horror fiction, horror film, horror western, list of holiday horror films, monster film, natural horror film, psychological horror, psychological horror film, religious horror, science fiction horror film, slasher film, space Western, space opera, surrealist cinema, vampire film, werewolf film, white savior film	Q775169, Q2447078, Q1077883, Q3641550, Q102260466, Q174526, Q5769663, Q16575965, Q157394,Q200092,Q319221, Q108280930, Q20649626, Q1065444,Q1342372,Q2973181, Q109626272, Q604725, Q109629396, Q123291714, Q10663882, Q531067,Q853630, Q4235011, Q468478, Q1190502, Q2137852, Q5258881, Q132803402	24	7139
4	Exploitation & Erotic	Bavarian porn, Nazi exploitation, Redsploitation, cannibal film, cartoon pornography, erotic film, erotic thriller, erotic thriller film, exploitation film, pink film, pornographic film, pornographic parody film, psycho-biddy, sex film, sexploitation film, soft-core pornography, women in prison film	Q84041, Q535518, Q66425231, Q1723850, Q4373044, Q599558, Q2254193,Q2439025,Q317309, Q109733294, Q1067324,Q909586, Q1194365, Q185529, Q16254232, Q7256286, Q2275499, Q2292320, Q2000136, Q583768	17	1381
5	Adventure & Fantasy	Christian fiction, bildungsroman, children's literature, robinsonade, swashbuckler film, sword and sorcery, sword and sorcery film, wuxia, wuxia film	Q100339660, Q223945, Q131539, Q279060, Q222639, Q1999690, Q130130466, Q754803, Q15858553	9	233
6	Romance	love, romance, romance film, romantic comedy, romantic comedy film, romantic drama, romantic drama film, romantic fantasy, romantic fiction	Q316, Q599510,Q738473, Q1047337,Q1054574,Q860626, Q226730,Q860626, Q118612349, Q3038946, Q117536422, Q930383, Q19765983	9	9402
7	Anime	adventure anime, drama anime, fantasy anime, mystery anime, romance anime, science fiction anime	Q104536870, Q104536994, Q104536896, Q104623091, Q104536771, Q103925569	6	53

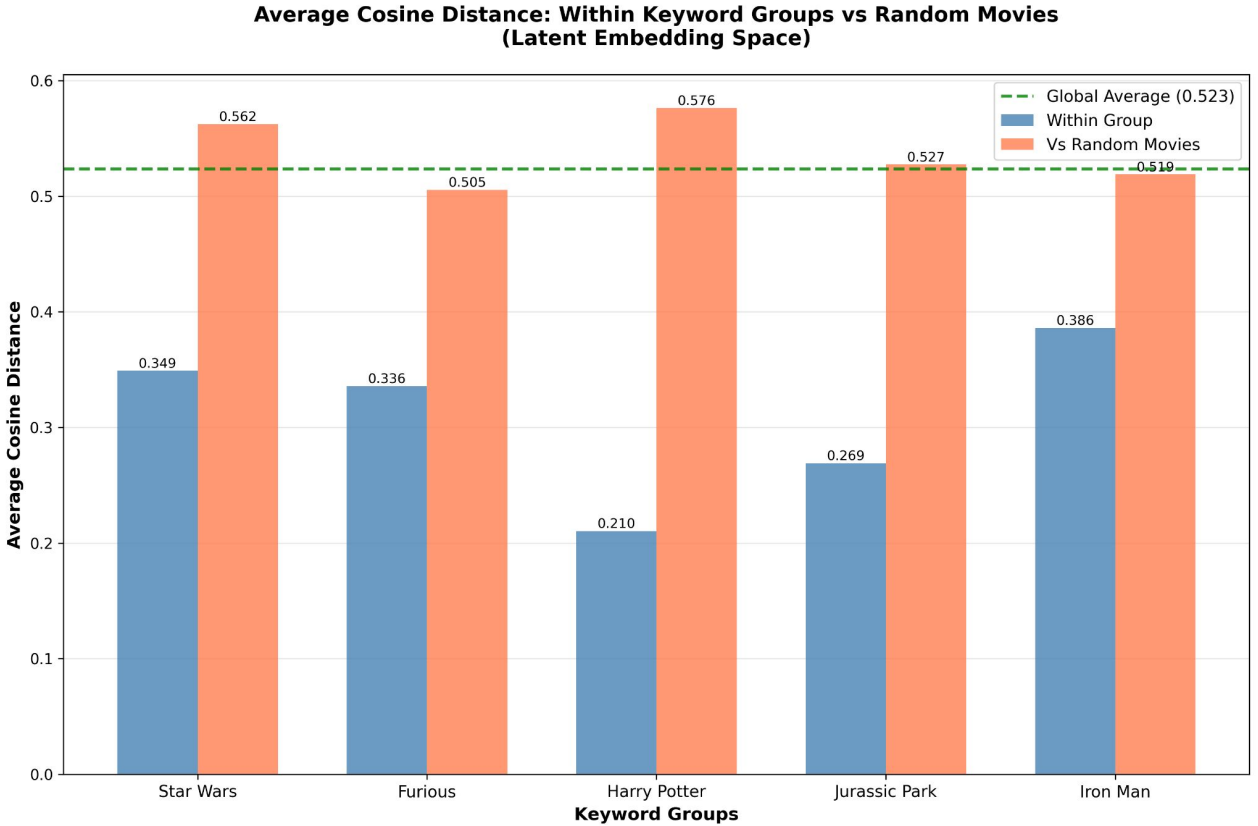
Table 1 – continued from previous page

Cluster	Cluster Name	Genres	QIDs	Genre Count	Total Movies
16	Crime & Thriller	action fiction, crime, crime comedy film, crime drama film, crime fiction, crime film, crime literature, crime thriller film, detective fiction, detective film, financial thriller, giallo, legal thriller, mystery fiction, mystery film, police film, police procedural, police procedural film, political thriller, political thriller film, psychological thriller, psychological thriller film, romantic thriller, spy film, suspense, suspense film, techno-thriller, thriller, thriller film, trial film, true crime	Q1762165, Q83267, Q1788980, Q113485322,Q130232,Q19367312, Q5937792, Q2484376,Q959790, Q20664530, Q116514801,Q12049743,Q1321123,Q19367312, Q186424, Q25533274, Q1864294, Q1332055, Q2490520, Q6585139, Q1128592,Q1200678, Q2101714, Q2321734, Q109733630, Q622291, Q109733333, Q242488,Q590103, Q107397740,Q109733304,Q52207399,Q63214877, Q7362831, Q2297927, Q9503, Q11304653,Q19367312, Q580850, Q182015, Q1141200,Q1200678,Q1341831,Q1342372,Q2116008,Q2439025,Q2484376,Q3072039, Q130232,Q3072039, Q3541116	31	13329
17	Documentary Experimental & Film History	Heimatfilm, Turksploitation, anthology film, compilation film, docufiction film, documentary film, environmental film, essay film, ethnofiction, experimental film, fan film, fiction, fiction film, film score, filmi music, found footage, miniseries, mockbuster, mockumentary, mondo film, nature documentary, propaganda film, pseudo-documentary, puppet film, rubble film, short film, silent film, sound film, soundtrack, talk show, television documentary	Q304538, Q1092361, Q336144, Q455315, Q472637, Q27406,Q37073,Q93204, Q112633637, Q11356864, Q3059309, Q459290,Q790192, Q2301591, Q8253, Q12912091, Q492264, Q544440, Q2484376,Q3272147, Q1259759, Q1941707, Q459435, Q1474387, Q1760864, Q1935609, Q8812380, Q10475300,Q3745424, Q882006, Q24862, Q226730,Q39818, Q848512, Q107736421, Q622812, Q7603925	31	4796
18	Japanese Media	J-pop, Japanese horror, Jidaigeki, adult animation, anime, anime film, ecchi, hentai, isekai, kaiju, magical girl, mecha, samurai cinema, tokusatsu, yakuza film	Q131578, Q2584671, Q629917, Q3249257, Q1107, Q20650540, Q219559, Q172067, Q53911753, Q1065444,Q846544, Q752321, Q4292083,Q858784, Q169672, Q275934, Q731194	15	353
19	Musical	musical, musical comedy, musical film, opera, opera film, operetta, vaudeville	Q2743, Q1548170, Q49451,Q842256, Q1344, Q16909344, Q170384, Q186286	7	3378

Data All Cleaned

Data All Cleaned

Data Sainty Check

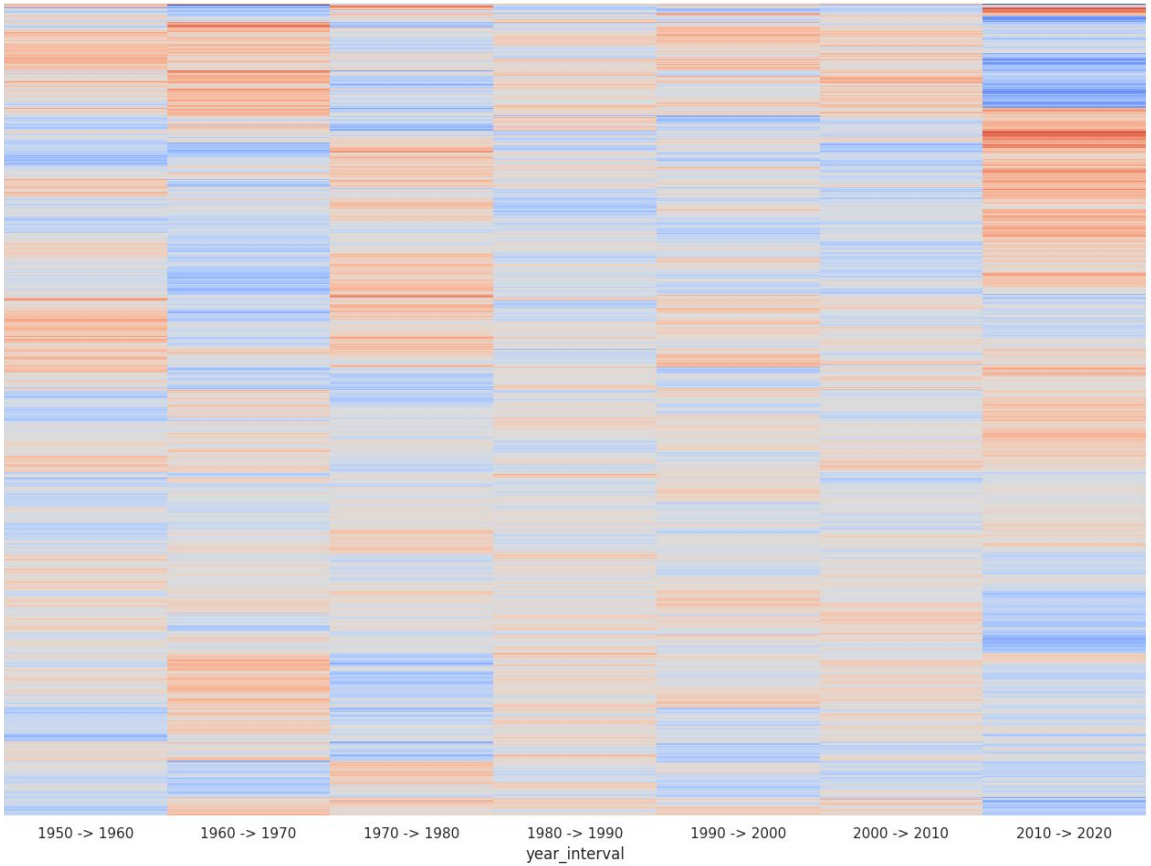


#	QID	Title	Cosine Distance	Cosine Similarity
1	Q102244	Harry Potter and the Chamber of Secrets	0.195199	0.804801
2	Q161687	Harry Potter and the Half-Blood Prince	0.206232	0.793768
3	Q102235	Harry Potter and the Order of the Phoenix	0.214408	0.785592
4	Q232009	Harry Potter and the Deathly Hallows - Part 2	0.216356	0.783644
5	Q102448	Harry Potter and the Prisoner of Azkaban	0.217598	0.782402
6	Q102225	Harry Potter and the Goblet of Fire	0.226034	0.773966
7	Q161678	Harry Potter and the Deathly Hallows - Part 1	0.243846	0.756154
8	Q3297713	Whorrey Potter and the Sorcerer's Balls	0.288847	0.711153
9	Q636906	Young Sherlock Holmes	0.324280	0.675720
10	Q841118	Epic Movie	0.324757	0.675243
11	Q3234275	Troll	0.328279	0.671721
12	Q204662	The Sword in the Stone	0.334512	0.665488
13	Q3231484	The Hounds of Baskerville	0.334758	0.665242
14	Q7620158	Storm Over the Nile	0.341511	0.658489
15	Q42689885	Tolkien	0.346885	0.653115
16	Q134430	Snow White and the Seven Dwarfs	0.349817	0.650183
17	Q4659631	A Snow White Christmas	0.349872	0.650128
18	Q13393950	My Little Pony: Equestria Girls	0.353233	0.646767
19	Q2247470	Snow White and the Seven Dwarfs	0.353462	0.646538
20	Q461768	The Spiderwick Chronicles	0.353912	0.646088
21	Q2002243	Snow White: The Fairest of Them All	0.354832	0.645168
22	Q369986	Halloweentown High	0.358626	0.641374
23	Q2417639	The Strange Affair of Uncle Harry	0.358653	0.641347
24	Q3203700	Fanny by Gaslight	0.359087	0.640913
25	Q74229	Hobbit	0.359391	0.640609
26	Q583483	Winx Club: The Secret of the Lost Kingdom	0.361421	0.638579
27	Q270061	7 Zwerge - Der Wald ist nicht genug	0.361610	0.638390
28	Q15713892	Pan	0.362400	0.637600
29	Q280400	Matilda	0.362798	0.637202

Top 30 Closest Neighbors for Skyfall (QID: Q4941)

#	QID	Title	Cosine Distance	Cosine Similarity
1	Q18602670	Spectre	0.218123	0.781877
2	Q212145	The World Is Not Enough	0.220027	0.779973
3	Q207916	Tomorrow Never Dies	0.230349	0.769651
4	Q151904	Casino Royale	0.233878	0.766122
5	Q591272	Casino Royale	0.248780	0.751220
6	Q19089	GoldenEye	0.251589	0.748411
7	Q181540	Quantum of Solace	0.253877	0.746123
8	Q107914	Diamonds Are Forever	0.260126	0.739874
9	Q320423	The Spy Who Loved Me	0.260874	0.739126
10	Q332368	A View to a Kill	0.266096	0.733904
11	Q21534241	No Time to Die	0.278656	0.721344
12	Q30931	Die Another Day	0.280897	0.719103
13	Q106440	Goldfinger	0.281409	0.718591
14	Q107894	On Her Majesty's Secret Service	0.281961	0.718039
15	Q102754	Dr. No	0.282595	0.717405
16	Q332330	For Your Eyes Only	0.283180	0.716820
17	Q180279	Never Say Never Again	0.285535	0.714465
18	Q107724	Thunderball	0.289128	0.710872
19	Q309289	The Man with the Golden Gun	0.291500	0.708500
20	Q334780	Moonraker	0.292210	0.707790
21	Q309086	Licence to Kill	0.292969	0.707031
22	Q27204	Live and Let Die	0.298403	0.701597
23	Q16991846	Escape Route	0.305203	0.694797
24	Q106571	From Russia with Love	0.307007	0.692993
25	Q35160593	Johnny English Strikes Again	0.307825	0.692175
26	Q5227843	The Parole Officer	0.311112	0.688888
27	Q107761	You Only Live Twice	0.311845	0.688155
28	Q272064	The Living Daylights	0.314873	0.685127
29	Q7699180	Ten Days in Paris	0.319941	0.680059
30	Q594402	Casino Royale	0.324638	0.675362

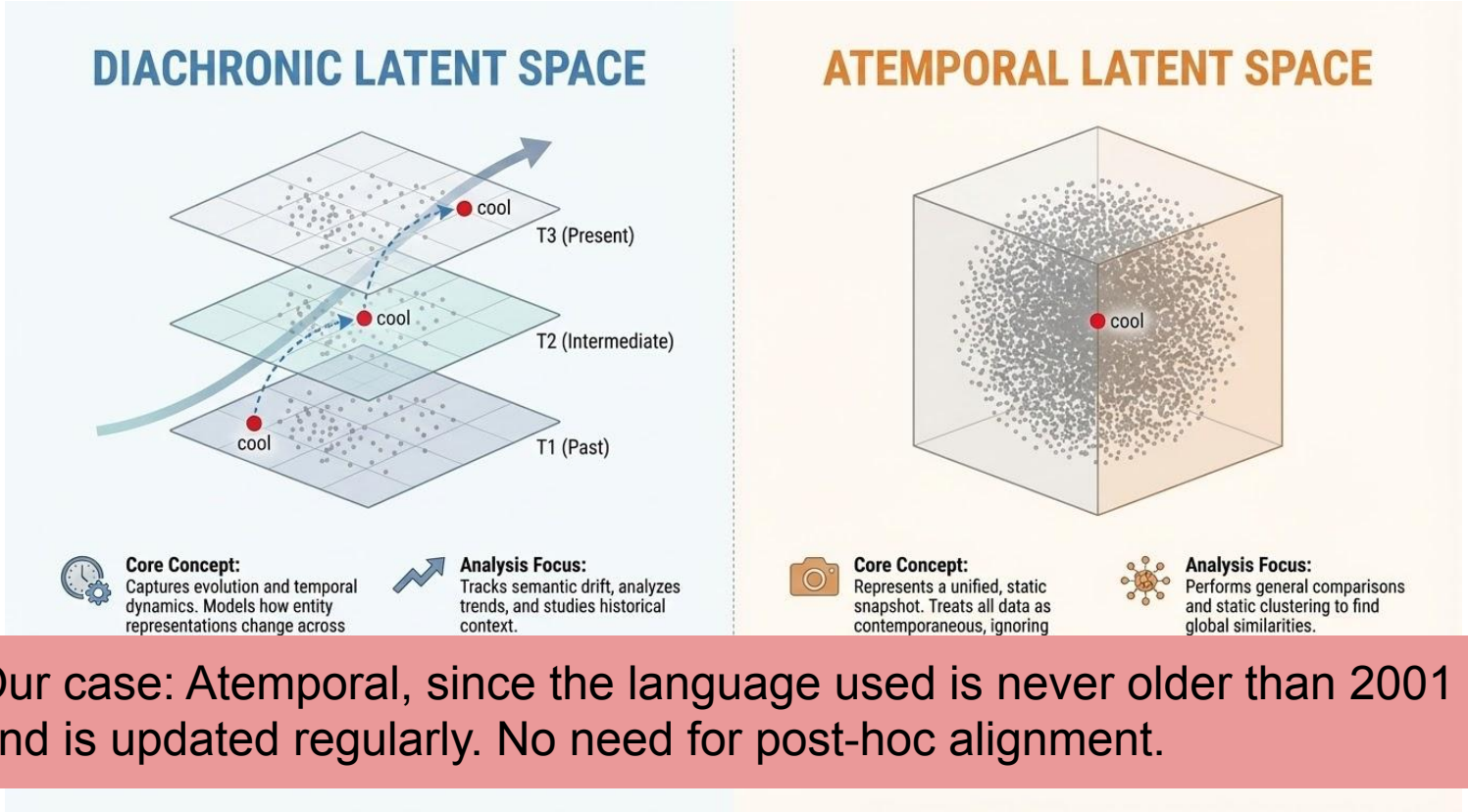
Romance movies



Literature Review - Methods

1. Visualization techniques:
    - Dimensionality Reduction
2. Is mean the best measure of “average”
3. Analysis of embeddings over time
    - To gain inspiration for plots
    - Transferable methods

Diachronic vs atemporal latent space for analysis



Our case: Atemporal, since the language used is never older than 2001 and is updated regularly. No need for post-hoc alignment.

Literature Review - Dimensionality Reduction

Why dimension reduction?

Dimensionality reduction is mainly used to

1. Enhance machine learning models - preserve essential features, reduce number of predictor variables, increased generalizability
2. Save storage space (especially for sparse data)
3. Reduce compute of distance measures (for repetitive comparisons, e.g. pairwise distances, RAG)

For us, dimensionality reduction is mainly used to visualize our analysis.

Literature Review - Dimensionality Reduction

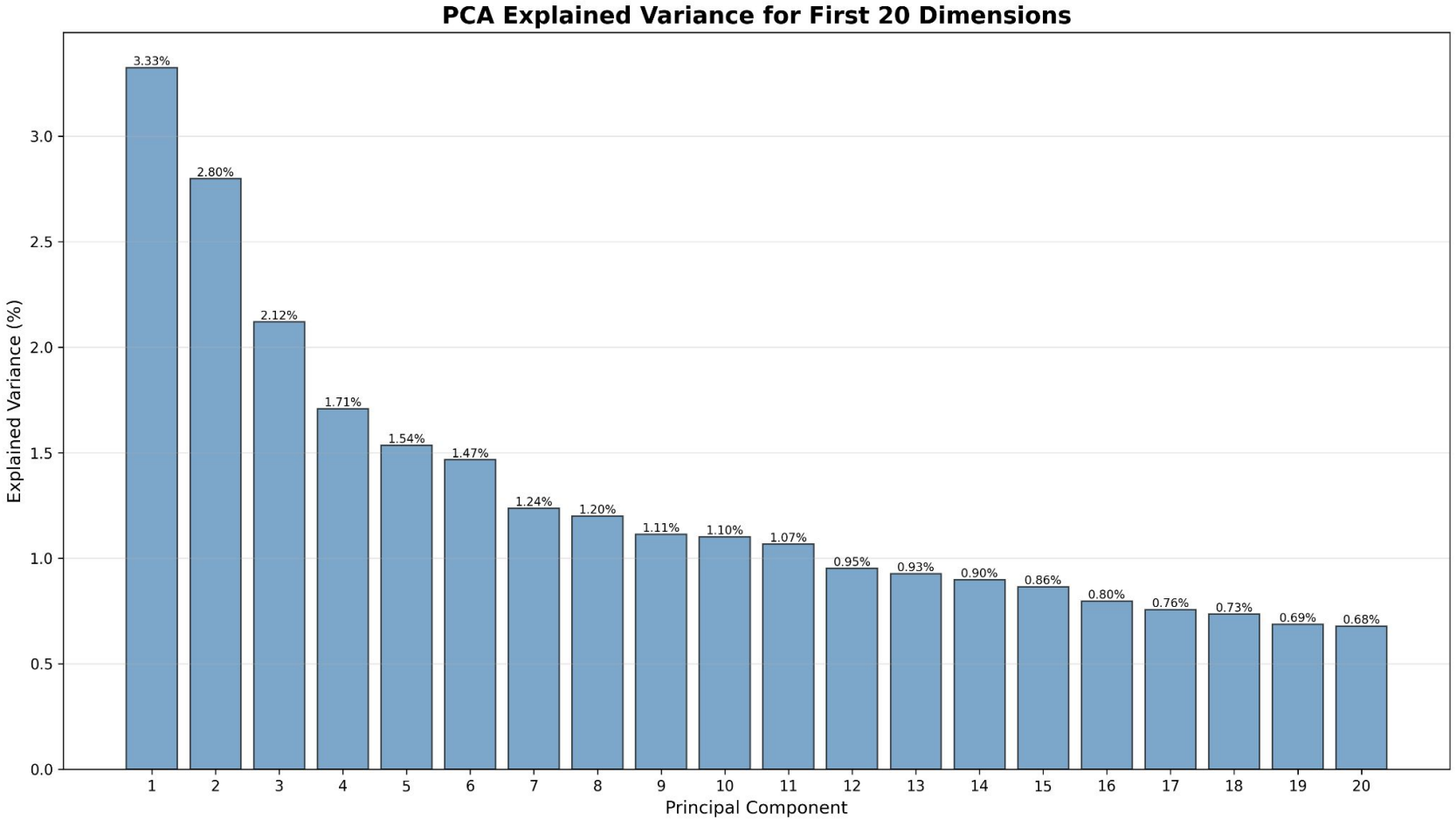
Method	Usage in embedding viz	Pros	Cons
PCA	Quick baseline, linear trends in movie space	Very fast; deterministic; globally interpretable axes (can relate PCs back to features or genres); robust and easy to tune.	Linear only; often fails to separate nuanced semantic clusters; can underperform on neighborhood preservation vs nonlinear methods.
t-SNE	“Island” plots to show tight clusters of similar movies	Excellent local neighborhood preservation; visually striking clusters; widely supported in tools and tutorials.	Poor global structure (distances between clusters are not meaningful); sensitive to hyperparameters and preprocessing; stochastic, slower and less scalable than UMAP; embedding new points post-hoc is awkward.
UMAP	Default nonlinear method for many embedding dashboards (including text)	Balances local and global structure better than t-SNE; faster and more scalable; can learn a transform to embed new movies; parameters more interpretable (n_neighbors, min_dist).	Still distorts global distances; results depend on parameter choice and preprocessing; stochastic and can create spurious small clusters or disconnected regions.
TriMAP	Emphasizing global layout of clusters	Designed to preserve global structure via triplet loss; can reveal coarse relationships between clusters better than t-SNE/UMAP in benchmarks.	Less widely implemented; can be less sharp locally than t-SNE/UMAP; some parameters to tune and less community intuition about defaults.
PaCMAP	Balanced local/global structure with robustness	Good local and global preservation in systematic evaluations; relatively robust to preprocessing and parameters; competitive runtime.	Newer method with fewer off-the-shelf integrations; behavior and tuning less familiar to practitioners.
PHATE	Smooth manifolds and progressions (e.g., narrative or temporal arcs)	Good at visualizing trajectories and continuous structures; used where “paths” between states are important.	Less standard for general text embedding plots; slower and more complex to configure than PCA/UMAP.
MDS	Classical similarity-preserving baseline	Conceptually simple distance-preserving objective; can be useful when pairwise similarity matrix is primary object.	Does not scale well; often inferior to modern manifold methods for high-dimensional embeddings; sensitive to noise.
Autoencoders	Custom, task-specific visualization mappings	Flexible; can be supervised or semi-supervised to align with labels (genres, ratings); can learn nonlinear projections and easily embed new points.	Requires training, tuning, and validation; risk of overfitting or learning artifacts; less standard metrics for visualization quality.

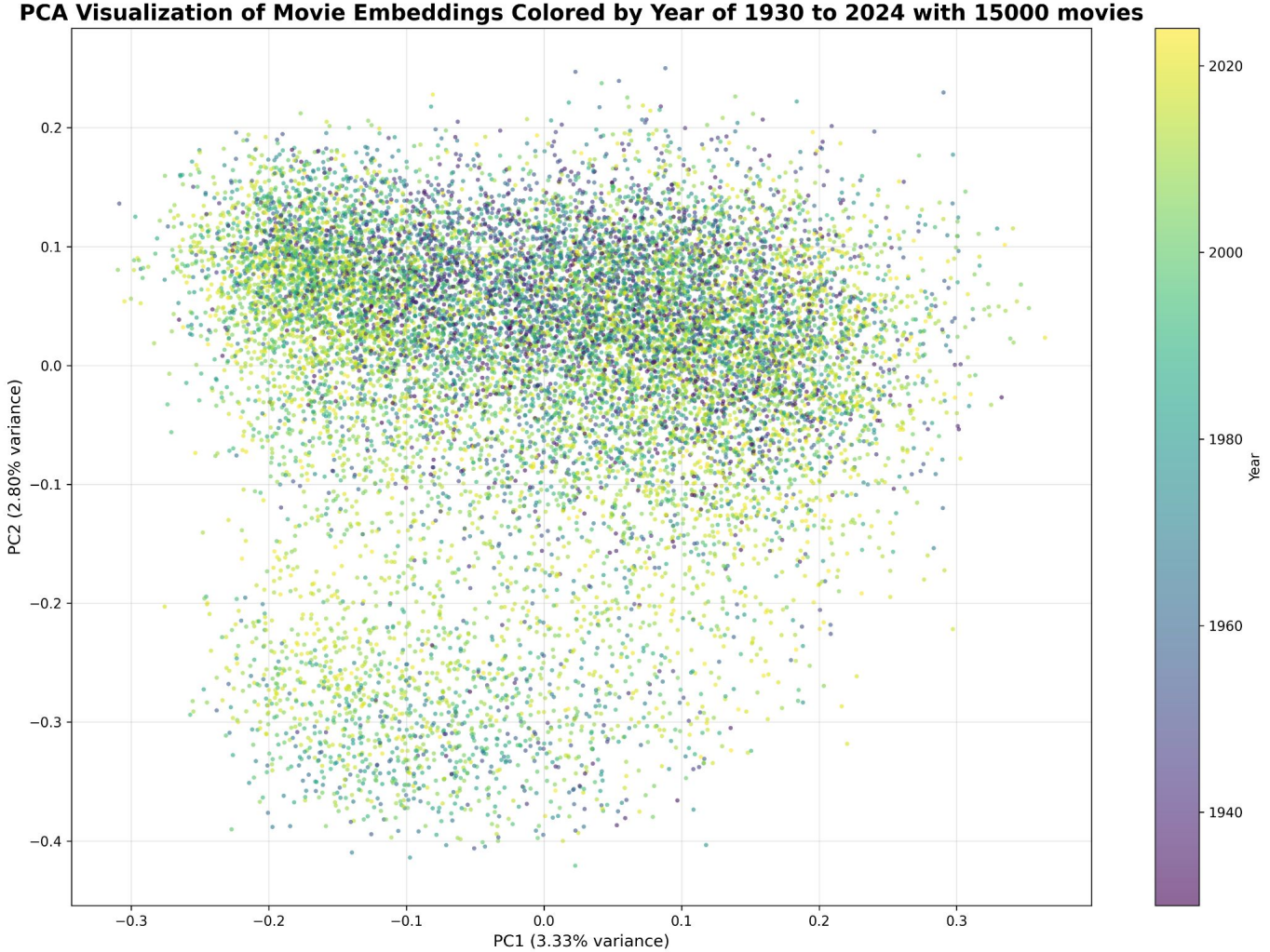
8	Fantasy & Sci-Fi	adaptation, alternate history, apocalyptic film, crossover fiction, dark fantasy, dieselpunk, dystopia, dystopian fiction, dystopian film, epic poem, fairy tale, fantasy, fantasy film, film adaptation, high fantasy, historical fiction, military science fiction, mythology, nonlinear narrative, post-apocalyptic fiction, post-apocalyptic film, psychological film, retrofuturism, science fantasy, science fiction, science fiction film, speculative fiction, speculative fiction film, steampunk, supernatural, supernatural fiction, supernatural film, technofantasy, urban fantasy	Q1213562, Q224989, Q47009776, Q21192427, Q200092,Q794912, Q2376899, Q180774, Q15062348, Q20443008,Q2484376, Q37484, Q699, Q132311, Q157394,Q188473,Q52162262,Q83440,Q859369, Q1257444, Q326439, Q1196408, Q904447, Q9134, Q2894685, Q197949, Q1341051, Q122360343, Q919036, Q1080374, Q24925, Q130232,Q1342372,Q471839, Q1782964,Q9326077, Q130232,Q2973181, Q223685, Q80837, Q7644030, Q2975633,Q3634883, Q20819599, Q1188977	34	6911
9	Family Youth & War	beach party film, buddy film, chick flick, children's film, coming-of-age film, family film, female buddy film, partisan film, social guidance film, social problem film, sport film, sports drama, submarine film, surf film, survival film, teen film, war film	Q4875794, Q663106, Q2075808, Q188473,Q2143665, Q102429885,Q15428604,Q18450913, Q1361932, Q5442753, Q3442060, Q7551149, Q7551315, Q1339864, Q117830868, Q1615638, Q7645884, Q15898171, Q1146335,Q130232, Q369747,Q645928	17	6513
10	Comedy	action comedy film, comedy, comedy drama, comedy film, comedy horror, comedy horror film, comedy of remarriage, comedy thriller, comics, commedia all'italiana, commedia sexy all'italiana, fantasy comedy, fantasy comedy film, farce, parody, parody film, science fiction comedy, screwball comedy film, sex comedy, sitcom, slapstick, slapstick film, stand-up comedy, tragicomedy, zombie comedy	Q1760610,Q2059703,Q2678111, Q40831, Q157394,Q859369, Q157443, Q224700, Q108466999,Q853630, Q5151497, Q16950433, Q1004, Q1115187, Q2991565, Q1159671,Q1637212, Q95440291, Q193979, Q170539, Q622548, Q761469, Q248583, Q2991560, Q170238, Q624771, Q17112331, Q145806, Q192881, Q1747837	25	19227
11	Adventure Music & Diverse	Eastern, Quinqui, adventure, adventure film, children's music, coming of age, gore, harem, jazz, martial art, mysticism, pop music, world music, yuri	Q1278107, Q3677206, Q1436734, Q147516,Q319221,Q471839,Q52207399, Q2389651, Q681737, Q1538137, Q690342, Q8341, Q11417, Q16861950, Q11399,Q37073, Q205049, Q320568	14	3557
12	Documentary Animation & Educational	Christian film, animated documentary, animated film, animation, collage film, concert film, dance film, educational film, interactive film, live action, live-action/animated film, medieval film, political film, rockumentary, screenlife, social film, television adaptation, television film	Q110324066,Q115122793, Q4765076, Q202866, Q11425, Q5145881, Q430525, Q3072043, Q596138, Q1635956, Q517386, Q25110269,Q7832972,Q870889, Q3327002, Q2973201, Q1800833, Q60753838, Q7551110, Q101716172, Q506240	18	773
13	Biographical	autobiography, biographical drama film, biographical film, biography, jukebox musical, slice of life	Q4184, Q104559206, Q2421031,Q645928, Q36279, Q643684, Q2561438	6	2502
14	Film Noir	French New Wave, film noir, neo-noir, noir fiction, queer film, tech noir	Q193541, Q185867, Q1494434,Q2421031,Q5897543, Q26221026, Q114079176,Q128149892, Q1894374	6	1194
15	Drama	docudrama, drama, drama film, drama television series, family drama, family drama film, fantasy drama, historical drama, historical drama film, historical film, history, legal drama, medical drama, melodrama, period drama film, period film, political drama, psychological drama, psychological drama film, war drama	Q622370, Q25372, Q130232,Q157394,Q369747,Q44342,Q471839,Q8434,Q959790, Q1366112, Q29621749, Q65224847, Q110024700, Q7168625, Q114413232,Q116514801,Q188473, Q113485322,Q16976178,Q17013749,Q474090,Q603291, Q309, Q643873, Q1786567, Q191489,Q1919632, Q542475,Q959790, Q116513497, Q7210294, Q108102008,	21	33391

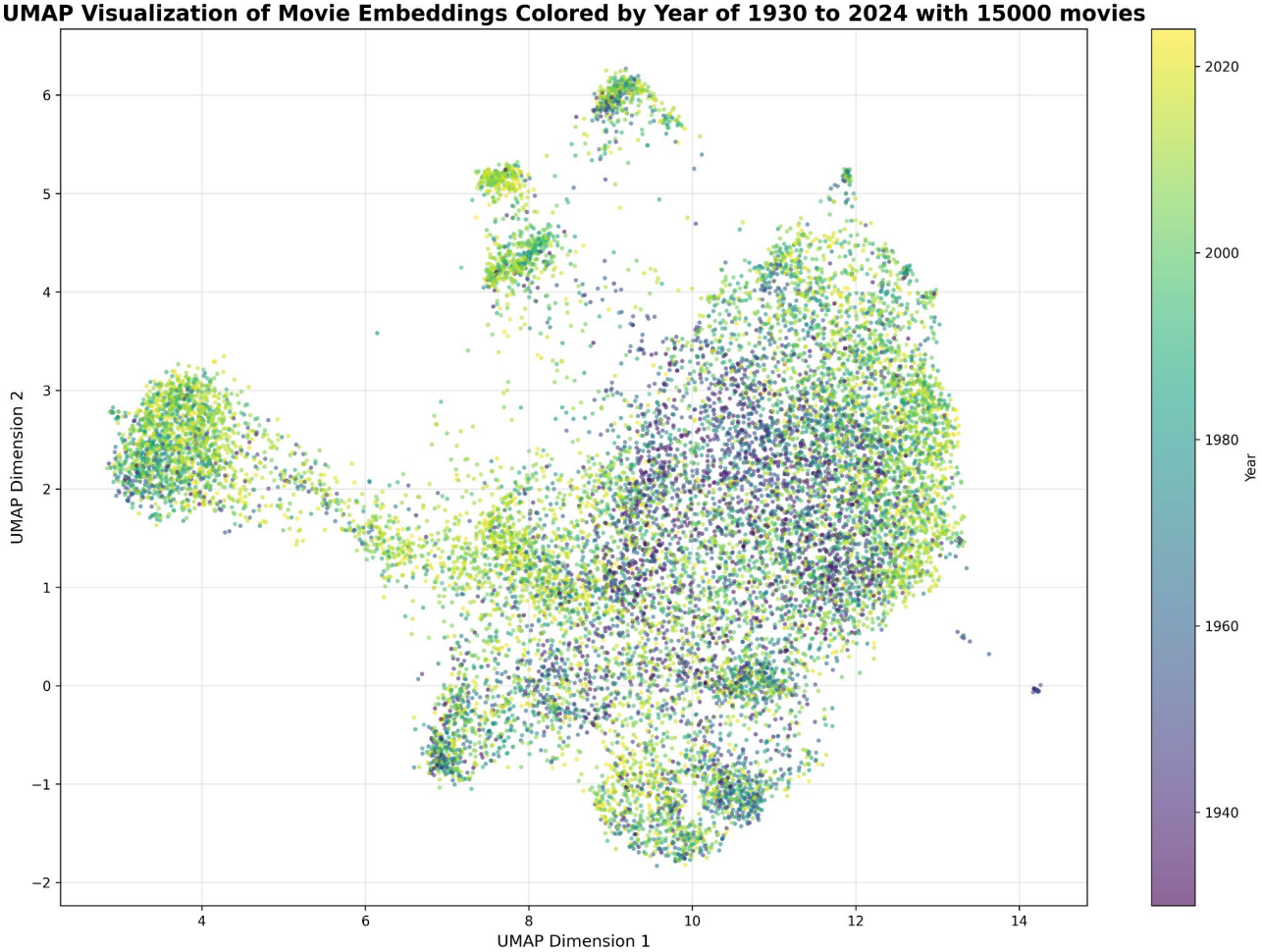
Literature Review - Dimensionality Reduction

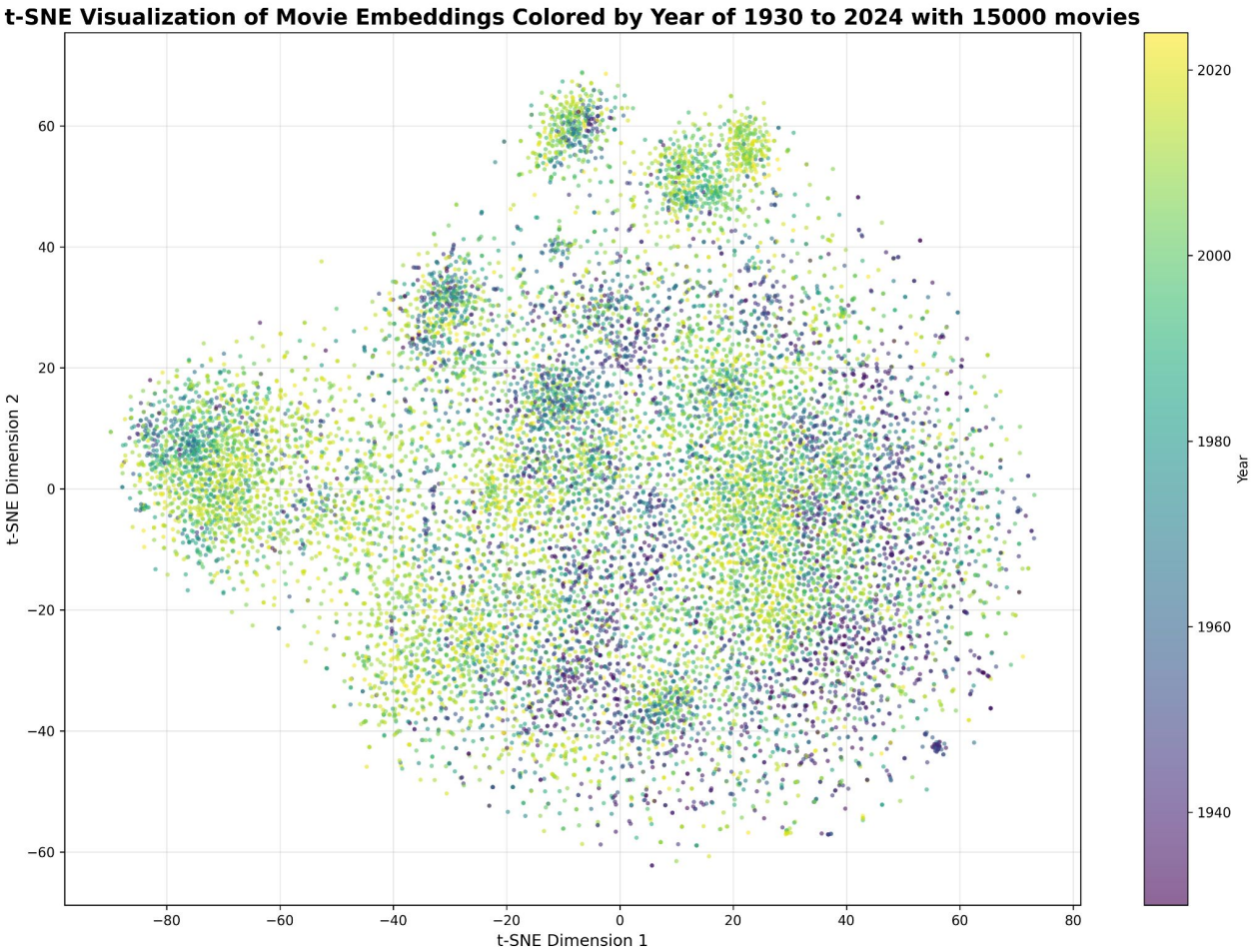
We want DR only for visualization

- Must be **sensitive** to shifts (e.g. action genre in 1940 vs 2025)
- Must be deterministic + global
  - t-SNE only shows relative distances, discards global structure
  - cannot compare 2 t-SNEs point to point


PCA Explained Variance for First 20 Dimensions
Explained Variance (%)
Principal Component
3.33%
2.80%
2.12%
1.71%
1.54%
1.47%
1.24%
1.20%
1.11%
1.10%
1.07%
0.95%
0.93%
0.90%
0.86%
0.80%
0.76%
0.73%
0.69%
0.68%


PCA Visualization of Movie Embeddings Colored by Year of 1930 to 2024 with 15000 movies
PC2 (2.80% variance)
PC1 (3.33% variance)
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Year
2020
2000
1980
1960
1940


UMAP Visualization of Movie Embeddings Colored by Year of 1930 to 2024 with 15000 movies
UMAP Dimension 2
UMAP Dimension 1
Year
6
5
4
3
2
1
0
−1
−2
4
6
8
10
12
14
2020
2000
1980
1960
1940


t-SNE Visualization of Movie Embeddings Colored by Year of 1930 to 2024 with 15000 movies
t-SNE Dimension 1
t-SNE Dimension 2
Year
2020
2000
1980
1960
1940
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60
80

Literature Review - Dimensionality Reduction

Method	distance/shift comparisons	Caveats
PaCMAP	Explicitly designed to preserve both global and local structure; ranks among the best for global-structure and triplet-based metrics while remaining robust to parameters and preprocessing. [1] [2] [3]	- Newer, fewer canonical settings - must fix hyperparameters and initialization across time slices.
TriMAP	Triplet-based loss prioritizes correct relative placement of clusters and often beats t-SNE/UMAP on global-structure scores; inter-cluster distances tend to be more meaningful. [4] [5]	- Less standard tooling - may show slightly weaker very-local structure than t-SNE.
PCA (linear)	Strong global structure and distance preservation in expectation; very stable, deterministic, and easy to apply jointly to all years so shifts are directly comparable. [6] [7]	- Linear only - may under-separate nonlinear genre/theme manifolds, so shifts may look subtle.
Classical MDS	Explicitly minimizes distortion of pairwise distances, so global geometry and between-cluster spacing are preserved as well as possible in 2D. [8] [9] [10]	- Can be slower and less scalable - like PCA, may not reveal fine nonlinear substructure.

Literature Review - Dimensionality Reduction

A quantifiable way to evaluate which DR method is good for us:

- Literature review
- Use genre/years as label + KNN for “classification”

Literature Review - “Average” embeddings

Mean has limitations:

- Sensitive to outliers
- Uses euclidean distance (we have decided on cosine distance for analysis)
- Not interpretable

Source: trust me bro

Literature Review - “Average” embeddings

Alternative 1: Geometric mean

- Robust to outliers in high-dim [1] [2]
- Works in any dimension and with any distance metric [3]
- Significantly improves clustering quality and robustness [1]

Literature Review - “Average” embeddings

Alternative 2: Medoids

- The most average entry in the dataset
- Interpretable (most average movie)
- Can use any distance metric (cosine distance)

Literature Review - “Average” embeddings

Alternative 3: Spherical K-means

- The most average entry in the dataset
- Interpretable (most average movie)
- Can use any distance metric (cosine distance)
- Basically K means but with different distance metric

Literature Review - Methods

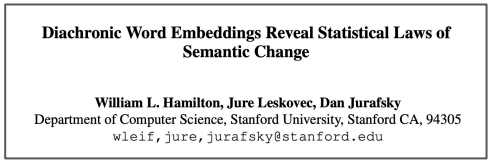
Core methodologies of temporal embeddings

1. Alignment based approaches
2. Probabilistic models
3. Contextualized and Transformer based models (modern standard)
4. Super advanced stuff

Literature Review - Alignment based Approaches

Analyse how word meaning changes over time

- Fit Word2Vec on the same word
- Different time periods
- **Align dimension** (orthogonal Procrustes)
- **Cosine Distance**



Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

William L. Hamilton, Jure Leskovec, Dan Jurafsky
Department of Computer Science, Stanford University, Stanford CA, 94305
wleif,jure,jurafsky@stanford.edu

Literature Review - Alignment based Approaches

Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

William L. Hamilton, Jure Leskovec, Dan Jurafsky
Department of Computer Science, Stanford University, Stanford CA, 94305
wleif,jure,jurafsky@stanford.edu

Need to find ways to "normalize"

embeddings over the years

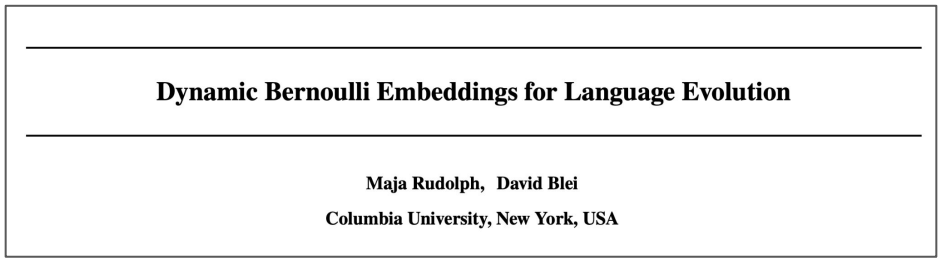
Word	Moving towards	Moving away	Shift start	Source
gay	homosexual, lesbian	happy, showy	ca 1950	(Kulkarni et al., 2014)
fatal	illness, lethal	fate, inevitable	<1800	(Jatowt and Duh, 2014)
awful	disgusting, mess	impressive, majestic	<1800	(Simpson et al., 1989)
nice	pleasant, lovely	refined, dainty	ca 1890	(Wijaya and Yeniterzi, 2011)
broadcast	transmit, radio	scatter, seed	ca 1920	(Jeffers and Lehiste, 1979)
monitor	display, screen	—	ca 1930	(Simpson et al., 1989)
record	tape, album	—	ca 1920	(Kulkarni et al., 2014)
guy	fellow, man	—	ca 1850	(Wijaya and Yeniterzi, 2011)
call	phone, message	—	ca 1890	(Simpson et al., 1989)

Table 2: Set of attested historical shifts used to evaluate the methods. The examples are taken from previous works on semantic change and from the Oxford English Dictionary (OED), e.g. using 'obsolete' tags. The shift start points were estimated using attestation dates in the OED. The first six examples are words that shifted dramatically in meaning while the remaining four are words that acquired new meanings (while potentially also keeping their old ones).

Literature Review - Probabilistic models

Dynamic Embeddings

- Capture how meaning of words change over time
- Random walk + Bernoulli thingy



Dynamic Bernoulli Embeddings for Language Evolution

Maja Rudolph, David Blei

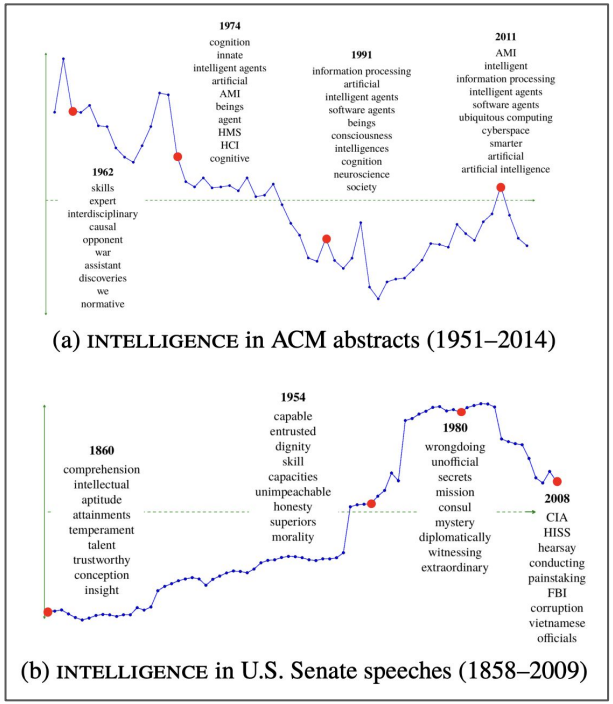
Columbia University, New York, USA

Literature Review - Probabilistic models

Plots

- 1D PCA against Time

words with largest drift (Senate)			
IRAQ	3.09	coin	2.39
tax cuts	2.84	social security	2.38
health care	2.62	FINE	2.38
energy	2.55	signal	2.38
medicare	2.55	program	2.36
DISCIPLINE	2.44	moves	2.35
text	2.41	credit	2.34
VALUES	2.40	UNEMPLOYMENT	2.34



COMPUTER (Senate)	
1858	1986
computer	computer
draftsman	software
draftsmen	computers
copyist	copyright
photographer	technological
computers	innovation
copyists	mechanical
janitor	hardware
accountant	technologies
bookkeeper	vehicles

BUSH (Senate)	
1858	1990
bush	bush
barberry	cheney
rust	nonsense
bushes	nixon
borer	reagan
eradication	george
grasshoppers	headed
cancer	criticized
tick	clinton
eradicate	blindness

DATA (ACM)				
1961	1969	1991	2011	2014
data	data	data	data	data
directories	repositories	voluminous	raw data	data streams
files	voluminous	raw data	voluminous	voluminous
bibliographic	lineage	repositories	data sources	raw data
formatted	metadata	data streams	data streams	warehouses
retrieval	snapshots	data sources	dws	dws
publishing	data streams	volumes	repositories	repositories
archival	raw data	dws	warehouses	data sources
archives	cleansing	dsms	marts	data mining
manuscripts	data mining	data access	volumes	marts

Potential research questions

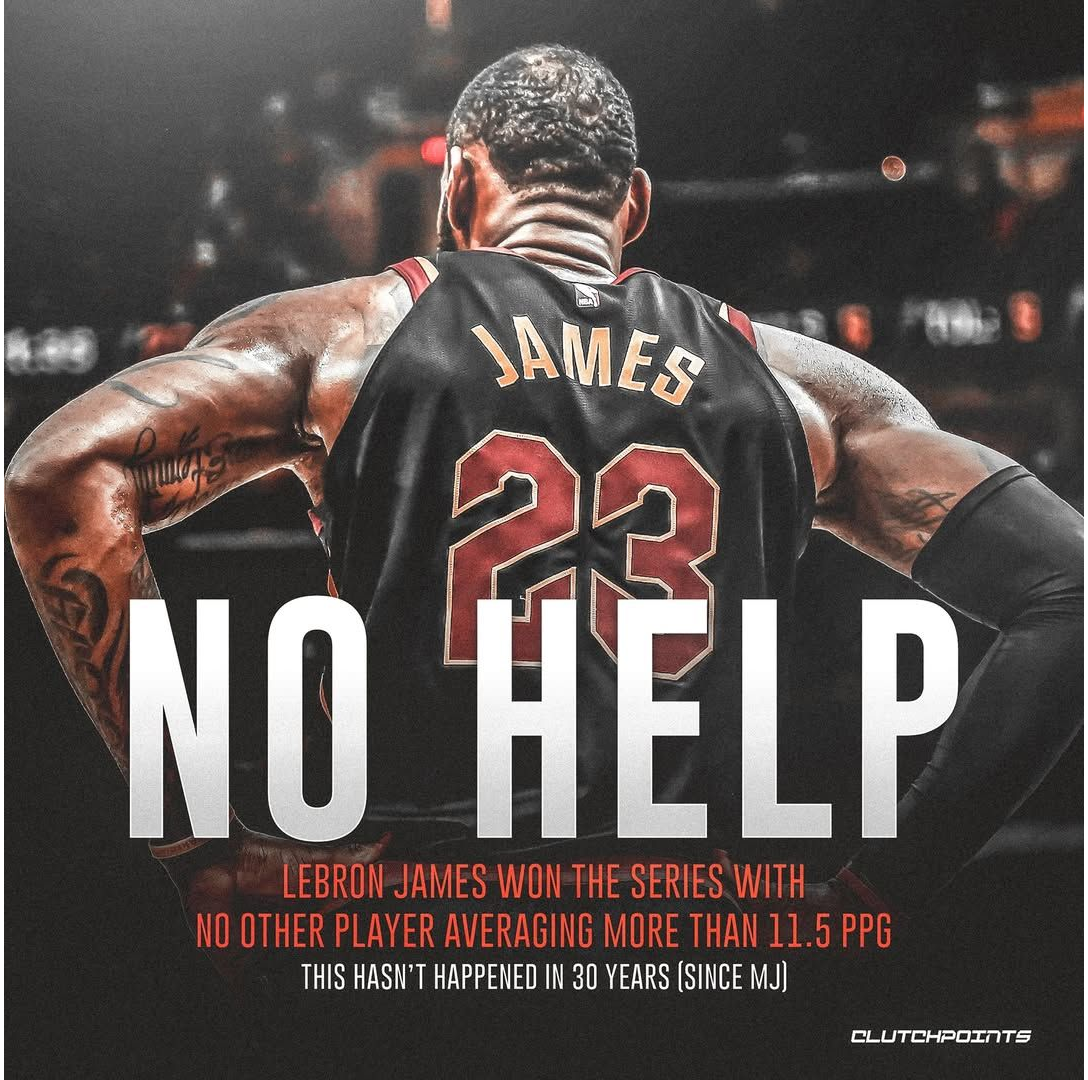
- Overall theme: Movie plots temporal analysis
  - Have genres moved together closer?
  - Are there shift towards less novel films?
  - What movies are trend setters, once a film appears over the next years many more appear…
  - Are director always doing the same kind of movies?

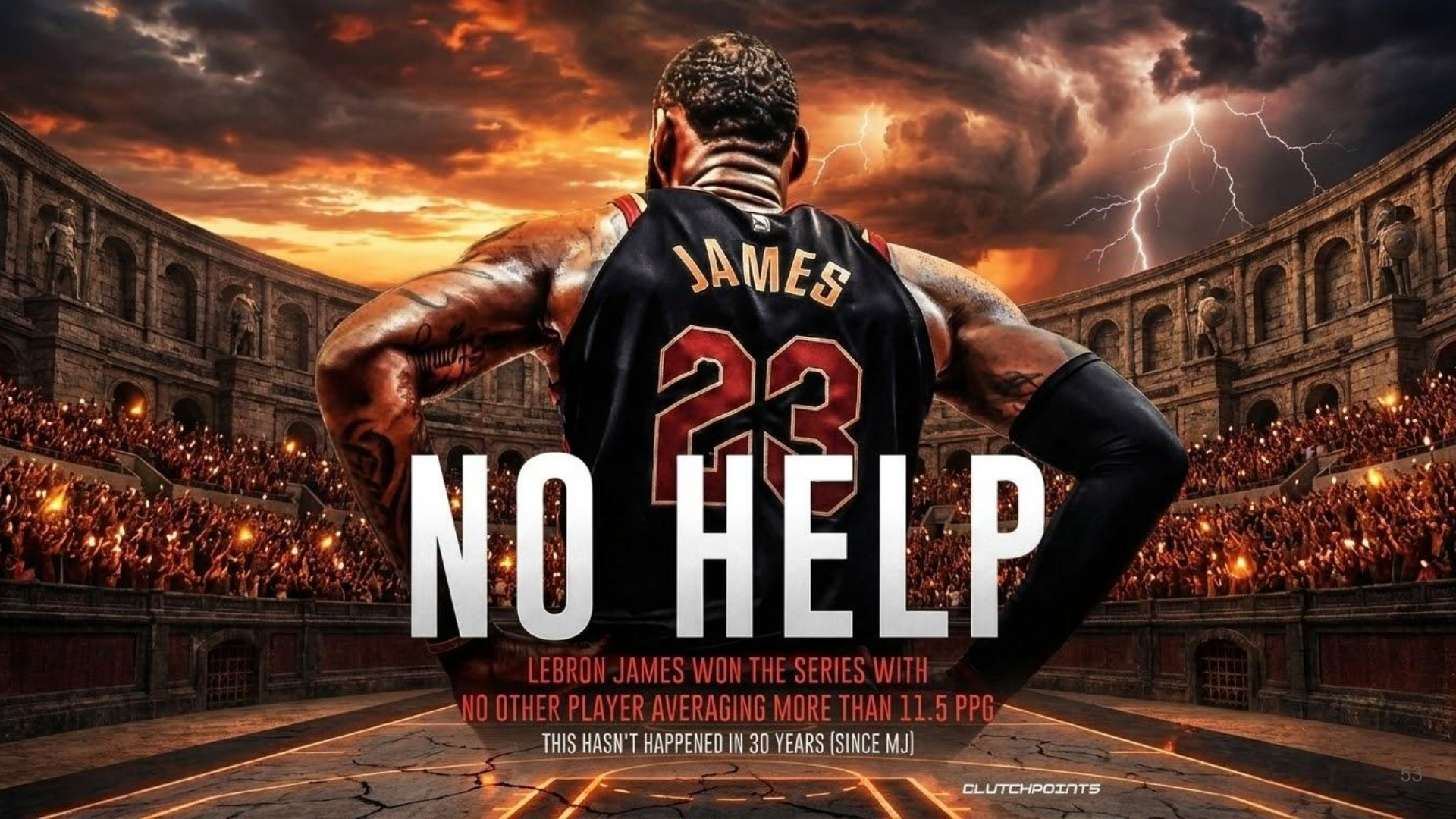
Literature Review - Transformer based models

- BiTimeBERT
- Changing pretraining
- MLM includes masking of time related tokens
- More of pretraining + finetuning tasks
- **Not really relevant**

Questions?

We're here to answer your queries.


JAMES
23
NO HELP
LEBRON JAMES WON THE SERIES WITH
NO OTHER PLAYER AVERAGING MORE THAN 11.5 PPG
THIS HASN'T HAPPENED IN 30 YEARS (SINCE MJ)
CLUTCHPOINTS


JAMES
23
NO HELP
LEBRON JAMES WON THE SERIES WITH
NO OTHER PLAYER AVERAGING MORE THAN 11.5 PPG
THIS HASN'T HAPPENED IN 30 YEARS (SINCE MJ)
CLUTCHPOINTS

TO-DO

- Define a cut off for text length → Bartol
- Have a full literature review → Ansel, Niklas
  - Dimension reduction methods
    - PCA/t-SNE
  - Latent space comparison methods → Måřtĩŋ Łœpëż Đę Įpĭñá
    - Average embeddings
    - Velocity
    - Acceleration
    - Distances/Similarity
  - How to use tabular data (not just embeddings)
  - See how we can adapt lecture methods to our own (no need to force it)
- Similar movies together ? → rough idea → Niklas
- Genre classification → Niklas
- Research questions → make slides → everyone writes the slides

Group 3

Status Update 10.12.2025

Outline

- Concept Word Extraction based on Embeddings
- PCA Dimension Concept Extraction
- Genre analysis failures
- Movie novelty analysis

Concept Word Extraction based on Embeddings 1/3

- Dense:
  - Input: Dense embedding vector (numpy array)
  - Process:
    - Normalizes the embedding
    - Computes cosine similarity with all concept embeddings
    - Returns top K by similarity
  - Requires: Pre-computed dense embedding
  - No text processing needed
- Sparse:
  - Input: Raw text + lexical weights dictionary
  - Process:
    - Projects lexical weights onto
    - Filters by Zipf frequency
    - Maps lemmas to concepts
  - Requires: Text, tokenizer, spaCy nlp model, lexical weights

Concept Word Extraction based on Embeddings 2/3

The narrative follows E.S., a Palestinian returning from a twelve-year exile in New York to a homeland that now feels alien and disjointed. Structured as a series of loosely connected vignettes rather than a conventional plot, the film depicts everyday life in Nazareth and Jerusalem with a blend of absurdist humor and political unease. The early “Nazareth Personal Diary” section portrays his family, local shopkeepers, and fishermen, revealing a quiet monotony shaped by statelessness and dwindling tourism. A brief middle interlude shows the frustrations of cultural expression, exemplified by a failed attempt to speak at a filmmaking conference. The final “Jerusalem Political Diary” adopts a sharper ideological tone, highlighting discrimination, paranoia, and resistance before ending with a somber scene of E.S.’s parents asleep as Israeli television broadcasts continue in the background.

An 18-year-old Hasidic woman, Shira Mendelman, is preparing for an arranged marriage she genuinely welcomes when her older sister dies in childbirth, disrupting the family’s plans. As Shira becomes increasingly involved in caring for her nephew, her mother proposes that Shira marry the widowed brother-in-law, Yochay, to keep the baby within the family. Both initially resist, and Shira’s uncertainty deepens after learning her sister might have wished for someone else to marry Yochay. When Yochay decides to move abroad and remarry, Shira reluctantly agrees to the engagement under family pressure, but a rabbi halts the match due to her lack of conviction. After grappling with grief, duty, and desire, Shira recognizes her true feelings and chooses to be with Yochay and his child, culminating in their marriage.

Concept Word Extraction based on Embeddings 3/3

Rank	Dense Concept	Dense Score	Sparse Concept	Sparse Score
1	arguing	0.544049	cinema	0.462526
2	filming	0.531092	sectional	0.327910
3	inside	0.525644	tourist	0.403241
4	socializing	0.523859	return	0.457992
5	reliving	0.523364	characteristic	0.362480
6	yelling	0.521545	female	0.425856
7	shouting	0.521472	police	0.416033
8	pretending	0.520833	feelings	0.471460
9	communicating	0.518198	shop	0.408851
10	someone	0.516860	friend	0.441347
11	backstage	0.516742	slut	0.410002
12	happening	0.515222	humor	0.433643
13	humans	0.513191	hombre	0.442704
14	halftime	0.513142	search	0.410907
15	altercation	0.512606	terrorist	0.422032
16	chattering	0.511847	tranquillity	0.469650
17	talks	0.511800	routine	0.414639
18	sweating	0.511625	discourse	0.422368
19	roleplaying	0.509357	life	0.427994
20	sneezing	0.509242	television	0.454755
21	sit-in	0.508355	scenery	0.469170
22	check-in	0.507920	light	0.465169
23	daydreaming	0.507528	period	0.374559
24	hurrying	0.506722	parent	0.362549
25	somebody	0.506603	election	0.382254
26	pleading	0.506445	theatre	0.446549
27	anecdote	0.506024	relation	0.410553
28	sit-down	0.505426	fashion	0.347873
29	haggling	0.504629	boat	0.406669
30	playback	0.504442	image	0.388053

Rank	Dense Concept	Dense Score	Sparse Concept	Sparse Score
1	mitzvah	0.526642	marriage	0.519790
2	marriage	0.519790	engagement	0.379622
3	yeshiva	0.505100	widow	0.456806
4	filming	0.494214	delay	0.414525
5	jeremiah	0.486933	husband	0.430518
6	bride	0.484931	times	0.355075
7	pleading	0.484276	death	0.414765
8	torah	0.483665	father	0.353841
9	wedding	0.482318	proposal	0.388070
10	shariah	0.481167	family	0.420231
11	shia	0.479988	mother	0.410383
12	married	0.479422	cinema	0.432136
13	happening	0.478301	wedding	0.482318
14	sharia	0.477847	girl	0.437378
15	begging	0.476959	house	0.399123
16	crying	0.475215	tragedy	0.443515
17	jeopardy	0.475045	dream	0.347910
18	bridesmaid	0.473968	sister	0.447530
19	birthing	0.472761	baby	0.414906
20	jew	0.471533	scenario	0.406843
21	wishing	0.471087	offer	0.400715
22	yemeni	0.470966	friend	0.411713
23	jealousy	0.470875	hombre	0.382067
24	someone	0.469350	life	0.391976
25	resolving	0.469068	son	0.328957
26	pregnancy	0.468715	year	0.353359
27	granddaughter	0.468101	possibility	0.447583
28	cheating	0.467979	prospect	0.394781
29	midwife	0.467928	country	0.341517
30	waiting	0.467740	idea	0.388355

PCA Analysis Results 1/2

Principal Component	Explained Variance	Explained Variance (%)	Eigenvalue
PC1	0.0353	3.53%	0.0183
PC2	0.0321	3.21%	0.0167
PC3	0.0208	2.08%	0.0108
PC4	0.0171	1.71%	0.0089
PC5	0.0154	1.54%	0.0080
	Total explained variance (first 10 PCs):		**0.1808 (18.08%)**

Table 1: Explained variance and eigenvalues for the first five principal components.

Top 10 selected concepts (highest cosine similarity):

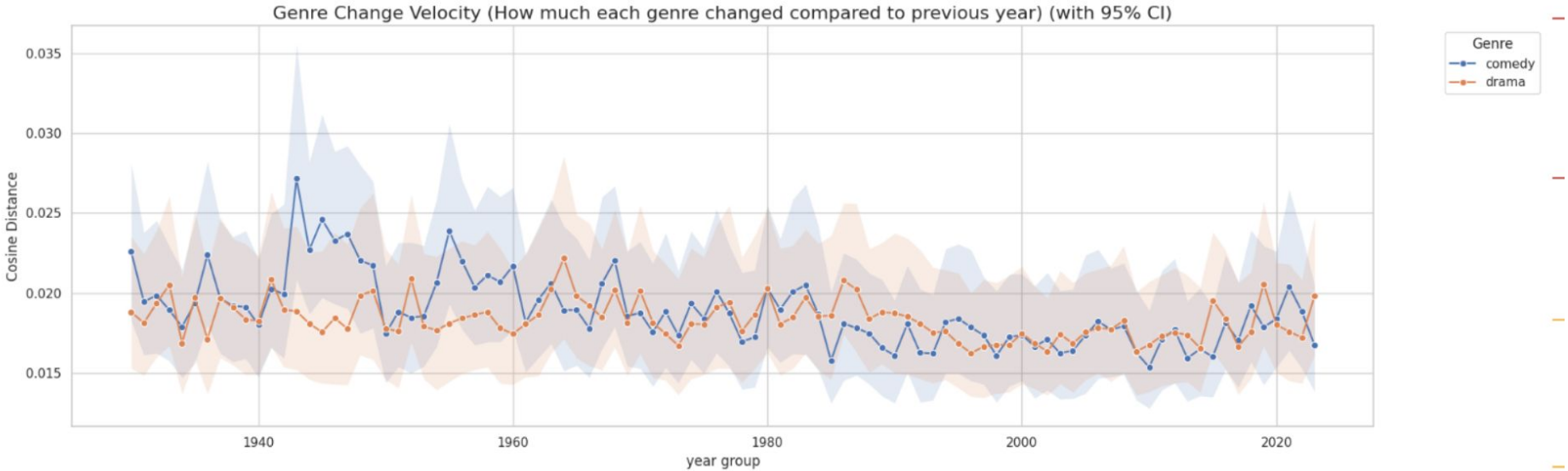
Rank	Concept	Similarity
1	someone	0.6816
2	inside	0.6798
3	somebody	0.6705
4	humans	0.6684
5	jealousy	0.6483
6	folks	0.6480
7	desperate	0.6443
8	face-off	0.6414
9	robbery	0.6413
10	stranger	0.6388

PCA Analysis Results 2/2

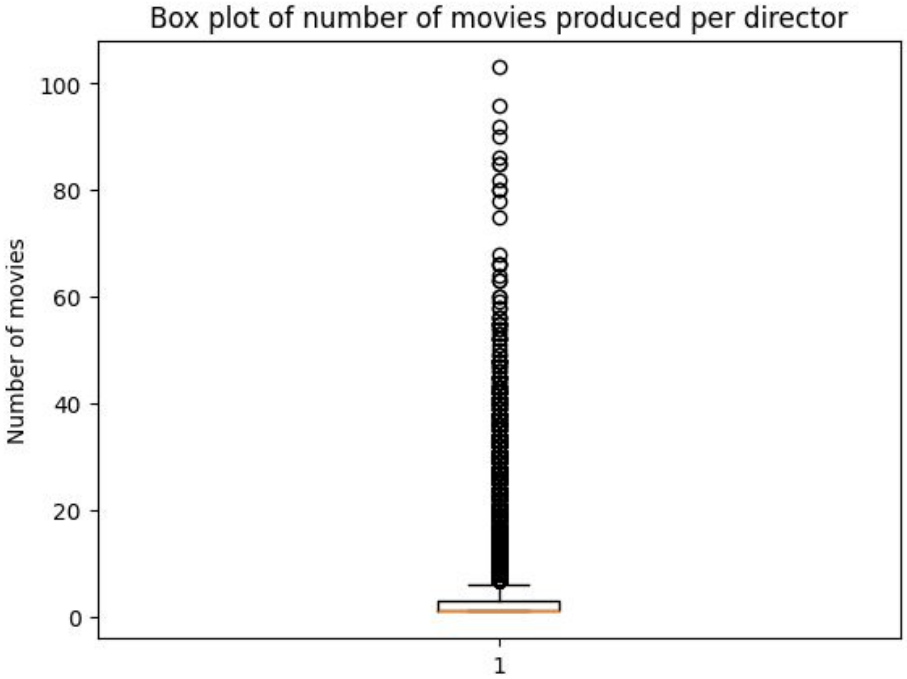
PC	Top 5 Positive Concepts		Top 5 Negative Concepts	
PC1	newark	0.1924	mantra	-0.1340
	america	0.1851	bollywood	-0.1015
	york	0.1850	love-song	-0.0783
	american	0.1848	karma	-0.0695
	hurricane	0.1845	love	-0.0677
PC2	suicide	0.2036	contemporaries	-0.0860
	death	0.1973	professionalism	-0.0819
	snake	0.1951	mid-sixties	-0.0802
	stupid	0.1911	contemporary	-0.0765
	raccoon	0.1780	mid-twenties	-0.0755
PC3	combat	0.1720	fiancee	-0.2790
	weaponry	0.1616	girlfriend	-0.2664
	samurai	0.1550	bride-to-be	-0.2624
	bombardment	0.1422	fiance	-0.2505
	wartime	0.1395	bride	-0.2503
PC4	infancy	0.2108	gangster	-0.2546
	childhood	0.1634	actor	-0.2402
	autumn	0.1632	gunman	-0.2182
	orphanage	0.1621	bollywood	-0.2137
	upbringing	0.1575	murderer	-0.1924
PC5	vocalist	0.1941	crime	-0.2339
	motherhood	0.1816	criminal	-0.2015
	actress	0.1788	extortion	-0.1916
	teens	0.1653	bribery	-0.1876
	co-ed	0.1635	businessman	-0.1833

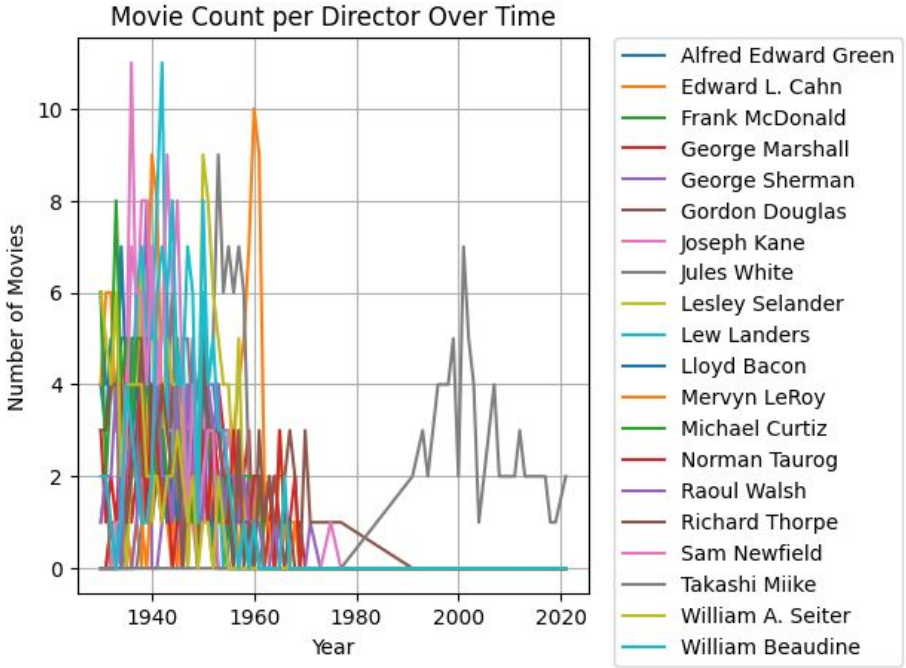
Only very vague ideas conferred in the PCA Dimensions. Not highly relevant not the main focus, would have been nice to interpret those results a bit more.

Genre drift analysis (same group size)



Directors analysis

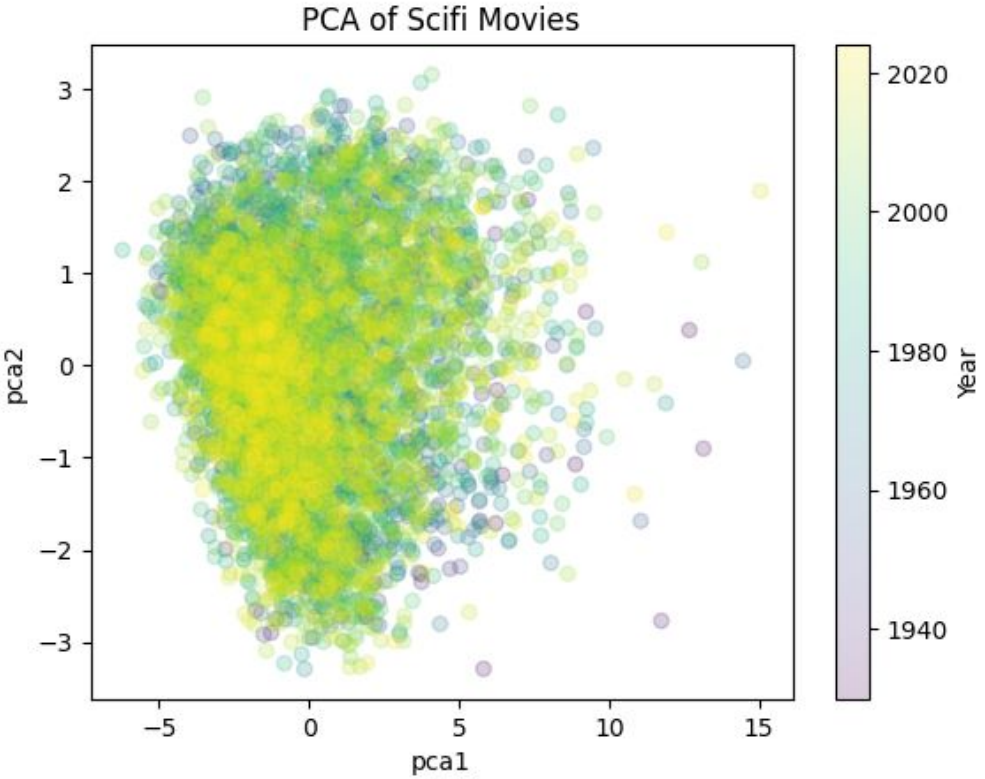


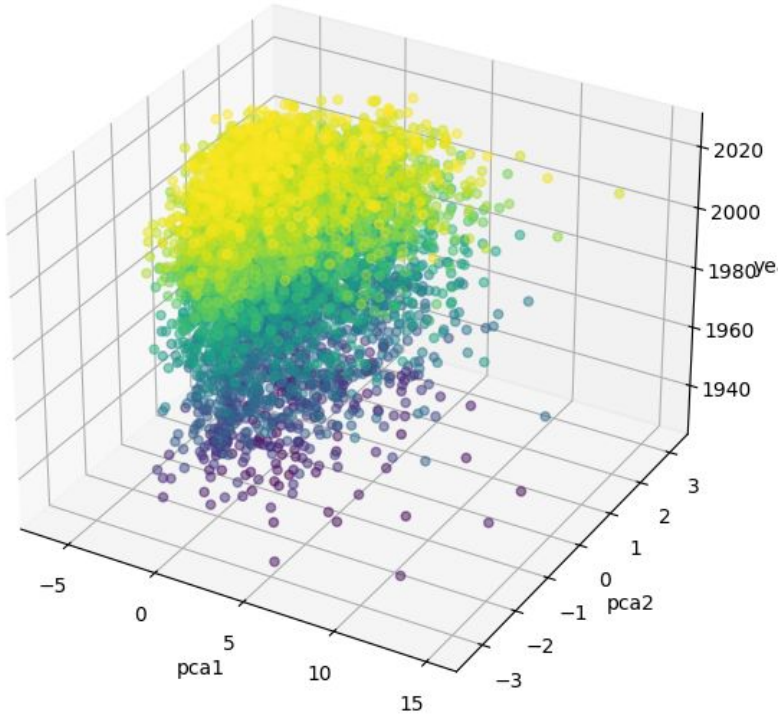


Last modification…

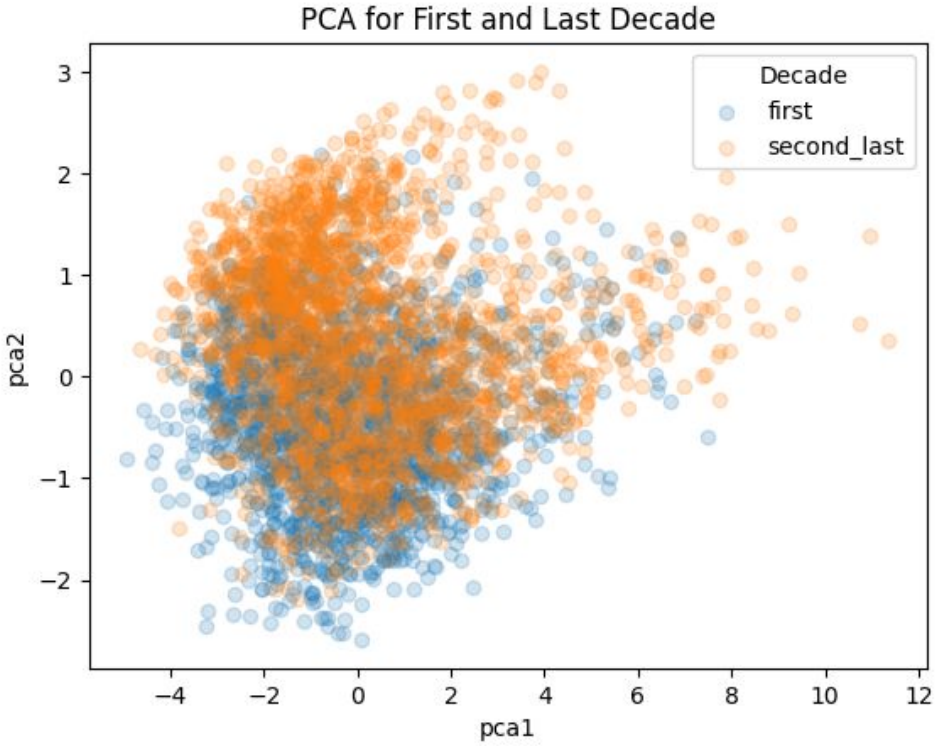
	vote_average	vote_count
count	87239.000000	87239.000000
mean	5.234452	261.671535
std	2.159972	1267.080997
min	0.000000	0.000000
25%	4.700000	3.000000
50%	5.800000	14.000000
75%	6.600000	69.000000
max	10.000000	38267.000000

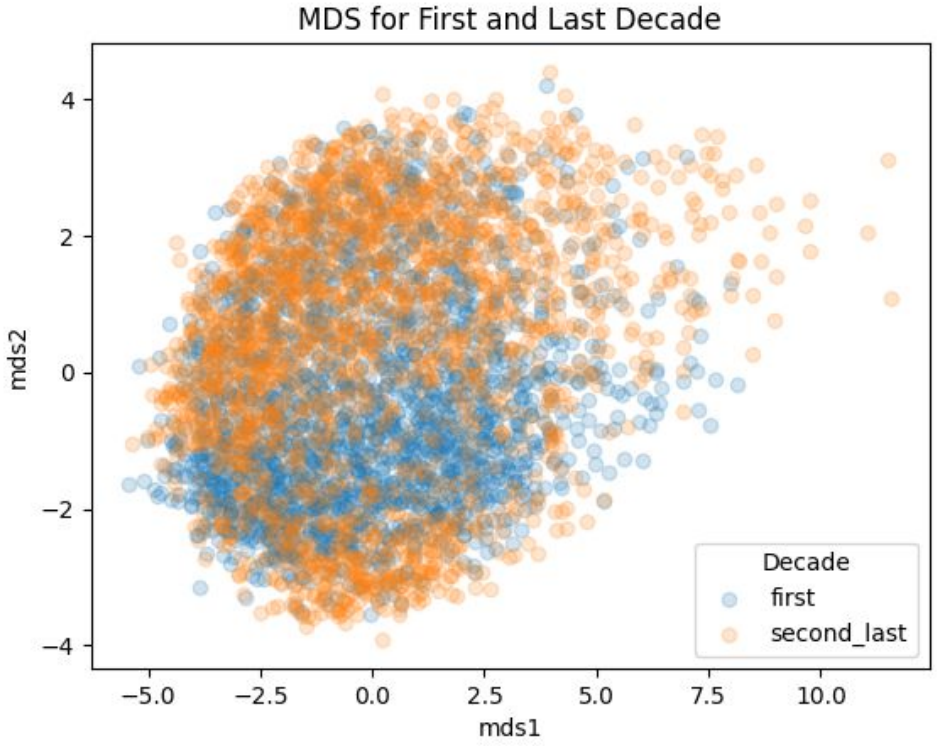
Scifi Genre Movie Analysis



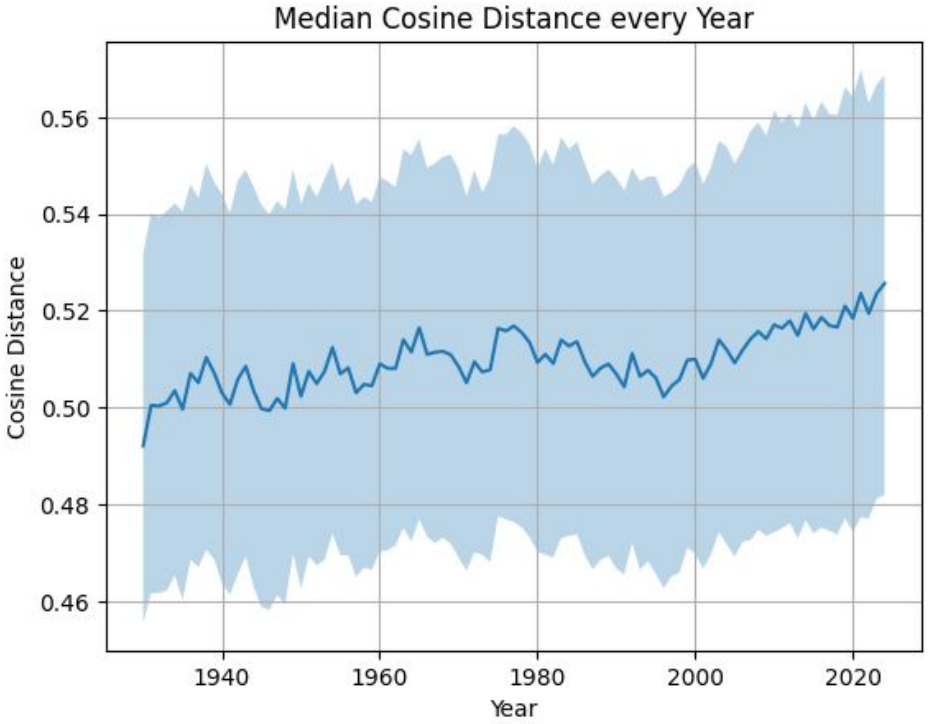


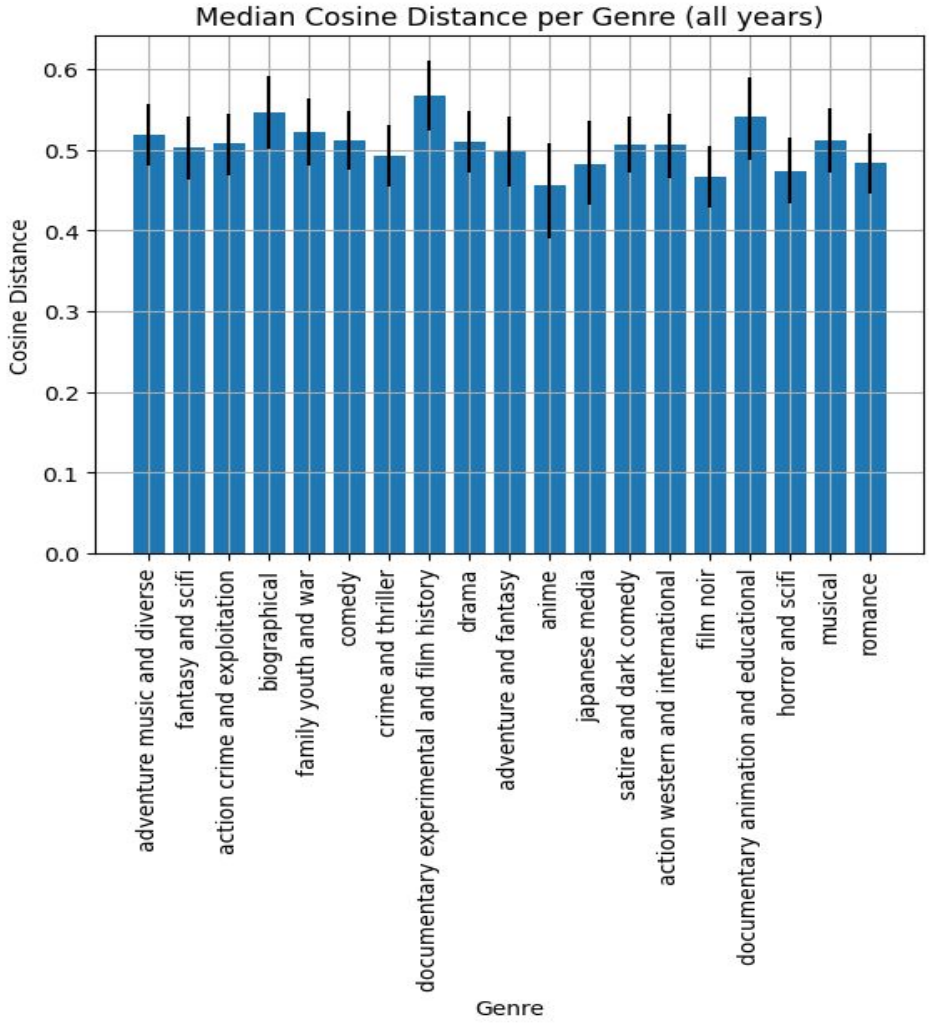
Decade Analysis



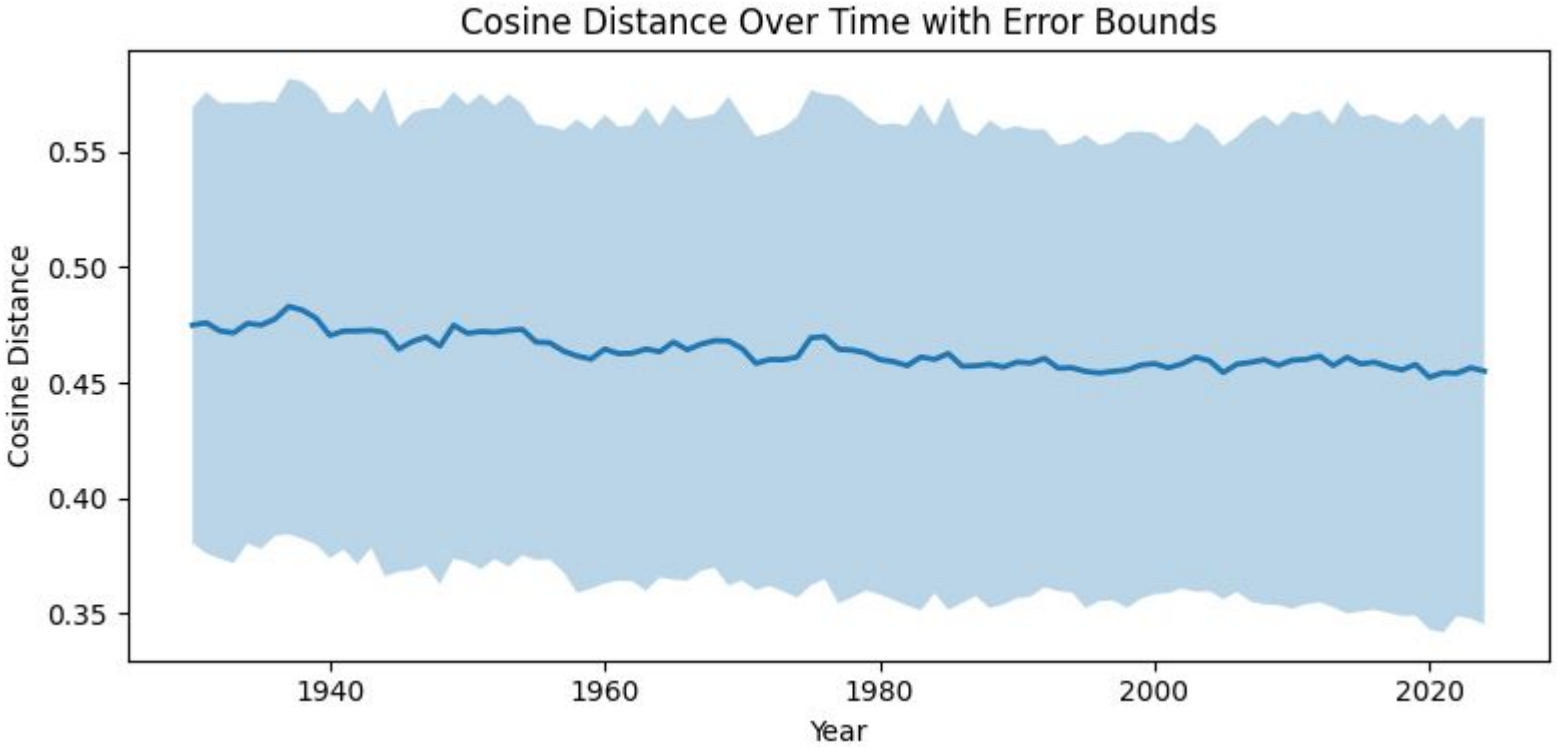


Intra Group Cosine Distance

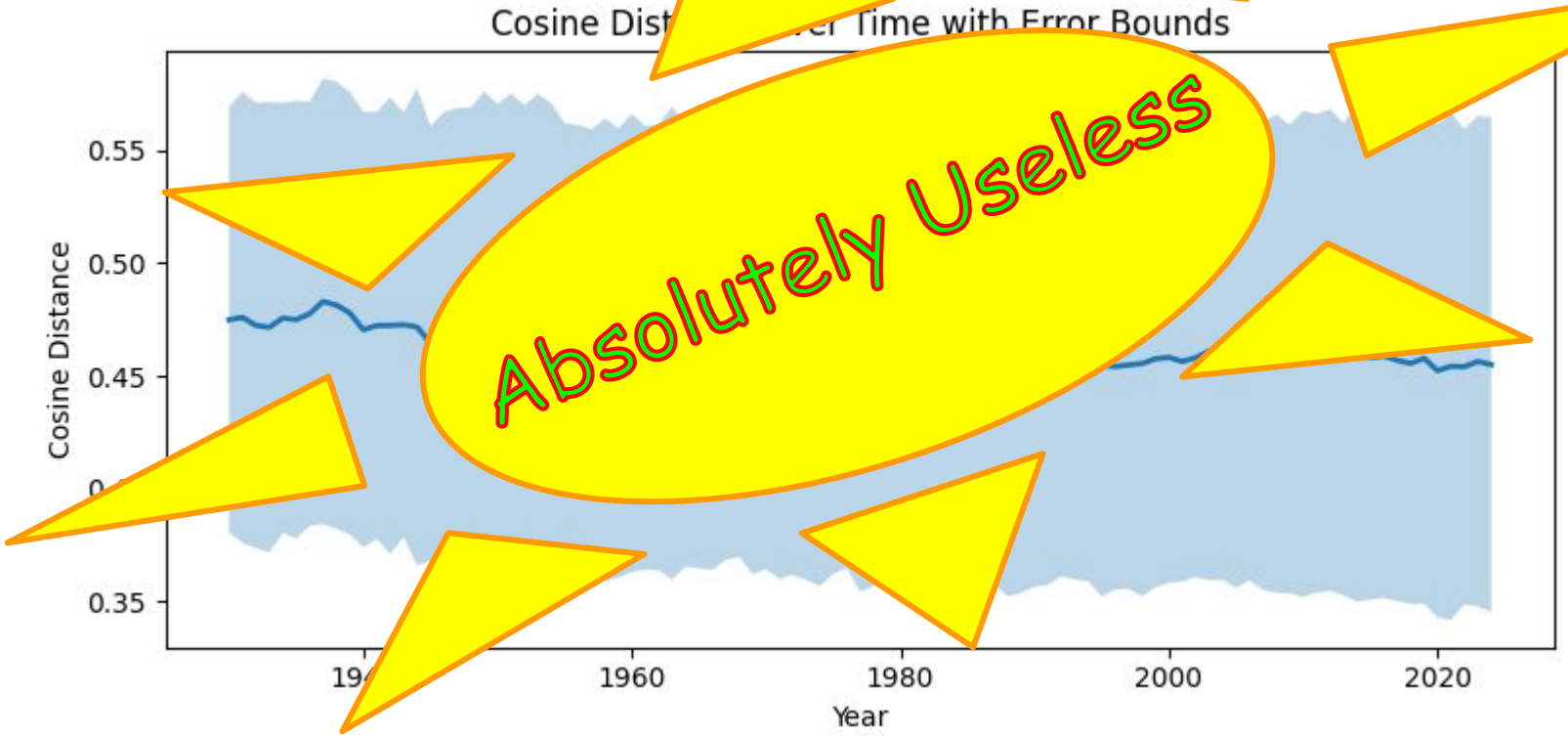




Cosine Distance to Anchor Movie (2024 Medoid)

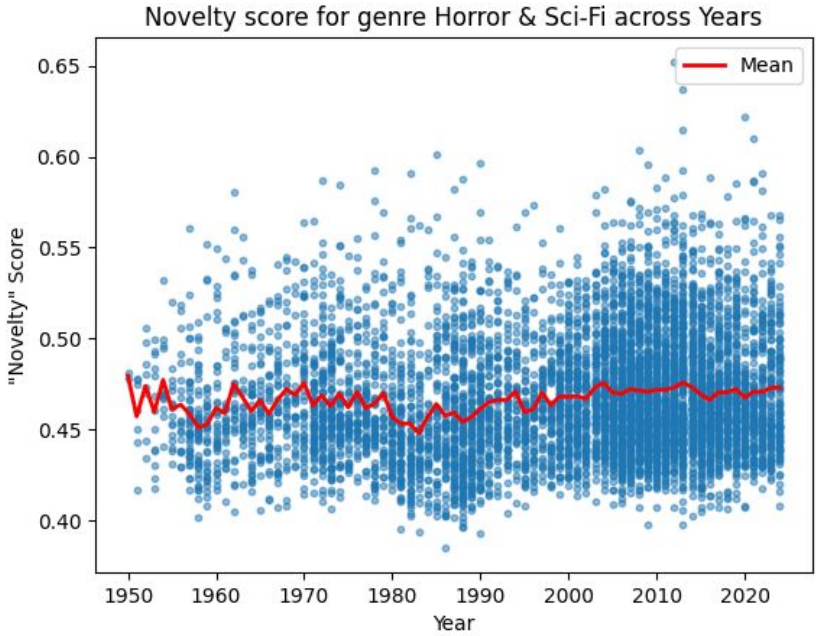


Cosine Distance to Ancho[illegible]ovie (2024 M[illegible]d)



Are movies getting less novel?

- Novelty score for a movie: Average cosine distance to all other movies released in the 10 years prior to the movie's release
- Results within Horror & Sci-Fi genre:



Alternative novelty measuring

- Possibly better novelty metric for a movie:
  - Average distance to the k nearest neighbours released before the movie

```
1. QID: Q1171321
   Title: Quatermass and the Pit
   Cosine Distance: 0.323483
   Cosine Similarity: 0.676517

2. QID: Q7717562
   Title: The Big Caper
   Cosine Distance: 0.332189
   Cosine Similarity: 0.667811

3. QID: Q1130297
   Title: Into the Blue
   Cosine Distance: 0.332843
   Cosine Similarity: 0.667157

4. QID: Q15300189
   Title: Extracted
   Cosine Distance: 0.332942
   Cosine Similarity: 0.667058

5. QID: Q31271369
   Title: Fast X
   Cosine Distance: 0.333034
   Cosine Similarity: 0.666966
```

Inception 5-NN

```
1. QID: Q320423
   Title: The Spy Who Loved Me
   Cosine Distance: 0.207398
   Cosine Similarity: 0.792602

2. QID: Q212145
   Title: The World Is Not Enough
   Cosine Distance: 0.235774
   Cosine Similarity: 0.764226

3. QID: Q207916
   Title: Tomorrow Never Dies
   Cosine Distance: 0.235821
   Cosine Similarity: 0.764179

4. QID: Q332368
   Title: A View to a Kill
   Cosine Distance: 0.235980
   Cosine Similarity: 0.764020

5. QID: Q4941
   Title: Skyfall
   Cosine Distance: 0.251589
   Cosine Similarity: 0.748411
```

GoldenEye 5-NN

While examining the data we found one more outlier …


Happy Birthday

14 Jan 2026


I DON'T KNOW WHERE YOU ARE
BUT HAPPY NEW YEAR
TO YOU
RD

IMDb ratings

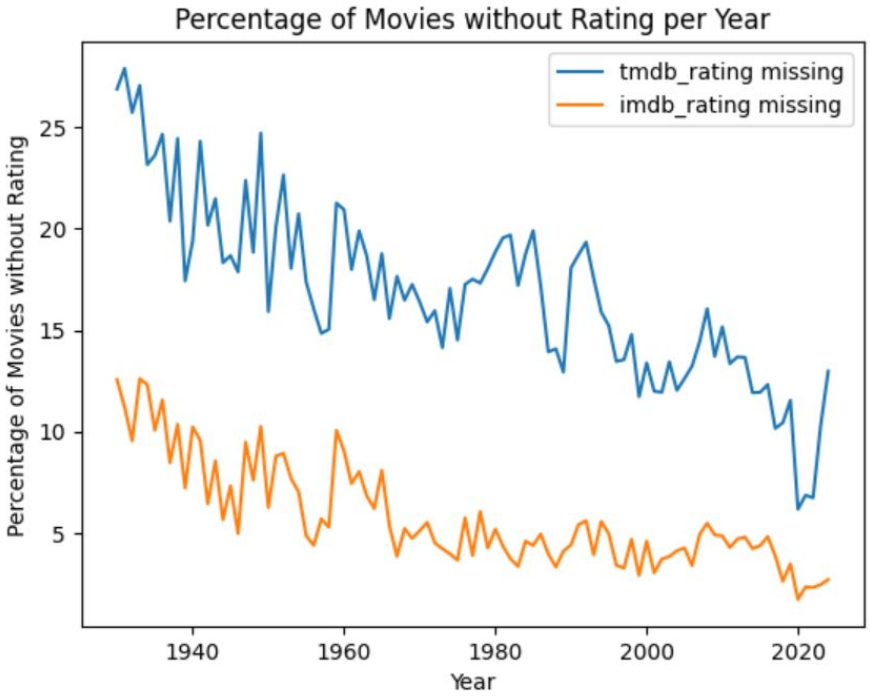
IMDb does not provide a free API to fetch ratings, but it provides data files that were last updated on March 18, 2024.

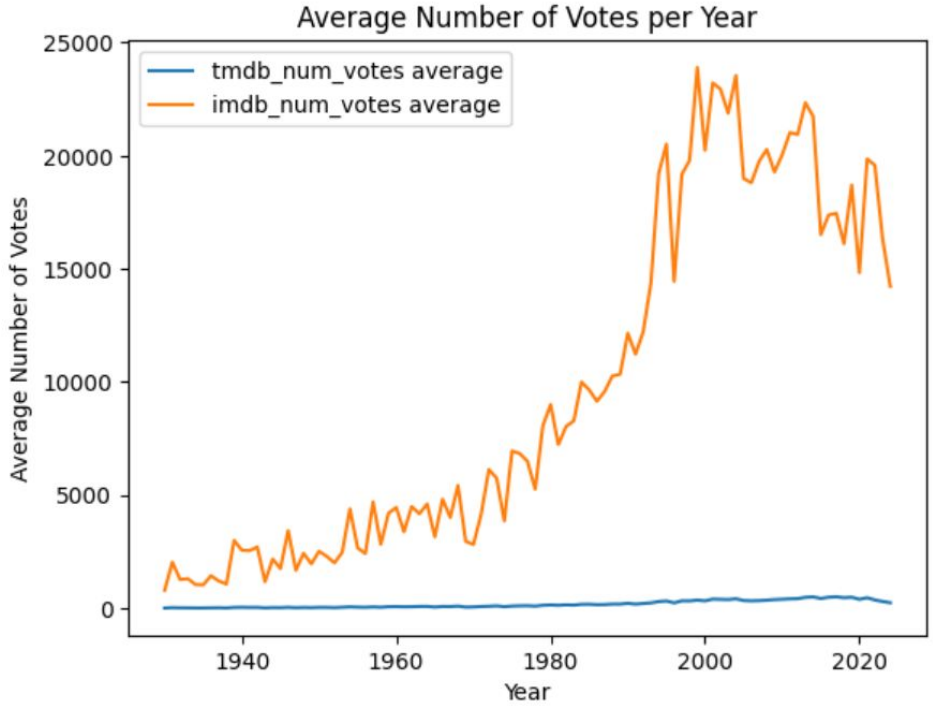
https://datasets.imdbws.com/

The target file contains:

- averageRating – weighted average of all the individual user ratings
- numVotes - number of votes the movie has received

Ratings comparison



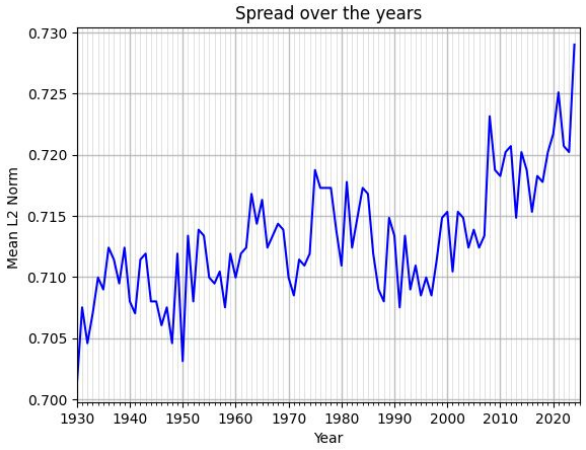


Movie Spread/Convergence Analysis

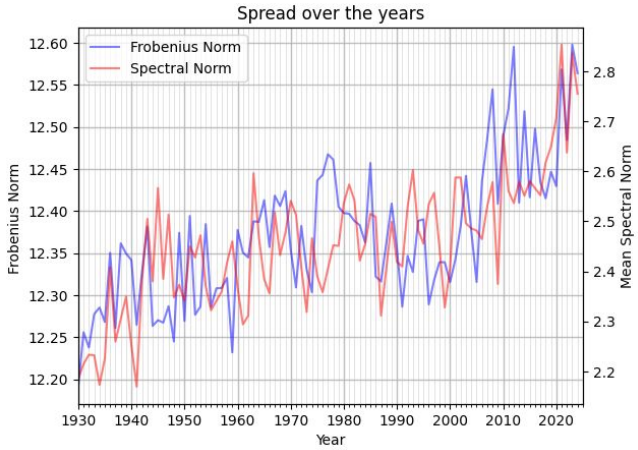
Question: Are all movies converging/becoming more similar?

- Total/Mean distance from mean embedding
- L2 norm of each movie to its year’s centroid (not medoid)
- Mean L2 norm for each year

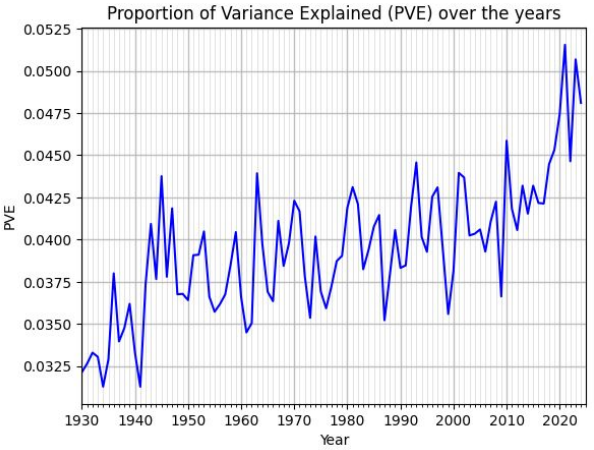
Movie Spread/Convergence Analysis



Mean L2 distance



Total Variance + strongest Singular value



PC1 explained variance

Movie Spread/Convergence Analysis

Question: Are all movies converging/becoming more similar?

- Mean L2 distance of each movie to centroid does not change much
- Total variance is increasing slowly
- Polarizing axis (?) is increasing faster

Interpretation:

- Movies are constantly diverse, but every year the outliers are getting further away

Movie Spread/Convergence Analysis

Question: What are these polarizing axis every year?

- Changes every year

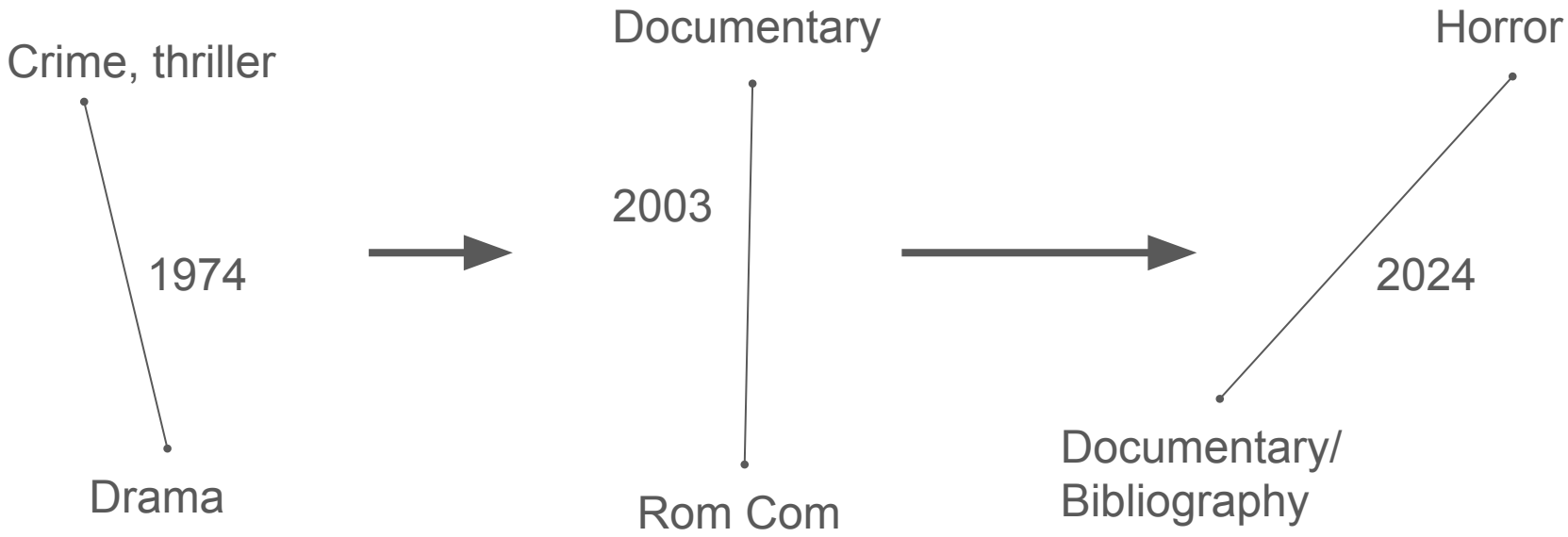


Table 1. Top and Bottom Performing Films by Country and Year

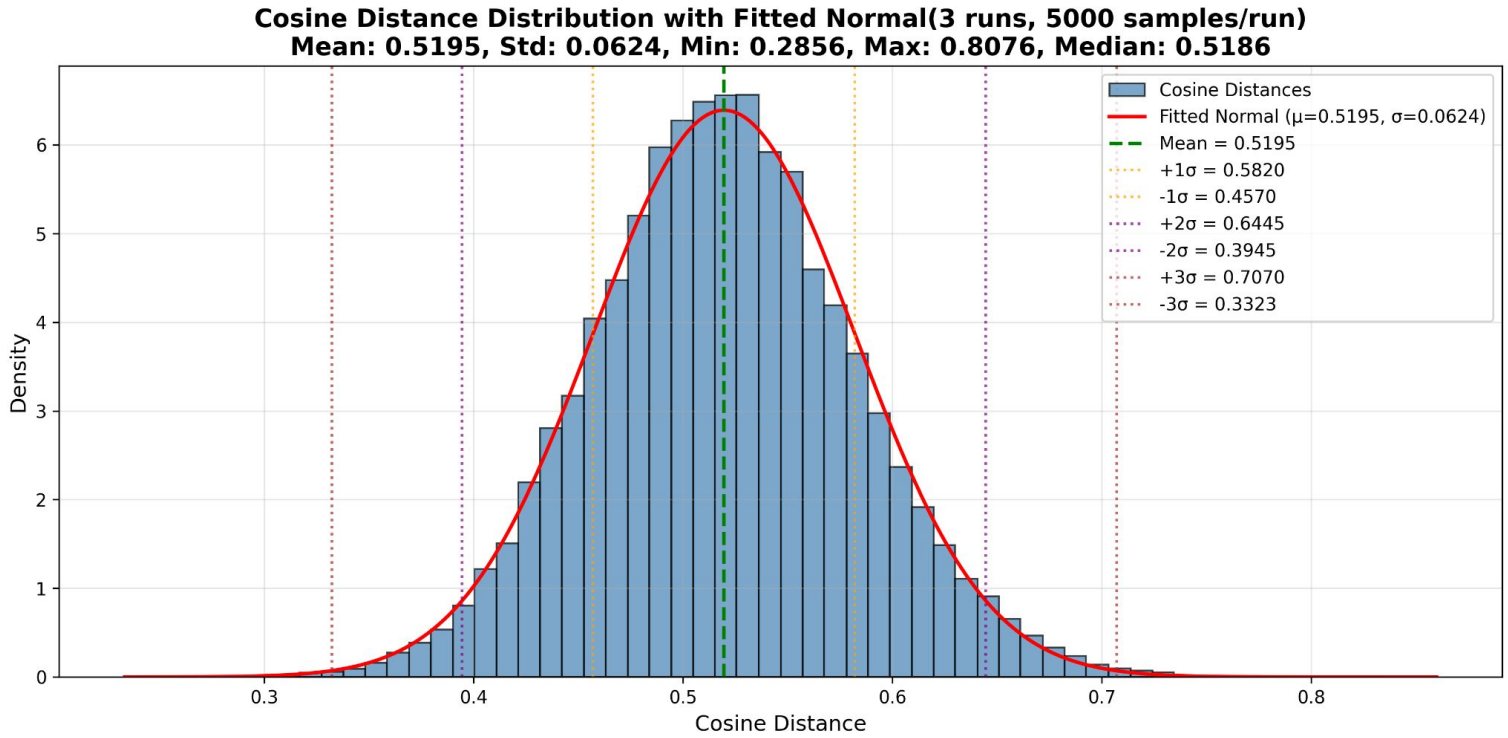
Year	Rank	US Films	German Films
1952	Top 1	Monsoon	Lockende Sterne
	Top 2	Strange Fascination	We're Dancing on the Rainbow
	Top 3	No Time for Flowers	The Colourful Dream
	Top 4	Just Across the Street	Rosen blühen auf dem Heidegrab
	Top 5	Everything I Have Is Yours	Das kann jedem passieren
	Bottom 5	Bend of the River	Toxi
	2	Denver and Rio Grande	The Condemned Village
	3	Red Skies of MontanaH	My Name is Niki
	4	The Savage	The Thief of Bagdad
	5	The Cimarron Kid	All Clues Lead to Berlin
2003	Top 1	Aileen: Life and Death of a Serial Killer	Beyond the Limits
	Top 2	Whole	Nikos the Impaler
	Top 3	Ghosts of the Abyss	Baltic Storm
	Top 4	DC 9/11: Time of Crisis	Debris documentar
	Top 5	Marion's Triumph	Wrong Turn
	Bottom 5	Sinbad: Legend of the Seven Seas	The Suit
	2	The One	Noi the Albino
	3	The Jungle Book 2	The Story of the Weeping Camel
	4	Cosmopolitan	A Little Bit of Freedom
	5	Flavors	Spring, Summer, Fall, Winter... and Spring
2024	Top 1	Terrifier 3	The Devil's Bath
	Top 2	It's What's Inside	Bird
	Top 3	Beetlejuice Beetlejuice	A Sacrifice
	Top 4	Smile 2	Cuckoo
	Top 5	The Thundermans Return	Santosh
	Bottom 5	The Firing Squad	Every You Every Me
	2	The True Story of Tamara de Lempicka and The Art of Survival	Spy vs. Spy
	3	Harvest	The Empire
	4	Waltzing with Brando	Rabia
	5	Putin	Harvest

Movie Spread/Convergence Analysis

Question: What are these polarizing axis every year?

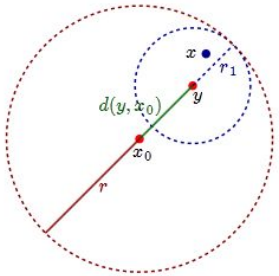
- Depends on individual dim contribution weight
- Hard to interpret
- Cannot really align (Orthogonal Procrustes)

Stats Regarding the general cosine distance



Epsilon Ball Method

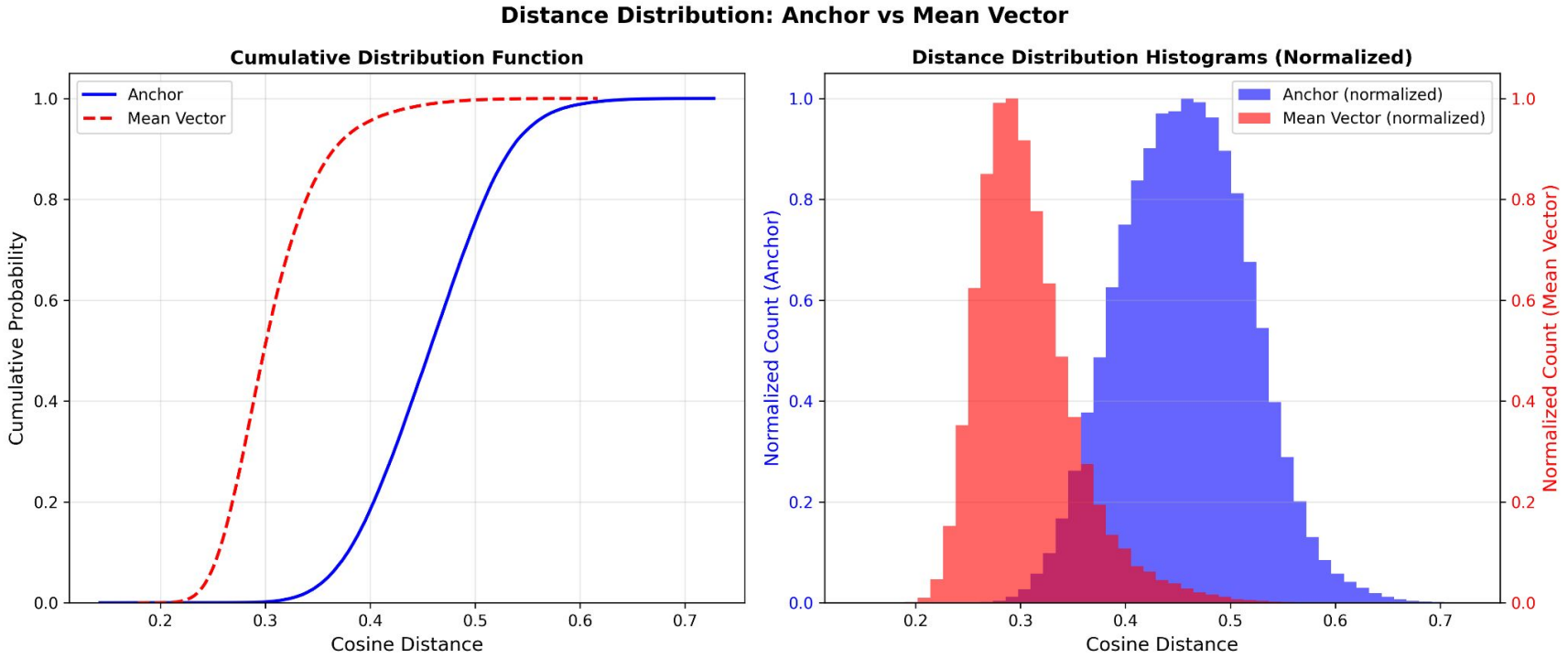
- The epsilon method finds all movies within a **specified distance (epsilon)** of an **anchor point** in the embedding space
- The anchor embedding is calculated from **one or more reference movies**
- Cosine distance is computed to measure similarity to the anchor; only movies within **the epsilon distance are included**
- Results are **sorted by distance** to the anchor
- Adjusting epsilon changes the strictness of similarity, enabling targeted semantic or temporal analysis of the resulting subset.



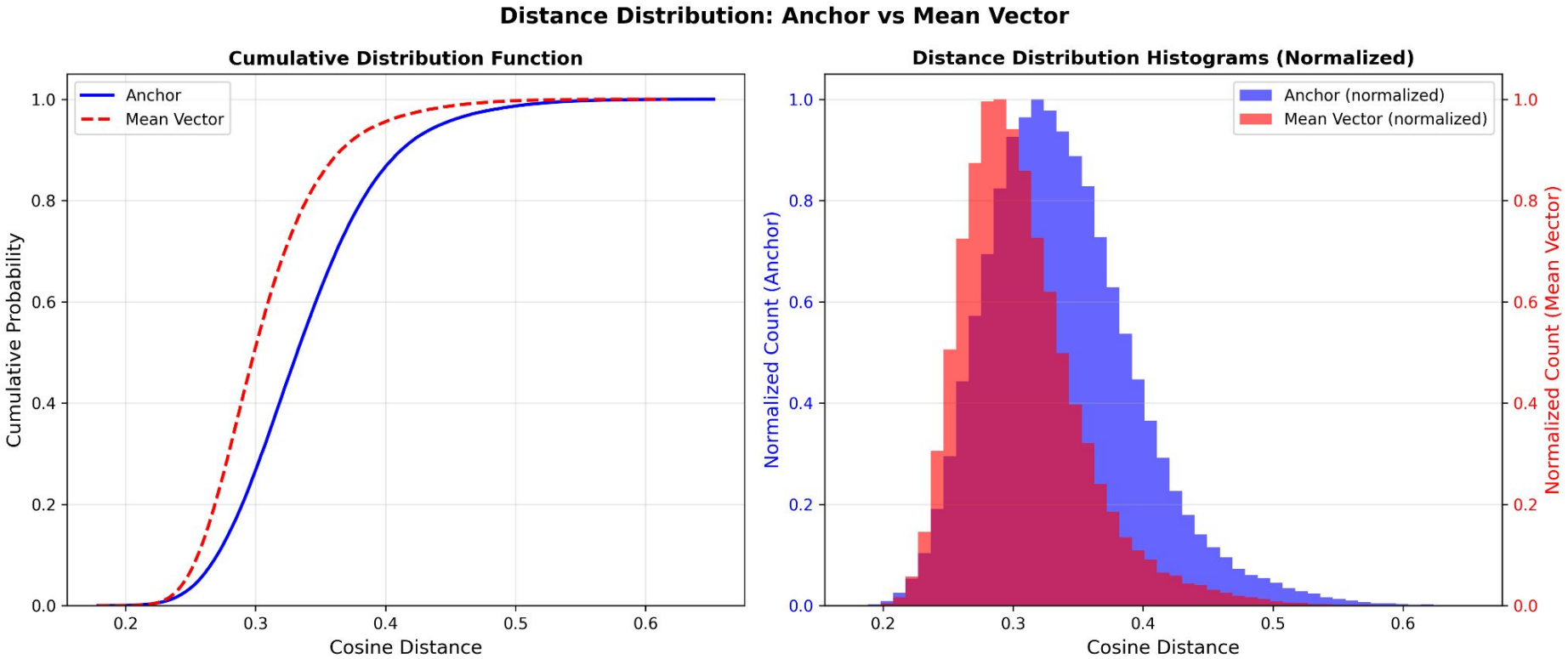
Kolmogorov-Smirnov (KS) Test

- The Kolmogorov-Smirnov (KS) test **compares two distributions** by measuring the **largest difference** between their **cumulative distribution functions**.
- It makes no assumptions about the distributions' shape and works for any continuous data.
- The KS statistic (D) ranges from **0 (identical) to 1 (very different)**; larger values mean greater differences.
- A p-value below 0.05 usually indicates the distributions are significantly different.
- In the epsilon ball context, the test checks **if anchor-based selections differ meaningfully from random or baseline selections**.

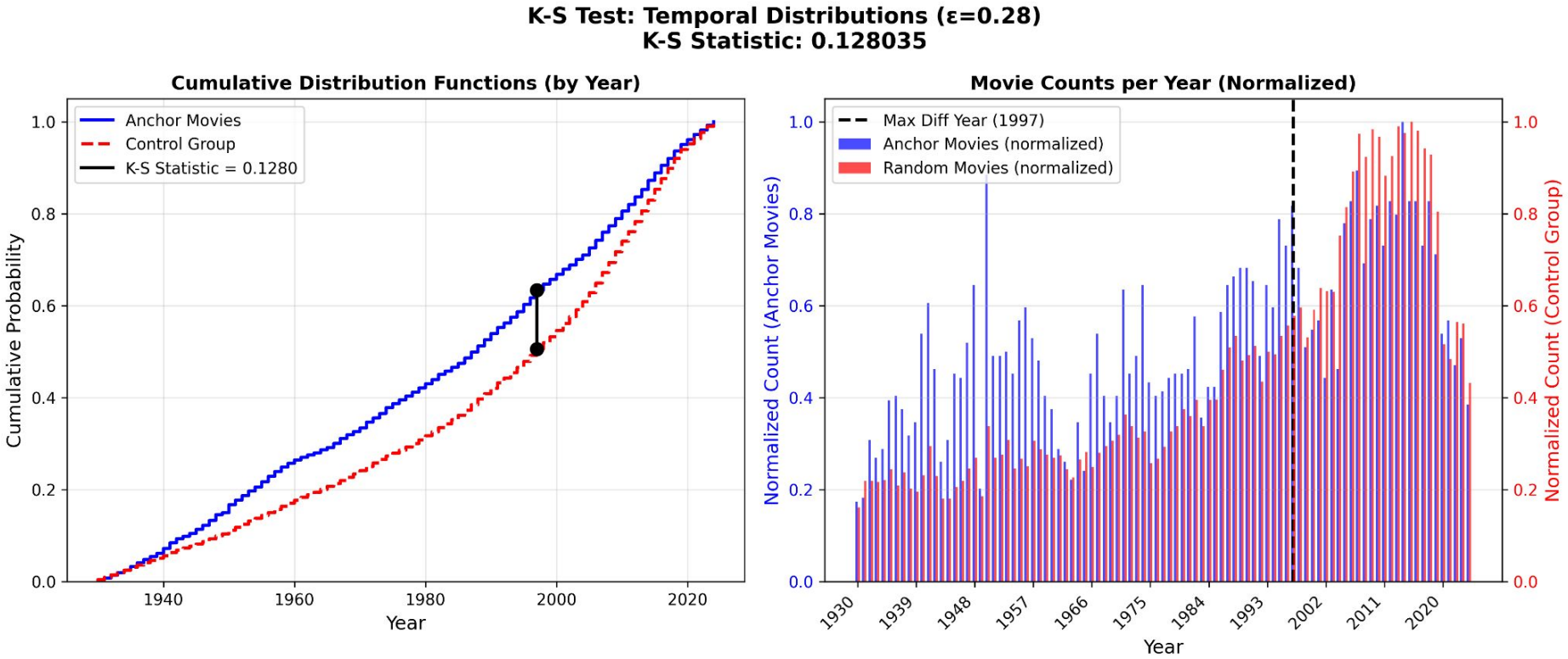
Stats first Distance Distribution: James Bond Movies



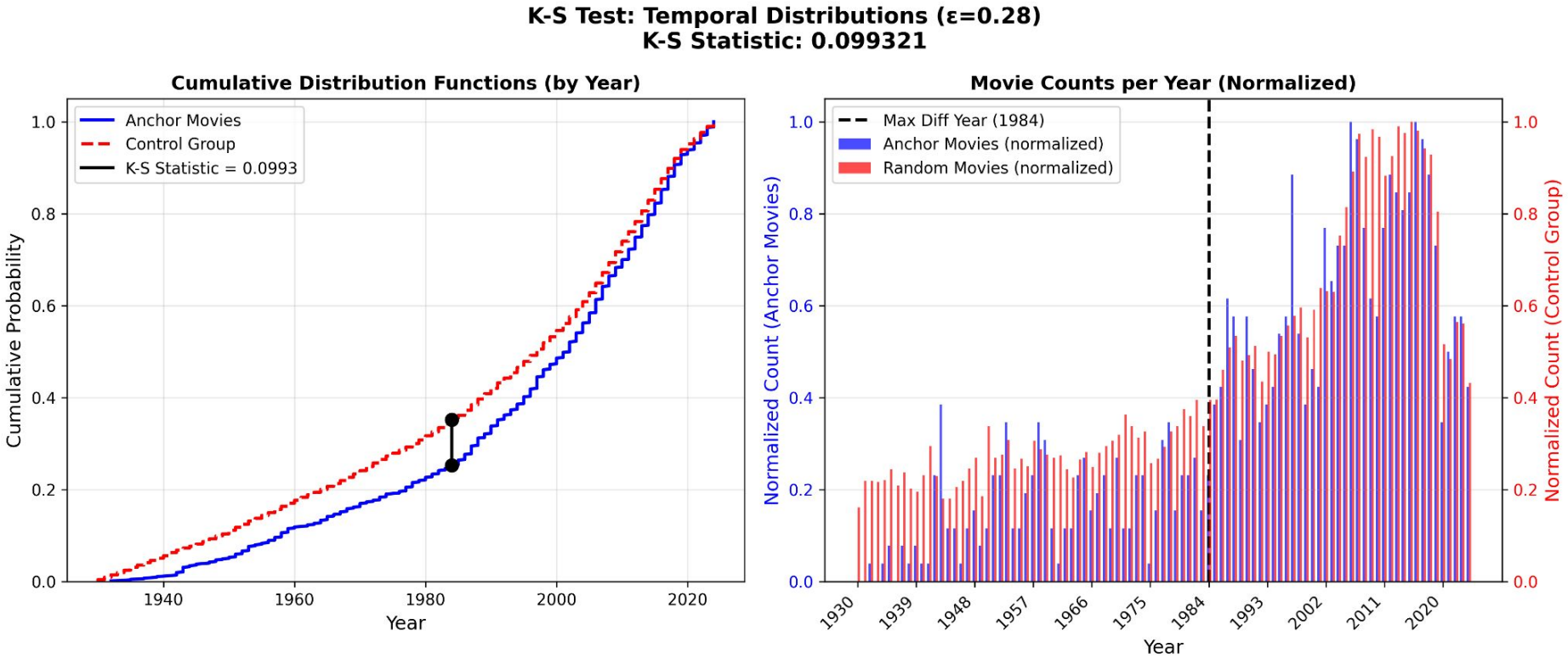
Stats first Distance Distribution: MOST AVERAGE Movie



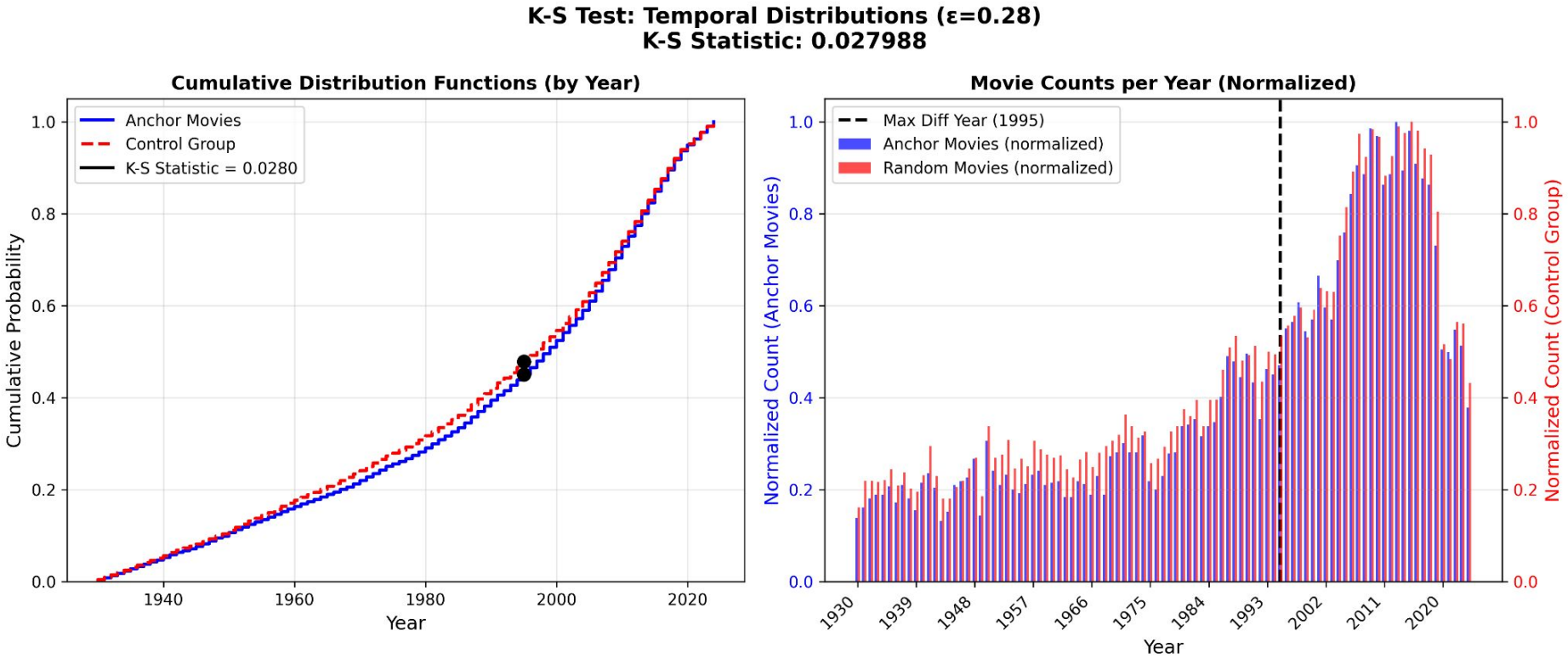
KS Test on the Temporal Distances in Ball Wild West



KS Test on the Temporal Distances in Ball War Movies (Iraq, Afghanistan)

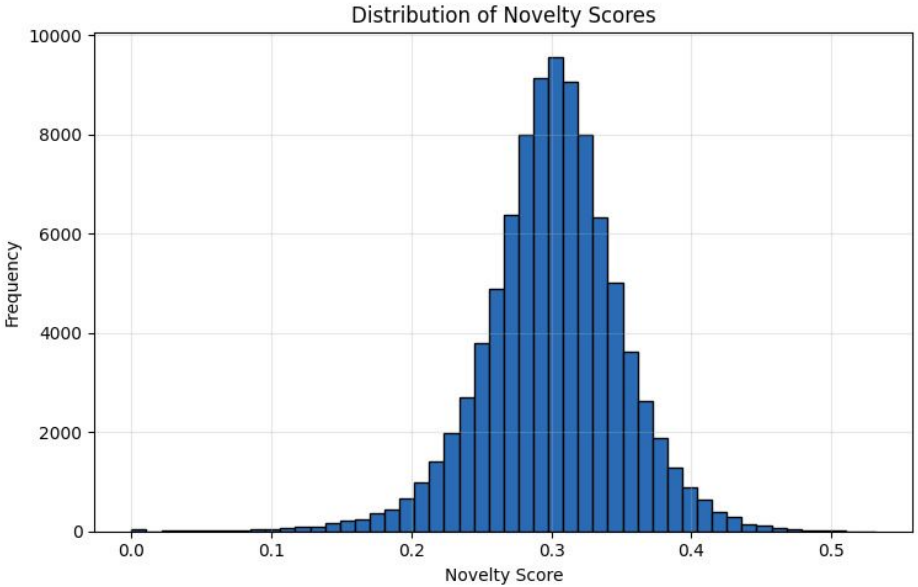


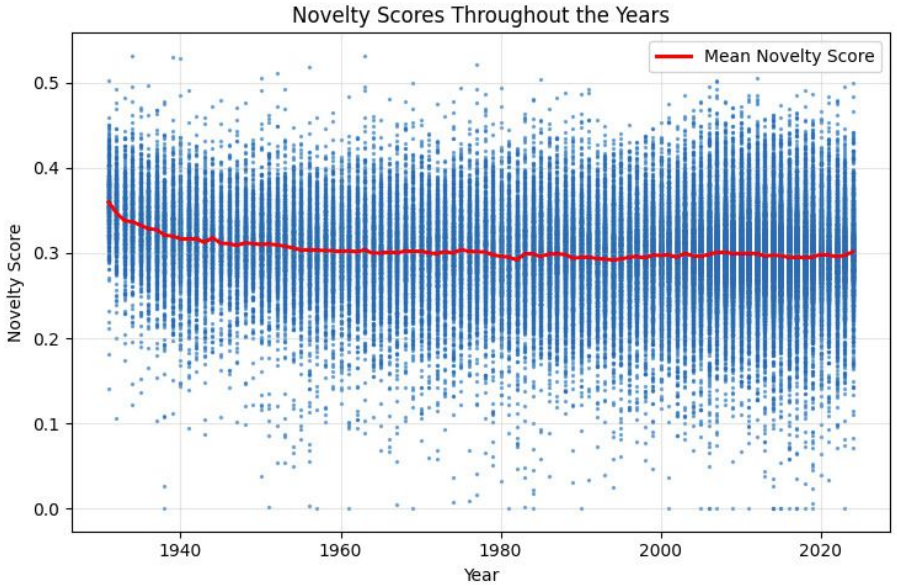
KS Test on the Temporal Distances in Ball Medoid Movie



Novelty score based on nearest neighbour

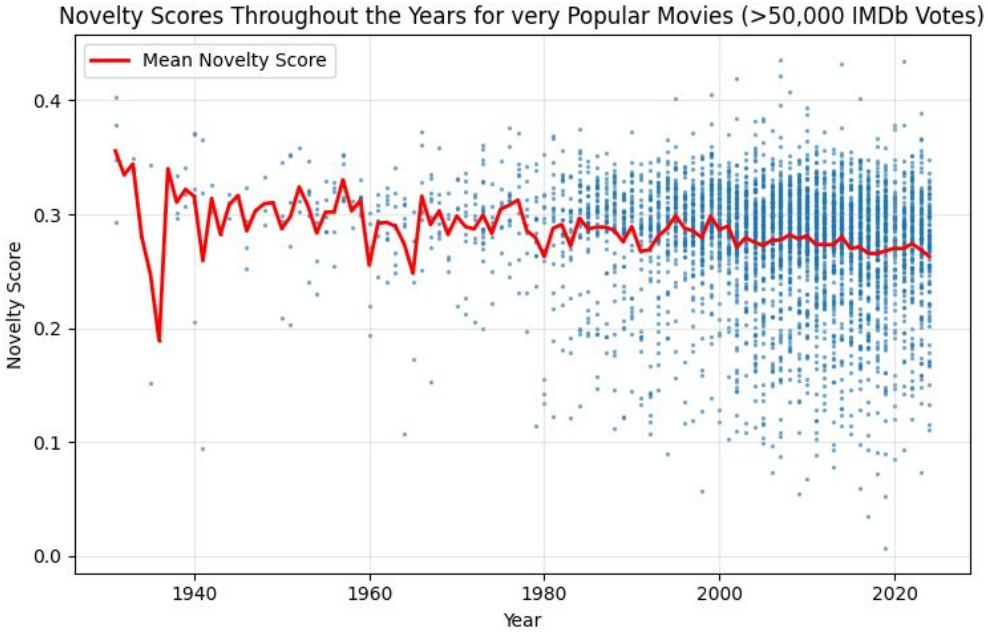
- Novelty score = cosine distance to the closest movie released prior to the movie's release year
- = 1 - "how similar a movie is to the most similar previously released movie"



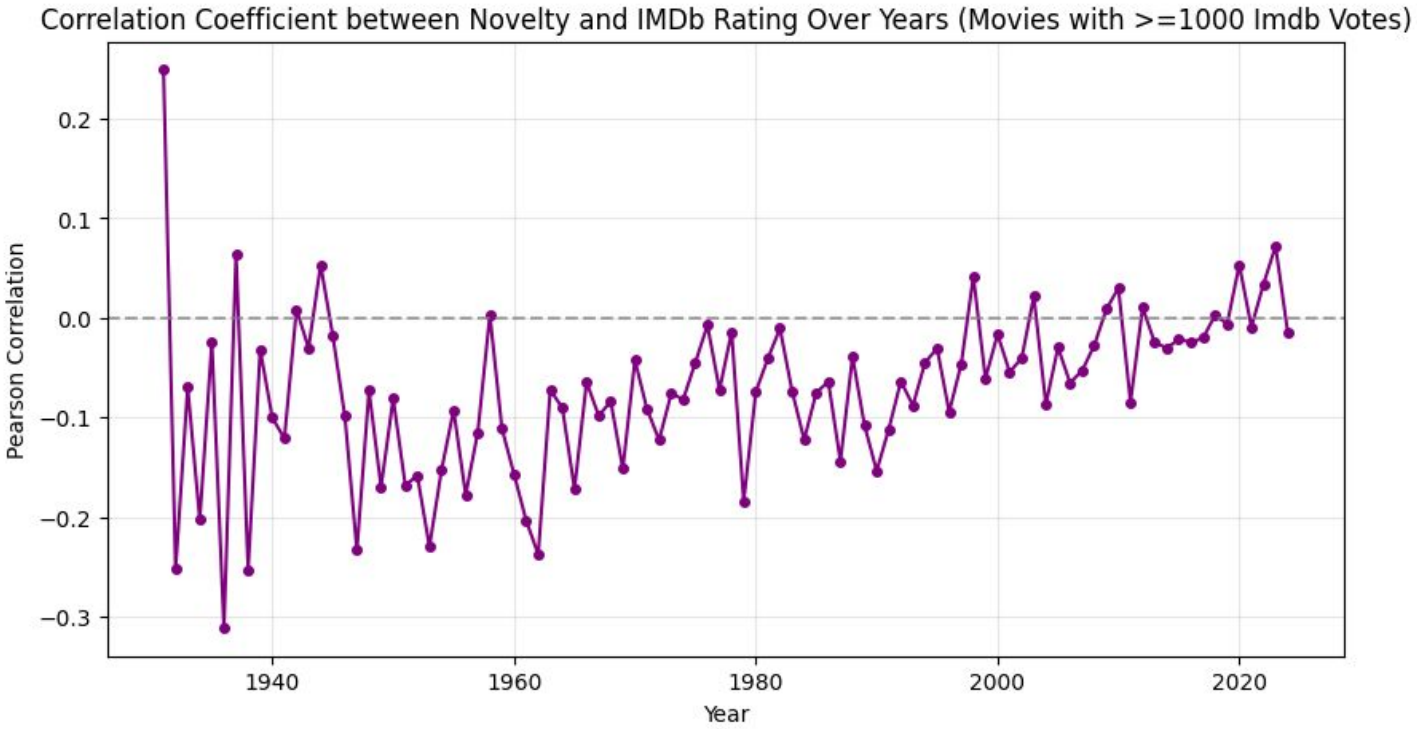


Novelty scores of “popular” movies

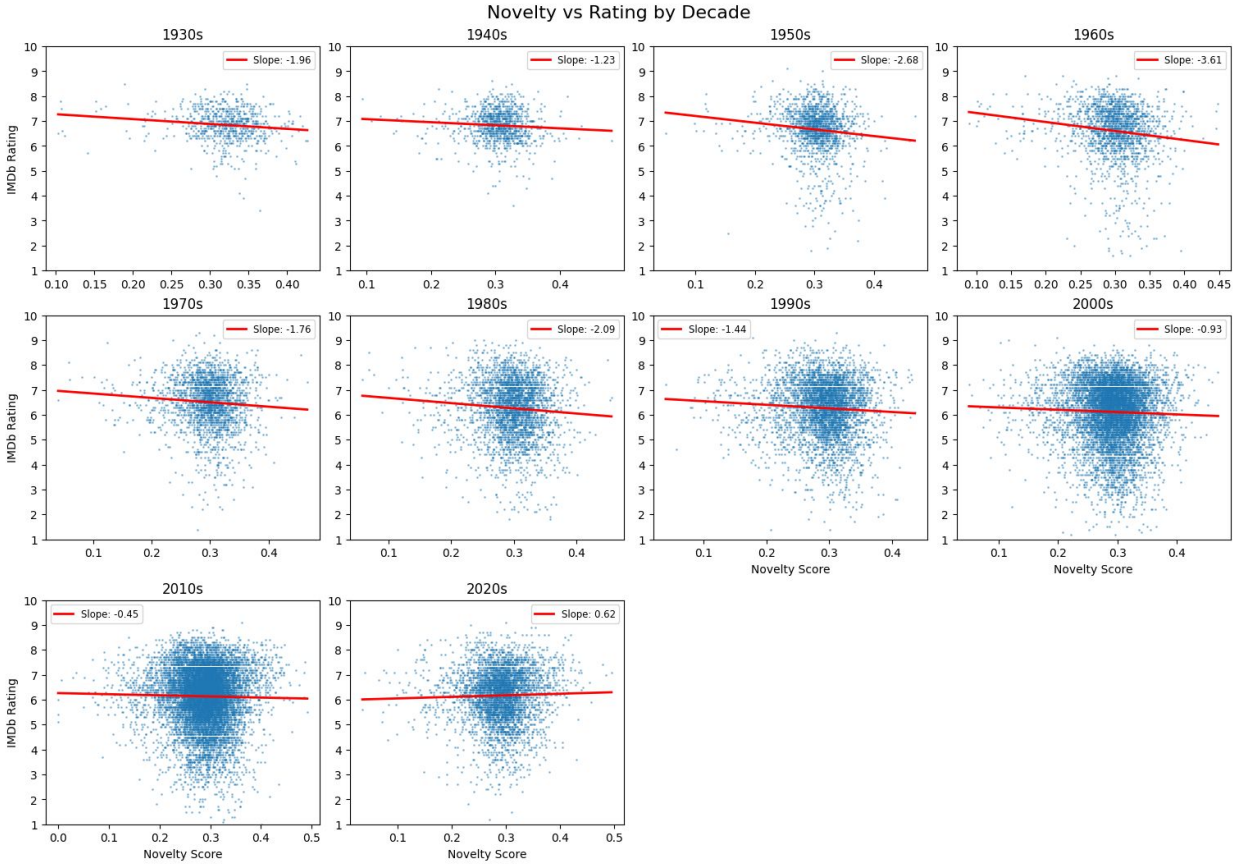
- Maybe only the very popular movies are getting less novel
- Popular movie = movie with more than 50 000 IMDb votes (only 2574 movies in our dataset)



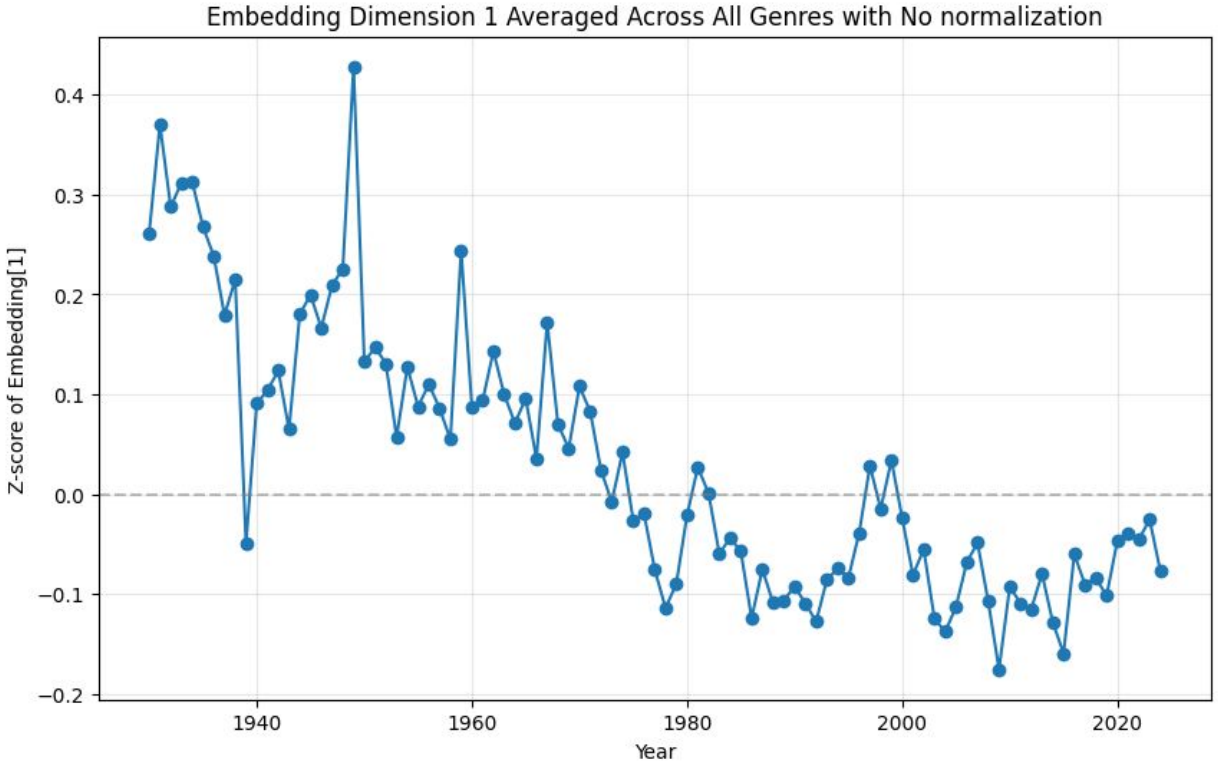
How does novelty relate to rating?



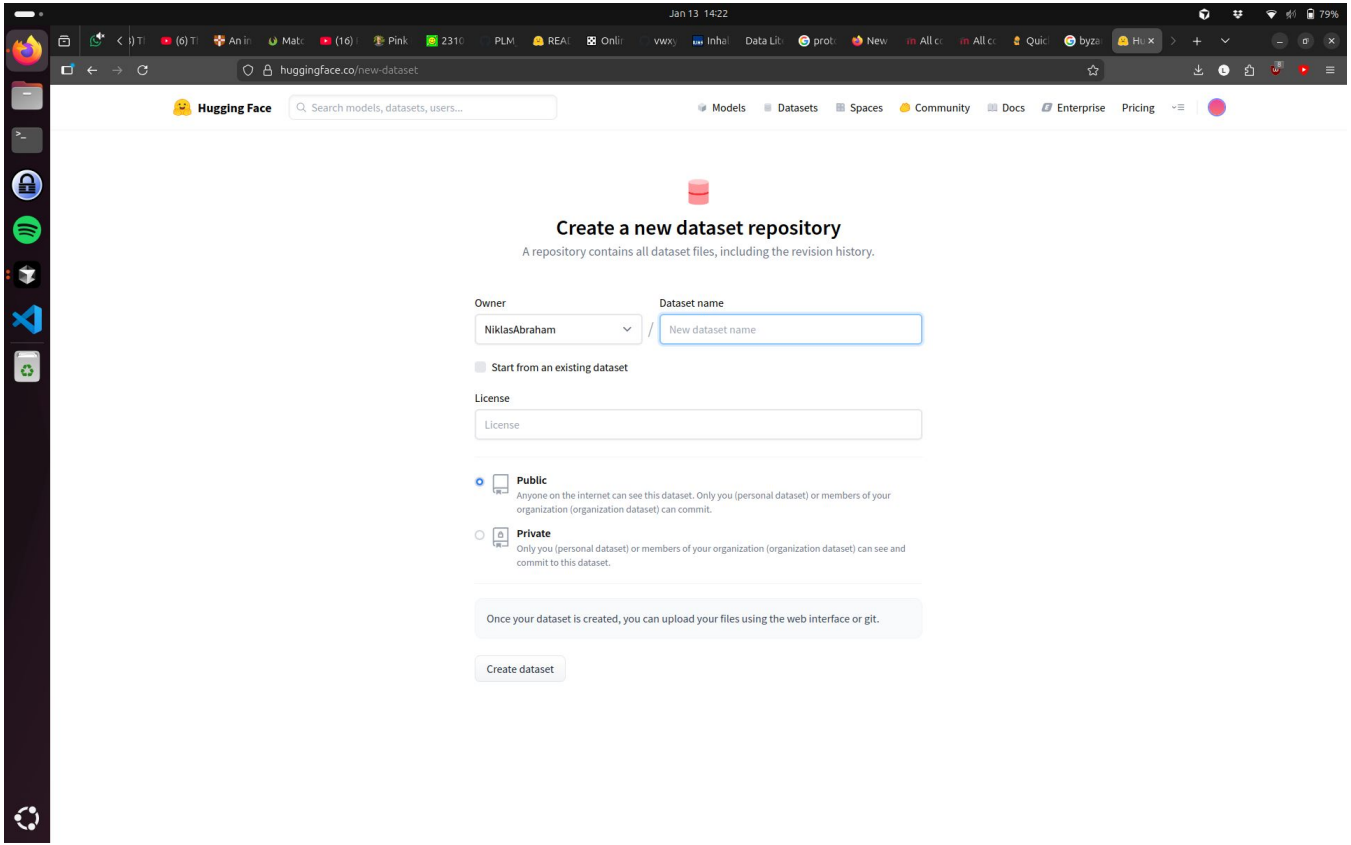
How does novelty relate to rating – historically?



Embeddings individual dimensions



Data Upload?


Jan 13 14:22
huggingface.co/new-dataset
Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Community
Docs
Enterprise
Pricing
Create a new dataset repository
A repository contains all dataset files, including the revision history.
Owner
NiklasAbraham
Dataset name
New dataset name
Start from an existing dataset
License
License
Public
Anyone on the internet can see this dataset. Only you (personal dataset) or members of your organization (organization dataset) can commit.
Private
Only you (personal dataset) or members of your organization (organization dataset) can see and commit to this dataset.
Once your dataset is created, you can upload your files using the web interface or git.
Create dataset