Data Dictionary — Movie Recommendation Optimizer (v1)

Datasets Overview (CSV/TSV)

1) IMDb official TSVs

title.basics.tsv — title metadata

title.crew.tsv — directors and writers

title.ratings.tsv — average rating and number of votes

2) MovieLens CSVs

movies.csv — MovieLens titles and year (often in title string)

links.csv — IDs mapping (movieId ↔ imdbId ↔ tmdbId)

ratings.csv — user–movie ratings (14.9M+ rows)

tags.csv — user–movie tag applications (2.0M+ rows)

3) Rotten Tomatoes CSVs

rotten\_tomatoes\_movies.csv — RT-wide catalog with audience & critic scores

rotten\_tomatoes\_top\_movies.csv — curated “top” movies with rich technical fields

rotten\_tomatoes\_movie\_reviews.csv — critic review texts & metadata

4) Reviews CSV

Movies\_Reviews\_modified\_version1.csv — labeled review text corpus (multi-class sentiment)

IMDb — title.basics.tsv

Primary Key: tconst (string, e.g., 'tt1234567')

Joins: to IMDb ratings/crew on tconst; to MovieLens links via imdbId → tconst = 'tt' + zfill(imdbId, 7)

Typical Shape: large; dynamic over time

Columns

tconst (string): IMDb title ID.

titleType (string): movie | short | tvSeries | tvMovie | video | etc.

primaryTitle (string): display title.

originalTitle (string): original-locale title.

isAdult (int: 0/1): adult content flag.

startYear (int or \N): release year.

endYear (int or \N): end year (for series).

runtimeMinutes (int or \N): runtime in minutes.

genres (string): pipe-separated genres (e.g., Action|Sci-Fi).

IMDb — title.crew.tsv

Primary Key: tconst

Joins: to basics/ratings via tconst

Columns

tconst (string): IMDb title ID.

directors (string): comma-separated nconst IDs.

writers (string): comma-separated nconst IDs.

IMDb — title.ratings.tsv

Primary Key: tconst

Joins: to basics/crew via tconst

Columns

tconst (string): IMDb title ID.

averageRating (float): mean user rating.

numVotes (int): votes count.

MovieLens — movies.csv (shape observed: 87,585 × 3)

Primary Key: movieId (int)

Joins: to links/ratings/tags on movieId

Columns

movieId (int): internal MovieLens movie identifier.

title (string): “Title (Year)” format commonly; year may be present in parentheses.

genres (string): pipe-separated genre list; may include '(no genres listed)'.

MovieLens — links.csv (shape observed: 87,585 × 3)

Primary Key: movieId (int)

Joins: to movies/ratings/tags on movieId; bridges to IMDb/TMDB

Columns

movieId (int): MovieLens ID.

imdbId (int): numeric IMDb ID (no 'tt' prefix). Use: tconst = 'tt' + str(imdbId).zfill(7).

tmdbId (float|int): TMDB ID; may be missing (NaN).

MovieLens — ratings.csv (shape observed: 14,922,255 × 4)

Columns

userId (int): anonymized user identifier.

movieId (int): MovieLens movie.

rating (float): typically 0.5..5.0 in 0.5 steps.

timestamp (int): Unix seconds; convert to UTC datetime for temporal analysis.

MovieLens — tags.csv (shape observed: 2,000,072 × 4)

Columns

userId (int): user identifier.

movieId (int): MovieLens movie.

tag (string): free-text tag (e.g., 'cult classic', 'mind-bending').

timestamp (int): Unix seconds.

Rotten Tomatoes — rotten\_tomatoes\_movies.csv (shape: 143,258 × 16)

Suggested Keys: id (slug) when available; otherwise (title, year) after title-normalization.

Columns

id (string): RT slug/identifier.

title (string): movie title.

audienceScore (float): audience score 0–100.

tomatoMeter (float): critic score 0–100.

rating (string): MPAA rating (e.g., PG-13, R).

ratingContents (string): content descriptors.

releaseDateTheaters (date): theatrical release date.

releaseDateStreaming (date): streaming release date.

runtimeMinutes (float): runtime in minutes.

genre (string): comma-separated genres.

originalLanguage (string): original language of release.

director (string): comma-separated director names.

writer (string): comma-separated writer names.

boxOffice (string): box office gross (may include '$', 'M').

distributor (string): distributor / studio.

soundMix (string): sound format (e.g., Dolby Atmos, Mono).

Rotten Tomatoes — rotten\_tomatoes\_top\_movies.csv (shape: 1,610 × 26)  
-----------------------------------------------------------  
Columns  
- Unnamed: 0 (int): index column; safe to drop.  
- title (string): movie title.  
- year (int): release year.  
- synopsis (string): plot summary.  
- critic\_score (float): Rotten Tomatoes critic score (0–100).  
- people\_score (float): Rotten Tomatoes audience score (0–100).  
- consensus (string): critics' consensus statement.  
- total\_reviews (int|string): number of critic reviews.  
- total\_ratings (int|string): number of audience ratings.  
- type (string): category or genre family (e.g., 'Action & Adventure').  
- rating (string): MPAA rating (PG, PG-13, R, etc.).  
- genre (string): genres listed.  
- original\_language (string): original language of release.  
- director (string): director(s) name(s).  
- producer (string): producer(s) name(s).  
- writer (string): writer(s) name(s).  
- release\_date\_(theaters) (date|string): theatrical release date.  
- release\_date\_(streaming) (date|string): streaming release date.  
- box\_office\_(gross\_usa) (string): formatted US gross (e.g., '$700.2M').  
- runtime (string): movie duration (e.g., '2h 14m').  
- production\_co (string): production company names.  
- sound\_mix (string): sound format (e.g., Dolby Atmos).  
- aspect\_ratio (string): display aspect ratio (e.g., 'Scope (2.35:1)').  
- view\_the\_collection (string): franchise/collection label.  
- crew (string): main cast and crew.  
- link (string): Rotten Tomatoes URL to the title page.

Note: 26 total columns; 20 were visible in the head() preview. The remaining ~6 will be confirmed and added in v1.1.

Columns (confirmed)

Unnamed: 0 (int): index column; safe to drop.

title (string): movie title.

year (int): release year.

synopsis (string): plot summary.

critic\_score (float): RT critic score 0–100.

people\_score (float): RT audience score 0–100.

consensus (string): critics' consensus statement.

total\_reviews (int|string): number of critic reviews.

total\_ratings (int|string): number of audience ratings.

type (string): category/genre family (e.g., 'Action & Adventure').

release\_date\_(theaters) (date|string): first US theatrical release.

release\_date\_(streaming) (date|string): first streaming availability.

box\_office\_(gross\_usa) (string): formatted gross (e.g., '$700.2M').

runtime (string): e.g., '2h 14m'.

production\_co (string): production company names.

sound\_mix (string): sound format.

aspect\_ratio (string): e.g., 'Scope (2.35:1)'.

view\_the\_collection (string): franchise/collection label.

crew (string): principal cast/crew list.

link (string): RT URL to the title page.

Columns (to confirm; placeholders)

col\_21 — to confirm

col\_22 — to confirm

col\_23 — to confirm

col\_24 — to confirm

col\_25 — to confirm

col\_26 — to confirm

Rotten Tomatoes — rotten\_tomatoes\_movie\_reviews.csv (shape: 58,770 × 11)

Columns

id (string): RT movie slug/identifier (links to RT title pages; join to RT movies via id).

reviewId (int): unique review identifier.

creationDate (date): review publication date.

criticName (string): critic name.

isTopCritic (bool): True if designated Top Critic.

originalScore (string): original score string (e.g., '3.5/4', 'B', etc.).

reviewState (string): 'fresh' | 'rotten' (critic's RT label).

publicationName (string): outlet/publication.

reviewText (string): review body (truncated in head(); full text present in data).

scoreSentiment (string): POSITIVE | NEGATIVE | (possibly NEUTRAL if present).

reviewUrl (string): source URL.

Labeled Reviews — Movies\_Reviews\_modified\_version1.csv (shape: 46,173 × 8)

Columns

Unnamed: 0 (int): index; safe to drop.

id (string|int): review/document ID.

review (string): review text.

label (string): sentiment/emotion label (e.g., anticipation, joy, anger).

genre (string): genre label (source-specific).

rating (float|int): numeric rating if provided.

movie (string): associated movie title.

split (string): dataset split (train|test|val).

Standardization & Joins

ID Harmonization

IMDb: tconst native; if starting from MovieLens, compute tconst as 'tt' + zfill(imdbId, 7).

TMDB: leverage MovieLens links.tmdbId to fetch TMDB fields (posters, popularity, providers).

Rotten Tomatoes: use id (slug) where present; otherwise a normalized (title, year) join with fallback fuzzy matching.

Dates & Times

Convert Unix timestamps (MovieLens ratings/tags) to timezone-aware UTC datetimes.

Normalize all date columns to ISO format (YYYY-MM-DD).

Genres

Build a canonical genre mapping (e.g., 'Sci-fi' → 'Science Fiction').

Split pipe/comma-separated genres into arrays; store normalized multi-hot encoding.

Text Fields

Keep raw text (synopsis, reviewText, consensus); compute derived features (TF‑IDF, BERT embeddings) in feature store.

Numeric Normalization

Cast scores to numeric; coerce invalid strings to NaN, then impute or drop.

Strip currency symbols from box office and parse to integers (USD).

Duplicates & Provenance

Deduplicate on stable IDs (tconst, tmdbId, RT id). If missing, use normalized (title, year).

Add boolean provenance columns: source\_imdb, source\_ml, source\_rt.

Quality Checks

Non-null keys after merges; row counts across joins; spot-check random titles for consistency.

Assert score ranges (0–100) and rating scales (0.5–5.0).