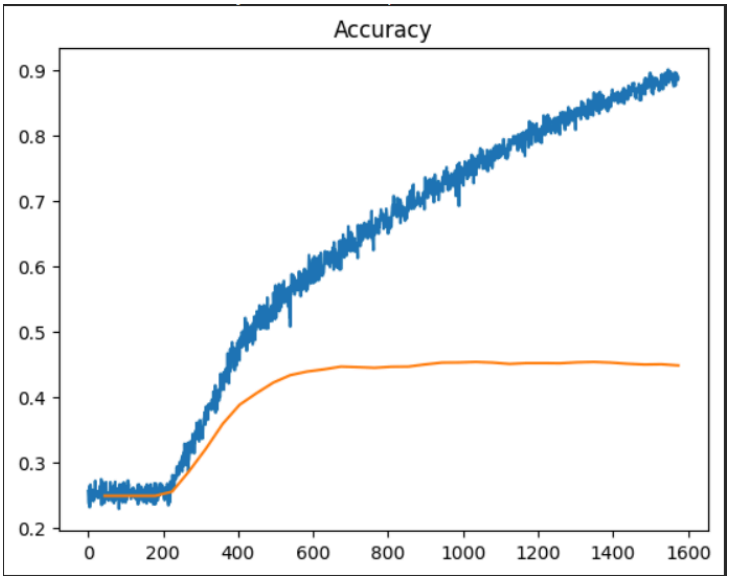
Assignment 3 : Movies Recommendation

1. Problem 1 : Train Embeddings

As for my setting for training the model, I have chose the hyperparameter n_factor = 200, because it has the best trade-off between accuracy, stability, and training time. It provides sufficient model capacity to represent complex movies–user interactions and should avoid the overfitting and underfitting.

I also set the hyperparameter n_epoch = 35, so it should have the best performance before overfitting. It has enough time for convergence, should have lower training Loss and Accuracy reaches the target > 0.44.



2. Problem 2 : Analyze Bias of Movie Embedding: Why do we use Bias ?

The movie bias allows the model to capture the overall popularity or unpopularity of a movie independently of individual user's preferences. Movies with a high bias are those that consistently receive high ratings, while movies with a low bias generally receive poor ratings.
Without the bias term, the model would be forced to use the embeddings to represent these global rating tendencies, which would reduce their ability to capture more fine-grained interactions between users and movies. By adding a bias term, the model becomes more accurate, converges faster, and produces better recommendations.

	title	bias
0	Shawshank Redemption, The (1994)	0.354043
1	Forrest Gump (1994)	0.342986
2	Dr. Strangelove or: How I Learned to Stop Worr...	0.321231
3	Princess Bride, The (1987)	0.320880
9721	I Know What You Did Last Summer (1997)	-0.249773
9722	Speed 2: Cruise Control (1997)	-0.253156
9723	Godzilla (1998)	-0.257874

3. Problem 3 : Similarity Search

The cosine similarity search produces results that are intuitively meaningful.For each of the selected movies, the system retrieves films that share similar genres, themes, and audience characteristics.

For The Matrix (1999) (ID 1938), the most similar movies are other science-fiction or cyber-action films, such as Matrix Reloaded, Dark City, or Equilibrium. These movies share elements like futuristic settings, dystopian themes, and high-intensity action.

For Scream (1996) (ID 1082), the top results consist of slasher or teen-horror movies, including titles like I Know What You Did Last Summer or other entries in the Scream franchise. These movies match well in tone, pacing, and genre conventions.

For Toy Story (1995) (ID 0), the system retrieves animated family films such as Toy Story 2, A Bug's Life, or other Pixar/Disney productions. These films target the same family-friendly audience and share similar visual style and themes.

The least similar movies tend to come from genres that are completely unrelated, such as documentaries, romance dramas, or films for adults when the selected movie is intended for children.This contrast demonstrates that the learned embeddings successfully capture meaningful semantic relationships between movies, separating genres and grouping films with similar narrative or stylistic characteristics.

Toy Story (1995)

Titles of selected movie_id is: Toy Story (1995)

	title	genre	similarity
0	Toy Story 2 (1999)	Adventure|Animation|Children|Comedy|Fantasy	0.629862
1	Shawshank Redemption, The (1994)	Crime|Drama	0.616015
2	Forrest Gump (1994)	Comedy|Drama|Romance|War	0.584013
9721	Catwoman (2004)	Action|Crime|Fantasy	-0.452020
9722	Return to the Blue Lagoon (1991)	Adventure|Romance	-0.452962
9723	Amityville II: The Possession (1982)	Horror	-0.469721

Scream (1996)

Titles of selected movie_id is: Scream (1996)

	title	genre	similarity
0	Scream 3 (2000)	Comedy|Horror|Mystery|Thriller	0.496061
1	Scream 2 (1997)	Comedy|Horror|Mystery|Thriller	0.476737
2	Funny Games (1997)	Drama|Horror|Thriller	0.419917
3	Jaws (1975)	Action|Horror	0.401635
9721	Caveman (1981)	Comedy	-0.342083
9722	Liam (2000)	Drama	-0.355303
9723	Big Top Pee-Wee (1988)	Adventure|Children|Comedy	-0.382491

The Matrix (1999)

Titles of selected movie_id is: Matrix, The (1999)

	title	genre	similarity
0	Snatch (2000)	Comedy|Crime|Thriller	0.678785
1	Gladiator (1992)	Action|Drama	0.677144
2	Saving Private Ryan (1998)	Action|Drama|War	0.671766
9721	Anaconda: The Offspring (2008)	Action|Horror|Sci-Fi|Thriller	-0.540754
9722	Master of Disguise, The (2002)	Comedy|Mystery	-0.555078
9723	Disaster Movie (2008)	Comedy	-0.558845

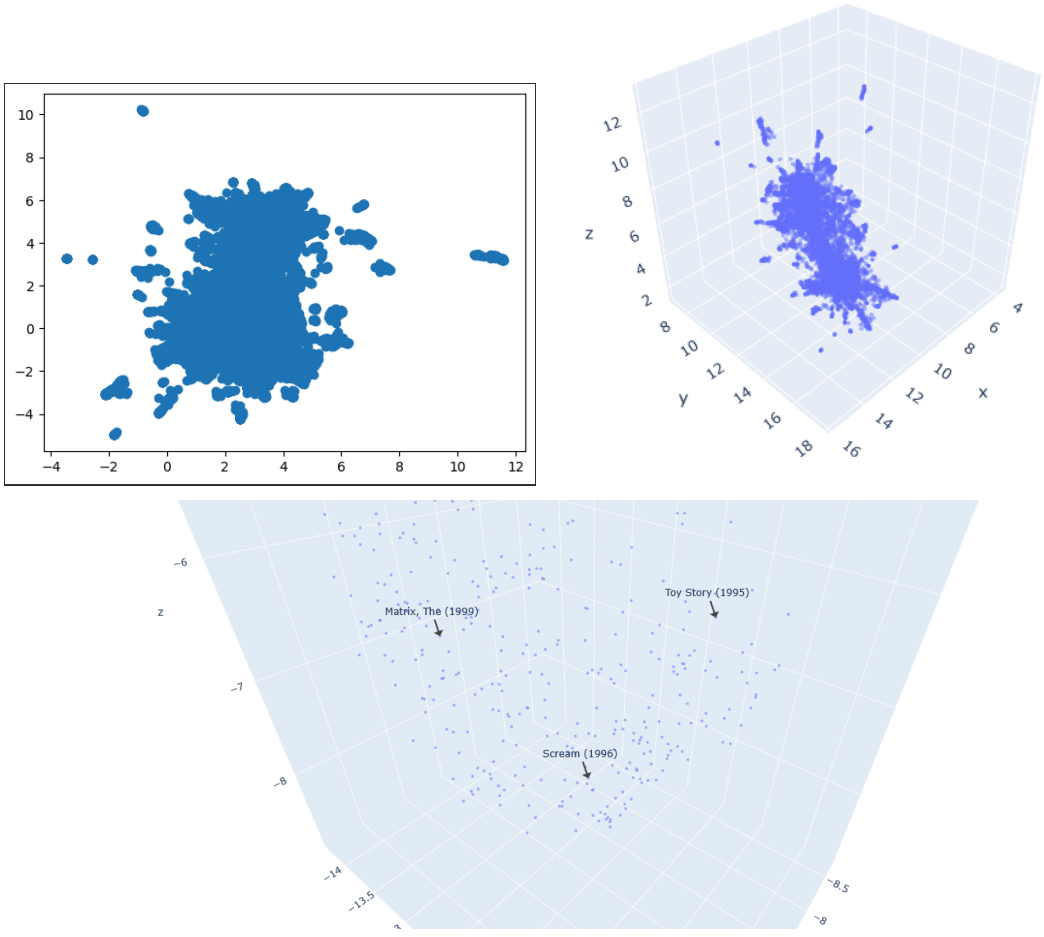
4. Problem 4 : Embedding Visualization

Even though the model receives no information about genres, actors, release year, or plot, it can still learn meaningful similarities between movies.
This happens because users tend to rate similar movies in similar ways.
For example :

- Users who like Toy Story often like other animated family movies.
- Users who rate The Matrix highly tend to enjoy other action/sci-fi movies.
- Horror fans consistently rate movies like Scream similarly.

During training, the matrix factorization model adjusts the movie embeddings so that:

- Movies watched and rated by similar groups of users get pushed closer together in the embedding space.
- Movies with completely different audiences get pushed farther apart.



The UMAP visualization clearly shows that the embedding model has learned meaningful structure from the rating matrix. Movies cluster naturally according to genre, target audience, and tone. Movies with unrelated audiences end up far apart. This confirms that matrix factorization effectively captures latent semantic relationships using only user's rating patterns.

5. Problem 5 : Interpreting Dimension

Dimension 50:
High values: movies like Congo (1995), Pretty Woman (1990), Lone Ranger, The (2013), Mission: Impossible II (2000), Nightmare on Elm Street (1984). These are mostly action/adventure films, some with romance or thriller elements, and a few comedies.
Low values: movies like Casino (1995), Big Short, The (2015), Panic Room (2002), Clockwork Orange (1971), Clerks (1994). These are often dramas, crime films, or darker/thriller-oriented movies.
The hypothesis for his dimension seems to represent the contrast between high-energy action/adventure films and serious/drama or crime/thriller films.

	title	Genres	Value
0	Congo (1995)	Action|Adventure|Mystery|Sci-Fi	0.572287
1	Pretty Woman (1990)	Comedy|Romance	0.504086
2	Lone Ranger, The (2013)	Action|Adventure|Western|IMAX	0.496388
3	Smokey and the Bandit II (1980)	Action|Comedy	0.483361
4	Mission: Impossible II (2000)	Action|Adventure|Thriller	0.481545
5	Nightmare on Elm Street, A (1984)	Horror|Thriller	0.469321
6	Mr. Mom (1983)	Comedy|Drama	0.441950
7	School of Rock (2003)	Comedy|Musical	0.436008
8	World Is Not Enough, The (1999)	Action|Adventure|Thriller	0.431771
9	Harry Potter and the Sorcerer's Stone (a.k.a. ...	Adventure|Children|Fantasy	0.425314
...	...	...	...
9714	Suspiria (1977)	Horror	-0.449088
9715	Happiness (1998)	Comedy|Drama	-0.449988
9716	V for Vendetta (2006)	Action|Sci-Fi|Thriller|IMAX	-0.454454
9717	Clerks (1994)	Comedy	-0.457842
9718	Clockwork Orange, A (1971)	Crime|Drama|Sci-Fi|Thriller	-0.467499
9719	Sideways (2004)	Comedy|Drama|Romance	-0.469008
9720	Panic Room (2002)	Thriller	-0.495790
9721	Mars Attacks! (1996)	Action|Comedy|Sci-Fi	-0.496441
9722	Big Short, The (2015)	Drama	-0.513866
9723	Casino (1995)	Crime|Drama	-0.528705

9724 rows × 3 columns

Dimension 120:
High values: movies like Outbreak (1995), RoboCop (1987), Broken Arrow (1996), Predator (1987), True Lies (1994). Most are action, sci-fi, or thriller films with high tension and excitement.
Low values: movies like Matrix Revolutions, The (2003), Donnie Darko (2001), Clockwork Orange (1971), The Revenant (2015), Shutter Island (2010). These include darker sci-fi, psychological thrillers, or intense dramas.
The hypothesis for this dimension seems to capture mainstream action/sci-fi intensity vs. darker or psychologically complex films.

	title	Genres	Value
0	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller	0.541891
1	RoboCop (1987)	Action|Crime|Drama|Sci-Fi|Thriller	0.517581
2	Broken Arrow (1996)	Action|Adventure|Thriller	0.507599
3	Predator (1987)	Action|Sci-Fi|Thriller	0.493076
4	Son in Law (1993)	Comedy|Drama|Romance	0.475938
5	True Lies (1994)	Action|Adventure|Comedy|Romance|Thriller	0.475819
6	Face/Off (1997)	Action|Crime|Drama|Thriller	0.474976
7	Speed (1994)	Action|Romance|Thriller	0.468088
8	Desperado (1995)	Action|Romance|Western	0.454805
9	Home Alone (1990)	Children|Comedy	0.454046
...	...	...	...
9714	Ocean's Twelve (2004)	Action|Comedy|Crime|Thriller	-0.451436
9715	Green Mile, The (1999)	Crime|Drama	-0.467814
9716	Dead Poets Society (1989)	Drama	-0.472763
9717	Underworld: Evolution (2006)	Action|Fantasy|Horror	-0.475096
9718	Beach, The (2000)	Adventure|Drama	-0.479978
9719	Shutter Island (2010)	Drama|Mystery|Thriller	-0.480602
9720	The Revenant (2015)	Adventure|Drama	-0.503373
9721	Clockwork Orange, A (1971)	Crime|Drama|Sci-Fi|Thriller	-0.513437
9722	Donnie Darko (2001)	Drama|Mystery|Sci-Fi|Thriller	-0.519439
9723	Matrix Revolutions, The (2003)	Action|Adventure|Sci-Fi|Thriller|IMAX	-0.524960

9724 rows × 3 columns

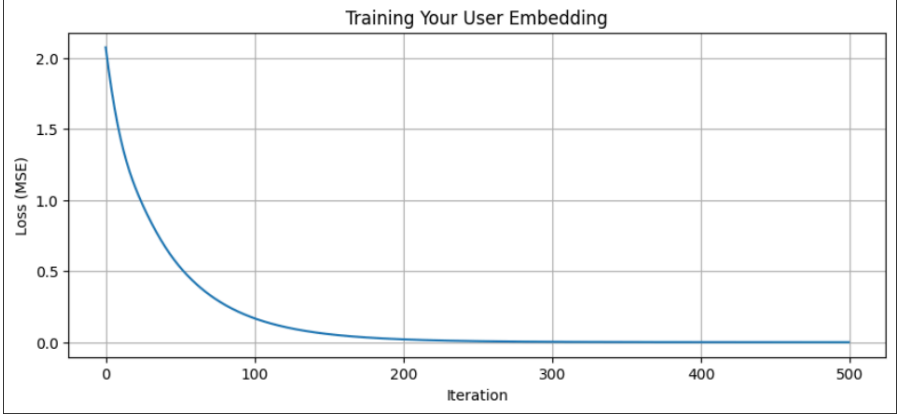
6. Problem 6: Train Your Own Personalized Embedding

For my movie rating I tried to select in priority movies that I really like, and some movie that i really didn't like :

```
my_ratings = {
  1: 4.0,   # Jumanji (1995)
  914: 5.0,   # Alien (1979)
  901: 4.5,   # Aliens (1986)
  4099: 4,   # Treasure Planet (2002)
  8954: 1,   # Dragon Ball Z Gaiden: The Plot to Destroy the Saiyans (1993)
  8998: 1,   # Doctor Who: The Waters of Mars (2009)
  9351: 2.5,   # Kubo and the Two Strings (2016)
  9302: 4,   # The Conjuring 2 (2016)
  9603: 3.5,   # Coco (2017)
  9702: 4,   # Mission: Impossible - Fallout (2018)
}
```

We can also see that our model have learned well, the Loss have almost reach 0 after 200 iterations :

```
Iteration 100/500, Loss: 0.1719
Iteration 200/500, Loss: 0.0205
Iteration 300/500, Loss: 0.0028
Iteration 400/500, Loss: 0.0004
Iteration 500/500, Loss: 0.0000
```



My Mean Absolute Error is 0.004, which is extremely low and the predicted ratings almost exactly match my actual ratings :

```
How well does your embedding fit your ratings?
                                                 Movie  Your Rating  Predicted Rating  Difference
                                        Jumanji (1995)          4.0              4.00       -0.00
                                         Alien (1979)          5.0              4.99       -0.01
                                        Aliens (1986)          4.5              4.50        0.00
                                    Treasure Planet (2002)          4.0              4.00       -0.00
Dragon Ball Z Gaiden: The Plot to Destroy the Saiyans (1993)          1.0              1.01        0.01
                      Doctor Who: The Waters of Mars (2009)          1.0              1.01        0.01
                           Kubo and the Two Strings (2016)          2.5              2.50        0.00
                                The Conjuring 2 (2016)          4.0              4.00       -0.00
                                         Coco (2017)          3.5              3.50        0.00
                       Mission: Impossible - Fallout (2018)          4.0              4.00       -0.00

Mean Absolute Error: 0.004
```

As for my Top 20 recommendation, those movies for some of them reflect my preference. For example, Groundhog Day and Jackass were probably influenced by Jumanji as they are comedy/adventure movies. However, many movies are not reflecting my preference, as I have put a 5 rating for the movie Alien which is a sci-fi/horror, but get none with those themes, when I could have gotten another Alien movie like Alien Covenant. All movie's predicted ratings are about 4.5, while I could have rated some of them at 2, like You've Got Mail (1998), which surprised me that it got on my recommendation as I don't have rated a movie similar to it. It is probably due to the genre comedy that appear in coco, but the genre romance doesn't belong in my recommendation :

```
================================================================================
YOUR TOP 20 RECOMMENDED MOVIES (that you haven't rated yet)
================================================================================
 Rank                                          Movie                                        Genres  Predicted Rating
    1                                Groundhog Day (1993)                       Comedy|Fantasy|Romance              4.91
    2                    Hunt for Red October, The (1990)                     Action|Adventure|Thriller              4.87
    3                                   Gladiator (2000)                        Action|Adventure|Drama              4.85
    4                             Jackass: The Movie (2002)                   Action|Comedy|Documentary              4.78
    5          Grave of the Fireflies (Hotaru no haka) (1988)                          Animation|Drama|War              4.78
    6     Brotherhood of the Wolf (Pacte des loups, Le) (2001)                       Action|Mystery|Thriller              4.75
    7                          Million Dollar Baby (2004)                                       Drama              4.74
    8                 Day the Earth Stood Still, The (1951)                        Drama|Sci-Fi|Thriller              4.74
    9                           What Lies Beneath (2000)                           Drama|Horror|Mystery              4.69
   10                                 Tombstone (1993)                           Action|Drama|Western              4.66
   11                  Bridge on the River Kwai, The (1957)                            Adventure|Drama|War              4.65
   12                                  Contact (1997)                                   Drama|Sci-Fi              4.61
   13                    Evil Dead II (Dead by Dawn) (1987)                  Action|Comedy|Fantasy|Horror              4.61
   14                      Untouchables, The (1987)                              Action|Crime|Drama              4.61
   15                        Black Hawk Down (2001)                                 Action|Drama|War              4.60
   16 For a Few Dollars More (Per qualche dollaro in più) (1965)                Action|Drama|Thriller|Western              4.60
   17                               Excalibur (1981)                                  Adventure|Fantasy              4.60
   18                               Inception (2010) Action|Crime|Drama|Mystery|Sci-Fi|Thriller|IMAX              4.59
   19                                Die Hard (1988)                             Action|Crime|Thriller              4.59
   20                            You've Got Mail (1998)                                   Comedy|Romance              4.59
```

The user that I feel like I share the most preference with is the 3rd user with the most similarity out of the 5, so user 13, I like most of his Top 5 ratings and some of his Top 5 lowest ratings. Theme that probably made me prefer this user is Action, as I like all the action movie he Top rated :

```
================================================================================
User 13 (Similarity: 0.238)
================================================================================

Top 5 Highest Ratings:
                                                                title                 genres  rating
                                                Seven (a.k.a. Se7en) (1995)       Mystery|Thriller     5.0
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)       Action|Adventure     5.0
                                                      Matrix, The (1999) Action|Sci-Fi|Thriller     5.0
                                                           Snatch (2000)  Comedy|Crime|Thriller     5.0
                        Crouching Tiger, Hidden Dragon (Wo hu cang long) (2000)   Action|Drama|Romance     5.0

Top 5 Lowest Ratings:
                             title         genres  rating
          Gone in 60 Seconds (2000)   Action|Crime     3.0
            Charlie's Angels (2000)  Action|Comedy     3.0
            Cruel Intentions (1999)          Drama     2.0
           Final Destination (2000) Drama|Thriller     2.0
Ready to Wear (Pret-A-Porter) (1994)         Comedy     1.0
```

Now we have my personal recommendation, I could say that I will probably like half of those movies, but will probably not appreciate the other half. We can notice that almost all of those movies are pre-2000, even though I rated many movies after 2000 with good rate :

```
================================================================================
MOVIES MOST ALIGNED WITH YOUR TASTE (by embedding similarity)
================================================================================
 Rank                                          Movie                       Genres  Similarity             Rated
    1                        Oblivion 2: Backlash (1996)                       Sci-Fi       0.316
    2                          Jackass: The Movie (2002)    Action|Comedy|Documentary       0.313
    3                                  Newsies (1992)             Children|Musical       0.305
    4                        Rookie of the Year (1993)               Comedy|Fantasy       0.302
    5                                  Repo Men (2010)       Action|Sci-Fi|Thriller       0.301
    6                                   Case 39 (2009)              Horror|Thriller       0.299
    7                       Protector, The (1985) Action|Comedy|Drama|Thriller       0.290
    8                        Goodbye Lover (1999)        Comedy|Crime|Thriller       0.290
    9                      What Lies Beneath (2000)         Drama|Horror|Mystery       0.286
   10          Day the Earth Stood Still, The (1951)        Drama|Sci-Fi|Thriller       0.286
   11                              Blue Chips (1994)                       Drama       0.286
   12                                   Alien (1979)                Horror|Sci-Fi       0.286 ✓ (You rated this)
   13                                   Steel (1997)                      Action       0.285
   14 Brotherhood of the Wolf (Pacte des loups, Le) (2001)      Action|Mystery|Thriller       0.284
   15                                  Selena (1997)                Drama|Musical       0.284
   16       Mortal Kombat: The Journey Begins (1995)             Action|Animation       0.279
   17              Hunt for Red October, The (1990)    Action|Adventure|Thriller       0.275
   18                       DOA: Dead or Alive (2006)             Action|Adventure       0.275
   19                     Million Dollar Baby (2004)                       Drama       0.274
   20                           Groundhog Day (1993)       Comedy|Fantasy|Romance       0.274

================================================================================
Notice: These movies define your 'type' in the embedding space!
================================================================================
```

Recommendation systems, like the embedding-based model we trained, learn to represent both users and movies in a shared latent space. Each user's embedding encodes our tastes and preferences as patterns in this multidimensional space (action vs. comedy vs. sci-fi, classic vs. modern style, etc.). Movie embeddings encode characteristics of each film, not just genres, but also patterns learned from other user's ratings.The predicted rating comes from the similarity (dot product) between our user embedding and movie embeddings, meaning the system recommends movies that are close to our "type" in the latent space.

My ratings are enough to position our embedding in the latent space relative to the pretrained movie embeddings. Once our embedding is trained, it automatically generalizes to movies we haven't rated, because similar movies are clustered together in the embedding space. Essentially, the model extrapolates our preferences from a small sample of movies based on the learned relationships in the embedding space.

The top recommendations were largely pre-2000 even though I rated many newer movies highly, indicating that the model is heavily influenced by the initial movies I rated and their nearby embeddings, while some recommendations don't perfectly match my taste because the model overgeneralizes from overlapping genres like comedy but ignores other important aspects such as romance, and with only 10 ratings it struggles to capture the full diversity of my preferences, lacks awareness of the nuanced reasons why I enjoy certain movies like story, tone, or director style, and may also predict very new or rare movies less accurately due to limited exposure in the training data.

This experiment shows that recommendation systems can quickly learn our general taste from a small number of ratings, but they cannot perfectly capture nuanced or recent preferences. They are most effective at identifying broad patterns and suggesting movies similar to what we have already rated. To improve accuracy, we would need to rate a more diverse set of movies across genres, eras, and styles.

Jean Baptiste Felici G20250339