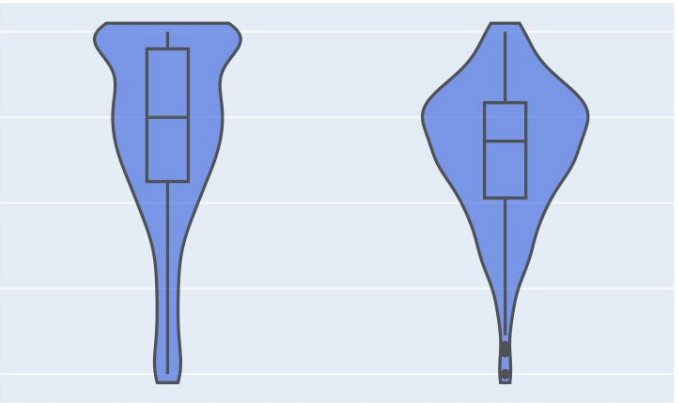
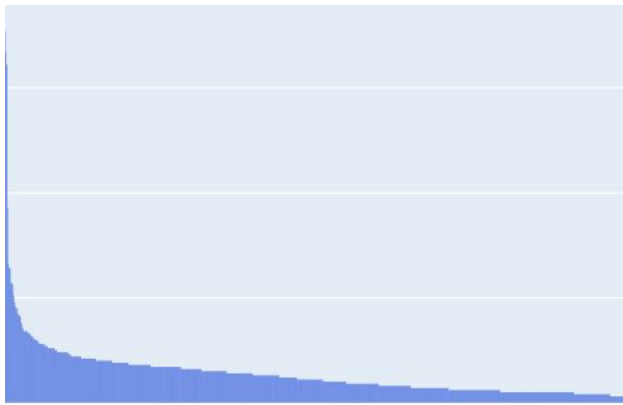
Web Science:
Amazon software recommendation

Andreas Bjerregaard

April, 2022





Outline

1 Data

Review data

	overall	reviewerID	asin	unixReviewTime
0	4.0	ARXU3FESTWMJJ	B00003JAU7	966988800
1	1.0	AE95Z3K6GVIC3	B00003JAU7	977443200
2	5.0	A1P4RH7KMJ1SV2	B000050HEI	978566400
3	5.0	A1P4RH7KMJ1SV2	B00003IRBV	980294400
4	4.0	A1P4RH7KMJ1SV2	B00003IRBU	980985600
...	...	...	...	...
10166	5.0	A24ET4BPOVKRHJ	B00L13X6QA	1530489600
10167	3.0	A3W4D8XOGLWUN5	B00UVTEJ7K	1531612800
10168	1.0	A2MKY8OUI8GZG1	B00HV9IM58	1531785600
10169	5.0	A3935GZFLPU28D	B015724V9Q	1532217600
10170	1.0	ACMXHJV1KCOSV	B015724V9Q	1533945600

10171 rows × 4 columns

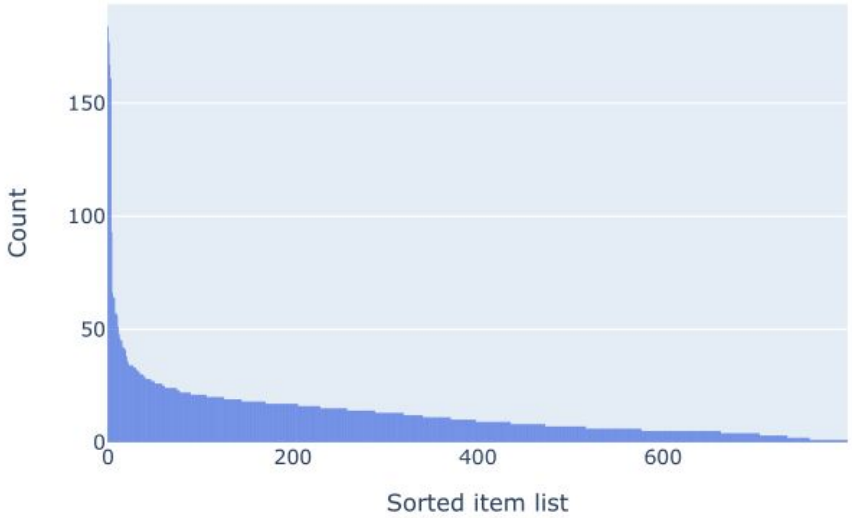
1711 rows × 4 columns

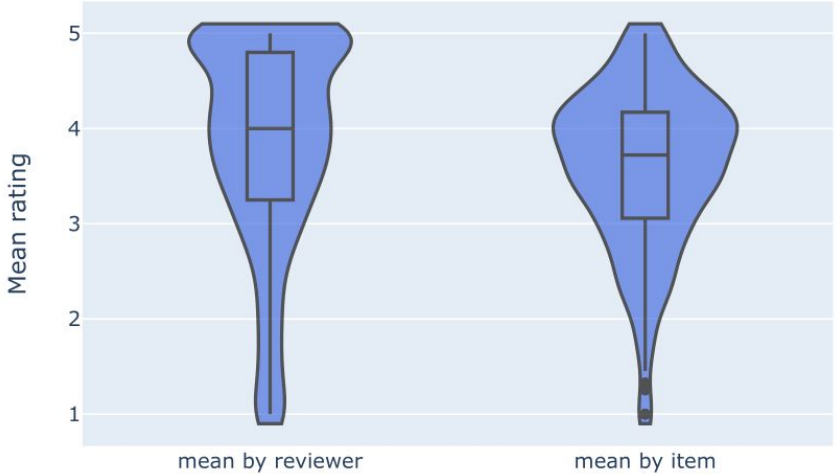
Metadata

category	description	title	brand	price	asin
[]	[This complete training program from Adobe Pre...	Learn Adobe Photoshop Lightroom 3 by Video	Peach Pit Press	$24.99	0321700945
[]	[This complete training program from Adobe Pre...	Learn Adobe Dreamweaver CS5 by Video: Core Tra...	Peach Pit Press	$35.23	0321719816
[]	[This complete training program from Adobe Pre...	Learn Adobe Flash Professional CS5 by Video: C...	Peach Pit Press		0321719824
are, Business & Office, Office Suites]	[Office 365 comes fully loaded with the latest...	Microsoft Office 365 Home | 1-year subscriptio...	Microsoft	$90.05	0763855553
Software, Education & Reference, Religion]	[Glo features five interactive browsing lenses...	NIV, GLO Premium, DVD: Multi-device	Immersion Digital		0982697813
...	...	...	...	...	...
re, Business & Office, Word Processing]	[Microsoft Office 2010 gives you powerful new ...	Microsoft Office Home and Student 2010 Family ...	Microsoft	$259.99	B01F7RJHIQ
[Software, Photography]	[Make your best work even better with Corel Af...	Corel AfterShot Pro 3 Photo Editing Software f...	Corel	$44.00	B01FFVDY9M
are, Digital Software, Video, Compositin...	[, onlineTV gives you access to hundreds ...	onlineTV Free [Download]	concept/design GmbH		B01H39M7ME
[Software, Photography]	[Get the power, creativity and control you nee...	Pinnacle Studio 20 Ultimate (Old Version)	Pinnacle Systems	$35.64	B01HAP47PQ
[Software, Photography]	[Create your best videos with the pro-quality ...	Pinnacle Studio 20 Plus (Old Version)	Pinnacle Systems	$34.99	B01HAP3NUG

801 rows × 6 columns

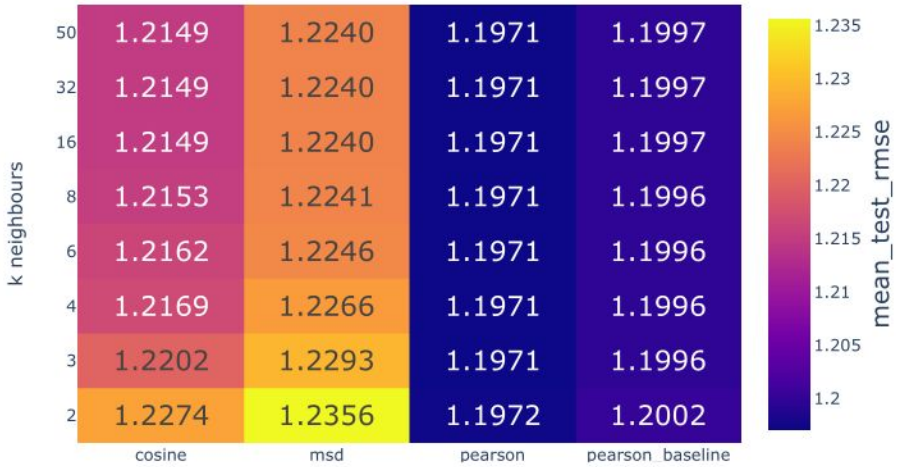
1 Data

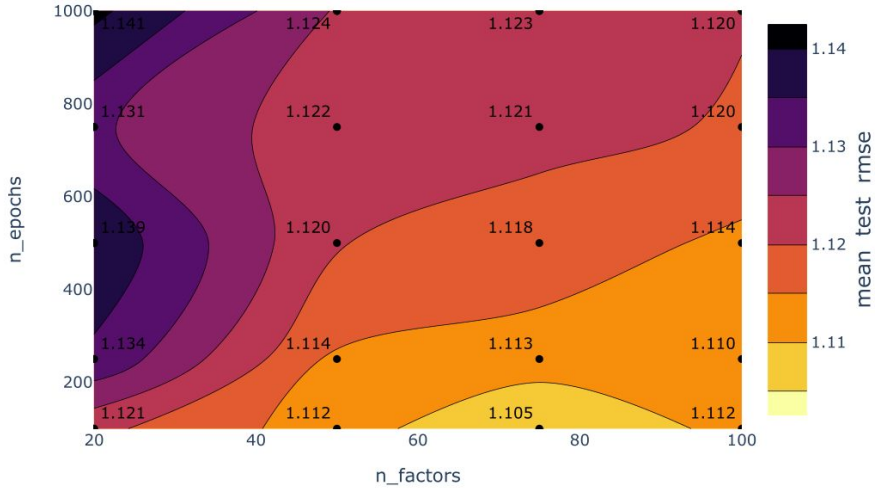

Count
150
100
50
0
0
200
400
600
Sorted item list


Mean rating
5
4
3
2
1
mean by reviewer
mean by item

2 Collaborative filtering

Model	Val RMSE	Test RMSE	Optimal hyperparameters
kNNWithMeans	1.1971	1.1497	`k_neighbours: 4, measure: 'pearson'`
SVD	1.1048	0.9972	`n_epochs: 100, n_factors: 75, lr_all: 0.15`

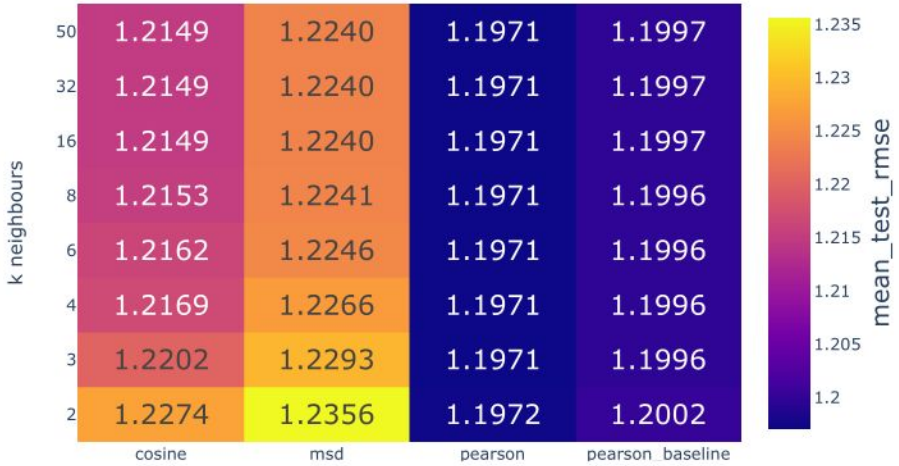


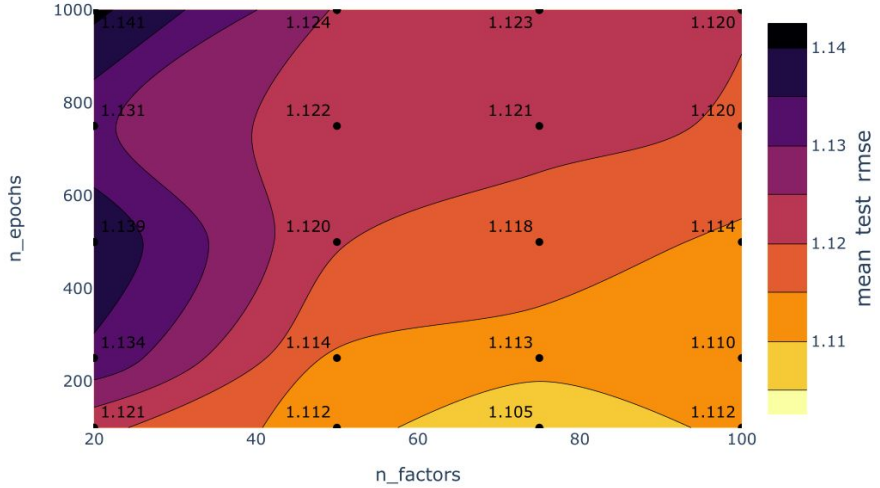


		P@5	MAP@5	MRR@5	HR@5
CF	KNNWithMeans	0.0023	0.0070	0.0070	0.0117
	SVD	0.0153	0.0172	0.0172	0.0766

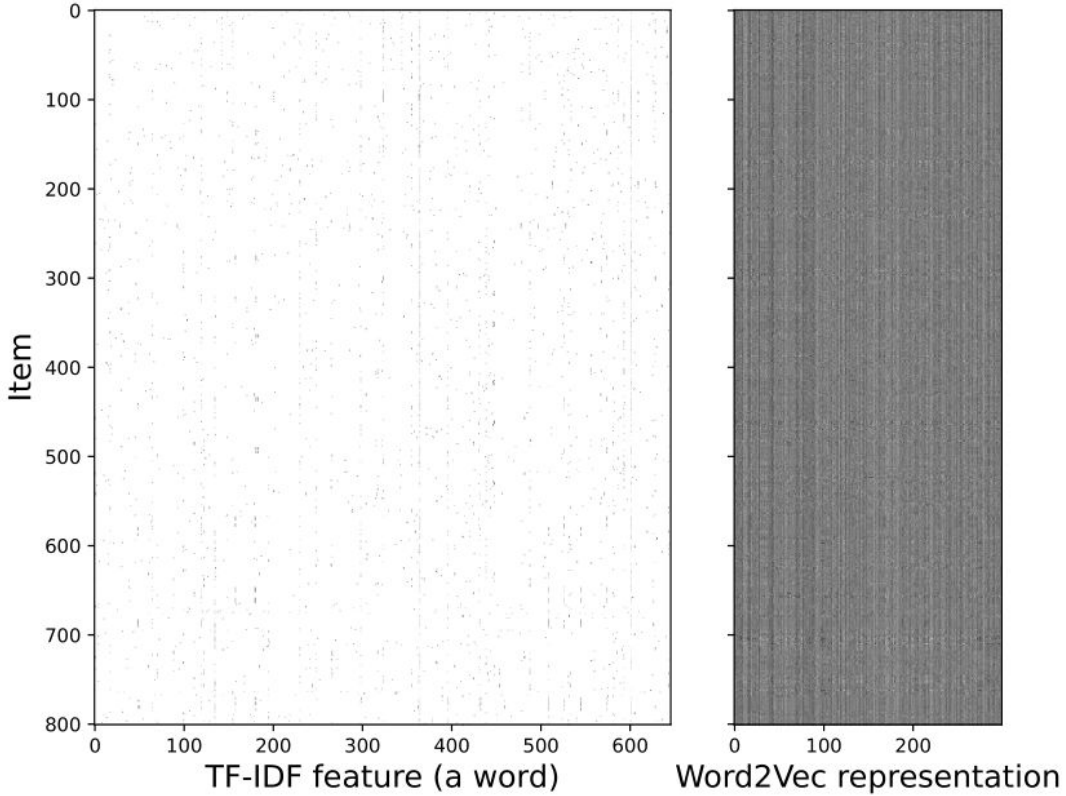
2 Collaborative filtering

Model	Val RMSE	Test RMSE	Optimal hyperparameters
kNNWithMeans	1.1971	1.1497	`k_neighbours: 4, measure: 'pearson'`
SVD	1.1048	0.9972	`n_epochs: 100, n_factors: 75, lr_all: 0.15`





3 Content-based systems

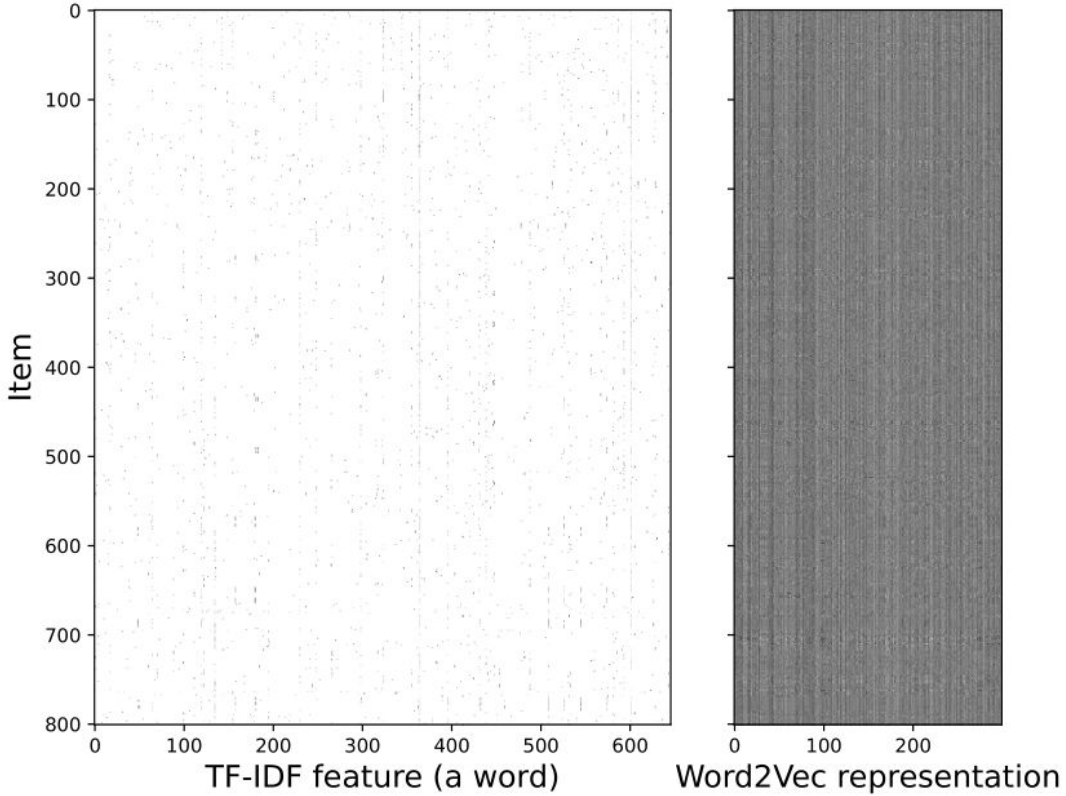


	TF-IDF			Word2Vec		
	Item1	**Item2**	**Item3**	**Item1**	**Item2**	**Item3**
Item1	1.00	0.06	0.06	1.00	0.66	0.64
Item2	0.06	1.00	0.74	0.66	1.00	0.92
Item3	0.06	0.74	1.00	0.64	0.92	1.00

	ASIN	**Title**
Item1	B001DSGXFY	Acronis True Image Home 2009 [OLD VERSION]
Item2	B01HAP47PQ	Pinnacle Studio 20 Ultimate (Old Version)
Item3	B01HAP3NUG	Pinnacle Studio 20 Plus (Old Version)

3 Content-based systems

		P@5	MAP@5	MRR@5	HR@5
CF	KNNWithMeans	0.0023	0.0070	0.0070	0.0117
	SVD	0.0153	0.0172	0.0172	0.0766
Content	TFIDF	0.0289	0.1039	0.1039	0.1444
	Word2Vec	0.0289	0.0808	0.0808	0.1444
	TFIDF-W2V	0.0309	0.0779	0.0779	0.1543
	TFIDF-W2V-all-text	**0.0393**	**0.1441**	**0.1441**	**0.1964**



	TF-IDF			Word2Vec		
	Item1	**Item2**	**Item3**	**Item1**	**Item2**	**Item3**
Item1	1.00	0.06	0.06	1.00	0.66	0.64
Item2	0.06	1.00	0.74	0.66	1.00	0.92
Item3	0.06	0.74	1.00	0.64	0.92	1.00

	ASIN	**Title**
Item1	B001DSGXFY	Acronis True Image Home 2009 [OLD VERSION]
Item2	B01HAP47PQ	Pinnacle Studio 20 Ultimate (Old Version)
Item3	B01HAP3NUG	Pinnacle Studio 20 Plus (Old Version)

4 Hybrid recommendations

Weighting

- $\alpha = \beta$
- $2\alpha = \beta$

Switching

- *on user*
- *on item*

Meta-level

- kNN on user encodings

4 Hybrid recommendations

Weighting

- $\alpha = \beta$
- $2\alpha = \beta$

Switching

- *on user*
- *on item*

Meta-level

- kNN on user encodings

		P@5	MAP@5	MRR@5	HR@5
CF	KNNWithMeans	0.0023	0.0070	0.0070	0.0117
	SVD	0.0153	0.0172	0.0172	0.0766
Content	TFIDF	0.0289	0.1039	0.1039	0.1444
	Word2Vec	0.0289	0.0808	0.0808	0.1444
	TFIDF-W2V	0.0309	0.0779	0.0779	0.1543
	TFIDF-W2V-all-text	**0.0393**	**0.1441**	**0.1441**	**0.1964**
Hybrid	RRF	0.0275	0.1117	0.1117	0.1373
	Normalized Sum	0.0330	0.1244	0.1244	0.1648
	Weighted Sum	0.0328	0.1254	0.1254	0.1642
	USwitch at 25	0.0367	0.1541	0.1541	0.1835
	ISwitch at 50	**0.0386**	**0.1592**	**0.1592**	**0.1929**
	ISwitch at 100	**0.0385**	**0.1592**	**0.1592**	**0.1923**
	Meta	0.0131	0.0408	0.0408	0.0655

5 Summary

Amazon software recommendation

CF/Content/Hybrid

Metrics

Advantages and limitations

Future directions

		P@5	MAP@5	MRR@5	HR@5
CF	KNNWithMeans	0.0023	0.0070	0.0070	0.0117
	SVD	0.0153	0.0172	0.0172	0.0766
Content	TFIDF	0.0289	0.1039	0.1039	0.1444
	Word2Vec	0.0289	0.0808	0.0808	0.1444
	TFIDF-W2V	0.0309	0.0779	0.0779	0.1543
	TFIDF-W2V-all-text	**0.0393**	**0.1441**	**0.1441**	**0.1964**
Hybrid	RRF	0.0275	0.1117	0.1117	0.1373
	Normalized Sum	0.0330	0.1244	0.1244	0.1648
	Weighted Sum	0.0328	0.1254	0.1254	0.1642
	USwitch at 25	0.0367	0.1541	0.1541	0.1835
	ISwitch at 50	**0.0386**	**0.1592**	**0.1592**	**0.1929**
	ISwitch at 100	**0.0385**	**0.1592**	**0.1592**	**0.1923**
	Meta	0.0131	0.0408	0.0408	0.0655

Results table

		P@5	MAP@5	MRR@5	HR@5	P@15	MAP@15	MRR@15	HR@15
CF	KNNWithMeans	0.0023	0.0070	0.0070	0.0117	0.0016	0.0085	0.0085	0.0245
	SVD	0.0153	0.0172	0.0172	0.0766	0.0092	0.0264	0.0264	0.1373
Content	TFIDF	0.0289	0.1039	0.1039	0.1444	0.0150	0.1133	0.1133	0.2250
	Word2Vec	0.0289	0.0808	0.0808	0.1444	0.0139	0.0879	0.0879	0.2092
	TFIDF-W2V	0.0309	0.0779	0.0779	0.1543	0.0153	0.0863	0.0863	0.2297
	TFIDF-W2V-all-text	**0.0393**	**0.1441**	**0.1441**	**0.1964**	**0.0176**	**0.1514**	**0.1514**	**0.2636**
Hybrid	RRF	0.0275	0.1117	0.1117	0.1373	0.0146	0.1202	0.1202	0.2192
	Normalized Sum	0.0330	0.1244	0.1244	0.1648	0.0159	0.1325	0.1325	0.2390
	Weighted Sum	0.0328	0.1254	0.1254	0.1642	0.0161	0.1337	0.1337	0.2414
	USwitch at 25	0.0367	0.1541	0.1541	0.1835	0.0161	0.1602	0.1602	0.2420
	ISwitch at 50	**0.0386**	**0.1592**	**0.1592**	**0.1929**	**0.0175**	**0.1670**	**0.1670**	**0.2624**
	ISwitch at 100	**0.0385**	**0.1592**	**0.1592**	**0.1923**	**0.0178**	**0.1674**	**0.1674**	**0.2665**
	Meta	0.0131	0.0408	0.0408	0.0655	0.0076	0.0448	0.0448	0.1146

Table 4: All implemented models and their ranking-based accuracy scores. The first three content-based models use just the item title, while the last one uses all the metadata text categories. The two last ones further use the three additional numerical features described in Section 3.1. Except for Meta, the hybrid models are all combinations on the SVD and TFIDF-W2V user-item scores.

Time-sorted test set performance

Figure 5: Moving average over mean hit rate for three recommender systems, window size 100.

Full SVD grid search

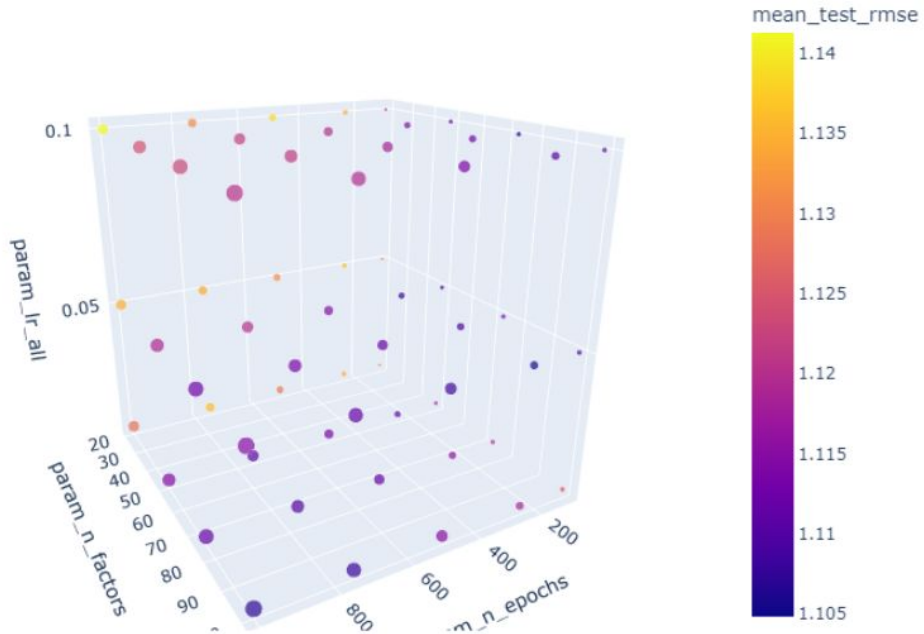


Figure 7: Scatter plot visualising the SVD hyperparameter grid search. Marker size proportional to training time.

9 Main points

Extensive preprocessing gives:

- more robust hyper params
- better performance

Segmenting pores from a single projection is possible

Real data is not simple, spherical and unchanging

… but results appear promising for efficient setups

even across real/synthetic media