INSTITUTO
DE INFORMÁTICA
UFRGS

From Timeout-based to Item-by-Item Analysis: Investigating Methodologies for Splitting User Sessions Originated from Shared Accounts in Online Platforms

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Engineering

Author: Matheus Toazza Tura

Advisor: Prof. Dr. Weverton Luis da Costa Cordeiro

Co-advisor: Prof. Dr. Renata de Matos Galante

May 26th 2021

UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

Summary

1. Overview
2. Privacy vs Accuracy Dilemma
3. Related Work
4. Objectives
5. Used Algorithms
6. Experiments

Overview

Basic concepts

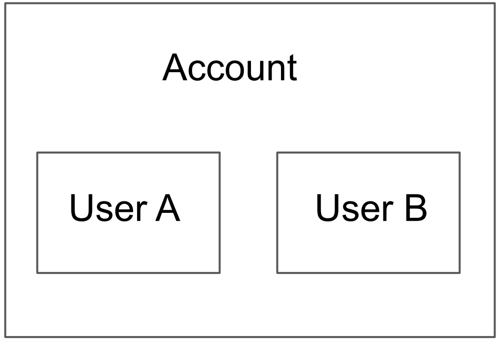
- User
- Account
- Item
- Clickstream
- Session

Basic concepts

- User
- Account
- Item
- Clickstream
- Session

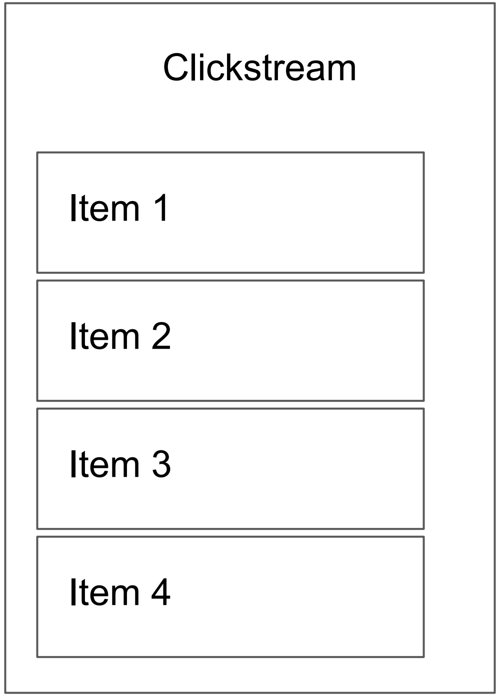
Basic concepts

- User
- Account
- Item
- Clickstream
- Session



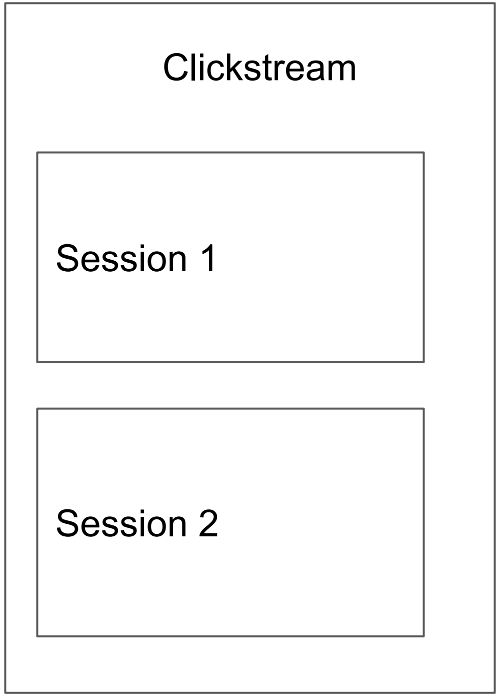
Basic concepts

- User
- Account
- Item
- Clickstream
- Session



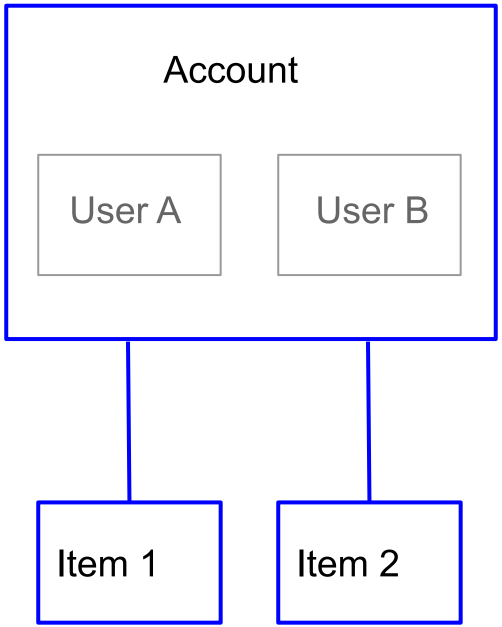
Basic concepts

- User
- Account
- Item
- Clickstream
- Session



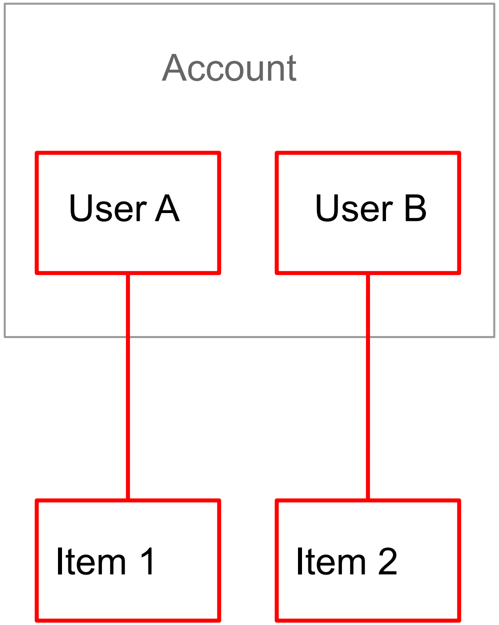
User/Account to Item association

- User
- Account
- Item
- Clickstream
- Session



User/Account to Item association

- User
- Account
- Item
- Clickstream
- Session



Recommender system

Recommender system

- **Implicit** (Click frequency, Ex.: Spotify)
- **Explicit** (Score based Ex.: Netflix)
- **Collaborative-Filtering** (user-Item similarity based)
- **Content-Based** (Item-item similarity based)

Privacy vs Accuracy

Privacy Concerns

Looming Privacy Concerns

Hybrid Warfare

- Elections
- Protests
- Fake News
- Geopolitical interests

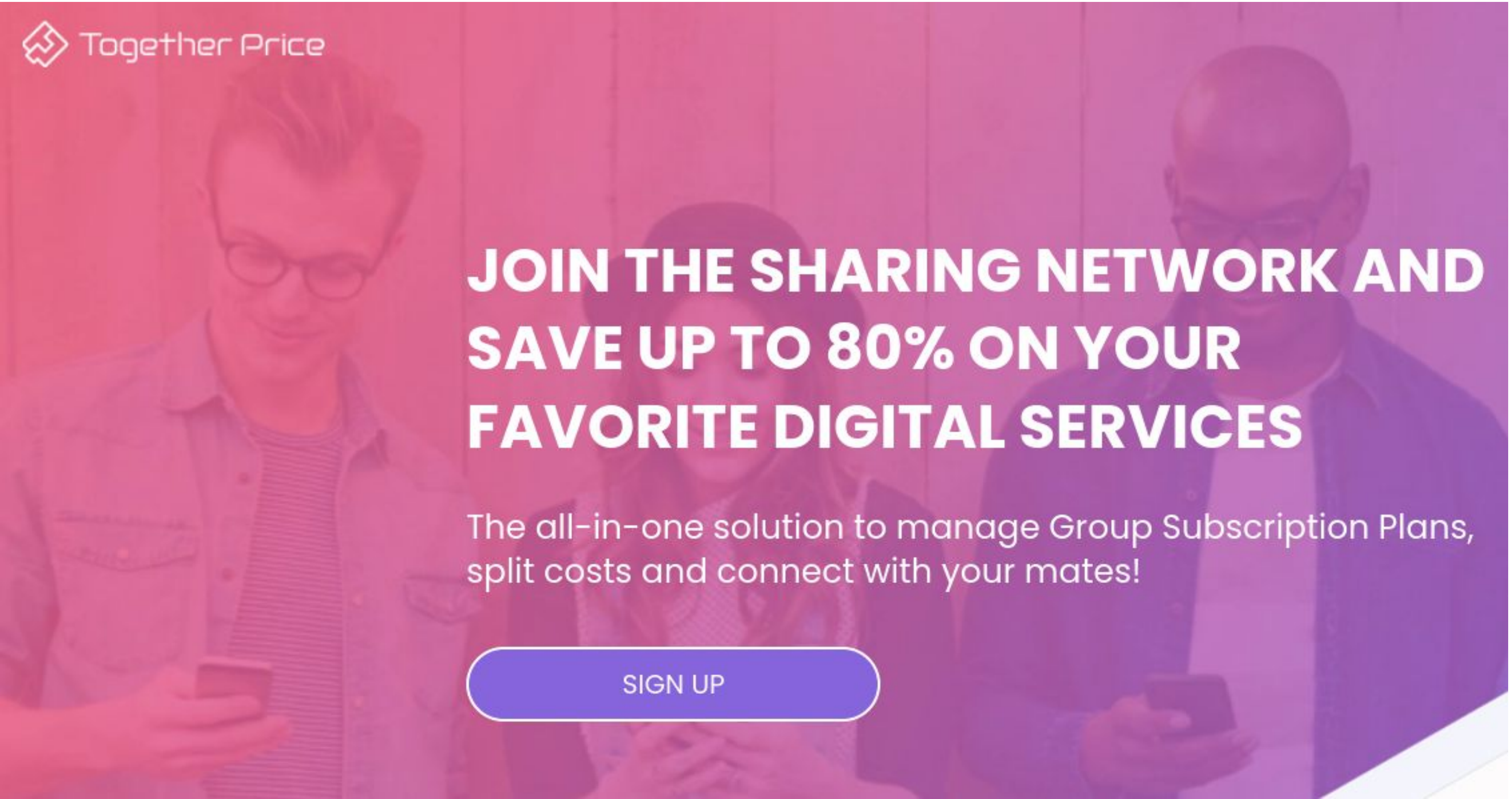
Looming Privacy Concerns

Marketing Abuse

- User profiling and behaviour abuse
- User data black market
- Abusive prices (Online travel booking)

Accuracy Concerns

Account sharing



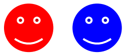
Accuracy problems

- Generality problem
- Dominance problem
- Presentation problem

VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Accuracy problems

- Generality problem



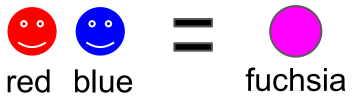
red blue

- Dominance problem
- Presentation problem

VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Accuracy problems

- Generality problem



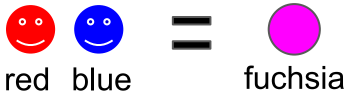
- Dominance problem

- Presentation problem

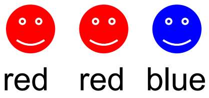
VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Accuracy problems

- Generality problem



- Dominance problem

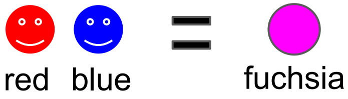


- Presentation problem

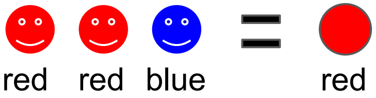
VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Accuracy problems

- Generality problem



- Dominance problem



- Presentation problem

VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Accuracy problems

- Generality problem

red blue = fuchsia

- Dominance problem

red red blue = red

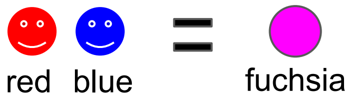
- Presentation problem

red

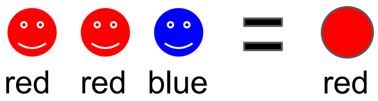
VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Accuracy problems

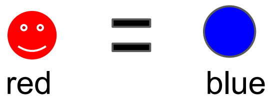
- Generality problem



- Dominance problem



- Presentation problem



VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts. 2015

Dilemma

Related Work

Timeout models

Clickstream Browsing activity

Session 1

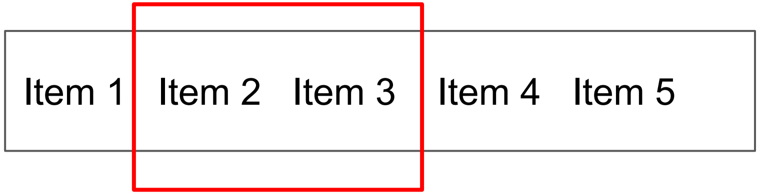
…
30 min
...

Session 2

HALFAKER, A. et al. User session identification based on strong regularities in inter-activity time.

Time-decay models

- Sliding Window
- Neural Network



SOTTOCORNOLA, G.; SYMEONIDIS, P.; ZANKER, M. Session-based news recommendations.

ZHANG, L.; LIU, P.; GULLA, J. Dynamic attention-integrated neural network for session-based news recommendation.

Profile-based models

- Explicit and Implicit recommendation systems

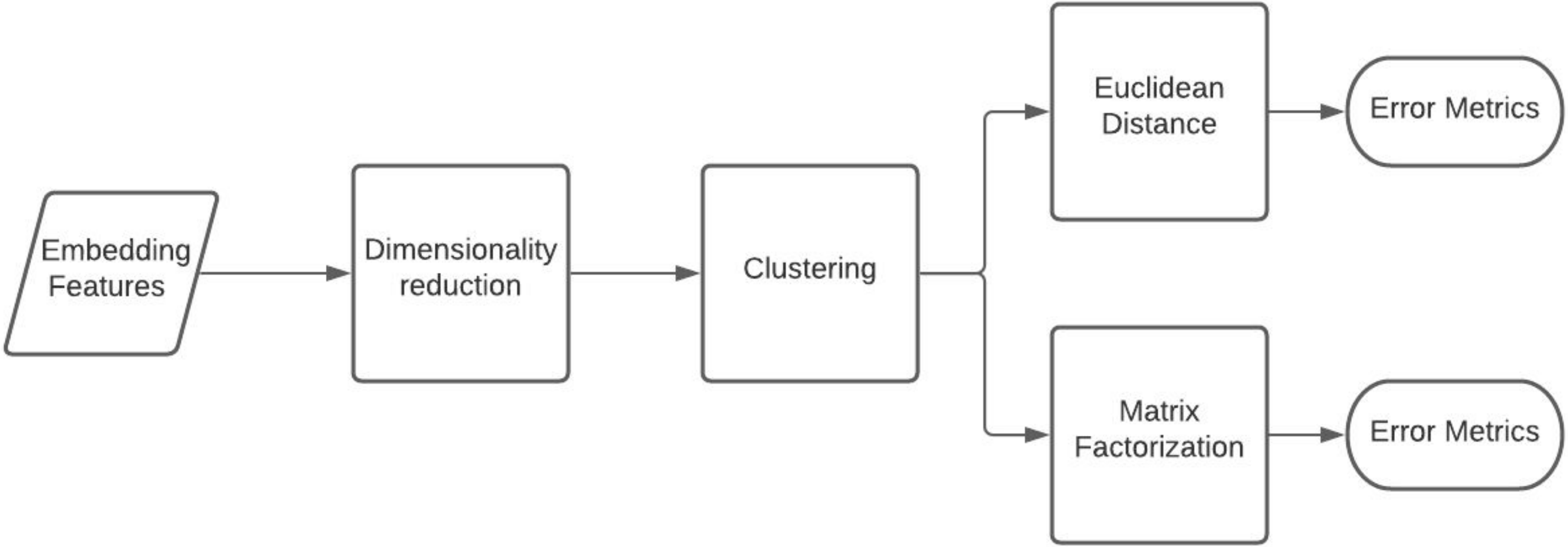
VERSTREPEN, K.; GOETHALS, B. Top-n recommendation for shared accounts.

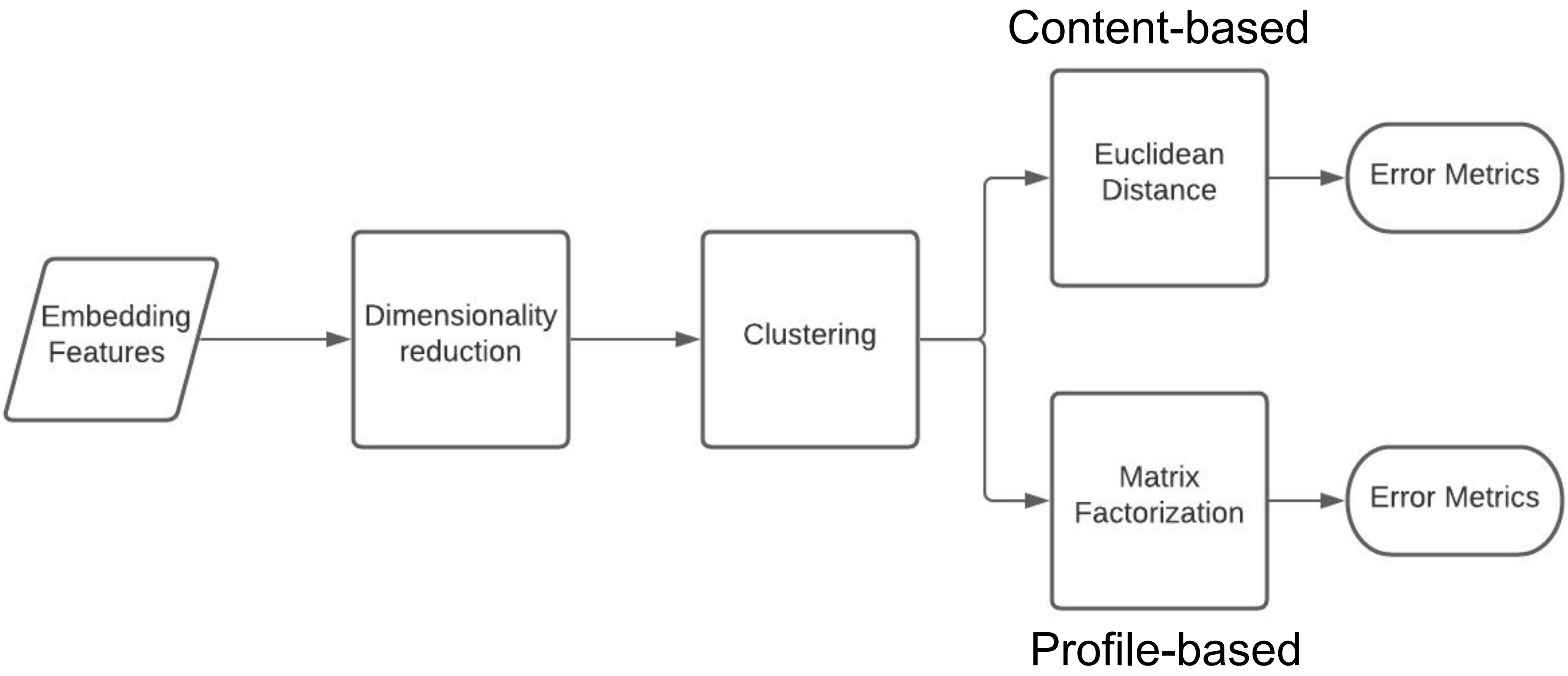
Objectives

- Content-only session splitting

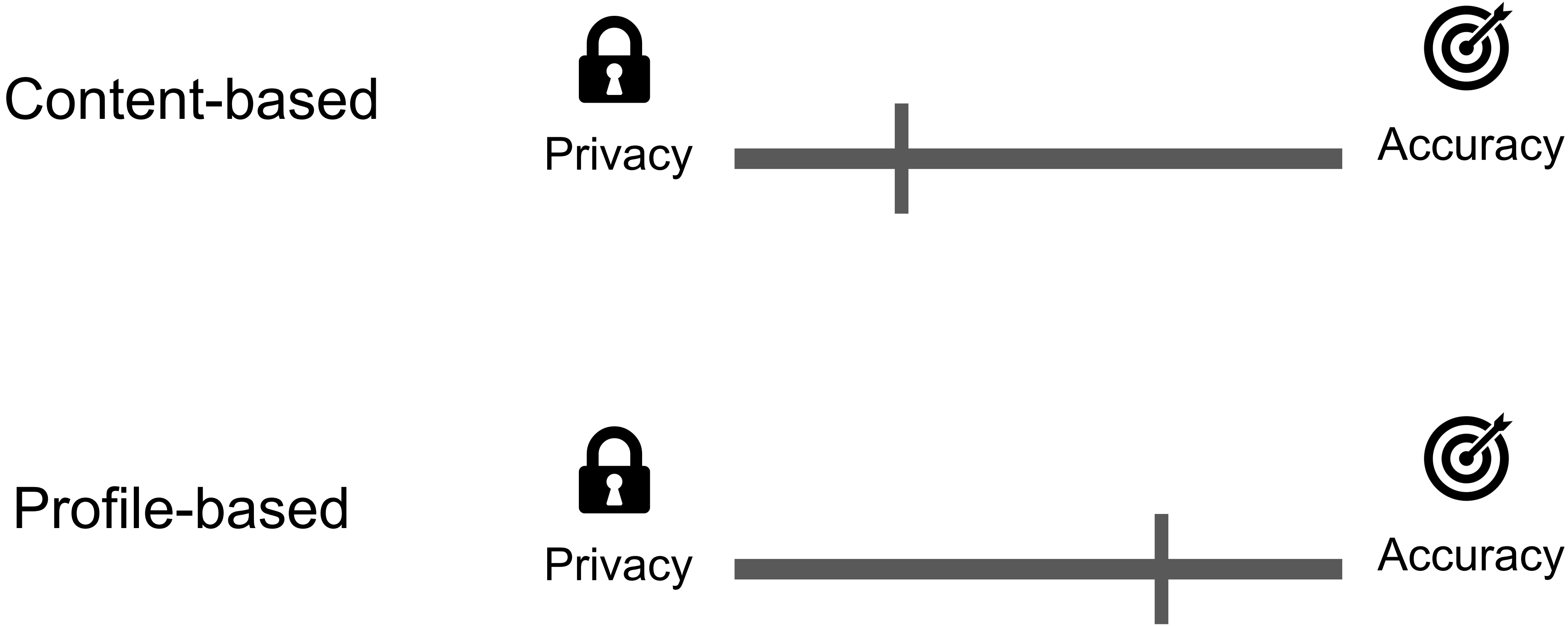
- Full data anonymization

Used algorithms


Embedding Features
Dimensionality reduction
Clustering
Euclidean Distance
Matrix Factorization
Error Metrics
Error Metrics

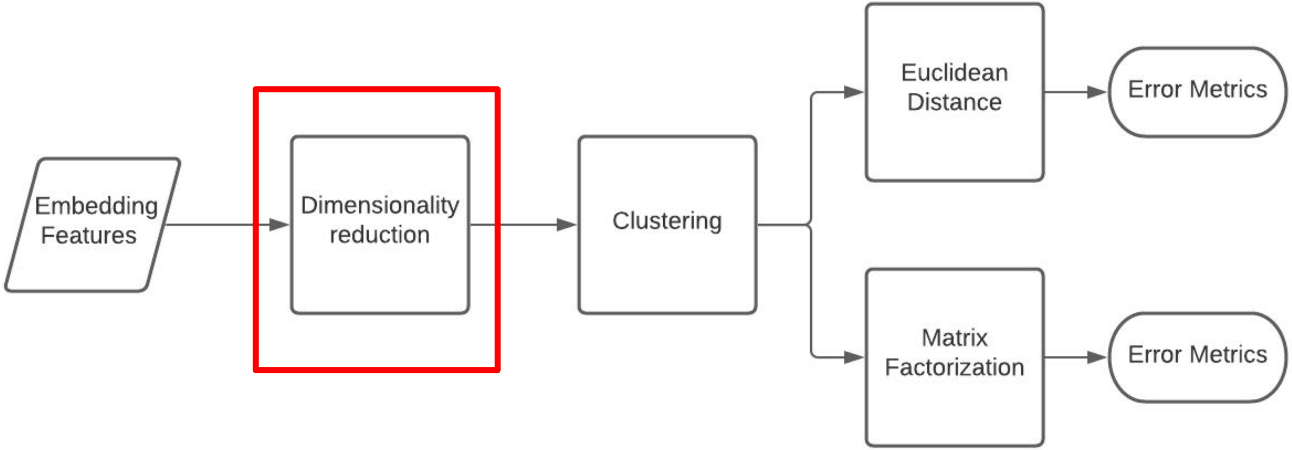


Content-based
Embedding Features
Dimensionality reduction
Clustering
Euclidean Distance
Error Metrics
Matrix Factorization
Error Metrics
Profile-based

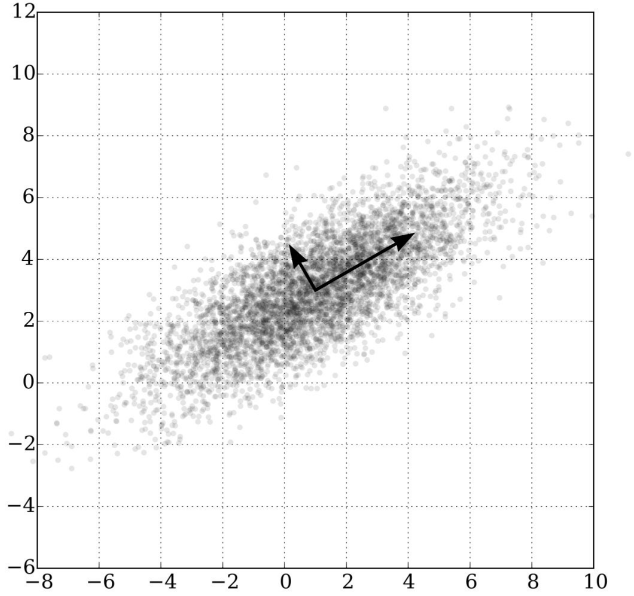


Content-based
Privacy
Accuracy
Profile-based
Privacy
Accuracy

Dimensionality Reduction



PCA - Principal Component Analysis

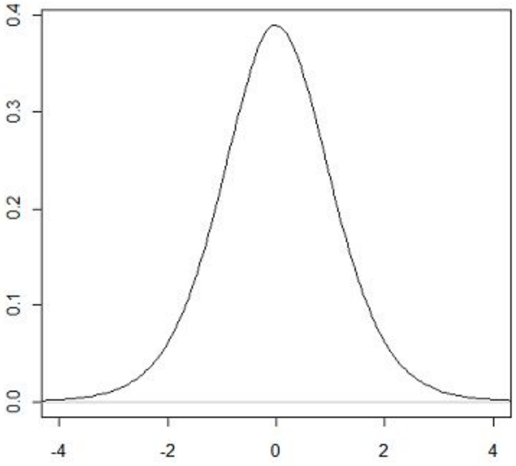


HOTELLING, H. Analysis of a complex of statistical variables into principal components. **1933**

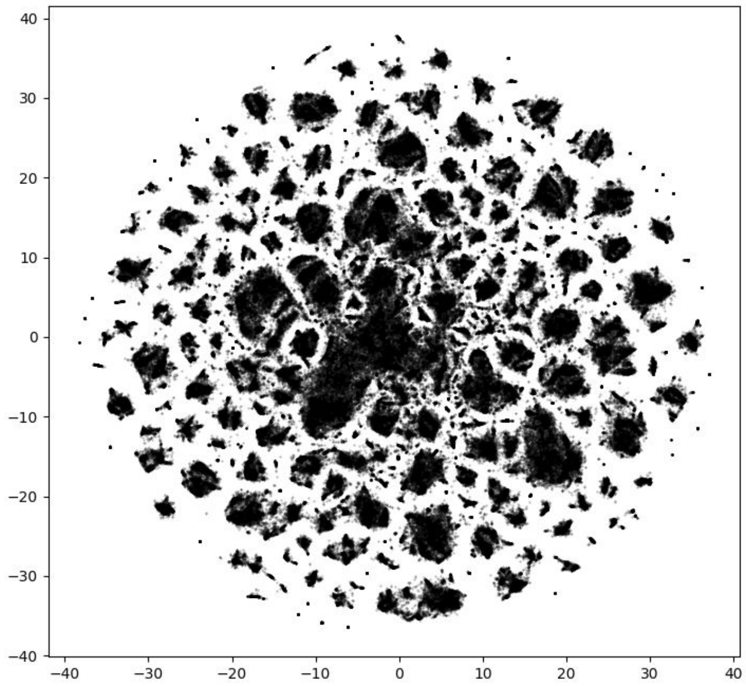
tSNE - t-Distributed Stochastic Neighbor Embedding

Gaussian $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

t-student



tSNE - t-Distributed Stochastic Neighbor Embedding



tSNE - t-Distributed Stochastic Neighbor Embedding

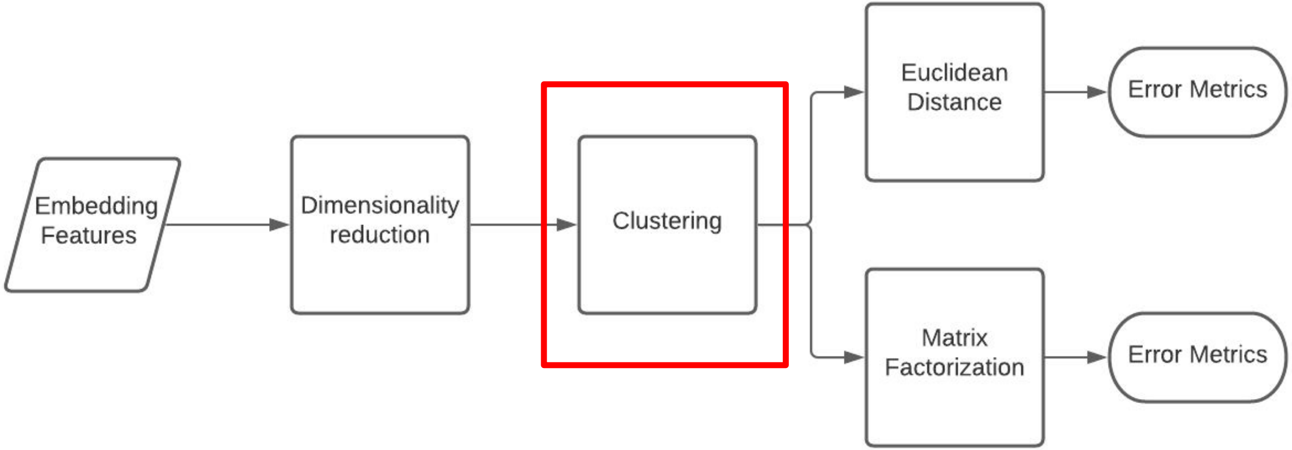
Table 4.1: tSNE coordinate results

article_id	x	y
...	...	...
69	10.950659	-26.211418
81	34.414822	-0.690890
84	35.335995	-1.578238
...	...	...

Data Sparsity and Scarcity

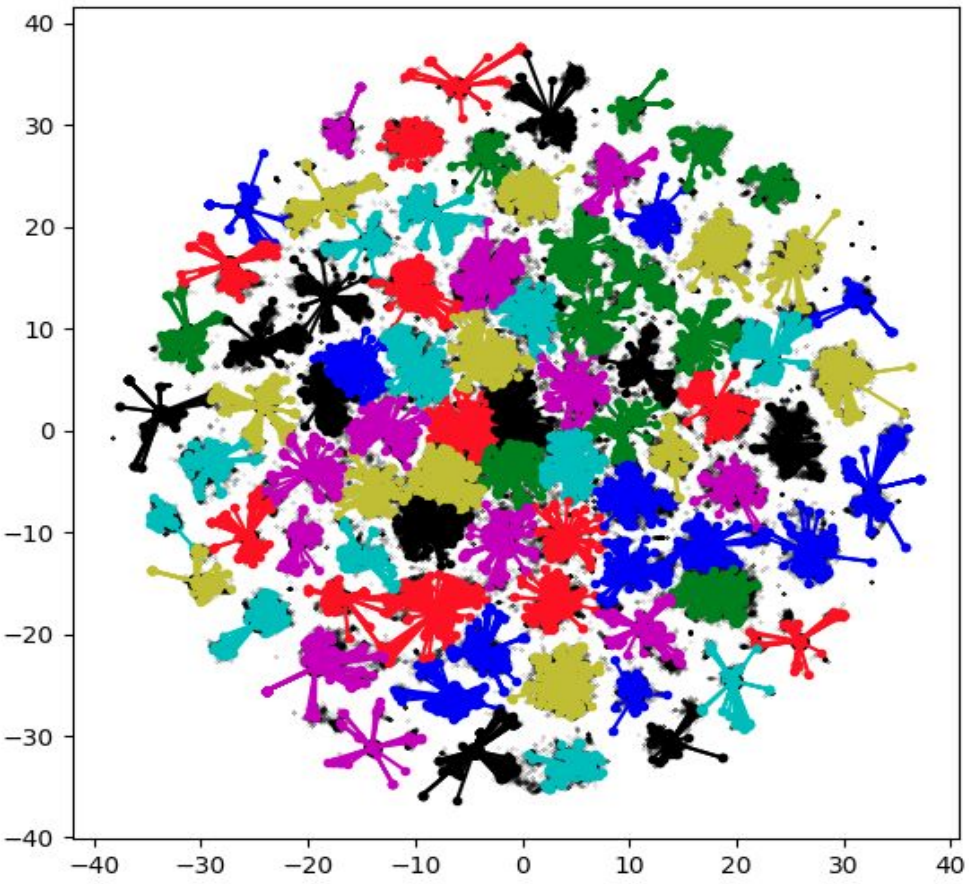
- Cold Start problem
- Curse of Dimensionality

Cluster Analysis



Affinity Propagation

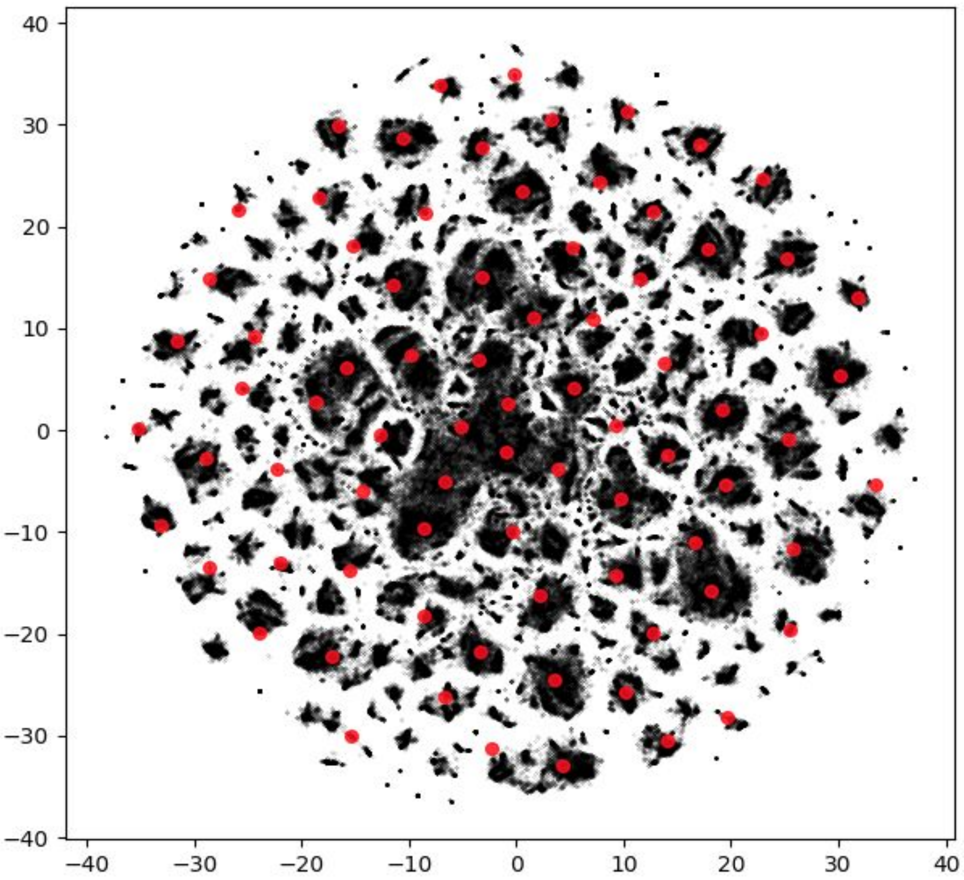
- Message passing algorithm
- Does not require number of clusters



FREY, B. J.; DUECK, D. Clustering by passing messages between data points. Science, v. 315, p. 2007, 2007.

Affinity Propagation

- Message passing algorithm
- Does not require number of clusters



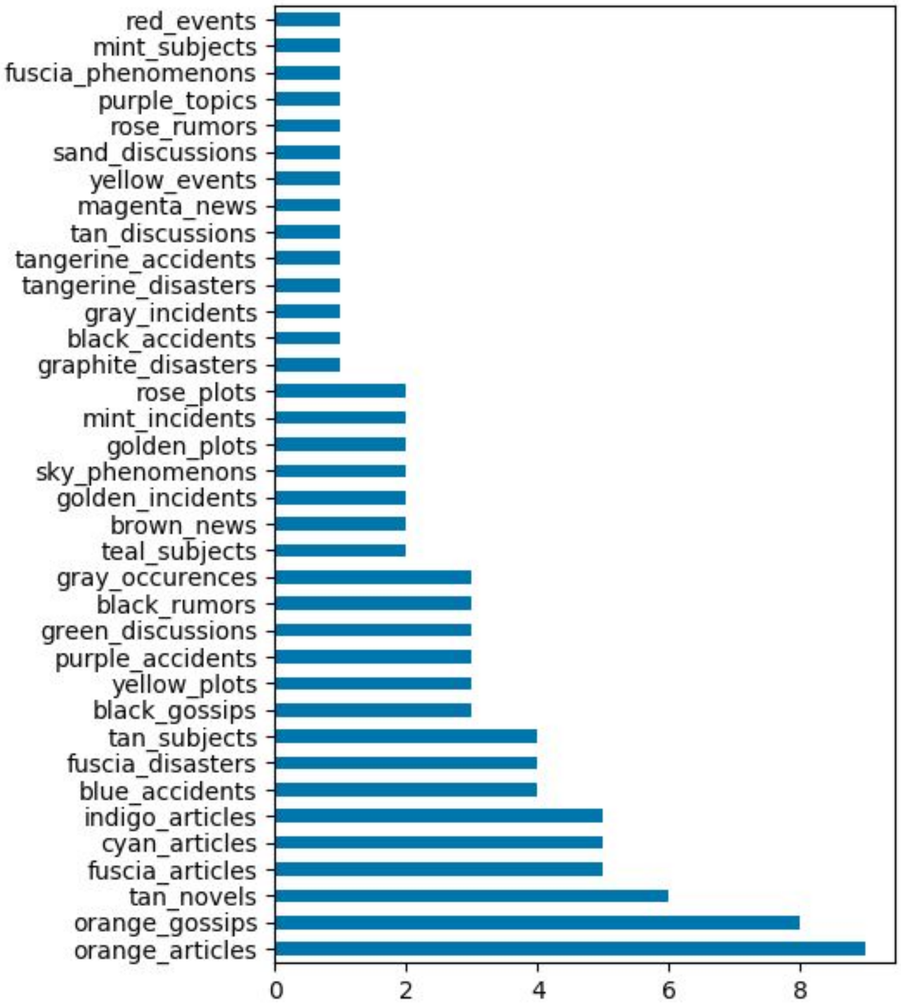
FREY, B. J.; DUECK, D. Clustering by passing messages between data points. Science, v. 315, p. 2007, 2007.

Affinity Propagation

Table 4.2: Labeled cluster centers

x	y	topic
...	...	...
19.699291	-28.117729	tangerine-events
-8.711248	21.906485	purple-occurrences
11.325381	3.893677	yellow-episodes
...	...	...

User topics frequency distribution



Experiments

Globo.com dataset

- Approximately 2.8 millions news clicks
- More than 300 thousand users
- 45 thousand news articles
- 250 dimensional feature matrix

MOREIRA, G. de S. P.; FERREIRA, F.; CUNHA, A. M. da.
News session-based recommendations using deep neural networks.

Globo.com dataset

Table 5.1: Globo dataset sample as Dataframe

click_timestamp	user_id	article_id	click_country	click_region	...
...	...	...	...	...	
1506826800026	59	234853	1	21	
1506826801702	79	159359	1	13	...
1506826804207	154	96663	1	25	
...	...	...	...	...	

Session Simulation

- Known user sessions concatenation
- It was considered that most users do not share account
- Sessions with more than 10 clicks

Session Simulation

Table 5.3: Simulated session sample with desired cut in gray

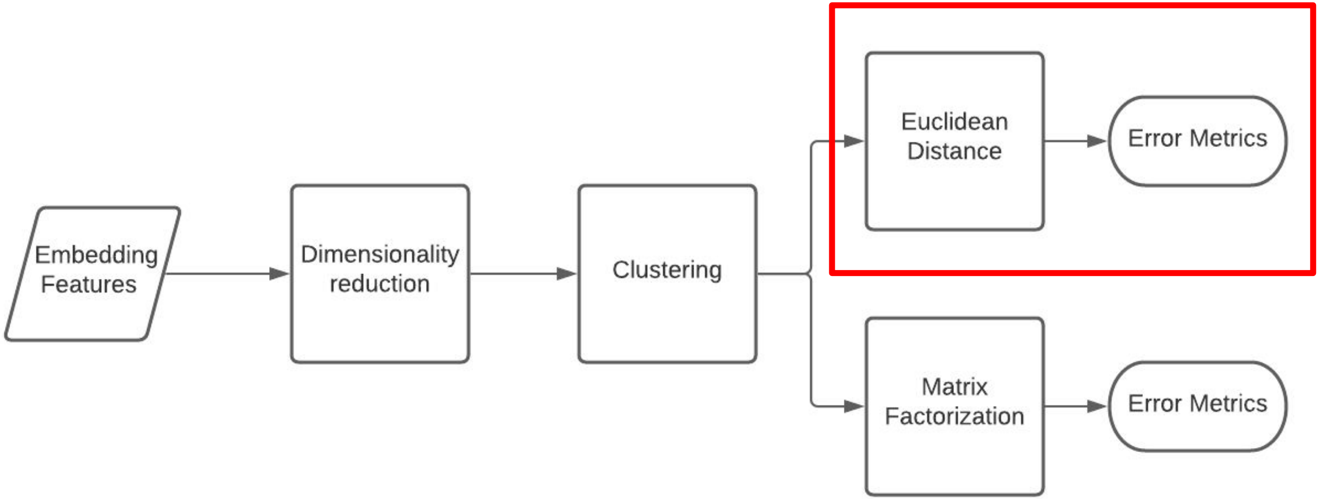
	click_timestamp	user_id	x_centroid	y_centroid	distance
...	...	...	...	...	...
11	1507987156229	6207	-8.487595	3.206146	43.517998
12	1507988999717	6207	3.864952	-3.751176	27.925744
13	1507989029717	6207	5.004131	3.442250	7.283071
14	1507061447189	143259	25.503668	-1.358489	21.054171
15	1507061615464	143259	18.530073	-15.994876	16.212799
16	1507061615464	143259	18.530073	-15.994876	0.000000
...	...	...	...	...	...

Session Simulation

Table 5.3: Simulated session sample with desired cut in gray

	click_timestamp	user_id	x_centroid	y_centroid	distance
...	...	...	...	...	...
11	1507987156229	6207	-8.487595	3.206146	43.517998
12	1507988999717	6207	3.864952	-3.751176	27.925744
13	1507989029717	6207	5.004131	3.442250	7.283071
14	1507061447189	143259	25.503668	-1.358489	21.054171
15	1507061615464	143259	18.530073	-15.994876	16.212799
16	1507061615464	143259	18.530073	-15.994876	0.000000
...	...	...	...	...	...

Euclidean Distance



Euclidean Distance

$$d(x,\ y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Euclidean Distance

Table 5.3: Simulated session sample with desired cut in gray

	click_timestamp	user_id	x_centroid	y_centroid	distance
...	...	...	...	...	...
11	1507987156229	6207	-8.487595	3.206146	43.517998
12	1507988999717	6207	3.864952	-3.751176	27.925744
13	1507989029717	6207	5.004131	3.442250	7.283071
14	1507061447189	143259	25.503668	-1.358489	21.054171
15	1507061615464	143259	18.530073	-15.994876	16.212799
16	1507061615464	143259	18.530073	-15.994876	0.000000
...	...	...	...	...	...

Euclidean Distance

Table 5.3: Simulated session sample with desired cut in gray

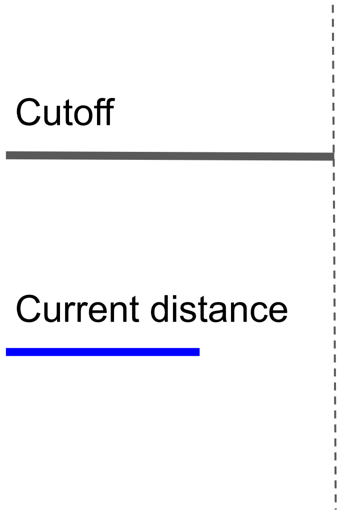
	click_timestamp	user_id	x_centroid	y_centroid	distance
...	...	...	...	...	...
11	1507987156229	6207	-8.487595	3.206146	43.517998
12	1507988999717	6207	3.864952	-3.751176	27.925744
13	1507989029717	6207	5.004131	3.442250	7.283071
14	1507061447189	143259	25.503668	-1.358489	21.054171
15	1507061615464	143259	18.530073	-15.994876	16.212799
16	1507061615464	143259	18.530073	-15.994876	0.000000
...	...	...	...	...	...

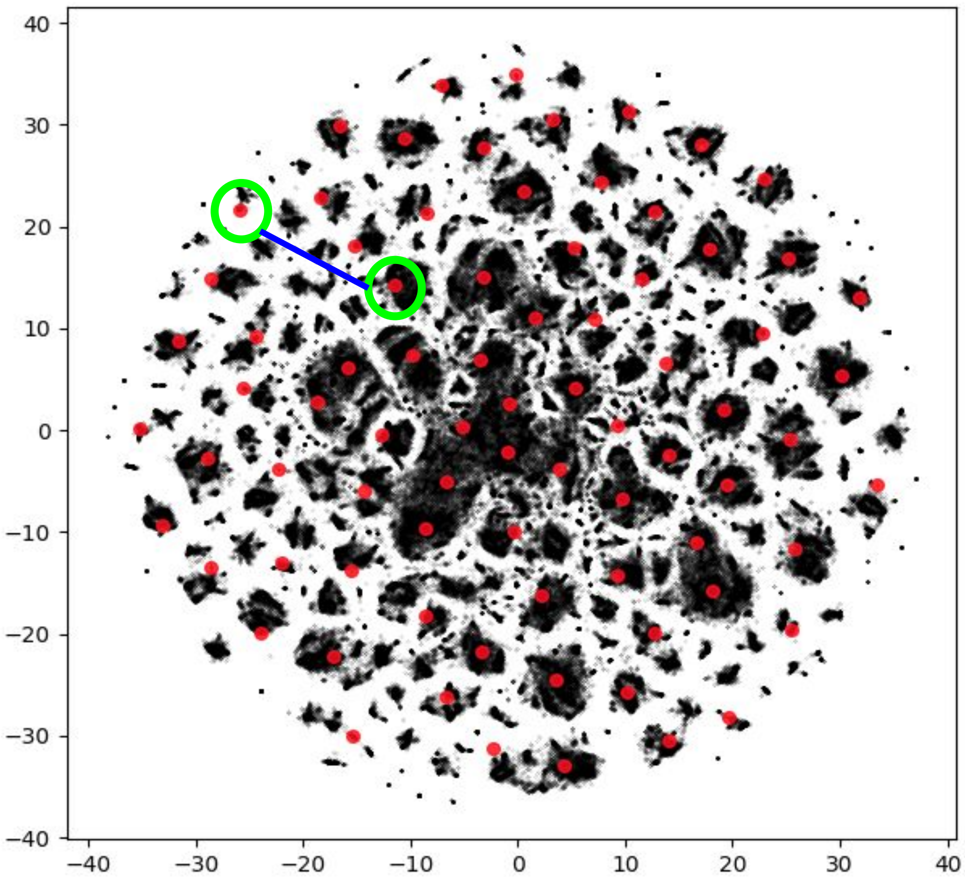
Euclidean Distance

Table 5.3: Simulated session sample with desired cut in gray

	click_timestamp	user_id	x_centroid	y_centroid	distance
...	...	...	...	...	...
11	1507987156229	6207	-8.487595	3.206146	43.517998
12	1507988999717	6207	3.864952	-3.751176	27.925744
13	1507989029717	6207	5.004131	3.442250	7.283071
14	1507061447189	143259	25.503668	-1.358489	21.054171
15	1507061615464	143259	18.530073	-15.994876	16.212799
16	1507061615464	143259	18.530073	-15.994876	0.000000
...	...	...	...	...	...

Cutoff heuristic

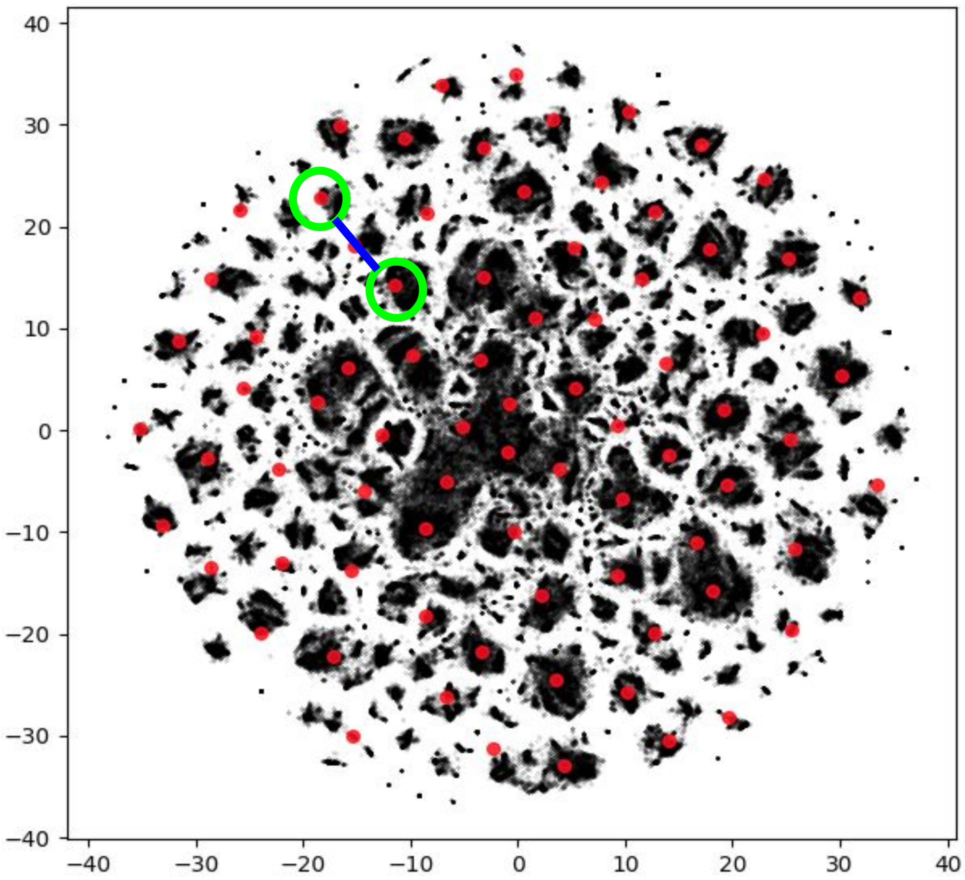




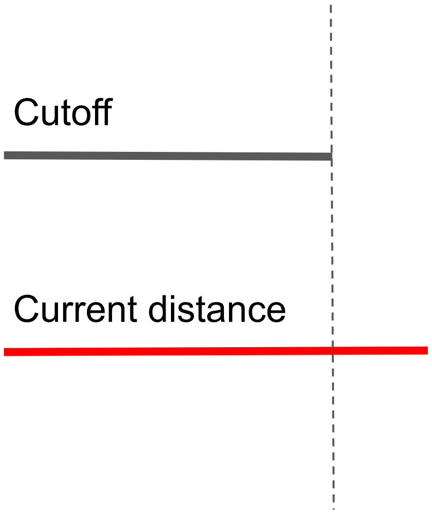
Cutoff heuristic

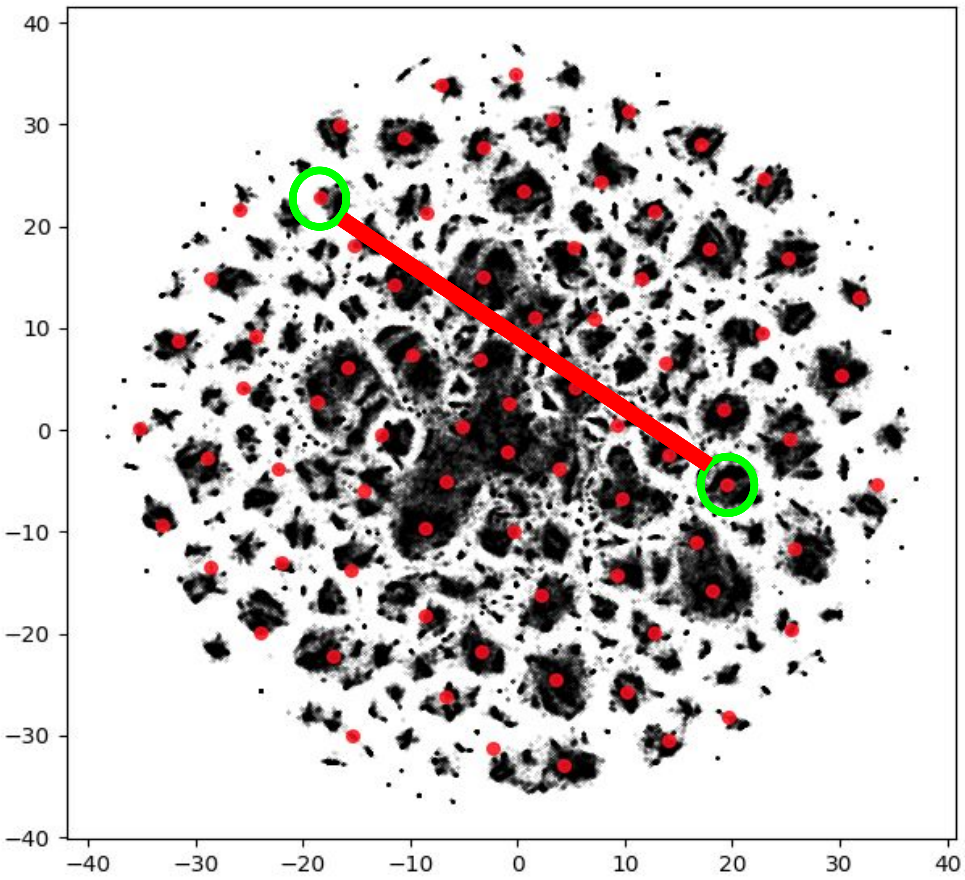
Cutoff

Current distance



Cutoff heuristic





Error metrics

Table 5.2: Error normalization map from desired cutoff index

index	user_id	error
0	111	1
1	111	0.66
2	111	0.33
3	888	0
4	888	0.16
5	888	0.33
6	888	0.5
7	888	0.66
8	888	0.83
9	888	1

Error metrics

Table 5.2: Error normalization map from desired cutoff index

	index	user_id	error
Session A	0	111	1
	1	111	0.66
	2	111	0.33
Session B	3	888	0
	4	888	0.16
	5	888	0.33
	6	888	0.5
	7	888	0.66
	8	888	0.83
	9	888	1

Does temporal ordering matters?

Temporal Disambiguation

Time ordered

timestamp	user_id	topic
A	111	teal_subjects
B	111	teal_subjects
C	111	fuscia_disasters
D	999	red_events
E	999	yellow_plots
F	999	orange_gossips

Temporal Disambiguation

Time shuffled

timestamp	user_id	topic
C	111	fuscia_disasters
A	111	teal_subjects
B	111	teal_subjects
F	999	orange_gossips
E	999	yellow_plots
D	999	red_events

Accuracy performance

Accuracy performance

- Filter: Sessions with more than 10 clicks
- 32 thousand sessions
- For each cutoff, an array with errors from each session split

Tuning

Error mean ↓

Error standard deviation ↓

Table 5.4: Error metrics, session ordered by timestamp 5.4a vs shuffled 5.4b

(a)

cutoff dist.	error mean	error std
1.00	0.933	$\boxed{0.053}$
5.21	0.932	0.054
9.42	0.922	0.071
13.64	0.917	0.078
17.85	0.901	0.098
22.07	0.877	0.126
26.28	0.857	0.151
30.5	0.812	0.191
34.71	0.765	0.230
38.92	0.681	0.266
43.14	0.629	0.286
43.35	$\boxed{0.611}$	0.301
51.57	0.630	0.311
55.78	0.707	0.320
60.00	0.814	0.287

(b)

cutoff dist.	error mean	error std
1.00	0.939	$\boxed{0.045}$
5.21	0.938	0.048
9.42	0.933	0.057
13.64	0.927	0.663
17.85	0.912	0.088
22.07	0.893	0.114
26.28	0.867	0.144
30.5	0.827	0.184
34.71	0.778	0.223
38.92	0.710	0.264
43.14	0.646	0.289
43.35	$\boxed{0.615}$	0.310
51.57	0.631	0.322
55.78	0.721	0.325
60.00	0.813	0.298

Table 5.4: Error metrics, session ordered by timestamp 5.4a vs shuffled 5.4b

(a)

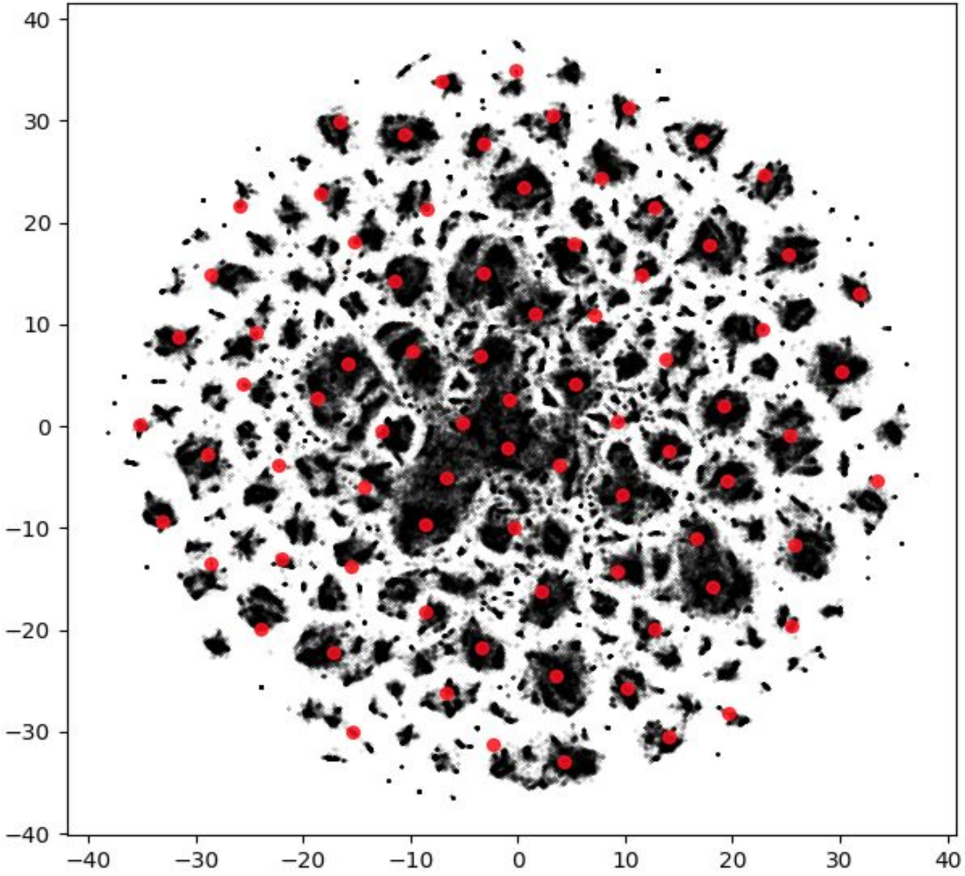
cutoff dist.	error mean	error std
1.00	0.933	0.053
5.21	0.932	0.054
9.42	0.922	0.071
13.64	0.917	0.078
17.85	0.901	0.098
22.07	0.877	0.126
26.28	0.857	0.151
30.5	0.812	0.191
34.71	0.765	0.230
38.92	0.681	0.266
43.14	0.629	0.286
43.35	0.611	0.301
51.57	0.630	0.311
55.78	0.707	0.320
60.00	0.814	0.287

(b)

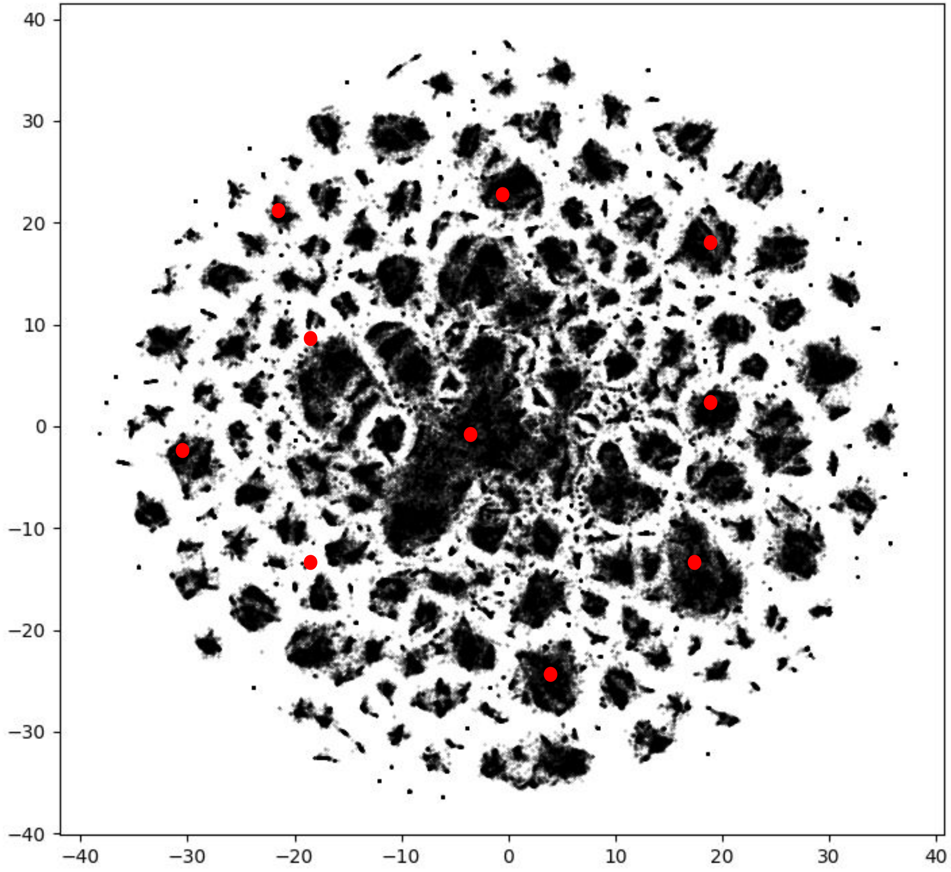
cutoff dist.	error mean	error std
1.00	0.939	0.045
5.21	0.938	0.048
9.42	0.933	0.057
13.64	0.927	0.663
17.85	0.912	0.088
22.07	0.893	0.114
26.28	0.867	0.144
30.5	0.827	0.184
34.71	0.778	0.223
38.92	0.710	0.264
43.14	0.646	0.289
43.35	0.615	0.310
51.57	0.631	0.322
55.78	0.721	0.325
60.00	0.813	0.298

How specific a topic can be without damaging accuracy?

Specificist



Generalist



Damping Factor

Table 5.5: damping factor effect on nº of clusters

damping	nº of clusters
0.5	4771
0.553	4716
0.606	2984
0.66	1146
0.766	91
0.82	152

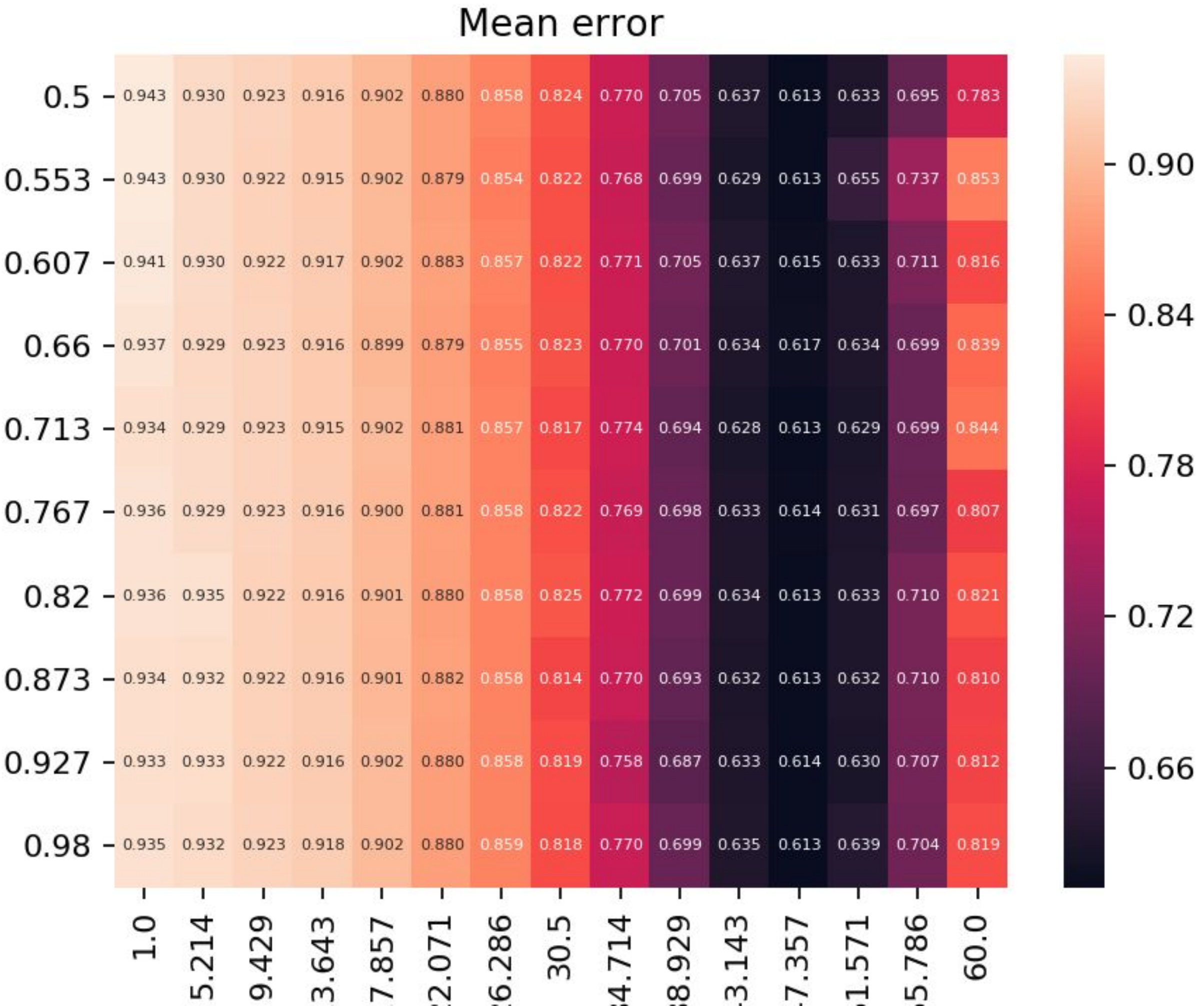
Damping Factor

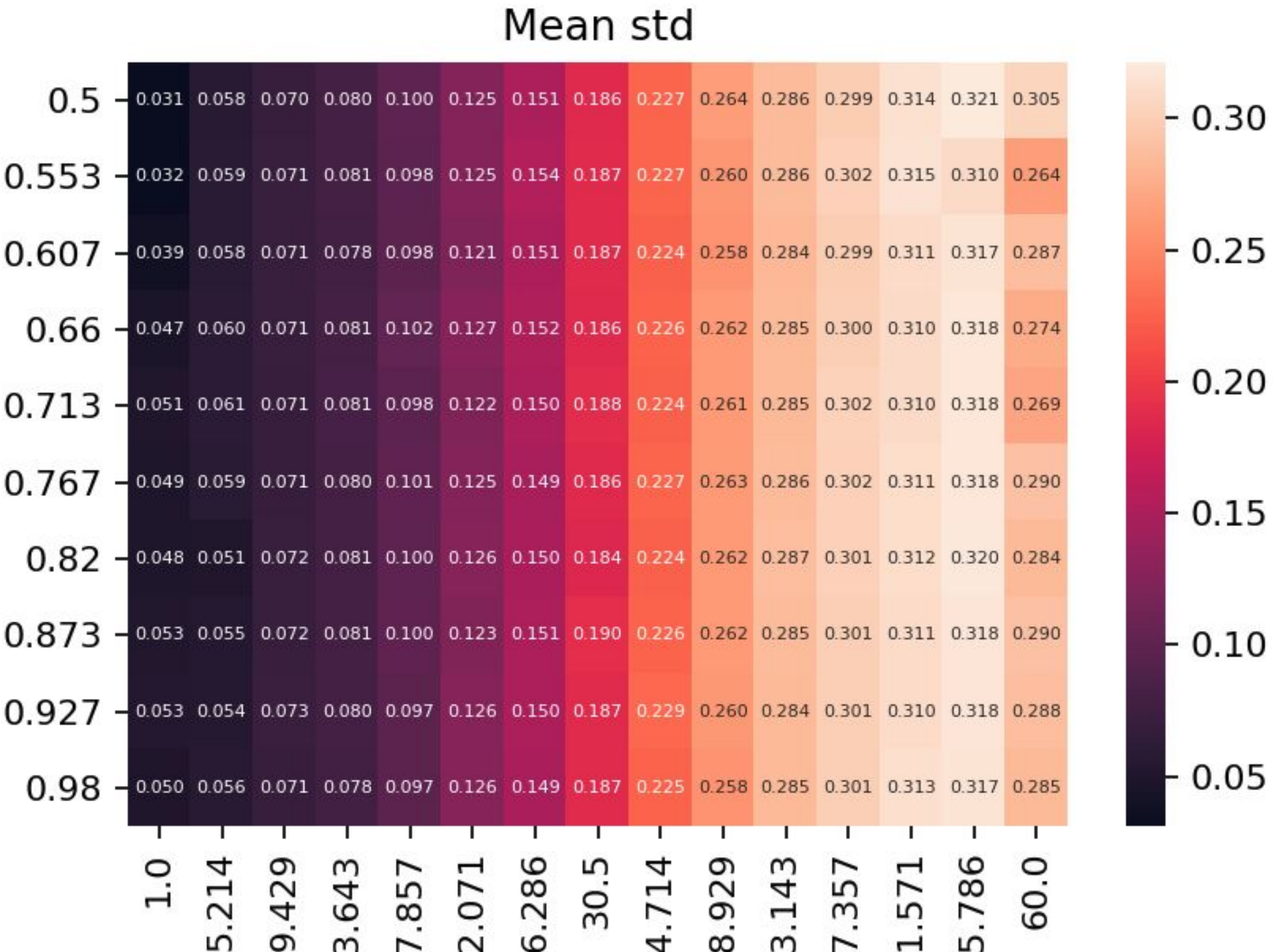
Table 5.5: damping factor effect on n° of clusters

damping	n° of clusters
0.5	4771
0.553	4716
0.606	2984
0.66	1146
0.766	91
0.82	152

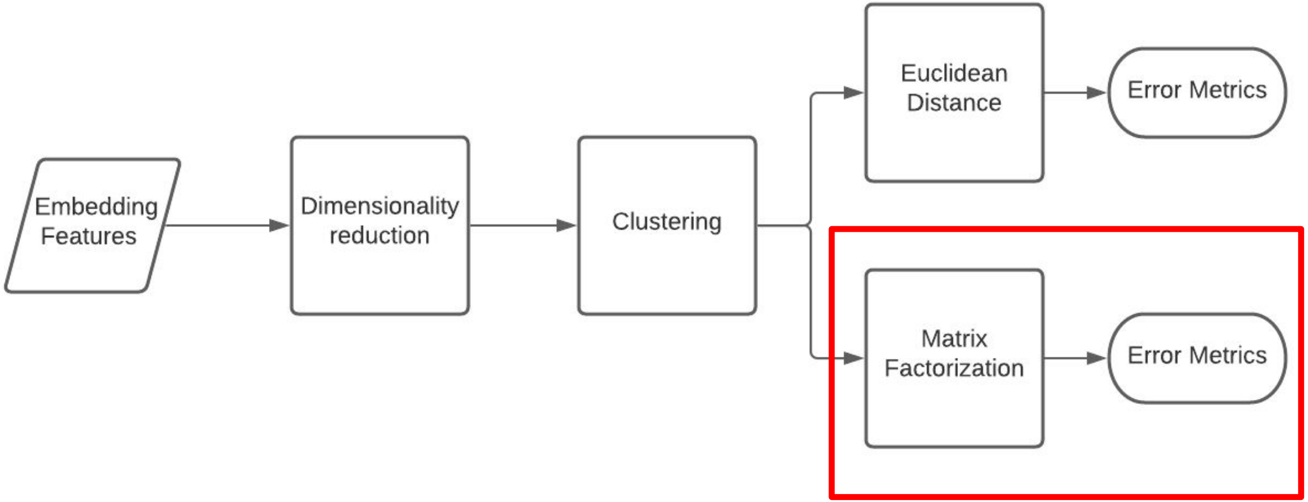
Specifism

Generalism


Mean error
0.5
0.553
0.607
0.66
0.713
0.767
0.82
0.873
0.927
0.98
1.0
5.214
9.429
3.643
7.857
2.071
6.286
30.5
4.714
8.929
3.143
7.357
1.571
5.786
60.0
0.90
0.84
0.78
0.72
0.66
0.943 0.930 0.923 0.916 0.902 0.880 0.858 0.824 0.770 0.705 0.637 0.613 0.633 0.695 0.783
0.943 0.930 0.922 0.915 0.902 0.879 0.854 0.822 0.768 0.699 0.629 0.613 0.655 0.737 0.853
0.941 0.930 0.922 0.917 0.902 0.883 0.857 0.822 0.771 0.705 0.637 0.615 0.633 0.711 0.816
0.937 0.929 0.923 0.916 0.899 0.879 0.855 0.823 0.770 0.701 0.634 0.617 0.634 0.699 0.839
0.934 0.929 0.923 0.915 0.902 0.881 0.857 0.817 0.774 0.694 0.628 0.613 0.629 0.699 0.844
0.936 0.929 0.923 0.916 0.900 0.881 0.858 0.822 0.769 0.698 0.633 0.614 0.631 0.697 0.807
0.936 0.935 0.922 0.916 0.901 0.880 0.858 0.825 0.772 0.699 0.634 0.613 0.633 0.710 0.821
0.934 0.932 0.922 0.916 0.901 0.882 0.858 0.814 0.770 0.693 0.632 0.613 0.632 0.710 0.810
0.933 0.933 0.922 0.916 0.902 0.880 0.858 0.819 0.758 0.687 0.633 0.614 0.630 0.707 0.812
0.935 0.932 0.923 0.918 0.902 0.880 0.859 0.818 0.770 0.699 0.635 0.613 0.639 0.704 0.819


Mean std
0.5
0.553
0.607
0.66
0.713
0.767
0.82
0.873
0.927
0.98
0.031 0.058 0.070 0.080 0.100 0.125 0.151 0.186 0.227 0.264 0.286 0.299 0.314 0.321 0.305
0.032 0.059 0.071 0.081 0.098 0.125 0.154 0.187 0.227 0.260 0.286 0.302 0.315 0.310 0.264
0.039 0.058 0.071 0.078 0.098 0.121 0.151 0.187 0.224 0.258 0.284 0.299 0.311 0.317 0.287
0.047 0.060 0.071 0.081 0.102 0.127 0.152 0.186 0.226 0.262 0.285 0.300 0.310 0.318 0.274
0.051 0.061 0.071 0.081 0.098 0.122 0.150 0.188 0.224 0.261 0.285 0.302 0.310 0.318 0.269
0.049 0.059 0.071 0.080 0.101 0.125 0.149 0.186 0.227 0.263 0.286 0.302 0.311 0.318 0.290
0.048 0.051 0.072 0.081 0.100 0.126 0.150 0.184 0.224 0.262 0.287 0.301 0.312 0.320 0.284
0.053 0.055 0.072 0.081 0.100 0.123 0.151 0.190 0.226 0.262 0.285 0.301 0.311 0.318 0.290
0.053 0.054 0.073 0.080 0.097 0.126 0.150 0.187 0.229 0.260 0.284 0.301 0.310 0.318 0.288
0.050 0.056 0.071 0.078 0.097 0.126 0.149 0.187 0.225 0.258 0.285 0.301 0.313 0.317 0.285
1.0
5.214
9.429
3.643
7.857
2.071
6.286
30.5
4.714
8.929
3.143
7.357
1.571
5.786
60.0
0.30
0.25
0.20
0.15
0.10
0.05

Matrix Factorization



Alternate Least Squares (Pyspark)

Table 5.6: ALS: Read frequency per topic

user_id	topic_alias	read_frequency
...	...	...
26340	purple-articles	7
26340	rose-stories	5
58277	cyan-contents	1
...	...	...

Alternate Least Squares (Pyspark)

Table 5.7: Alternating Least Squares measurements

user_id	user_id_to_predict	prediction_mean	prediction_std
413	413	0.83287	0.2799
45502	413	0.86549	0.25900
12897	12897	0.84201	0.3291
16695	12897	0.91515	0.27717
2930	2930	0.79370	0.34552
9261	2930	0.87126	0.2551
20001	20001	0.8777	0.24580
6344	20001	0.91090	0.2383
62025	62025	0.75506	0.35445
19864	62025	0.47764	0.27196
43017	43017	0.83342	0.27467
21356	43017	0.81031	0.28030
3391	3391	0.85998	0.2949
681	3391	0.87366	0.30739
11521	11521	0.61621	0.3742
59193	11521	1.01384	0.5024
11359	11359	0.88445	0.2596
23036	11359	0.91379	0.24409

Conclusion

- Event-driven nature of news
- Looming need for privacy
- Need for Anonymized Recommender systems
- Need for Anonymized Data reliability

Thank you

Questions?