Data Analytics I| Assignment 3

Author: Prateek Pandey
ID: 2022201035

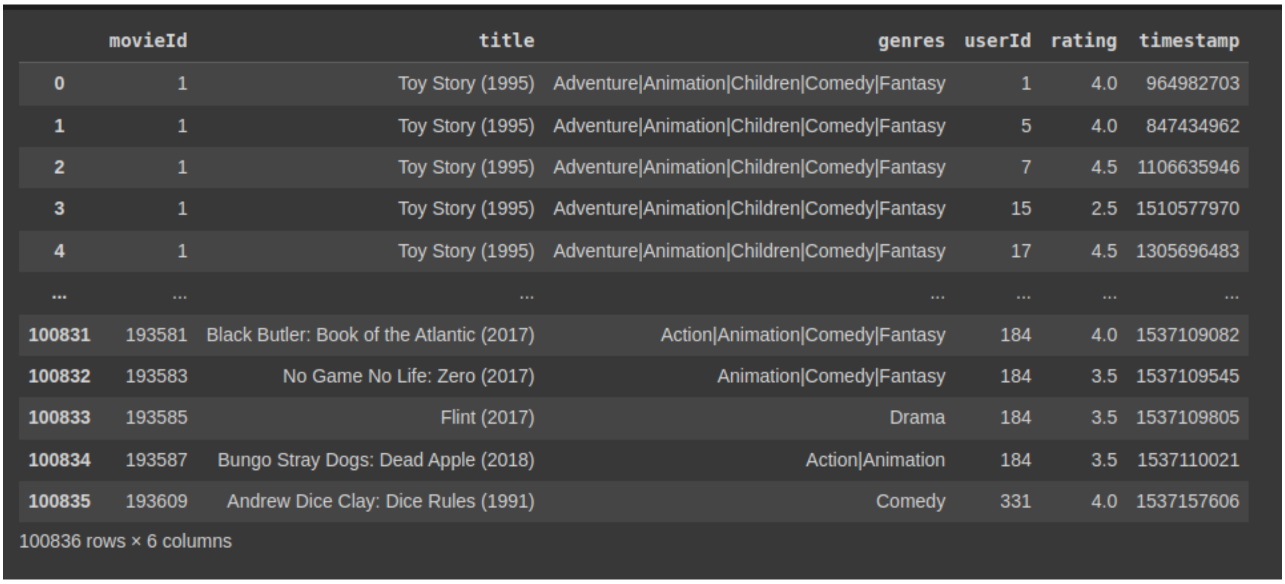
October 15, 2023

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY

HYDERABAD

Contents

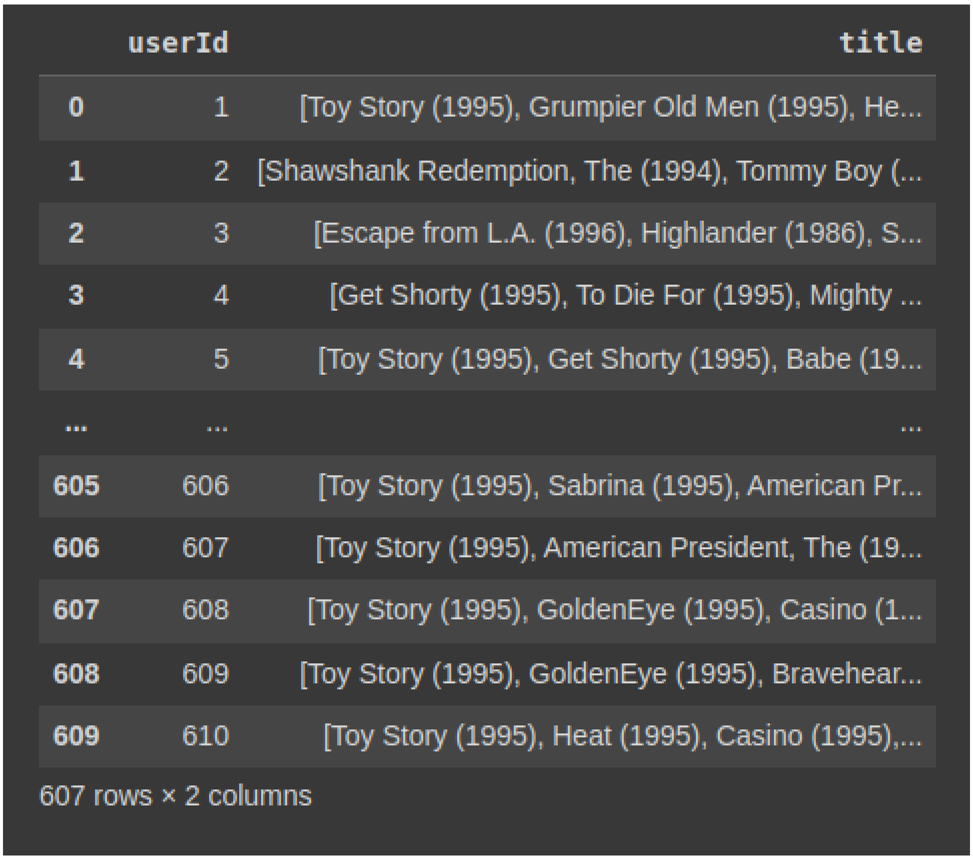
- After Joining Movies and Ratings Table



	movieId	title	genres	userId	rating	timestamp
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1	4.0	964982703
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	5	4.0	847434962
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	7	4.5	1106635946
3	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	15	2.5	1510577970
4	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	17	4.5	1305696483
...	...	...	...	...	...	...
100831	193581	Black Butler: Book of the Atlantic (2017)	Action|Animation|Comedy|Fantasy	184	4.0	1537109082
100832	193583	No Game No Life: Zero (2017)	Animation|Comedy|Fantasy	184	3.5	1537109545
100833	193585	Flint (2017)	Drama	184	3.5	1537109805
100834	193587	Bungo Stray Dogs: Dead Apple (2018)	Action|Animation	184	3.5	1537110021
100835	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy	331	4.0	1537157606

100836 rows × 6 columns

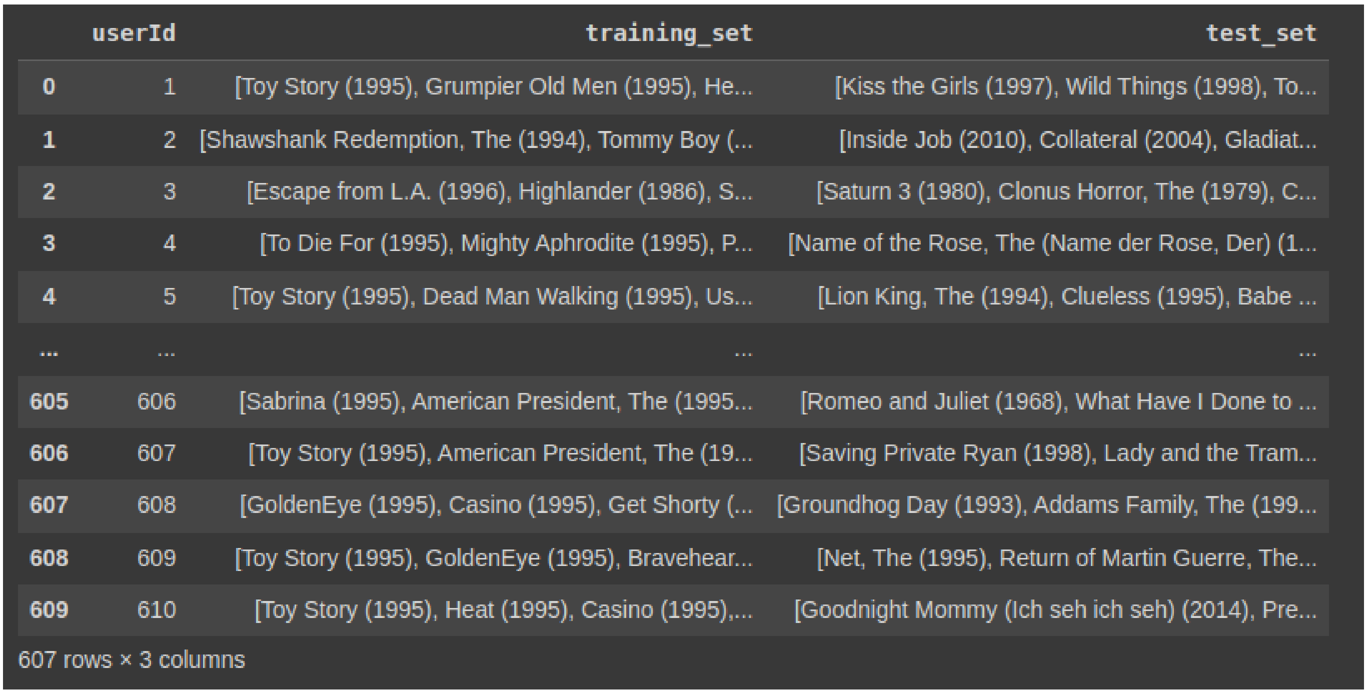
- After Considering only those movies who have been rated > 2 and users who have rated more than 10 movies



	userId	title
0	1	[Toy Story (1995), Grumpier Old Men (1995), He...
1	2	[Shawshank Redemption, The (1994), Tommy Boy (...
2	3	[Escape from L.A. (1996), Highlander (1986), S...
3	4	[Get Shorty (1995), To Die For (1995), Mighty ...
4	5	[Toy Story (1995), Get Shorty (1995), Babe (19...
...	...	...
605	606	[Toy Story (1995), Sabrina (1995), American Pr...
606	607	[Toy Story (1995), American President, The (19...
607	608	[Toy Story (1995), GoldenEye (1995), Casino (1...
608	609	[Toy Story (1995), GoldenEye (1995), Bravehear...
609	610	[Toy Story (1995), Heat (1995), Casino (1995),...

607 rows × 2 columns

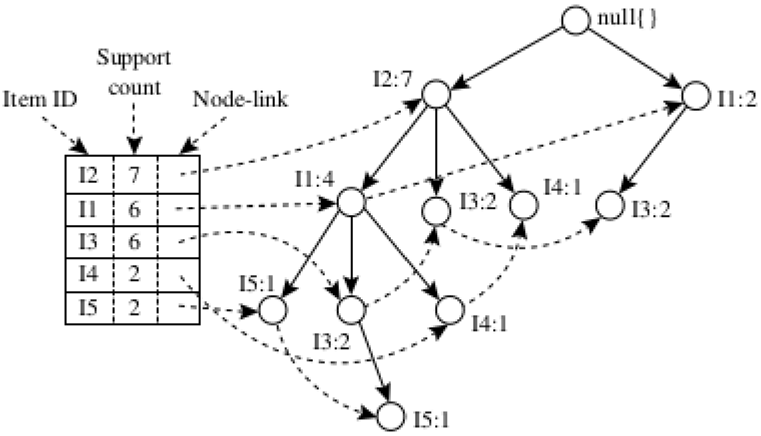
- Dividing the data set into 80% training set and 20% test set for each user



	userId	training_set	test_set
0	1	[Toy Story (1995), Grumpier Old Men (1995), He...	[Kiss the Girls (1997), Wild Things (1998), To...
1	2	[Shawshank Redemption, The (1994), Tommy Boy (...	[Inside Job (2010), Collateral (2004), Gladiat...
2	3	[Escape from L.A. (1996), Highlander (1986), S...	[Saturn 3 (1980), Clonus Horror, The (1979), C...
3	4	[To Die For (1995), Mighty Aphrodite (1995), P...	[Name of the Rose, The (Name der Rose, Der) (1...
4	5	[Toy Story (1995), Dead Man Walking (1995), Us...	[Lion King, The (1994), Clueless (1995), Babe ...
...	...	...	...
605	606	[Sabrina (1995), American President, The (1995...	[Romeo and Juliet (1968), What Have I Done to ...
606	607	[Toy Story (1995), American President, The (19...	[Saving Private Ryan (1998), Lady and the Tram...
607	608	[GoldenEye (1995), Casino (1995), Get Shorty (...	[Groundhog Day (1993), Addams Family, The (199...
608	609	[Toy Story (1995), GoldenEye (1995), Bravehear...	[Net, The (1995), Return of Martin Guerre, The...
609	610	[Toy Story (1995), Heat (1995), Casino (1995),...	[Goodnight Mommy (Ich seh ich seh) (2014), Pre...

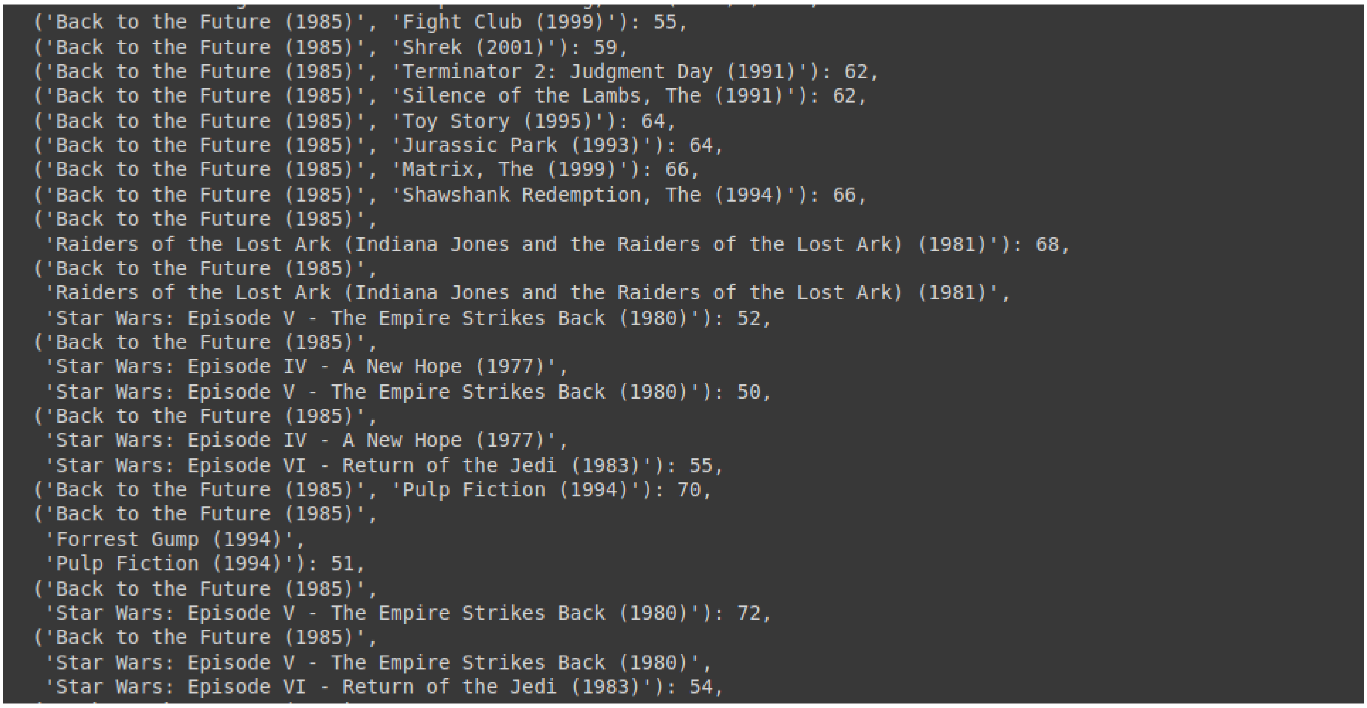
607 rows × 3 columns

- Using FP-Growth Algorithm we first created an FP-TREE



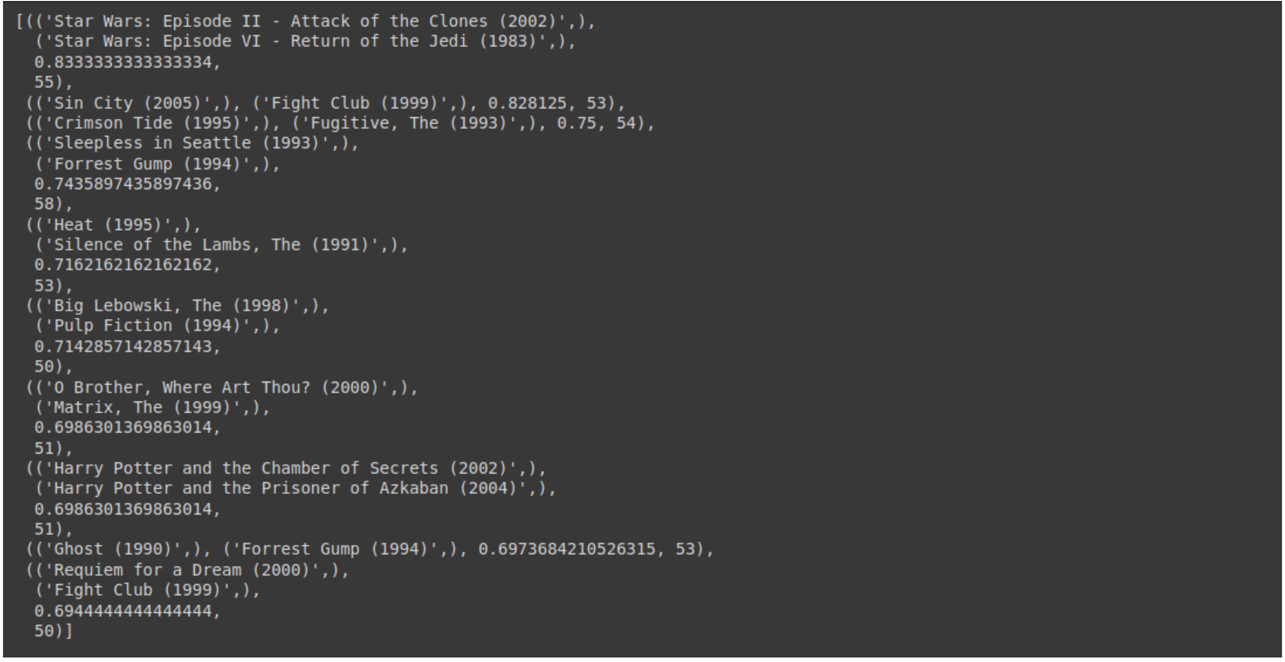
Item ID	Support count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	

- After Mining Frequent Patterns on the FP-Tree resulting Frequent Patterns are



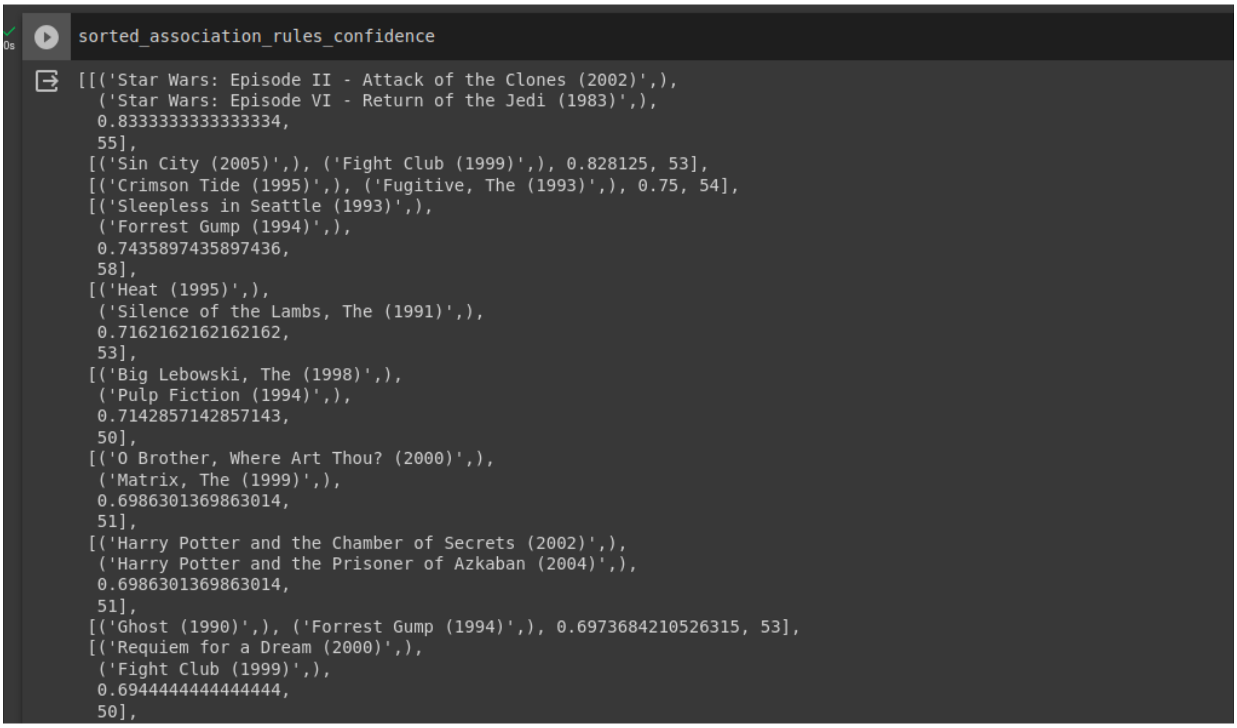
```
 ('Back to the Future (1985)', 'Fight Club (1999)'): 55,
 ('Back to the Future (1985)', 'Shrek (2001)'): 59,
 ('Back to the Future (1985)', 'Terminator 2: Judgment Day (1991)'): 62,
 ('Back to the Future (1985)', 'Silence of the Lambs, The (1991)'): 62,
 ('Back to the Future (1985)', 'Toy Story (1995)'): 64,
 ('Back to the Future (1985)', 'Jurassic Park (1993)'): 64,
 ('Back to the Future (1985)', 'Matrix, The (1999)'): 66,
 ('Back to the Future (1985)', 'Shawshank Redemption, The (1994)'): 66,
 ('Back to the Future (1985)',
  'Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)'): 68,
 ('Back to the Future (1985)',
  'Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)',
  'Star Wars: Episode V - The Empire Strikes Back (1980)'): 52,
 ('Back to the Future (1985)',
  'Star Wars: Episode IV - A New Hope (1977)',
  'Star Wars: Episode V - The Empire Strikes Back (1980)'): 50,
 ('Back to the Future (1985)',
  'Star Wars: Episode IV - A New Hope (1977)',
  'Star Wars: Episode VI - Return of the Jedi (1983)'): 55,
 ('Back to the Future (1985)', 'Pulp Fiction (1994)'): 70,
 ('Back to the Future (1985)',
  'Forrest Gump (1994)',
  'Pulp Fiction (1994)'): 51,
 ('Back to the Future (1985)',
  'Star Wars: Episode V - The Empire Strikes Back (1980)'): 72,
 ('Back to the Future (1985)',
  'Star Wars: Episode V - The Empire Strikes Back (1980)',
  'Star Wars: Episode VI - Return of the Jedi (1983)'): 54,
```

- Mining all Association Rules



```
[(('Star Wars: Episode II - Attack of the Clones (2002)',),
  ('Star Wars: Episode VI - Return of the Jedi (1983)',),
  0.8333333333333334,
  55),
 (('Sin City (2005)',), ('Fight Club (1999)',), 0.828125, 53),
 (('Crimson Tide (1995)',), ('Fugitive, The (1993)',), 0.75, 54),
 (('Sleepless in Seattle (1993)',),
  ('Forrest Gump (1994)',),
  0.7435897435897436,
  58),
 (('Heat (1995)',),
  ('Silence of the Lambs, The (1991)',),
  0.7162162162162162,
  53),
 (('Big Lebowski, The (1998)',),
  ('Pulp Fiction (1994)',),
  0.7142857142857143,
  50),
 (('O Brother, Where Art Thou? (2000)',),
  ('Matrix, The (1999)',),
  0.6986301369863014,
  51),
 (('Harry Potter and the Chamber of Secrets (2002)',),
  ('Harry Potter and the Prisoner of Azkaban (2004)',),
  0.6986301369863014,
  51),
 (('Ghost (1990)',), ('Forrest Gump (1994)',), 0.6973684210526315, 53),
 (('Requiem for a Dream (2000)',),
  ('Fight Club (1999)',),
  0.6944444444444444,
  50)]
```

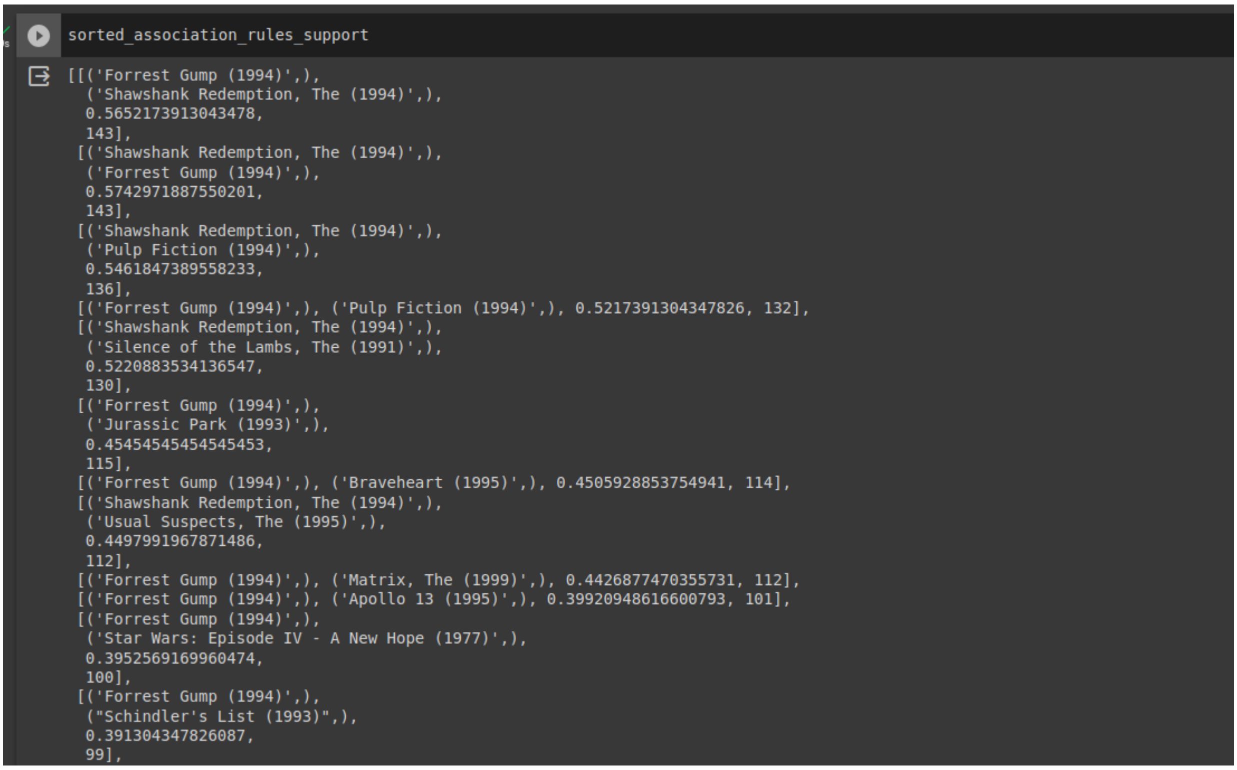
- Recommendation-
  - After Mining Association Rules over the frequent patterns
    - Sorting these rules based on confidence



```
sorted_association_rules_confidence
```

```
[[('Star Wars: Episode II - Attack of the Clones (2002)',),
  ('Star Wars: Episode VI - Return of the Jedi (1983)',),
  0.8333333333333334,
  55],
 [('Sin City (2005)',), ('Fight Club (1999)',), 0.828125, 53],
 [('Crimson Tide (1995)',), ('Fugitive, The (1993)',), 0.75, 54],
 [('Sleepless in Seattle (1993)',),
  ('Forrest Gump (1994)',),
  0.7435897435897436,
  58],
 [('Heat (1995)',),
  ('Silence of the Lambs, The (1991)',),
  0.7162162162162162,
  53],
 [('Big Lebowski, The (1998)',),
  ('Pulp Fiction (1994)',),
  0.7142857142857143,
  50],
 [('O Brother, Where Art Thou? (2000)',),
  ('Matrix, The (1999)',),
  0.6986301369863014,
  51],
 [('Harry Potter and the Chamber of Secrets (2002)',),
  ('Harry Potter and the Prisoner of Azkaban (2004)',),
  0.6986301369863014,
  51],
 [('Ghost (1990)',), ('Forrest Gump (1994)',), 0.6973684210526315, 53],
 [('Requiem for a Dream (2000)',),
  ('Fight Club (1999)',),
  0.6944444444444444,
  50],
```

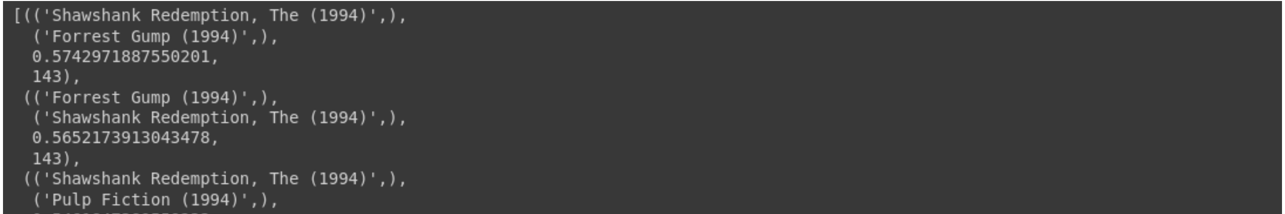
- Sorting these rules based on Support



```
sorted_association_rules_support
```

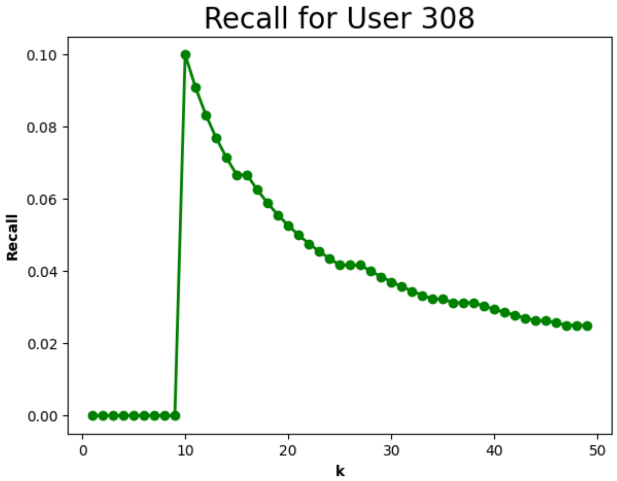
```
[[('Forrest Gump (1994)',),
  ('Shawshank Redemption, The (1994)',),
  0.5652173913043478,
  143],
 [('Shawshank Redemption, The (1994)',),
  ('Forrest Gump (1994)',),
  0.5742971887550201,
  143],
 [('Shawshank Redemption, The (1994)',),
  ('Pulp Fiction (1994)',),
  0.5461847389558233,
  136],
 [('Forrest Gump (1994)',), ('Pulp Fiction (1994)',), 0.5217391304347826, 132],
 [('Shawshank Redemption, The (1994)',),
  ('Silence of the Lambs, The (1991)',),
  0.5220883534136547,
  130],
 [('Forrest Gump (1994)',),
  ('Jurassic Park (1993)',),
  0.45454545454545453,
  115],
 [('Forrest Gump (1994)',), ('Braveheart (1995)',), 0.4505928853754941, 114],
 [('Shawshank Redemption, The (1994)',),
  ('Usual Suspects, The (1995)',),
  0.4497991967871486,
  112],
 [('Forrest Gump (1994)',), ('Matrix, The (1999)',), 0.4426877470355731, 112],
 [('Forrest Gump (1994)',), ('Apollo 13 (1995)',), 0.39920948616600793, 101],
 [('Forrest Gump (1994)',),
  ('Star Wars: Episode IV - A New Hope (1977)',),
  0.3952569169960474,
  100],
 [('Forrest Gump (1994)',),
  ("Schindler's List (1993)",),
  0.391304347826087,
  99],
```

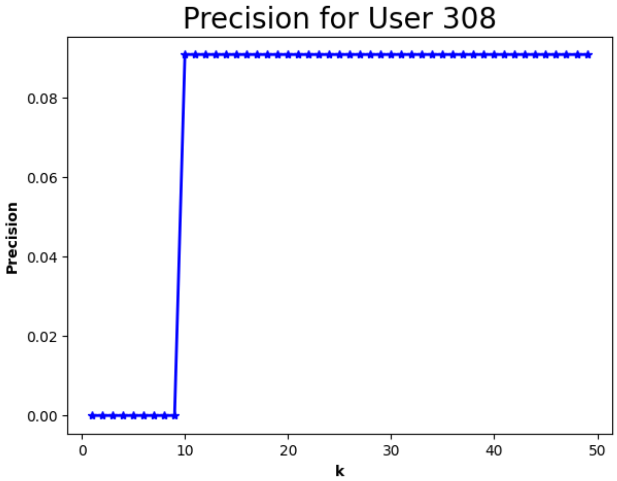
- Taking Intersection of both these top 100 sorted rules by support and confidence to get common list

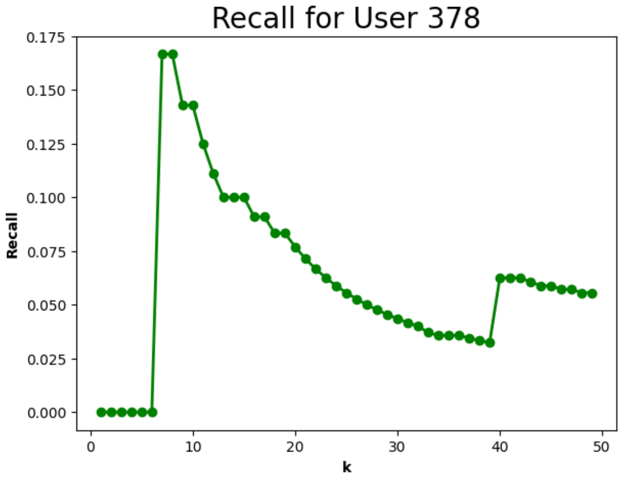


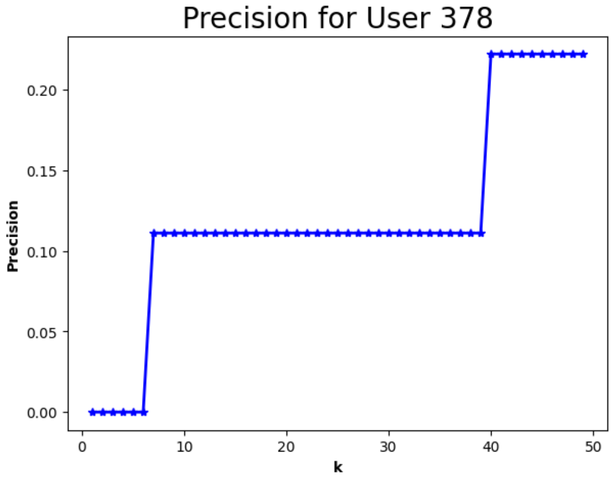
```
[(('Shawshank Redemption, The (1994)',),
  ('Forrest Gump (1994)',),
  0.5742971887550201,
  143),
 (('Forrest Gump (1994)',),
  ('Shawshank Redemption, The (1994)',),
  0.5652173913043478,
  143),
 (('Shawshank Redemption, The (1994)',),
  ('Pulp Fiction (1994)',),
```

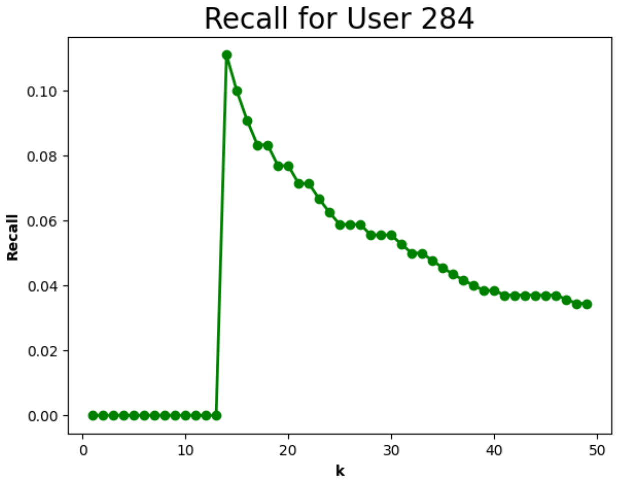
- Precision and Recall Graph of Random 25 users
- Given rule X → Y, with X representing a movie from the training set, set Y comprises recommended movies. By aggregating the movies on the right side of N rules, denoted as R, the resulting set serves as recommendations. The overlap between R and the test set is termed the hit set. The recall is determined by the ratio of the hit set to the test set, while the precision is calculated as the ratio of the hit set to the recommendation set.

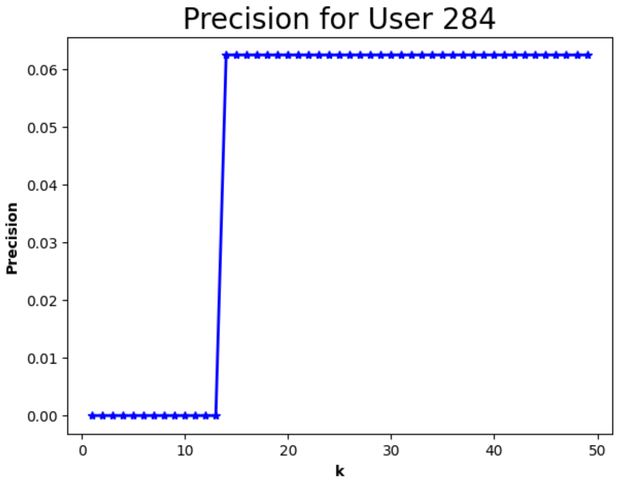


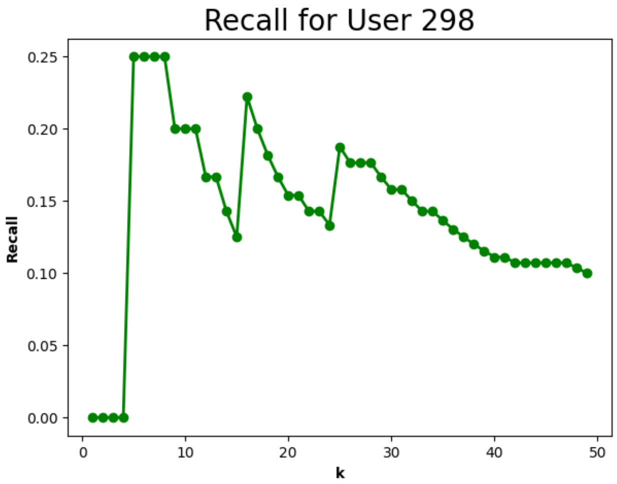


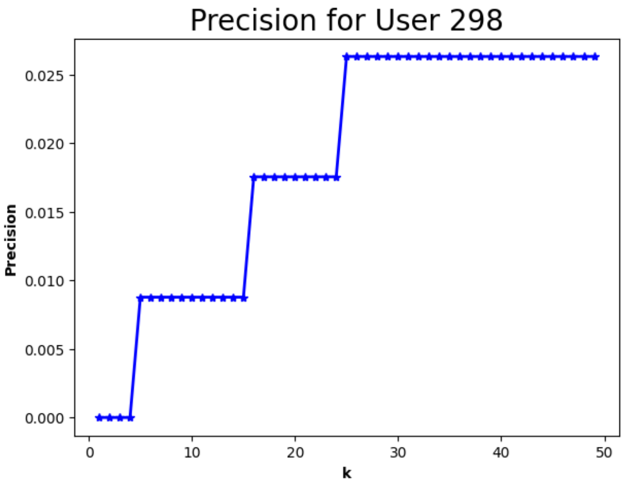


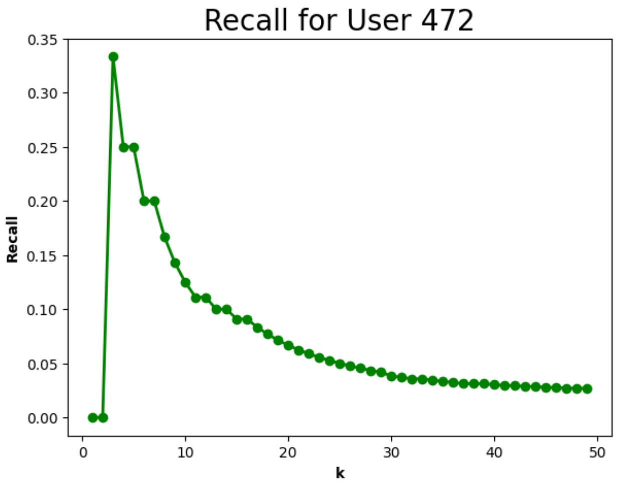


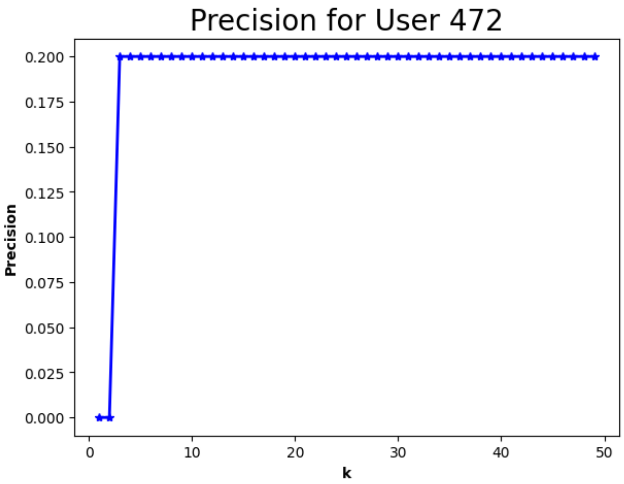


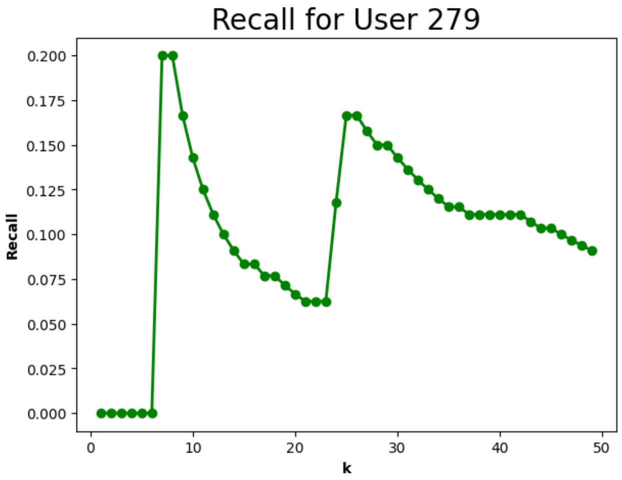


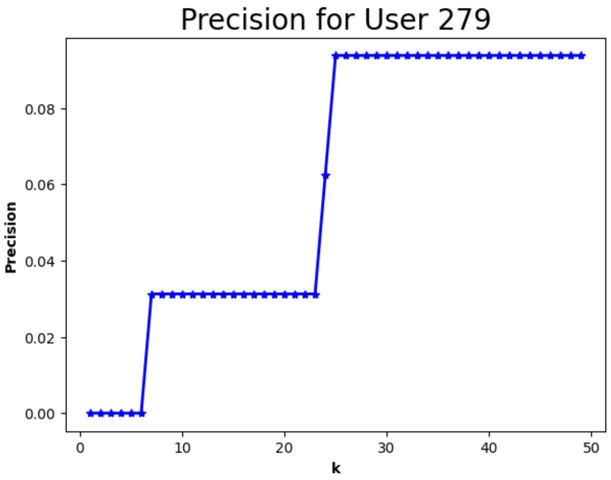

Recall for User 298
Recall
k
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50

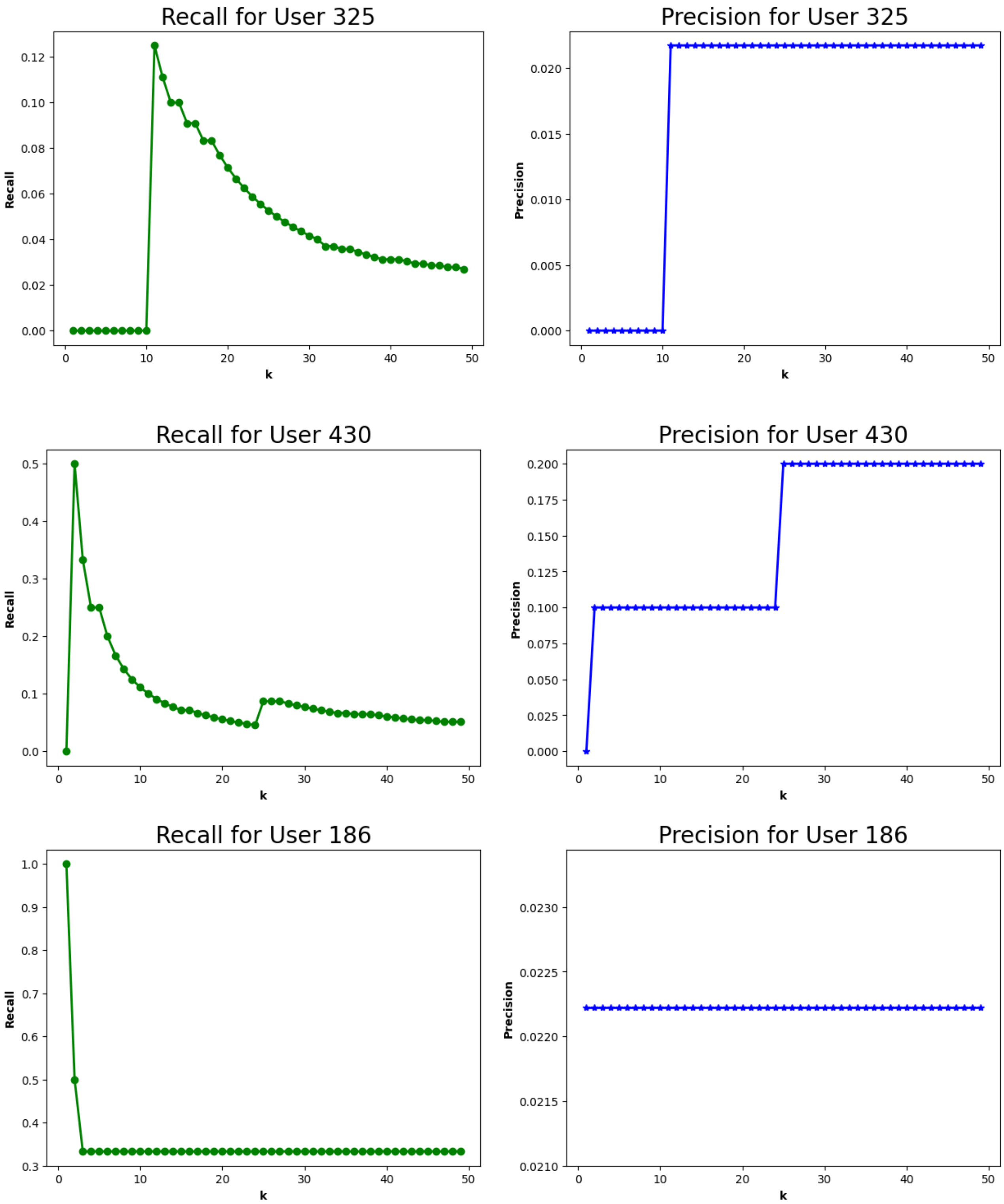

Precision for User 298
Precision
k
0.000
0.005
0.010
0.015
0.020
0.025
0
10
20
30
40
50

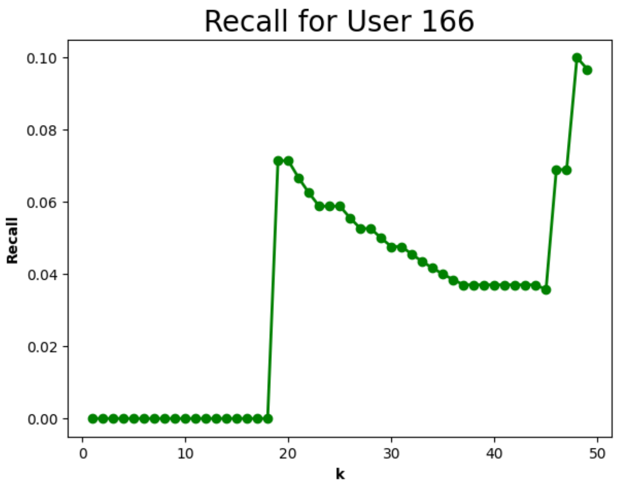

Recall for User 472
Recall
k
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50


Precision for User 472
Precision
k
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0
10
20
30
40
50

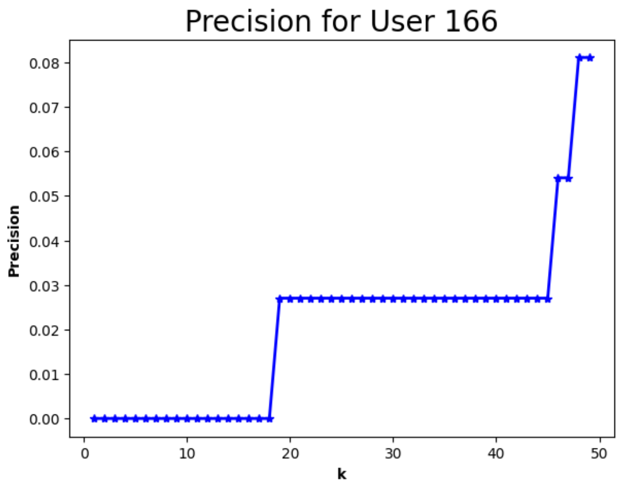

Recall for User 279
Recall
k
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0
10
20
30
40
50


Precision for User 279
Precision
k
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
50

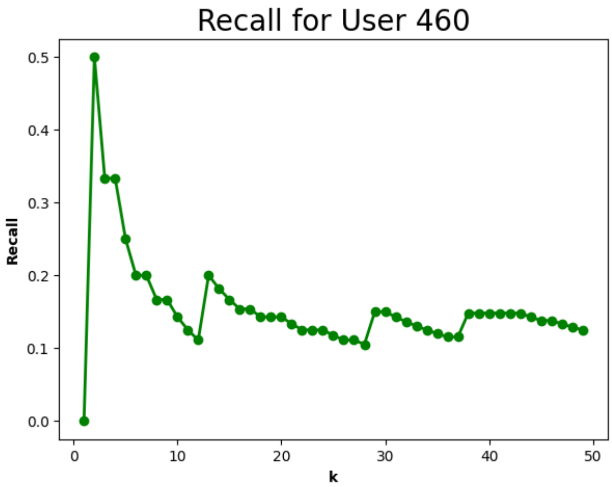

Recall for User 325
Recall
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30
40
50
k
Precision for User 325
Precision
0.000
0.005
0.010
0.015
0.020
0
10
20
30
40
50
k
Recall for User 430
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50
k
Precision for User 430
Precision
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0
10
20
30
40
50
k
Recall for User 186
Recall
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
10
20
30
40
50
k
Precision for User 186
Precision
0.0210
0.0215
0.0220
0.0225
0.0230
0
10
20
30
40
50
k



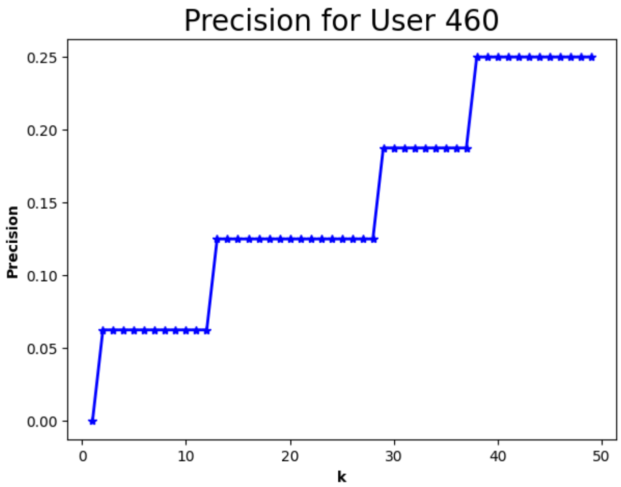
Recall for User 166



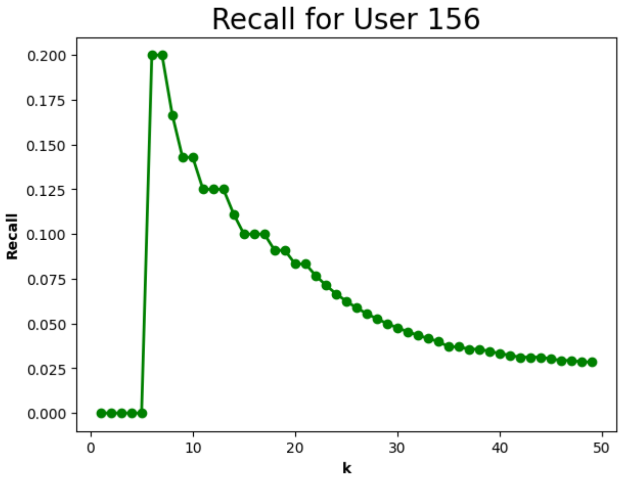
Precision for User 166



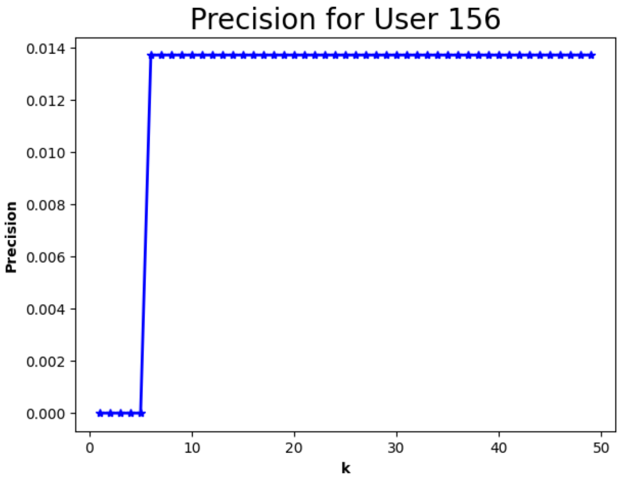
Recall for User 460



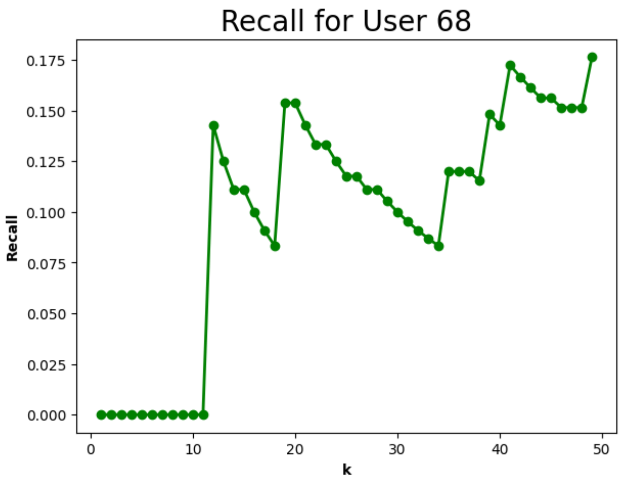
Precision for User 460

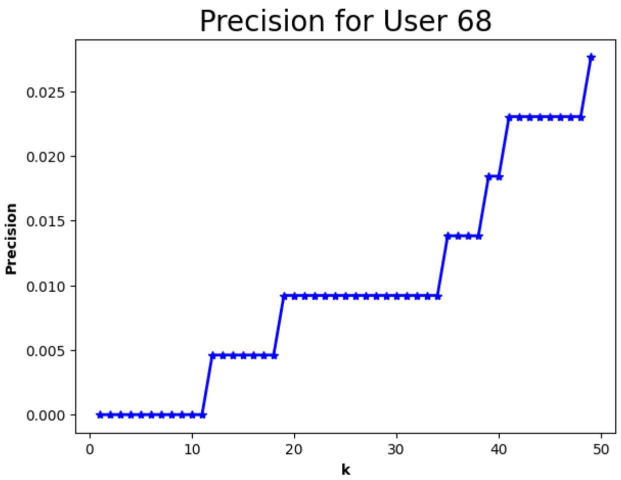


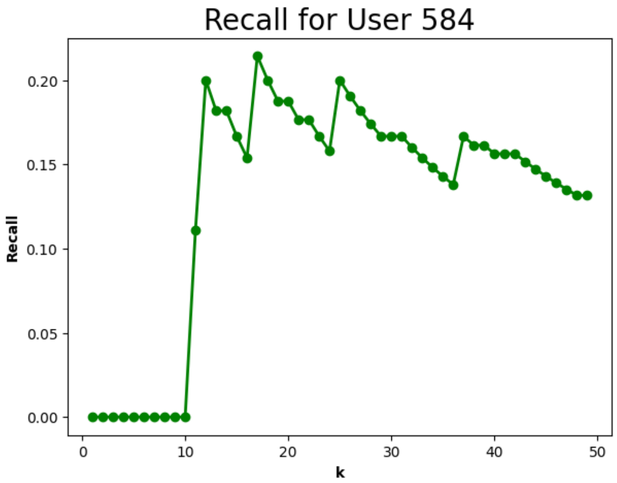
Recall for User 156

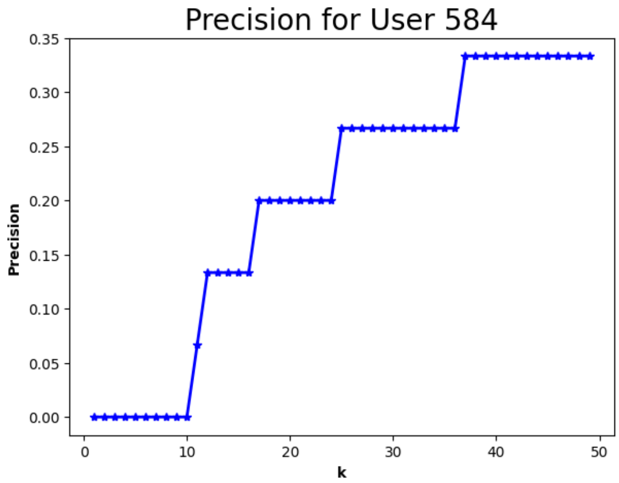


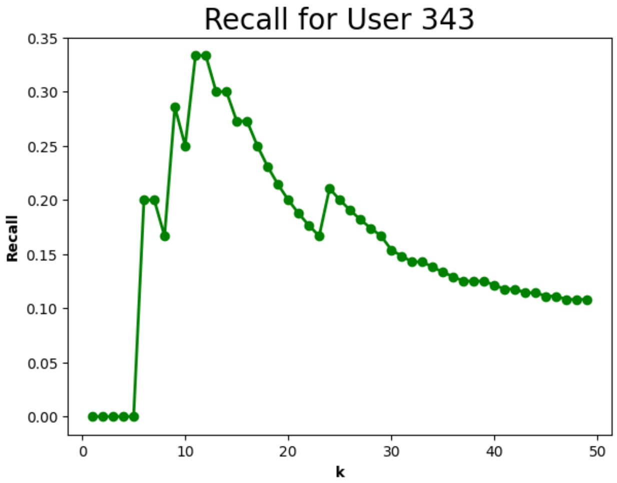
Precision for User 156

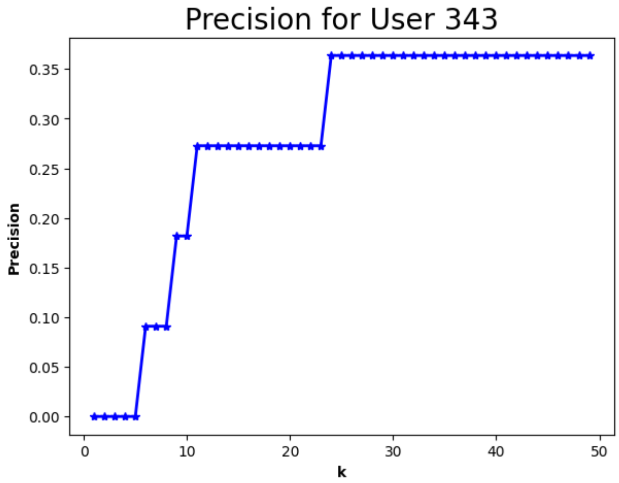

Recall for User 68
Recall
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
k
0
10
20
30
40
50

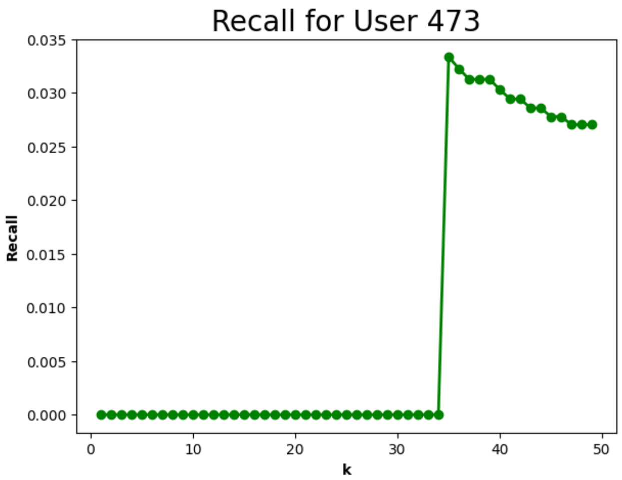

Precision for User 68
Precision
0.000
0.005
0.010
0.015
0.020
0.025
k
0
10
20
30
40
50

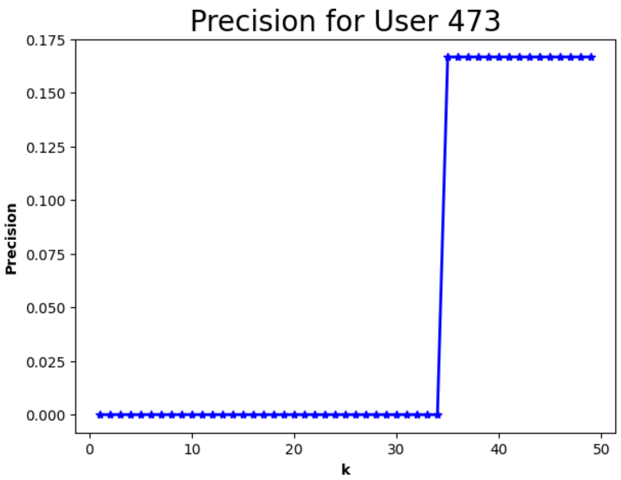

Recall for User 584
Recall
0.00
0.05
0.10
0.15
0.20
k
0
10
20
30
40
50

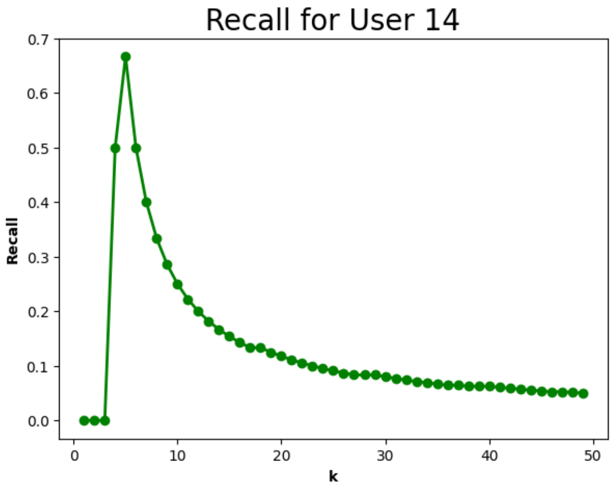

Precision for User 584
Precision
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
k
0
10
20
30
40
50

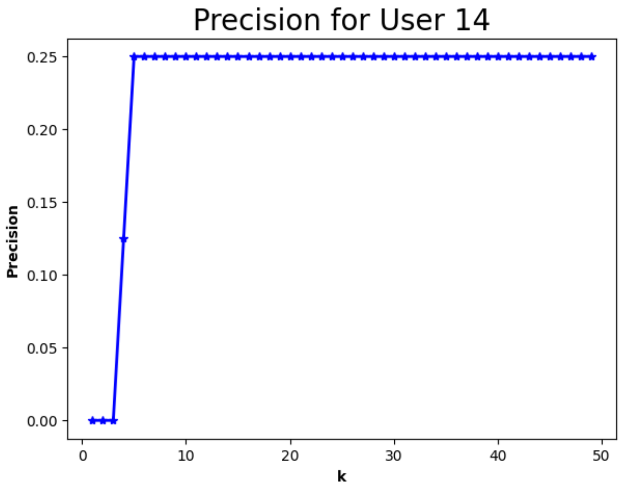

Recall for User 343
Recall
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
k
0
10
20
30
40
50

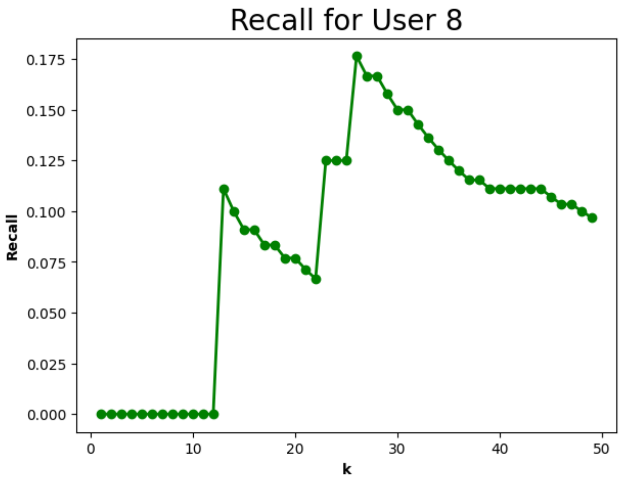

Precision for User 343
Precision
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
k
0
10
20
30
40
50

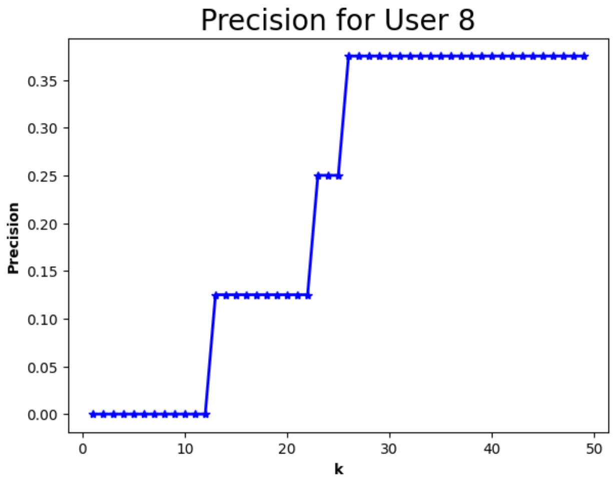

Recall for User 473
Recall
k
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0
10
20
30
40
50

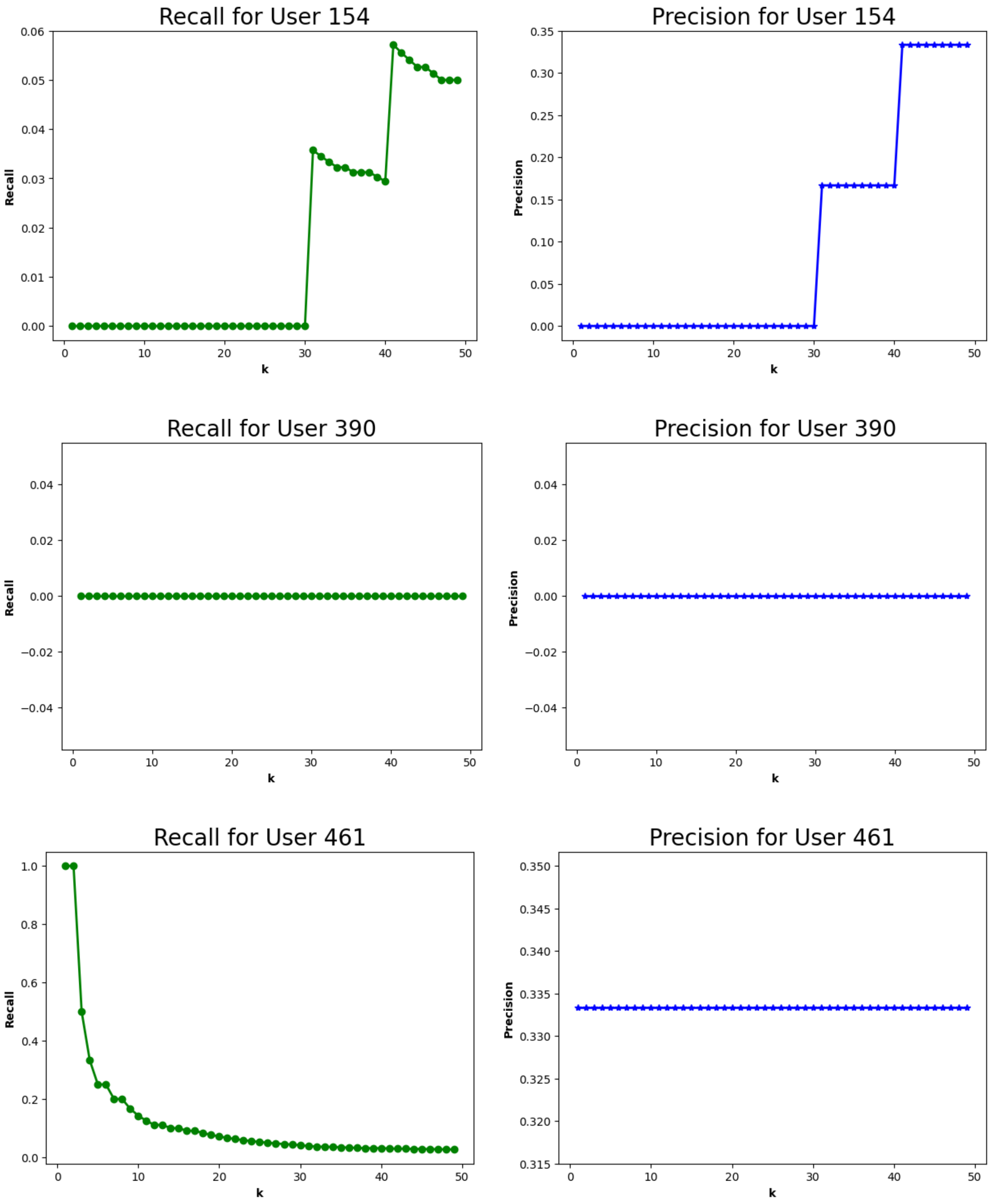

Precision for User 473
Precision
k
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0
10
20
30
40
50

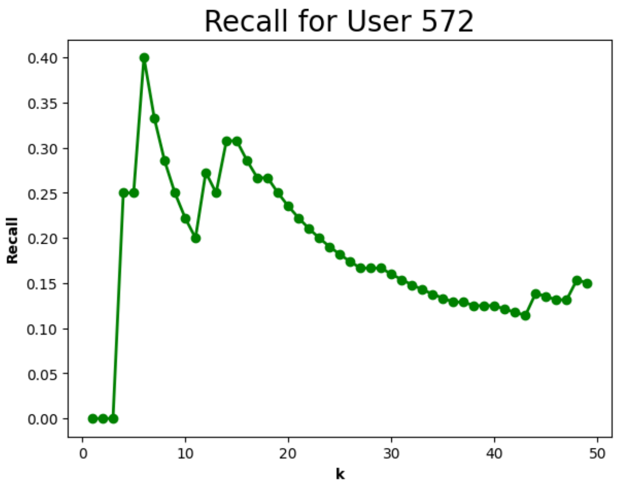

Recall for User 14
Recall
k
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

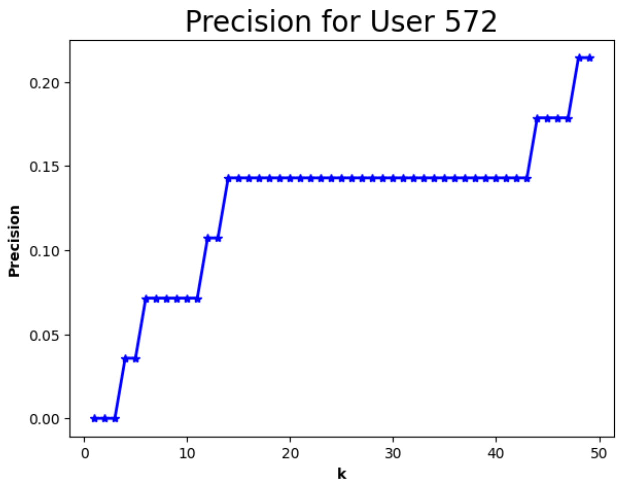

Precision for User 14
Precision
k
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50

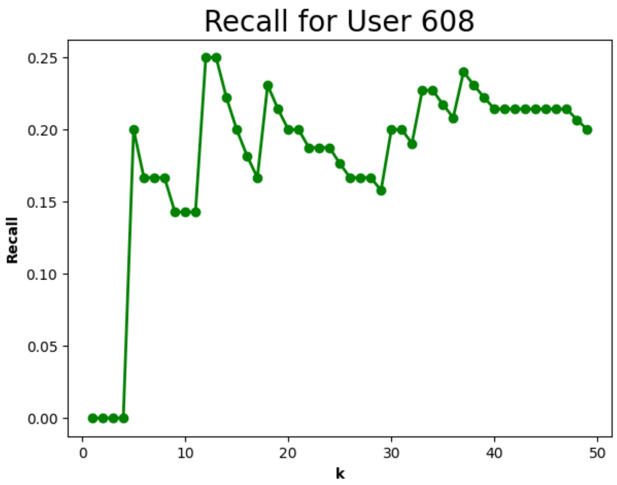

Recall for User 8
Recall
k
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0
10
20
30
40
50

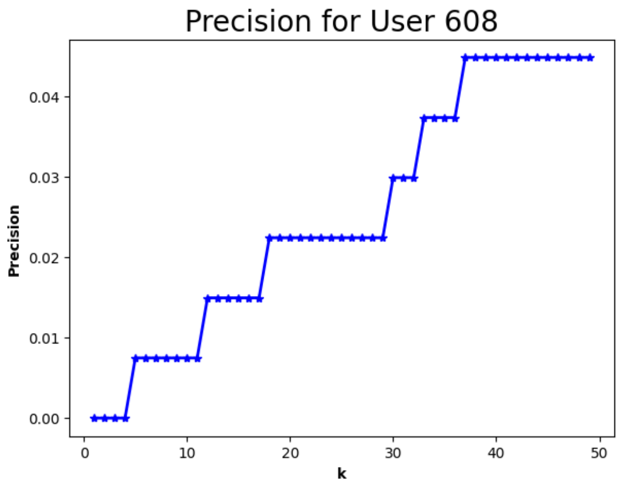

Precision for User 8
Precision
k
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50

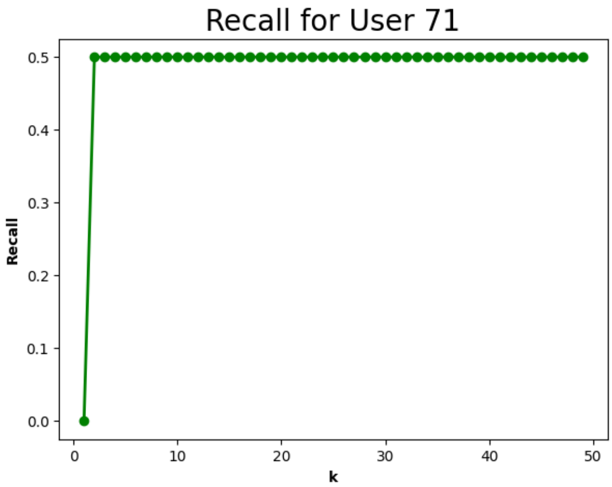

Recall for User 154
Recall
k
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
10
20
30
40
50
Precision for User 154
Precision
k
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50
Recall for User 390
Recall
k
−0.04
−0.02
0.00
0.02
0.04
0
10
20
30
40
50
Precision for User 390
Precision
k
−0.04
−0.02
0.00
0.02
0.04
0
10
20
30
40
50
Recall for User 461
Recall
k
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Precision for User 461
Precision
k
0.315
0.320
0.325
0.330
0.335
0.340
0.345
0.350
0
10
20
30
40
50

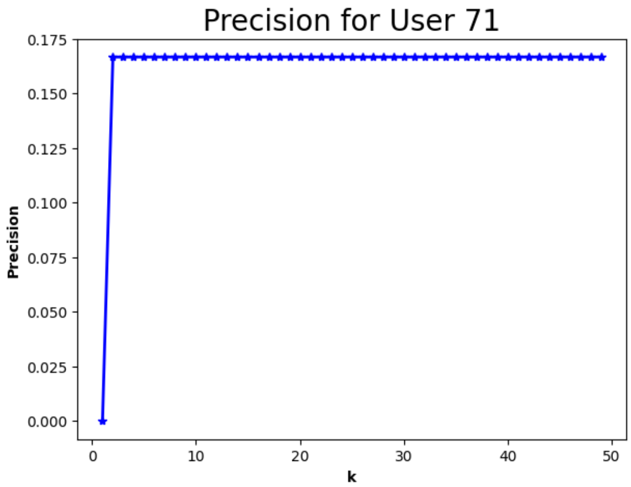

Recall for User 572
Recall
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
10
20
30
40
50
k

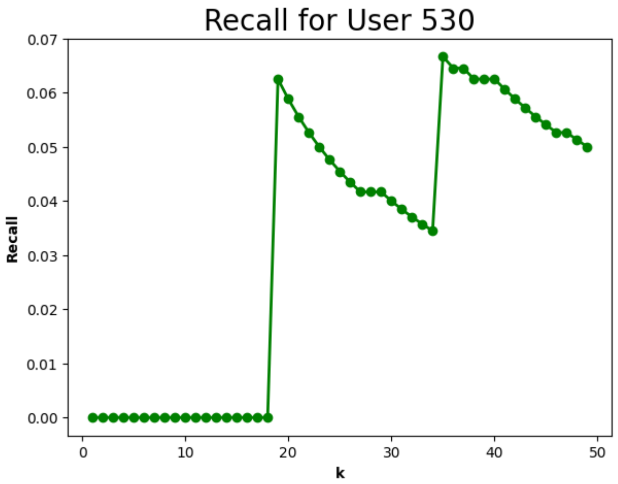

Precision for User 572
Precision
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40
50
k

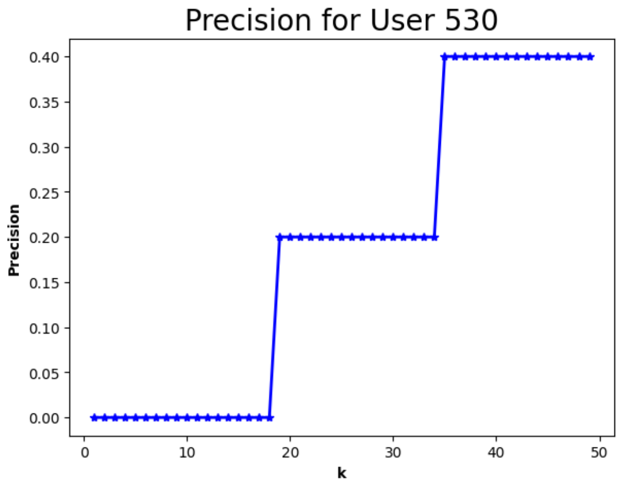

Recall for User 608
Recall
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50
k

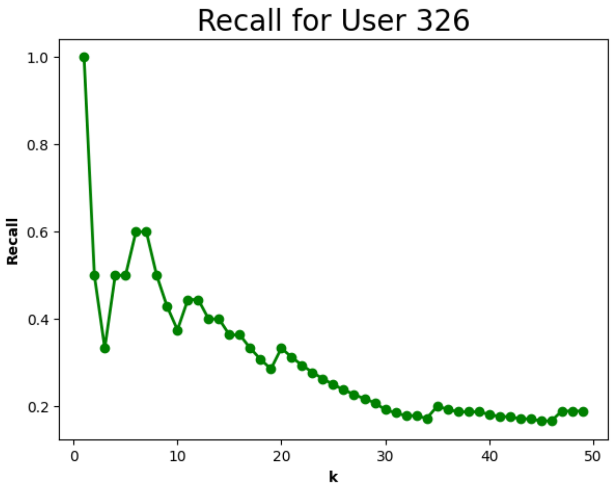

Precision for User 608
Precision
0.00
0.01
0.02
0.03
0.04
0
10
20
30
40
50
k

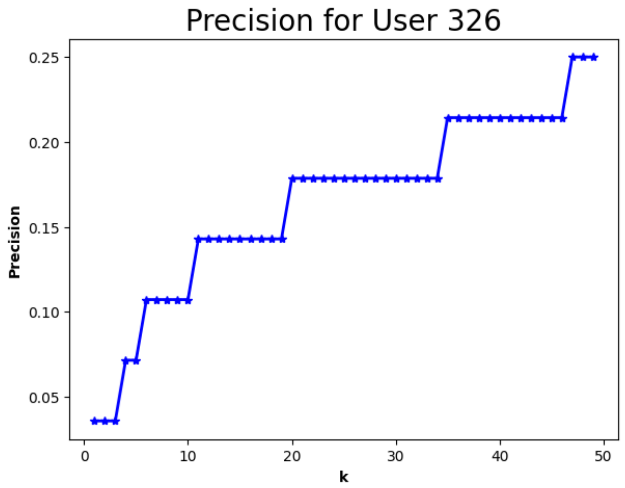

Recall for User 71
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50
k

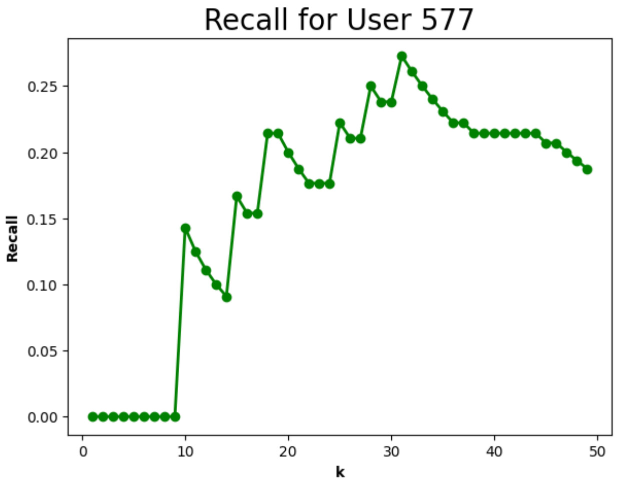

Precision for User 71
Precision
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0
10
20
30
40
50
k

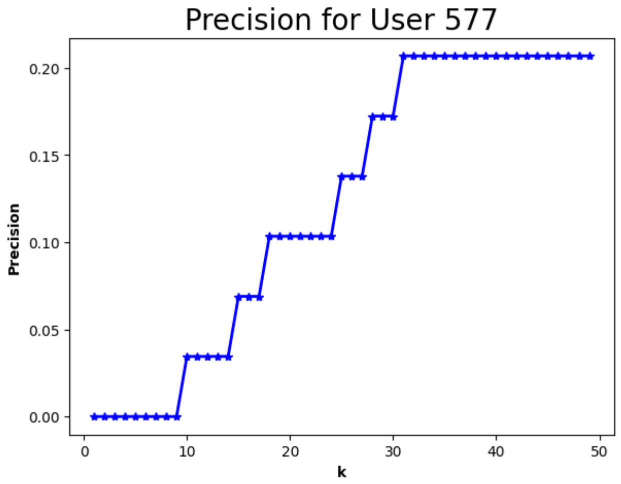

Recall for User 530
Recall
k
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
10
20
30
40
50

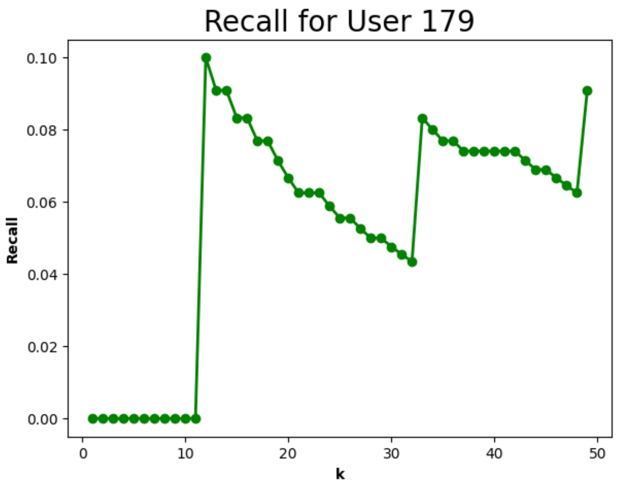

Precision for User 530
Precision
k
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
10
20
30
40
50

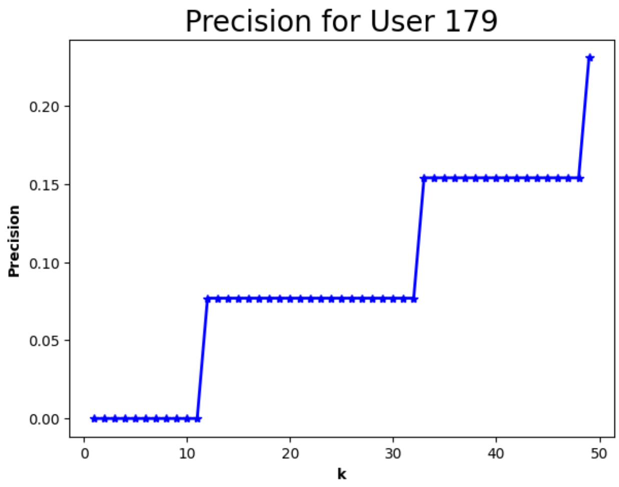

Recall for User 326
Recall
k
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

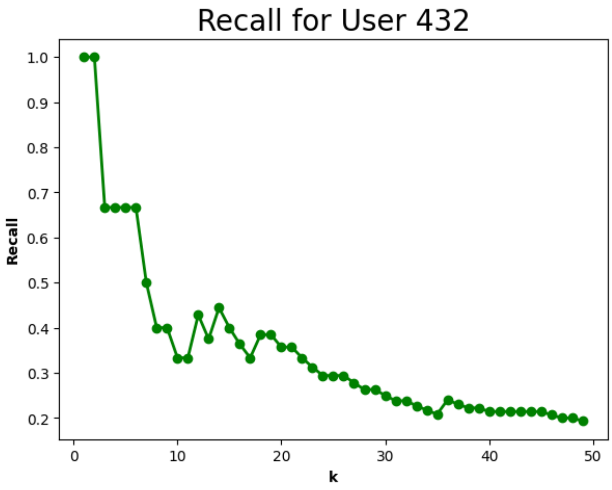

Precision for User 326
Precision
k
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50

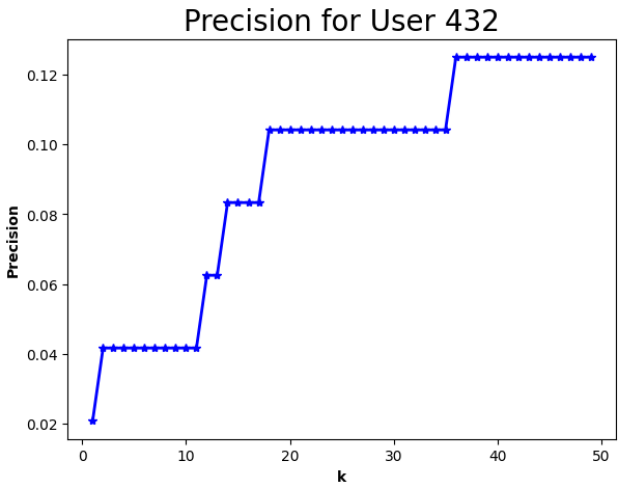

Recall for User 577
Recall
k
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50

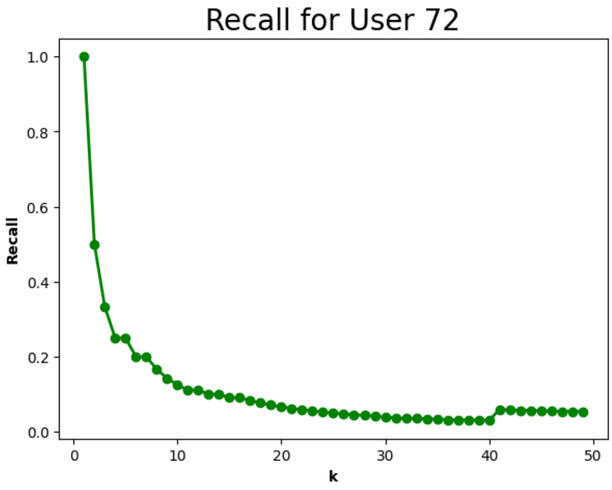

Precision for User 577
Precision
k
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40
50

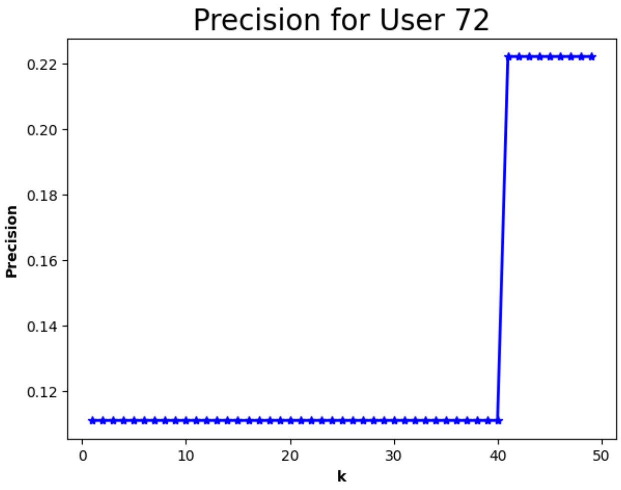

Recall for User 179
Recall
k
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50


Precision for User 179
Precision
k
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40
50


Recall for User 432
Recall
k
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
10
20
30
40
50


Precision for User 432
Precision
k
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30
40
50


Recall for User 72
Recall
k
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


Precision for User 72
Precision
k
0.12
0.14
0.16
0.18
0.20
0.22
0
10
20
30
40
50