Analysing Youtube Datasets Structure and Relation

Haihan Jiang, Ruyuan Zuo
Case Western Reserve University
Department of Computer and Data Science

Background and motivation

Youtube and video website alike has recommendation system written to give personalized recommendations

How does properties of the video itself affect the rank of the videos

- We want to understand how are videos related
- What is the order for the recommended videos
- What are some assumptions does the order entail

Is a personalized recommendation system narrow the view of its users, and how?

Problem Statement

1. Our question is to understanding the structure of the youtube video network
    - SQL
    - Visualization
2. How was recommendation algorithm was done
    - Study pageRank as an naive example
3. Learn the underlying assumption between recommendations

Problem applications and Importance

Applications

- PageRank Algorithm in the SQL queries
- Helps to understand what the what is the importance of the videos that you are viewing
- Gives a understanding of the problem recommendation AND select videos that are more important to view/or avoid

Challenges:

- Large dataset and long run-time

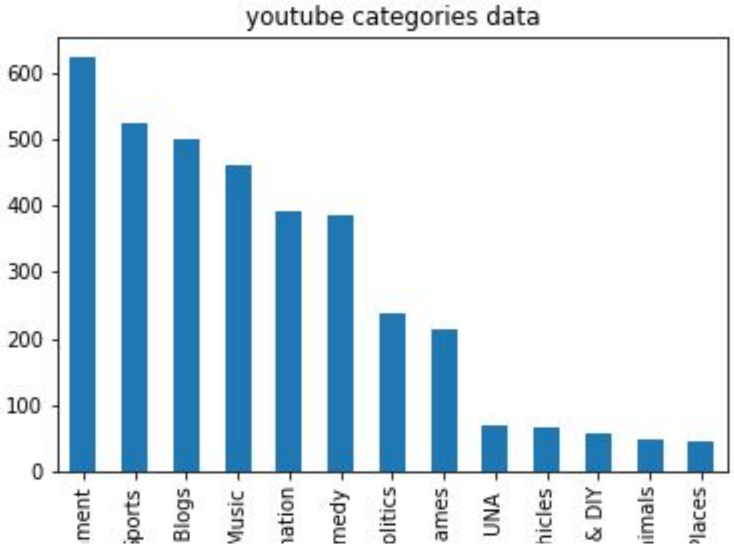
Input/output workflow

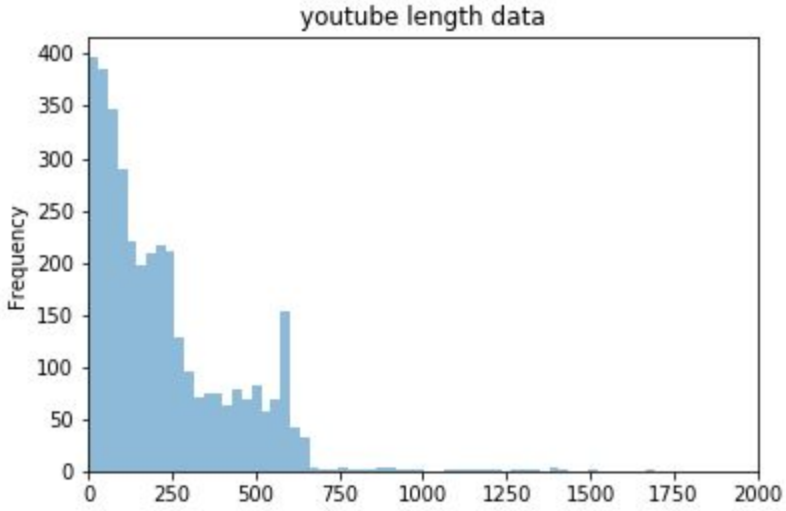
- Raw input is a txt file
- Using a parser written by Java to structured the data and input to SQL
- Using python to perform pageRank calculation
- Validate the assumption with Gelphi visualization and degree informations using NetworkX packages
- Input the rank to SQL again

Problem formulation: Dataset

- 3636 entries, scripting layer = 1
- A public available datasource from from Simon Fraser University British Columbia, Canada
  - http://netsg.cs.sfu.ca/youtubedata/
- Categories
  - video ID an 11-digit string, which is unique
  - uploader a string of the video uploader's username
  - age an integer number of days between the date when the video was uploaded
  - category a string of the video category chosen by the uploader
  - length an integer number of the video length
  - views an integer number of the views
  - rate a float number of the video rate
  - ratings an integer number of the ratings
  - comments an integer number of the comments
  - related IDs up to 20 strings of the related video IDs

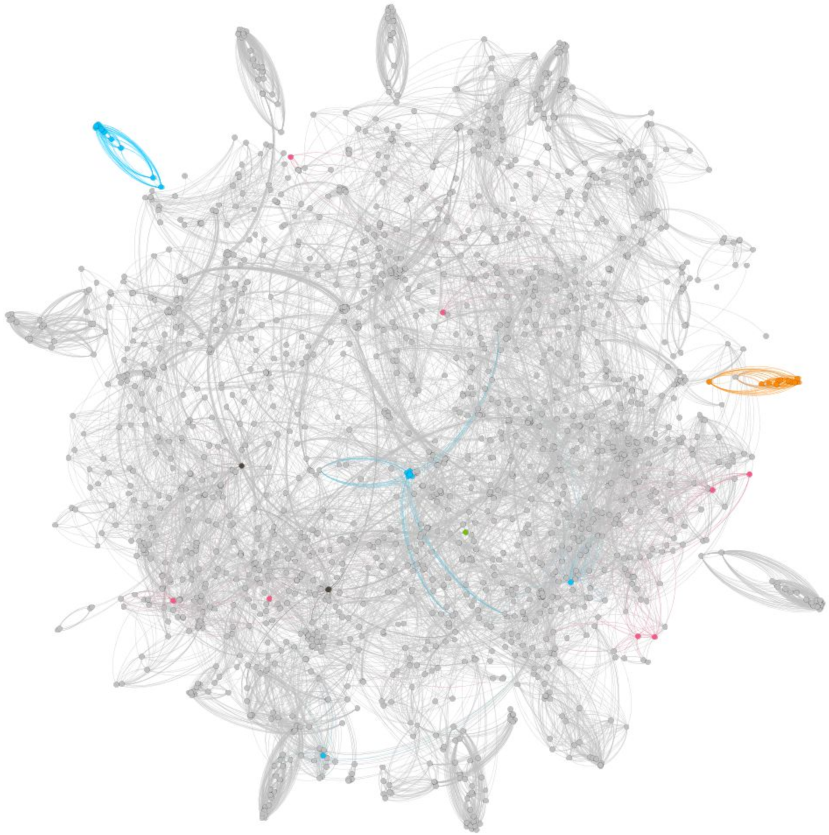
How the dataset looks like



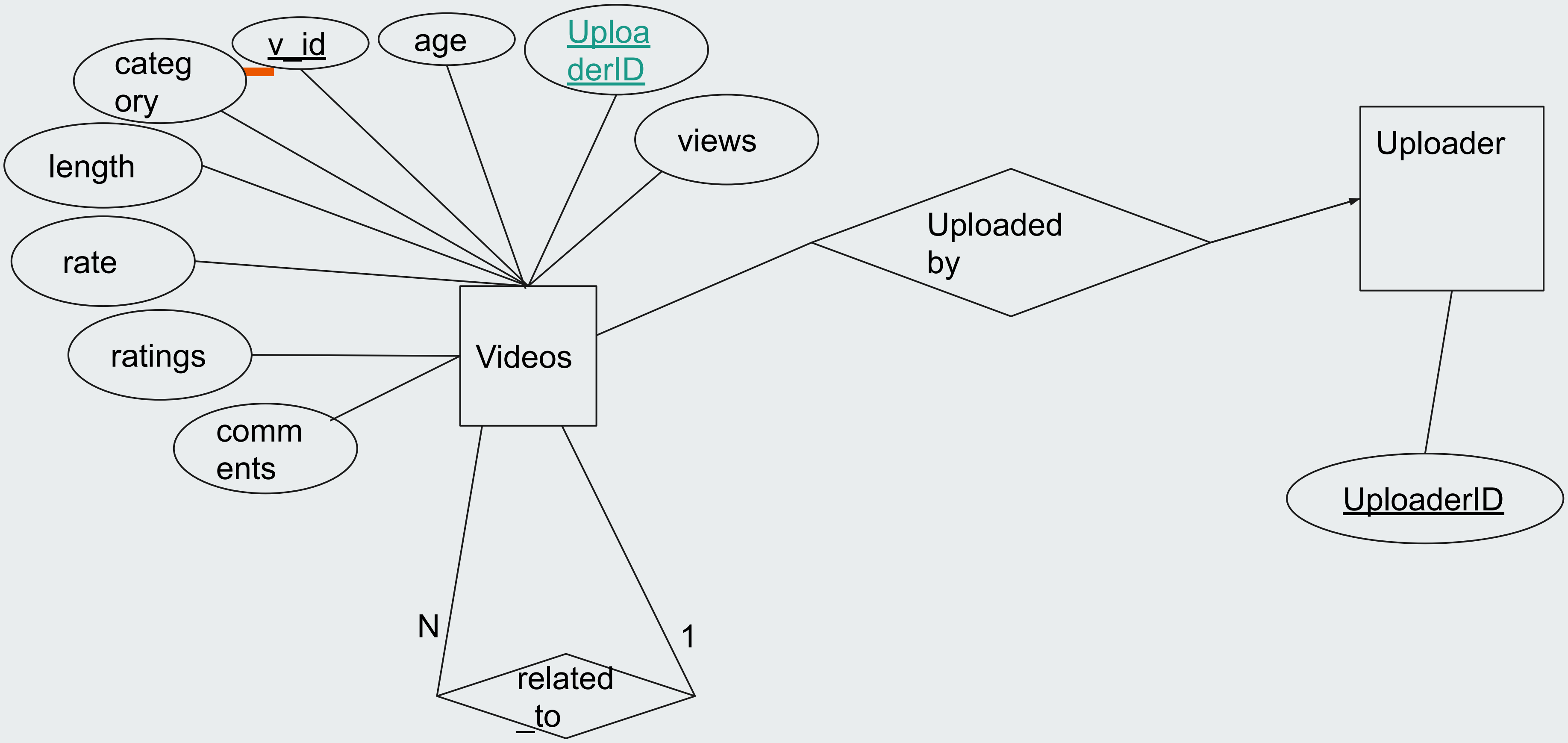


Data Distribution

	videoID	uploader	age	category	length	views	rate	ratings	comments
Max Age	Ho-DLeAdZj4	Kiyaz	**7423.0**	Sports	77.0	1491.0	4.0	3.0	3.0
Max Len	wewuXQpTHTU	mlee71	614.0	News & Politics	**3440.0**	128.0	2.67	3.0	2.0
Max Rate	4q5jSGOcZb8	EA	742.0	Gadgets & Games	92.0	1059.0	**5.0**	6.0	13.0
Max ratings	UmAfQ-GgtCQ	DiGiTiLsOuL	632.0	News & Politics	330.0	1842033.0	4.13	**9279.0**	5854.0
Max Views	4c_Grdrx7t0	horsetak	0.0	UNA	210.0	**24133454.0**	1.93	713.0	132.0
Max related	**yNhDgOzo9n8**	ilovesonic247	524.0	Music	200.0	5239.0	4.55	29.0	18.0

ER- Diagram



categ
ory
v_id
age
Uploa
derID
views
length
rate
ratings
comm
ents
Videos
Uploaded
by
Uploader
UploaderID
N
1
related
_to

MySql Database

- Data Cleaning
- Reading Data (Implemented Java)
- Connection with the database

vid	uploaderID	age	category	length	views	rate	ratings	comments
_-lcaXabZ8I	jhjmonnee	741	Sports	9	1732	0	0	1
_yUo9crgm3U	kylaalee	726	Music	218	7372	4.12	17	3
_Zksb5iiM7Q	Harokin	532	FilmandAnimation	449	14865	4.81	57	10
-6qIAMcOeSk	beyazitliprens	667	FilmandAnimation	411	214	0	0	1
-B-BceISUSw	FergusonWellman	625	NewsandPolitics	239	325	0	0	0
-eW6NY2G_YE	MihaelKeehl2	738	FilmandAnimation	471	5083	4.39	18	3
-hYi6CpNYGw	Londer	451	Sports	308	4065	4	5	4
-jzNwO-0wbQ	omarfoh	743	Music	110	346	5	4	2
-KcBW63qW3I	yatsubame	712	Comedy	356	25418	4.53	15	3
-QrG5aspGv4	usingmycomputer	741	Sports	9	21719	4.2	5	14
-U7JTglWpXM	Cloudywolf1337	742	FilmandAnimation	560	6894	5	26	39
-UIiz06lYQI	SGE4LIFE	743	Sports	361	23	0	0	0
-vFRkOd3GjY	jrc0912	743	Sports	28	329	0	0	0
01xAOGHxY5E	franxd	742	Entertainment	41	1544	0	0	1
048Y216L6YM	IndieMusicBlog	741	Music	318	166	0	0	0

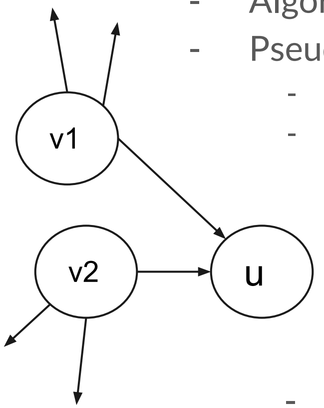
SQL Queries

- select vid from videos where ratings>1 & ratings <5 ;
- select ratings from videos where category = "Sports" & length<5;

vid
1HxER8CPamY
1lYXY_eTNY8
1umiJrKfpdk
20Er3QOq8ks
26h06irOVIE
26ZRpwosc2s
2aDGS2ObyS8
2AYRl_uRW5U
2B9TW4NG1vw
2cidEv9vchE
2EfuhjnvIsA

ratings
0
17
57
0
0
18
5
4
15

PageRank - Algorithm



- Algorithm: Google Search Engine -- Larry Page (1998)
- Pseudocode
  - Assign weight matrix for directed graph
  - Calculate the individual node according to the sum of the probability of the current node over total out node

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

  - Iterative until convergence
  - Added in damping factors to force convergence

Each iteration requires a whole matrix multiplication, so the complexity is the size of the table, which is $O(n^2)$ time complexity, where n is the size of nodes in the network

Example PageRank

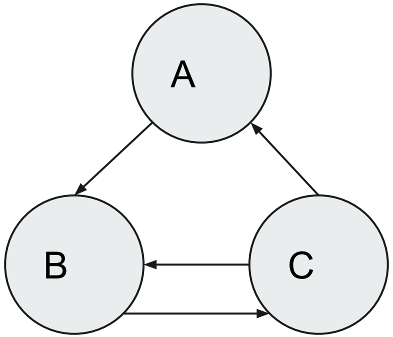


Table:

Itr = 0	Itr = 1	Itr = converge
1/3	1/6	...
1/3	1/2	...
1/3	1/3	...

Iteration 1
p(a)=⅓ / 2
p(b)=⅓ / 1 + ⅓ /2 = 1/2
p(c)=⅓ / 1 = 1/3

PageRank output

- Top five node
    - ALjw5wrhzU: 9.248e-05,
    - y0_XLRcKH_Y:: 9.207e-05,
    - Er3K59aVJmM: 9.050e-05,
    - LclVYYHxoEA:: 8.693e-05,
    - 8240cBUuP-c: 8.447e-05

	videoID	uploader	age	category	length	views	rate	ratings	comments
38846	Er3K59aVJmM	SensNetworkDotCom	739.0	Sports	140.0	13562.0	5.0	13.0	90.0

```
1  new_list[new_list['relatedID'] == 'Er3K59aVJmM']
```

	videoID	uploader	age	category	length	views	rate	ratings	comments	relatedID
388	bk5WiqF0AVM	SensNetworkDotCom	494.0	Sports	104.0	8276.0	5.00	2.0	9.0	Er3K59aVJmM
548	Bp0tGnQgeZw	SensNetworkDotCom	494.0	Sports	9.0	296.0	0.00	0.0	0.0	Er3K59aVJmM
594	tD9Jr7GP67c	SensNetworkDotCom	739.0	Sports	57.0	5300.0	4.44	9.0	38.0	Er3K59aVJmM
30599	DXXtJ_Xi79c	NYR21135	741.0	Sports	88.0	46.0	0.00	0.0	0.0	Er3K59aVJmM
30619	vN6_8Bj3Mw8	SyntaxZ	739.0	Sports	27.0	1910.0	1.50	2.0	2.0	Er3K59aVJmM
30639	95EjlJdbxlw	eassona	740.0	Sports	55.0	441.0	5.00	2.0	0.0	Er3K59aVJmM
30676	NaN	jedi24	646.0	Sports	60.0	722.0	0.00	0.0	0.0	Er3K59aVJmM
30697	LPZ_XlHWdh0	GRiSHoCKEY	741.0	Sports	201.0	397.0	5.00	1.0	0.0	Er3K59aVJmM
30705	nPD-h2cMYaM	Disengage	477.0	Sports	62.0	9818.0	5.00	8.0	7.0	Er3K59aVJmM
30737	rtWos0ZsK8Q	GRiSHoCKEY	642.0	Sports	60.0	3098.0	3.00	1.0	0.0	Er3K59aVJmM
30769	CrURQGlRKgo	Bush6	434.0	Entertainment	62.0	14629.0	4.46	13.0	21.0	Er3K59aVJmM
30783	XqkoTmKlgoA	ghollett	741.0	Sports	88.0	165.0	5.00	1.0	1.0	Er3K59aVJmM
30802	ZQUuzpbCirc	ghollett	727.0	Sports	68.0	3574.0	4.20	5.0	0.0	Er3K59aVJmM
30829	ZYtEzFSGpAw	PandoraJam	740.0	Sports	171.0	1187.0	5.00	2.0	7.0	Er3K59aVJmM
30840	PlcqPqm9zJ0	hockeyfightsdotcom	739.0	Sports	53.0	5089.0	5.00	3.0	0.0	Er3K59aVJmM
30859	kQI5rbh7_G4	hockeyfightsdotcom	739.0	Sports	60.0	5227.0	5.00	1.0	0.0	Er3K59aVJmM
30878	lleUaVdSDbw	hockeyfightsdotcom	739.0	Sports	72.0	5142.0	5.00	2.0	0.0	Er3K59aVJmM
30897	g9p3R_JwXb0	hockeyfightsdotcom	739.0	Sports	60.0	7303.0	5.00	1.0	1.0	Er3K59aVJmM
34161	Hx6I5Hu2Nk8	SensNetworkDotCom	740.0	Sports	452.0	3063.0	4.86	7.0	9.0	Er3K59aVJmM
34187	u4q7_YsKQxQ	daesharacor	740.0	Sports	73.0	1382.0	5.00	3.0	2.0	Er3K59aVJmM
34227	Y524JXOSR6M	SensNetworkDotCom	739.0	Sports	99.0	8327.0	4.20	10.0	79.0	Er3K59aVJmM
34254	d51Xr5qK7Lw	afilmby	741.0	Sports	109.0	613.0	0.00	0.0	0.0	Er3K59aVJmM
38677	wRIlZXbKBQY	alexygd	740.0	Sports	53.0	61.0	0.00	0.0	0.0	Er3K59aVJmM
									59.0	Er3K59aVJmM
									4.0	Er3K59aVJmM
									19.0	Er3K59aVJmM
									0.0	Er3K59aVJmM
38894	qyvbQswbN-E	SyntaxZ	737.0	Sports	273.0	3037.0	5.00	6.0	9.0	Er3K59aVJmM
38932	4x4eitLuA7w	News99forYou	738.0	Sports	220.0	11563.0	4.35	17.0	35.0	Er3K59aVJmM
38954	gmuwtCeUfkA	RHSproductions	738.0	Sports	284.0	21934.0	4.91	68.0	130.0	Er3K59aVJmM
43652	guXjmXZ0Jmo	afilmby	740.0	Sports	120.0	12855.0	3.50	18.0	45.0	Er3K59aVJmM
43671	DZT2PlyE5Vg	KILLERCONVIC	738.0	Sports	273.0	44585.0	4.93	83.0	145.0	Er3K59aVJmM
43912	nup3zVwqEVc	sport24greece2	739.0	Sports	242.0	11093.0	5.00	12.0	25.0	Er3K59aVJmM
47270	X6wUPNyXqs4	jedi24	341.0	Sports	81.0	10291.0	4.14	7.0	7.0	Er3K59aVJmM
47420	IZCE0lZ0uXM	SensNetworkDotCom	494.0	Sports	126.0	13240.0	4.83	12.0	8.0	Er3K59aVJmM

Network Analysis

- Nodes: 32752
- Edges: 67911
- Average degree: 2.073
- SCC: 30903
- WCC: 172

- avg_clust: 0.16543
- All at Gadgets & Games category

```
In [185]:  1  bet_cen
Out[185]: {'IqlxYO7YCI8': 0.0,
 'VnLVtz4Vq18': 0.0,
 'krT9Pjy9d8s': 0.0,
 '1umiJrKfpdk': 0.0,
 '1gxK1e5MSYg': 0.0,
 '2aDGS2ObyS8': 0.0,
 'FyuYJsBavBs': 0.0,
 'vURuMxGC53A': 0.0,
 'uG1Q5LhqpsM': 0.0,
 'FkKWCBWVwQg': 0.0,
 'PmRHEQaCFsw': 0.0,
 'N4DdAIc_0tY': 0.0,
 'AYNFCy6hvFQ': 0.0,
 '2rwktobtv9s': 0.0,
 'U0raaoN6I6M': 0.0,
 'OW_Azt-ZFvI': 0.0,
 'SQI9xPF9rdk': 0.0,
 'VdHsMJRszck': 0.0,
 '4q5jSGOcZb8': 0.0,
 'xXfmxQO2xz0': 0.0}

In [183]:  1  top_clo_cen
Out[183]: {'IqlxYO7YCI8': 1.0,
 'VnLVtz4Vq18': 1.0,
 'krT9Pjy9d8s': 1.0,
 '1umiJrKfpdk': 1.0,
 '1gxK1e5MSYg': 1.0}

In [184]:  1  top_eig_cent
Out[184]: {'IqlxYO7YCI8': 0.22412950080863098,
 'VnLVtz4Vq18': 0.22412950080863098,
 'krT9Pjy9d8s': 0.22412950080863098,
 '1umiJrKfpdk': 0.22412950080863098,
 '1gxK1e5MSYg': 0.22412950080863098}
```

PageRank in SQL

- After calculation of the pageRank scores, we input the rank probabilities back to SQL to add a column of the ranks for each videos

vid	uploaderID	age	category	length	views	rate	ratings	comments	pageRank
1umiJrKfpdk	EA	742	Gadgets & Games	77	1043	7	9	7	6.96531
4q5jSGOcZb8	EA	742	Gadgets & Games	92	1059	5	6	13	6.96531
AYNFCy6hvFQ	EA	742	Gadgets & Games	74	454	5	3	13	6.96531
OW_Azt-ZFvI	EA	742	Gadgets & Games	291	227	3.67	3	5	6.96531
SQI9xPF9rdk	EA	742	Gadgets & Games	68	1518	4.79	14	17	6.96531
U0raaoN6I6M	EA	742	Gadgets & Games	61	1128	4.67	9	6	6.96531
vURuMxGC53A	EA	742	Gadgets & Games	105	884	5	5	9	6.96531

Future Work & Conclusion

Future Works

- Need consider all four layers of data
  - Was not able to do because the data size
- Real website uses much complicated algorithms than PageRank

Conclusion

- Using pageRank algorithm to study the relationships between videos than using SQL is much more effective method in terms of learning about clustering and connections
- Videos has higher rank will induces a strong preference for recommending the same categories

Q&A