Project Statement for Milestone 4

Kongqueror
Ross Kugler, Huy Ky, Ben Bordon

Overview:

At the end of Milestone 4 of the project, teams should have developed a prototype of an end-to-end application with a graphical user interface. Teams should have integrated the graphical user interface with a NoSQL database and Hadoop/Spark framework. In cases where teams are focused on a research topic, they should have validated their findings and compared them against the existing methods.

Teams should have updated their database schema and algorithms for robustness and scalability, and should have validated the scalability of their solution using a large dataset on distributed instances of a NoSQL database and Hadoop/Spark.

All team members are expected to make a significant contribution to the project milestone tasks.

Project Report Topics:

The report should cover the following subtopics and answer the questions listed:

1. User Interface and Data Visualization:
    a. Describe the user interface and data visualization components of the software. The user interface should be interactive, taking inputs from a user and providing outputs, i.e., results, to the user. The user interface should include data visualization.

    Home Page: does a connection check with MongoDB and Spark. Has an overview of the dataset, including number of videos, number of edges. Has simple figures, including a pie-chart of video categories, and bar chart of top creators.

    Network Statistics: allow users to view degree distributions and categorized statistics. User selects a sample size and runs on-demand queries. The result is then presented to user in tables, bar graphs, and pie charts

    Top-K Queries: user chooses between Top Categories, Most Viewed Videos, and Highest Rated Videos. User selects parameters (K, minimum ratings) and requests on-demand queries. The results are shown with tables and bar graphs.

    Range Queries: Filter videos using multiple criteria. Users send on-demand queries filter by duration, views, and category. Results are visualized using graphs and detailed in tables.

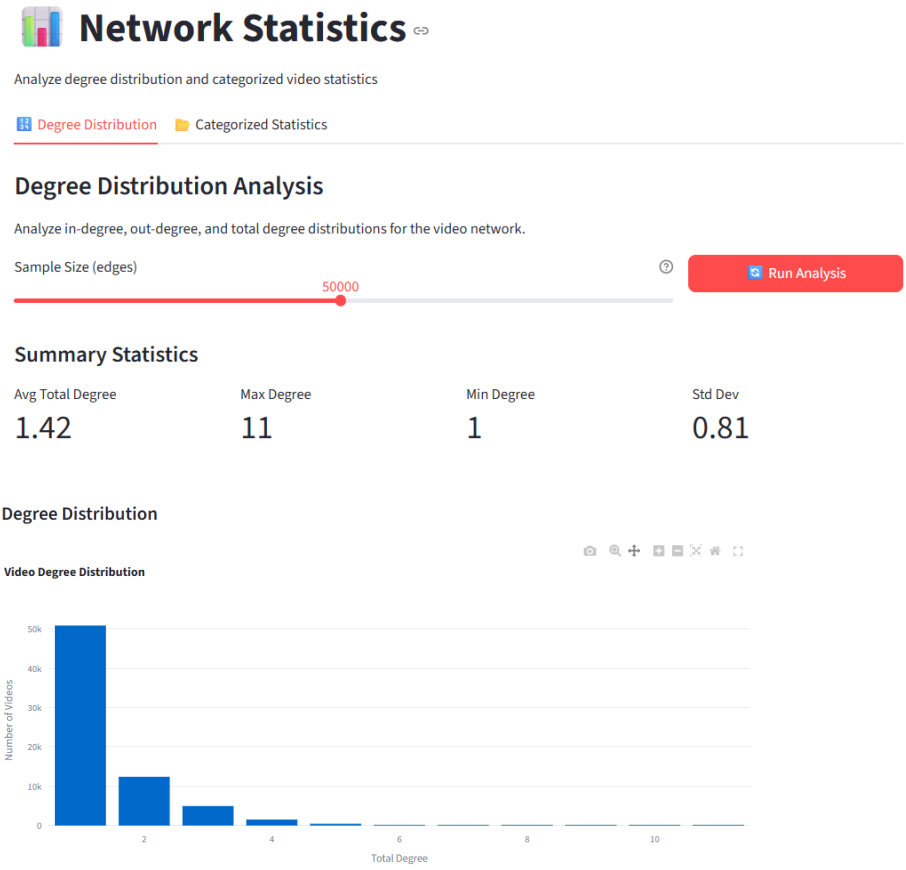
    Graph Analytics: Compute PageRank scores and analyze influence.

    Pattern Search: The tool allows the user to find patterns in the network. The three patterns we have implemented are related pairs, where video A recommends video B; video chains, where video (a) → (b) → (c); and common recommendations,

where video (a) → (c) ← (b). In addition, the user can filter by category, set a maximum number of results, and configure the sample size (edges). After running the algorithm, the tool outputs the results in a table and in a graph visualizer. The user can also save their results into MongoDB.

2. User Queries and Results:
   a. Provide all user queries and their results to describe the overall functionality of the software.

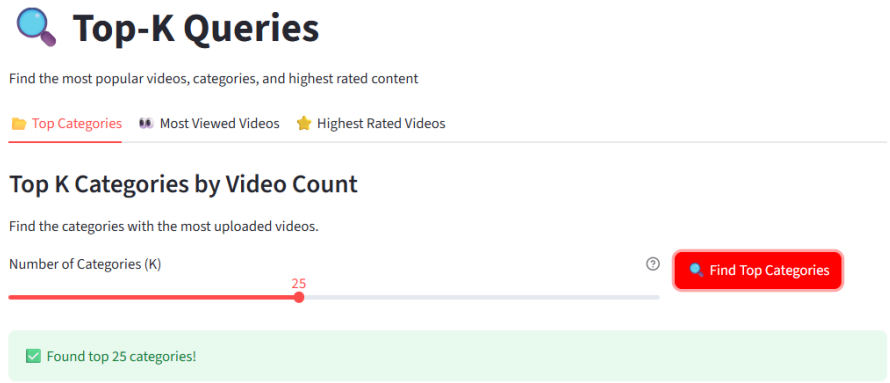
Network Statistics:



Top 10 Videos by Degree

Video	Total Degree	In-Degree	Out-Degree
Q-9ry38AhbU	11	10	1
xpcUxwpOQ_A	10	7	3
xaHEa_EMw54	10	9	1
diZFDhZ8eII	9	5	4
G_inMOZ8zPU	9	8	1
f-mlcoabhpc	8	7	1
trkl77YDfv8	8	5	3
5F5aCUNE4Z8	8	4	4
vrEQBj-CRQA	8	7	1
NBMupEuo0Zo	8	4	4

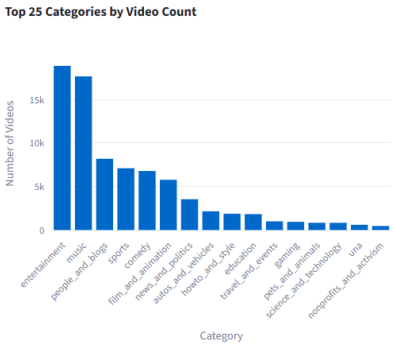
Top-K queries:



Results Table

	Category	Video Count
1	entertainment	18871
2	music	17659
3	people_and_blogs	8192
4	sports	7103
5	comedy	6785
6	film_and_animation	5772
7	news_and_politics	3539
8	autos_and_vehicles	2149
9	howto_and_style	1875
10	education	1827

Distribution Chart



Top Categories | Most Viewed Videos | Highest Rated Videos

Top K Most Viewed Videos

Find the videos with the highest view counts.

Number of Videos (K)

15

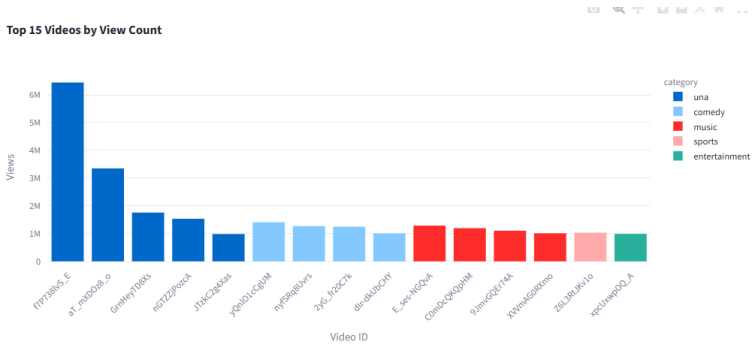
5 100

Find Most Viewed

Found top 15 most viewed videos!

Results Table

Video	views	category	rating	num_ratings
f7P738lvS_E	6446580	una	4.34	3904
aT_mXDOz8_o	3349815	una	4.67	4370
GrnHeyTD8Xs	1757440	una	4.69	2432
nGTZZjPozcA	1534401	una	4.34	1463
yQnlO1cCgUM	1410488	comedy	4.38	895
E_ses-NGQvA	1288157	music	4.86	1202
nyfSRq8Uvrs	1270697	comedy	4.84	402
2yG_fr20C7k	1249627	comedy	4.10	990
C0mDcQKQpHM	1195238	music	4.80	600
9JmvGQEr74A	1104160	music	4.76	834
Z6L3RtJKv1o	1031723	sports	4.53	730
XVVmAG0RXmo	1010363	music	4.90	757
dIr-dkUbCHY	1005276	comedy	3.05	185
xpcUxwpOQ_A	988775	entertainment	4.86	7654
JTzkC2g4Xas	982604	una	4.63	180



Top Categories | Most Viewed Videos | Highest Rated Videos

Top K Highest Rated Videos

Find the videos with the highest ratings.

Number of Videos (K)

10

Minimum Ratings Required

500

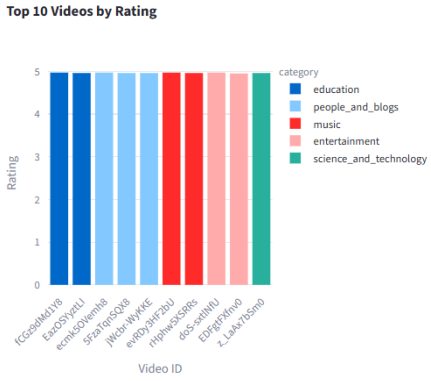
Find Top Rated

Found top 10 highest rated videos (min 500 ratings)!

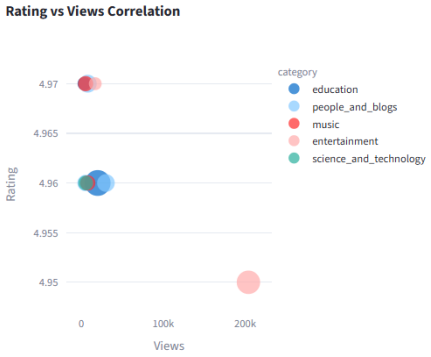
Results Table

Video	rating	num_ratings	views	category
fCGz9dMd1Y8	4.97	533	3414	education
ecmk5OVemh8	4.97	960	7804	people_and_blogs
evRDy3HF2bU	4.97	698	5130	music
doS-sxtlNfU	4.97	502	16914	entertainment
5FzaTqnSQX8	4.96	943	29991	people_and_blogs
rHphw5XSRRs	4.96	698	7832	music
EazOSYyztLI	4.96	2023	19899	education
z_LaAx7bSm0	4.96	660	5249	science_and_technology
jWcbr-WyKKE	4.96	802	5068	people_and_blogs
EDFgtFXfnv0	4.95	1662	204052	entertainment

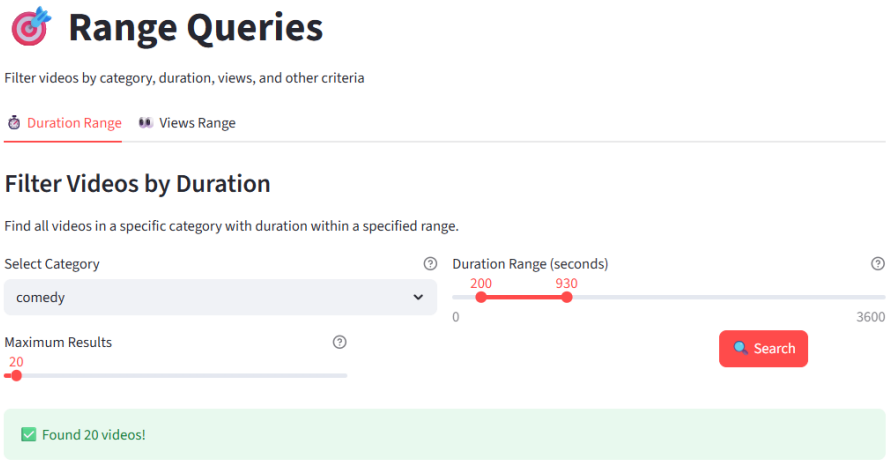
Rating Distribution



Rating vs Views

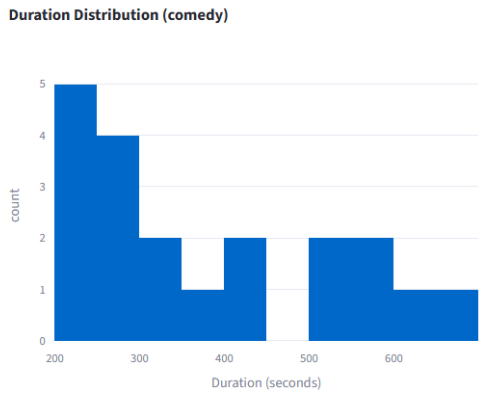


Range queries:



Results: comedy videos (200-930 sec)

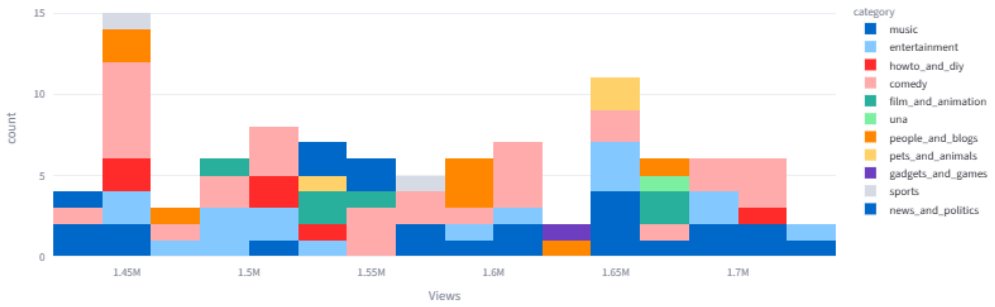
Video	Duration	Views	Rating
MVbv6r_tKnE	4:57	775457	4.75
ik2iR8mlZyM	4:08	571747	4.70
P11J07k51ss	4:16	571279	4.87
Wtposdpspn8	3:21	482292	4.80
zn6c53V1LrE	4:15	444657	4.70
Daq2GhJWmml	9:30	415215	4.70
EI_VtKutD30	4:22	405018	4.19
6ihpfM0KXMU	10:24	398407	4.85
tXhak_gPQg0	8:59	393957	4.57
Kv9S6YPQ8gw	5:58	391924	4.42
wJPVVk9UU9s	3:33	371730	4.83

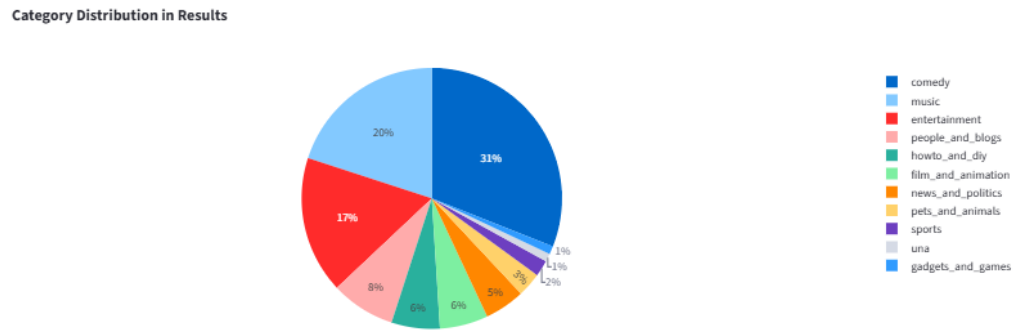


Range Query (View Count All categories 450,000 to 1740000 view)

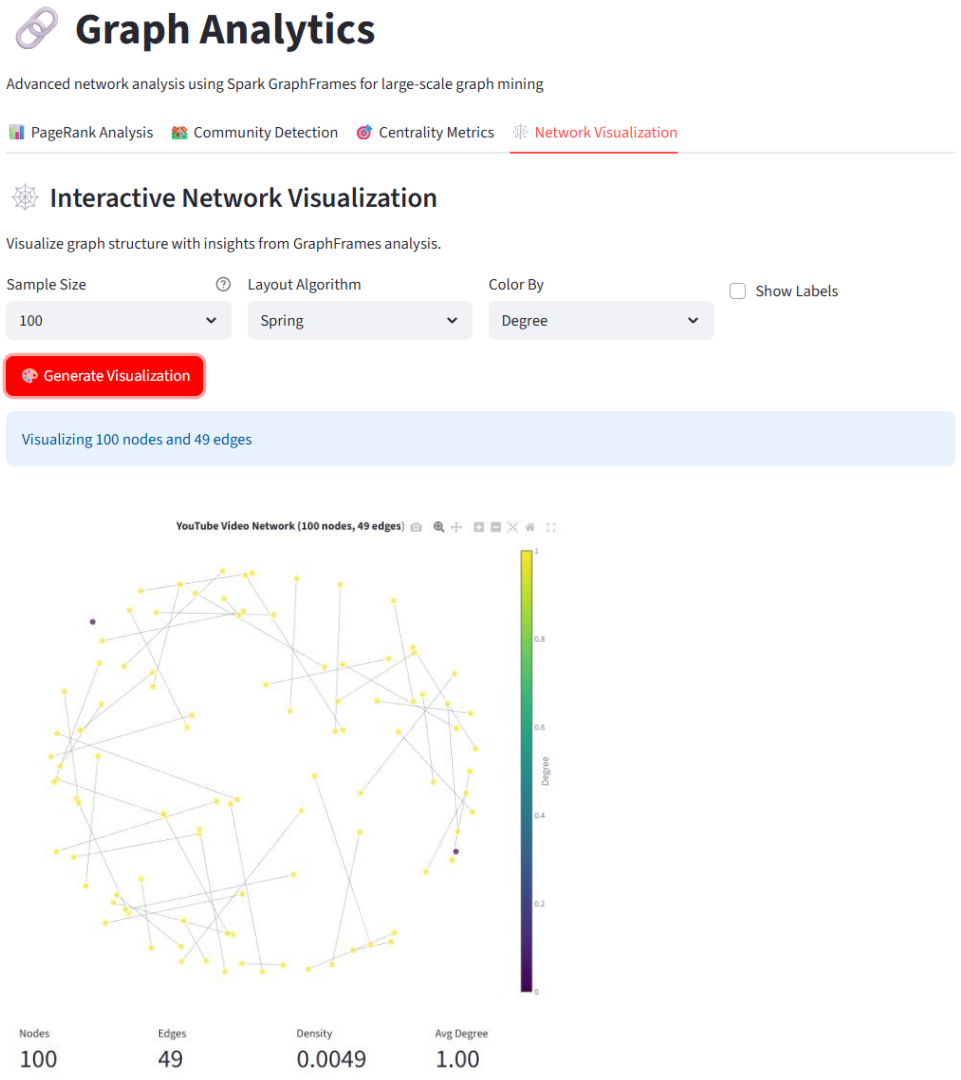
Results: Videos in all categories with 450,000-1,740,000 views

	Video ID	Views	Rating	# Ratings	Category	Duration (s)
1	TcreQQqyNtU	1727996	4.46	424	music	204
2	w2xUzv6iZWo	1726429	0.00	0	entertainment	243
3	MVc2HICdLQ	1719291	4.01	7815	music	139
4	STQ3nhXuuEM	1713974	4.61	2741	howto_and_diy	39
5	oc4oP_ITIqMc	1713532	4.25	5806	comedy	44
6	0hdeIubQMDw	1710477	3.68	3471	comedy	25
7	cvHeVdlZQPs	1707028	4.67	8973	comedy	187
8	FFHCBASrpY	1701447	4.83	4089	music	252
9	OKQ37AdIr3Y	1693426	2.67	154	entertainment	302
10	Ud_-AuBmp6Q	1683810	4.69	1121	music	238





Graph Analytics:



PageRank Analysis | Community Detection | Centrality Metrics | Network Visualization

Centrality Metrics

Measure video importance using various centrality algorithms.

Sample Size (edges)

30000

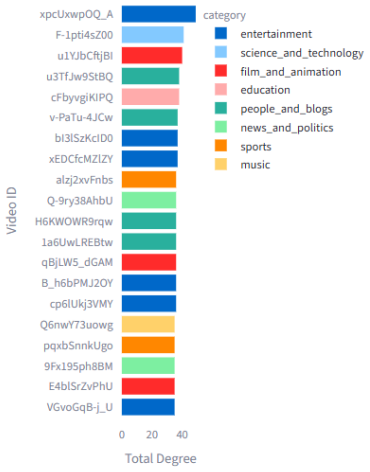
Calculate Centrality Metrics

Analyzing network: 20941 nodes, 30000 edges

Centrality metrics calculated!

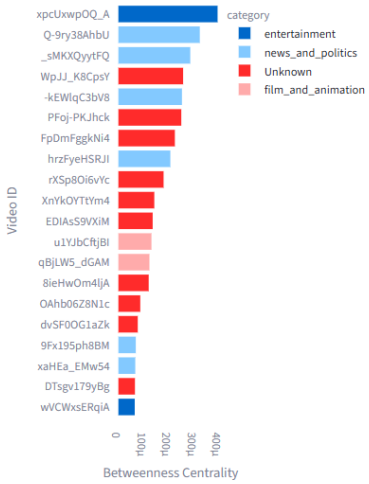
Top Videos by Degree Centrality

Highest Degree Centrality



Top Videos by Betweenness Centrality

Highest Betweenness (Bridge Videos)



Video	Total Degree	Category
xpcUxwpOQ_A	49	entertainment
F-1pti4sZ00	41	science_and_technology
u1YJbCftjBI	40	film_and_animation
cFbyvgiKIPQ	38	education
u3TfJw9StBQ	38	people_and_blogs
xEDCfcMZIZY	37	entertainment
bI3lSzKcID0	37	entertainment
v-PaTu-4JCw	37	people_and_blogs
Q-9ry38AhbU	36	news_and_politics

Advanced network analysis using Spark GraphFrames for large-scale graph mining

PageRank Analysis | Community Detection | Centrality Metrics | Network Visualization

Community Detection

Identify clusters of highly connected videos using Label Propagation algorithm.

Sample Size (edges) 50000

Max Iterations 5

Detect Communities

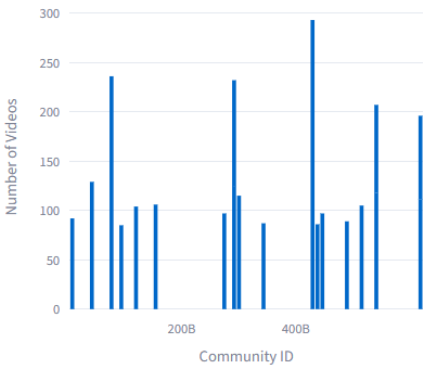
Running community detection algorithm...

Found 2372 communities!

Community Size Distribution

Community Size Categories

Top 20 Largest Communities



PageRank Analysis | Community Detection | Centrality Metrics | Network Visualization

PageRank Analysis

Identify the most influential videos in the network based on link structure.

Sample Size (edges) 30000

Reset Probability 0.15

Max Iterations 5

Run PageRank Analysis

Running PageRank algorithm on video network...

Loaded 30,000 edges for analysis

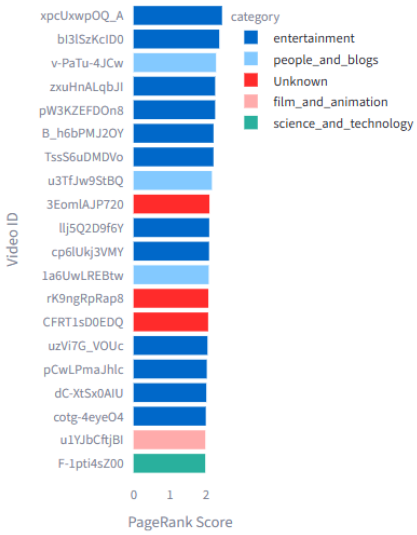
Identified 20,941 unique videos

Computing PageRank (max 5 iterations)...

✅ PageRank completed successfully!

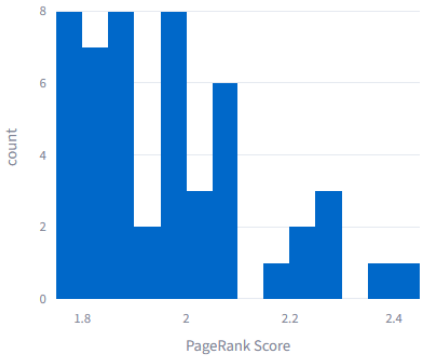
🏆 Top 20 Most Influential Videos

Videos with Highest PageRank Scores



📊 PageRank Distribution

PageRank Score Distribution (Top 50)

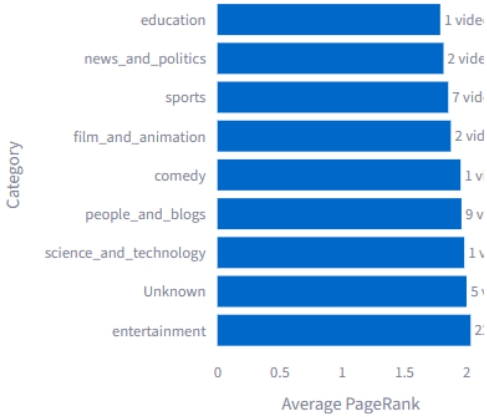


📁 PageRank by Category

Average PageRank by Category

📁 PageRank by Category

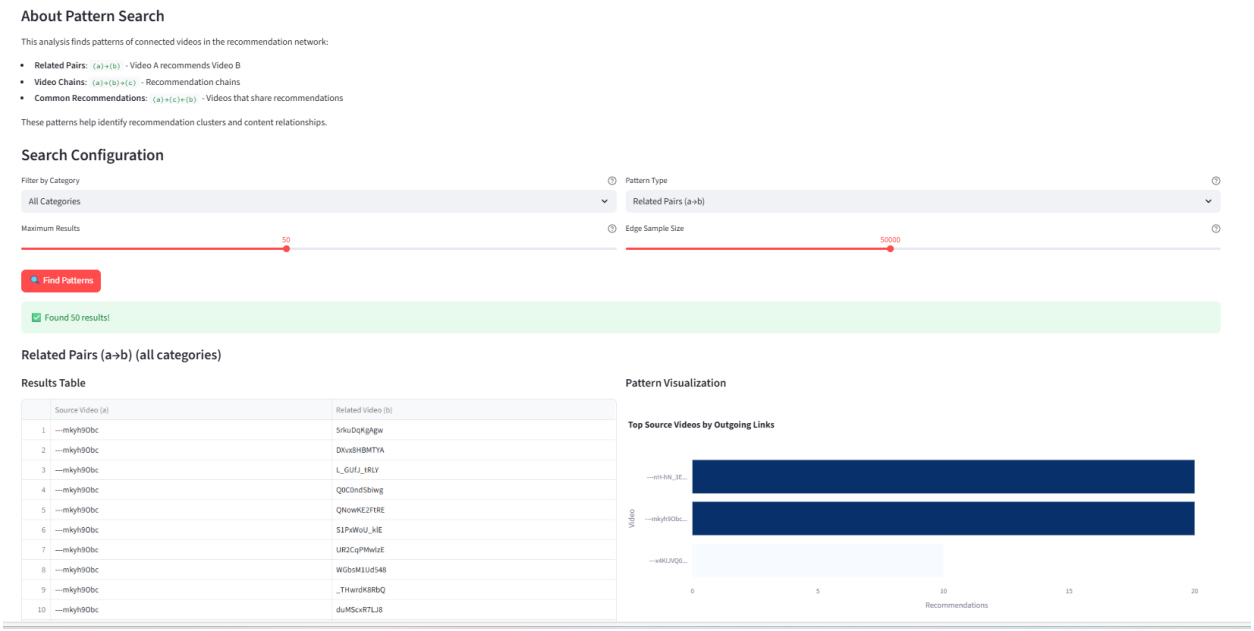
Average PageRank by Category



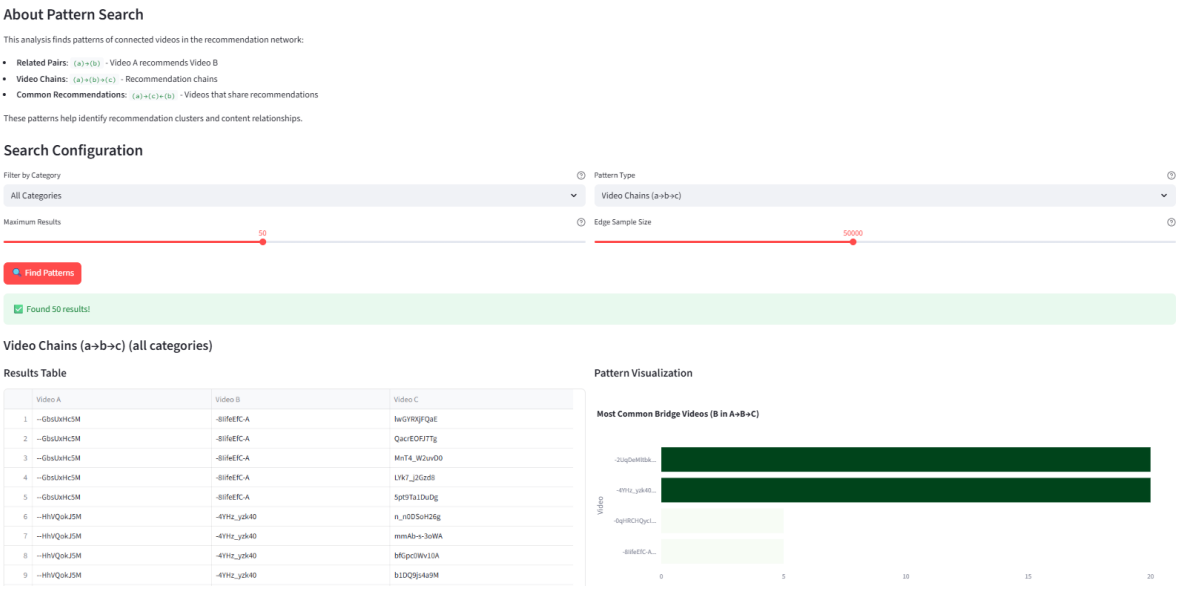
📋 Top 50 Videos by PageRank

Video	PageRank Score	Category	Uploader
xpcUxwpOQ_A	2.444577	entertainment	meepmeepmeepow
bI3lSzKcID0	2.368115	entertainment	MzViSyon
v-PaTu-4JCw	2.278786	people_and_blogs	bootylicious0714
pW3KZEFDOn8	2.254238	entertainment	MzViSyon
zxuHnALqbJI	2.254238	entertainment	MzViSyon
B_h6bPMJ2OY	2.210008	entertainment	tacksy
TssS6uDMDVo	2.208477	entertainment	idance4me22
u3TfJw9StBQ	2.163861	people_and_blogs	11Rohit11
3EomlAJP720	2.096532	Unknown	Unknown
llj5Q2D9f6Y	2.090472	entertainment	evr716
cp6lUkj3VMY	2.086060	entertainment	fayakoxxx
1a6UwLREBtw	2.077284	people_and_blogs	zoczoc1234
rK9ngRpRap8	2.066103	Unknown	Unknown
CFRT1sD0EDQ	2.060772	Unknown	Unknown
uzVi7G_VOUc	2.042572	entertainment	lanefan37

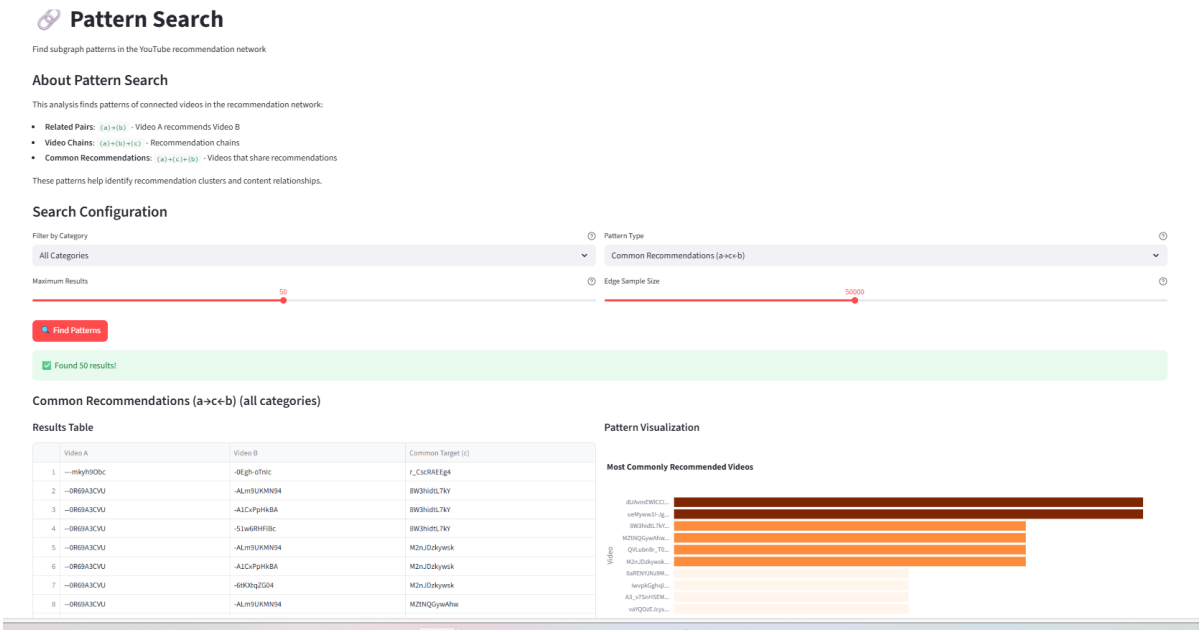
Pattern Search (Related Pairs)



Pattern Search (Video Chains)



Pattern Search (Common Recommendations)



3. Scalability:

a. Describe how data is stored in a cluster configuration of a NoSQL database. Provide performance benchmarks for data ingestion/query in a cluster deployment. How do the results compare to the performance benchmarks on the reduced dataset or non-cluster deployment?

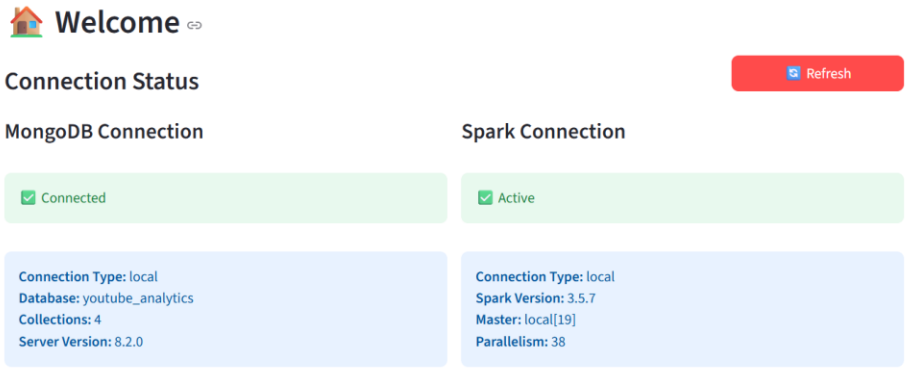
Our NoSQL database is stored locally with replica sets (through MongoDB config). We have not done sharding or cluster deployment yet.

b. Describe how the algorithms are run on a cluster configuration of Hadoop/Spark. Provide performance benchmarks for algorithm in a cluster deployment. How do the results compare to the performance benchmarks on the reduced dataset or non-cluster deployment?
Our algorithms can run in Spark's standalone mode with multiple worker nodes based on number of cores of the system.
For testing purposes and simplicity, we are running our application in local mode. Our app will run in Standalone mode for the actual demo.

c. What hardware (computers with CPU, memory and disk storage) did the team use to test the scalability of the solution?

We all tested on our local computers. They have different numbers of logical clusters. These are the number of worker threads:
Ross: master: local[15]
Ben: master: local[7]
Harry: master: local[19]

Our Home Page shows the configuration:



4. Source Code:
   a. Provide the source code of your application prototype, including user interface, data ingestion, data query, and analytics algorithms, in a ZIP file. Make sure to **exclude** the dataset.

   Included in submission and repo: https://github.com/rk3026/Youtube_Analyzer