Youtube for Content Creators

Analytics and Recommendations for the Budding Youtuber

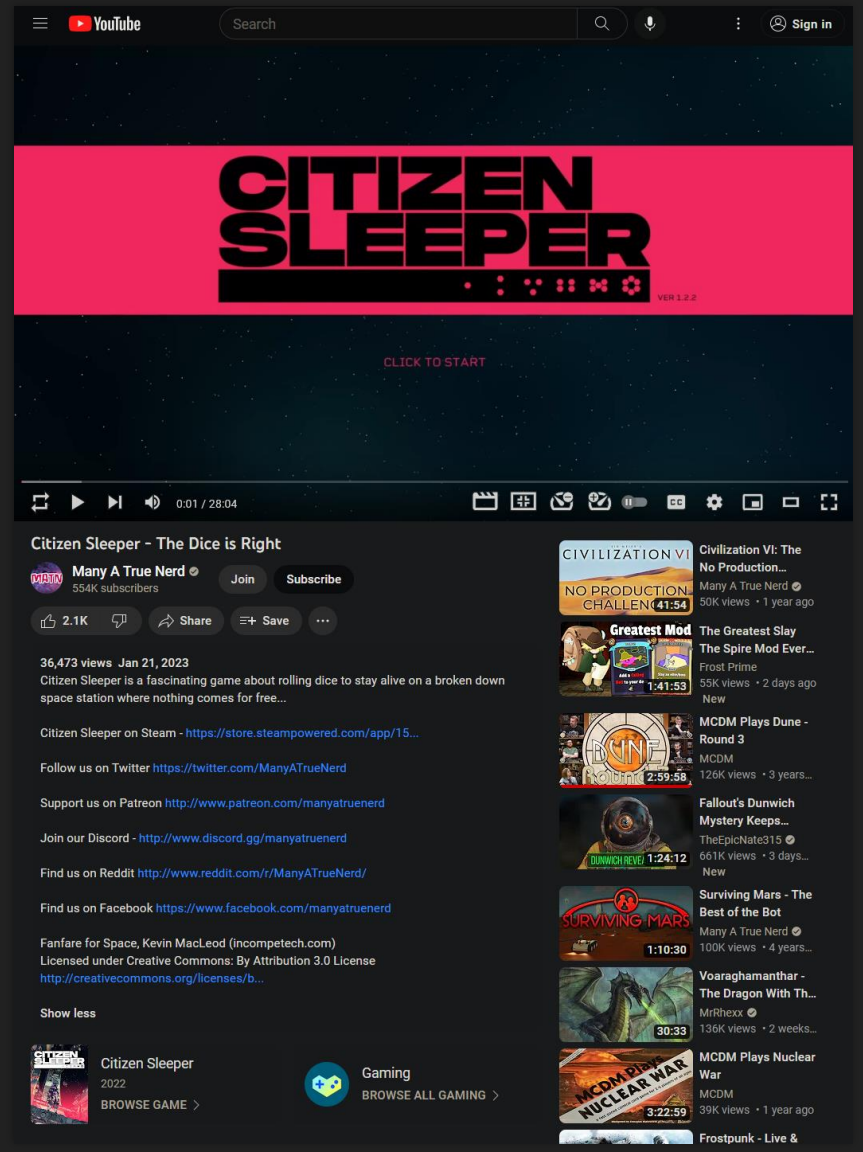
Kapilan Mahalingam
January, 2023

Motivation

- Video content creation is difficult
  - Most youtube gamers, and twitch streamers struggle to increase viewership
  - Time limited, making videos is a lot of work
  - Record number of new games being released, difficult to keep up
- Objective
  - Leverage youtube and steam to take a data driven approach to games
  - Content based and Audience based recommendation systems

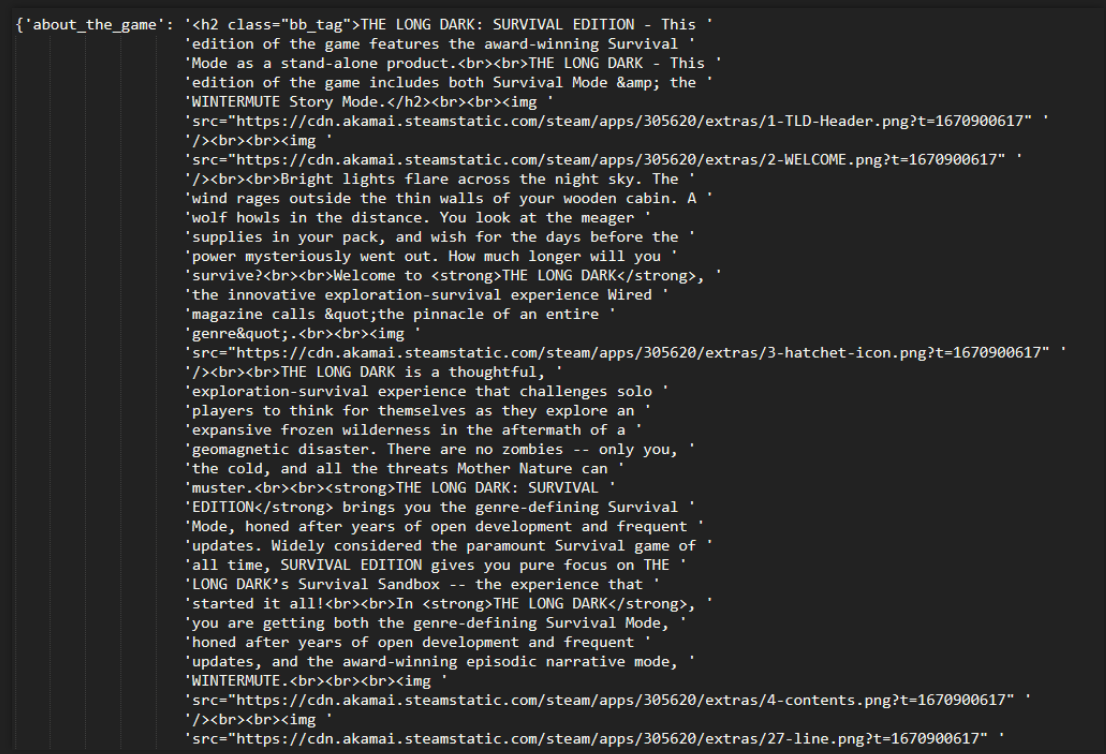
Sample Youtube Data Point

- Youtube Data
  - List of Videos on the Channel
  - Youtuber subscriber count, channel age; content mix
  - Data per video
- Sources
  - Youtube API
    - Youtube's API is terrible
  - Direct Scraping
    - Scraping youtube is also terrible
  - Scraped all videos for about ~20 channels (~120,000 videos)



Sample Steam Data Point

- Game data from Steam
  - Game Identification is terrible
    - Third party python libraries
    - Team Fortress (?!) API
    - Web scraping search suggestions
    - Steamspy
    - Multilayered ID extraction
  - Steam API
    - Surprisingly was actually pretty good



```
{'about_the_game': '<h2 class="bb_tag">THE LONG DARK: SURVIVAL EDITION - This '
                   'edition of the game features the award-winning Survival '
                   'Mode as a stand-alone product.<br><br>THE LONG DARK - This '
                   'edition of the game includes both Survival Mode &amp; the '
                   'WINTERMUTE Story Mode.</h2><br><br><img '
                   'src="https://cdn.akamai.steamstatic.com/steam/apps/305620/extras/1-TLD-Header.png?t=1670900617" '
                   '/><br><br><img '
                   'src="https://cdn.akamai.steamstatic.com/steam/apps/305620/extras/2-WELCOME.png?t=1670900617" '
                   '/><br><br>Bright lights flare across the night sky. The '
                   'wind rages outside the thin walls of your wooden cabin. A '
                   'wolf howls in the distance. You look at the meager '
                   'supplies in your pack, and wish for the days before the '
                   'power mysteriously went out. How much longer will you '
                   'survive?<br><br>Welcome to <strong>THE LONG DARK</strong>, '
                   'the innovative exploration-survival experience Wired '
                   'magazine calls &quot;the pinnacle of an entire '
                   'genre&quot;.<br><br><img '
                   'src="https://cdn.akamai.steamstatic.com/steam/apps/305620/extras/3-hatchet-icon.png?t=1670900617" '
                   '/><br><br>THE LONG DARK is a thoughtful, '
                   'exploration-survival experience that challenges solo '
                   'players to think for themselves as they explore an '
                   'expansive frozen wilderness in the aftermath of a '
                   'geomagnetic disaster. There are no zombies -- only you, '
                   'the cold, and all the threats Mother Nature can '
                   'muster.<br><br><strong>THE LONG DARK: SURVIVAL '
                   'EDITION</strong> brings you the genre-defining Survival '
                   'Mode, honed after years of open development and frequent '
                   'updates. Widely considered the paramount Survival game of '
                   'all time, SURVIVAL EDITION gives you pure focus on THE '
                   'LONG DARK’s Survival Sandbox -- the experience that '
                   'started it all!<br><br>In <strong>THE LONG DARK</strong>, '
                   'you are getting both the genre-defining Survival Mode, '
                   'honed after years of open development and frequent '
                   'updates, and the award-winning episodic narrative mode, '
                   'WINTERMUTE.<br><br><br><img '
                   'src="https://cdn.akamai.steamstatic.com/steam/apps/305620/extras/4-contents.png?t=1670900617" '
                   '/><br><br><img '
                   'src="https://cdn.akamai.steamstatic.com/steam/apps/305620/extras/27-line.png?t=1670900617" '
```

Preprocessing

- Significant error checking
  - Used a secondary CNN to consistency check the data for the primary CNN
  - Lots of data cleaning, stripping html, url's, emojis, non-ascii chars, unnecessary capitalization
- Pruned based on several criteria
  - Common enough Game
  - Game details on Steam
  - Livestreams
  - Consolidating Versions

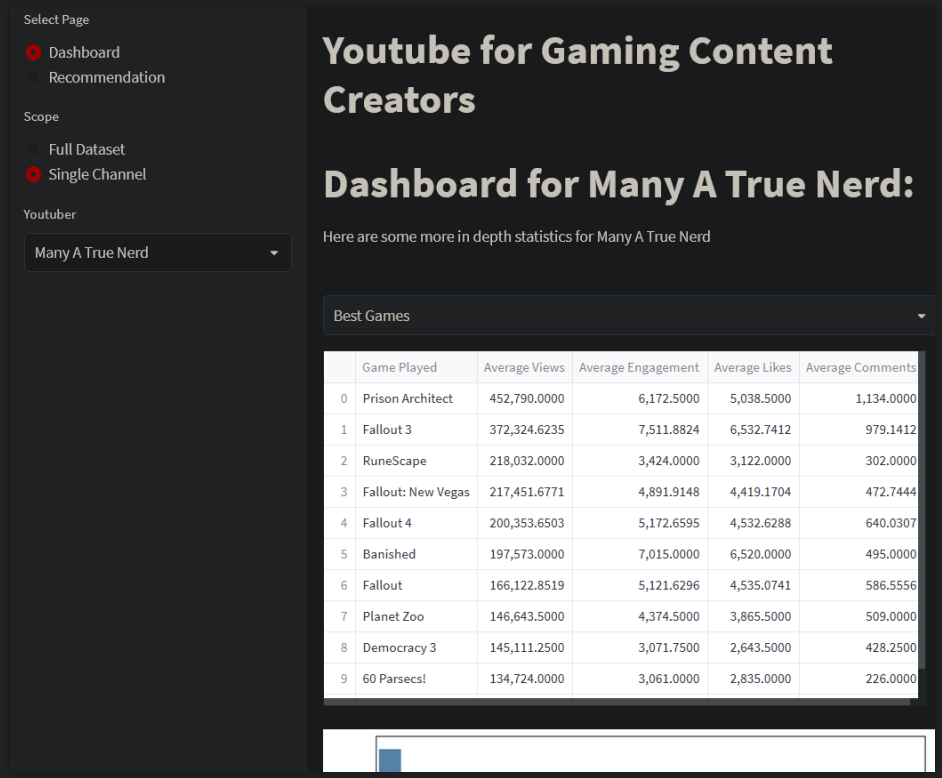
Model

Dumped data into a SQL database, and tried two different models

- Content Based Model:
  - Similar games based on description, genres, publishers, developers, etc.
  - Ran both on youtuber's 'top' videos, as well as their entire library
- Collaborative models:
  - Harder due to the paucity of users, and how specialized some were
  - Difficult to judge recommendations
- Focused on matrix factorization methods rather than CNN's for "computational" reasons

Deployment I

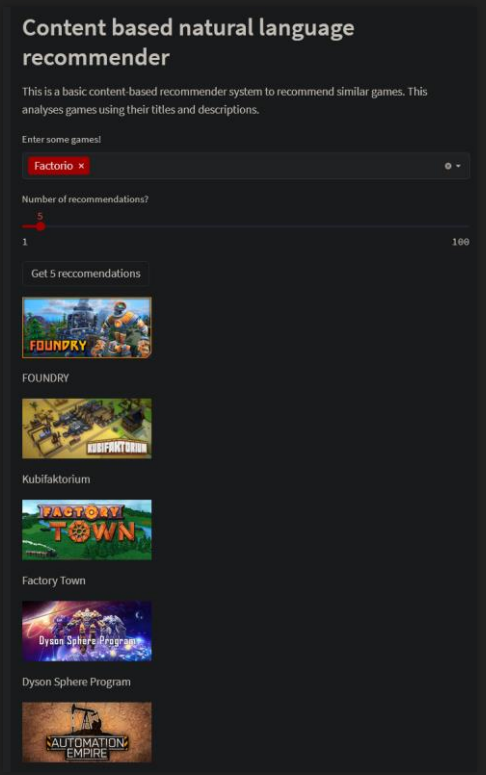
- Deployed Webapp on streamlit
  - Dashboard
    - Youtube Gaming, in general
      - Views, per capita views, engagement, popular games/genres, etc.
    - Channel analyses
      - Graphs, summary statistics, trends, correlation matrices and the like

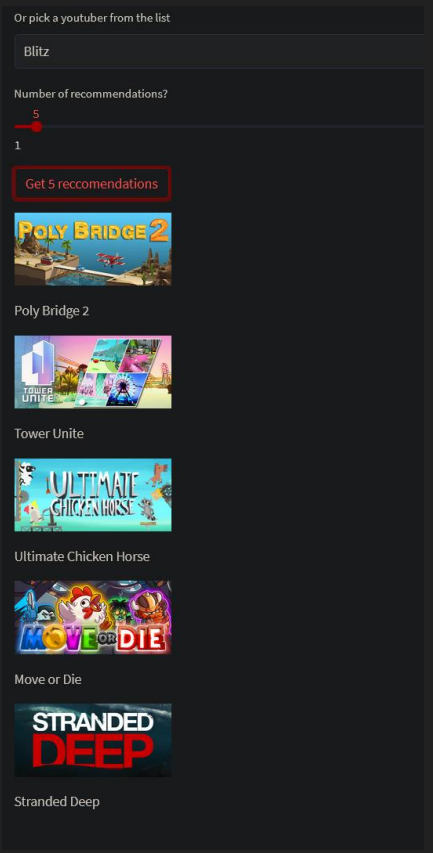


	Game Played	Average Views	Average Engagement	Average Likes	Average Comments
0	Prison Architect	452,790.0000	6,172.5000	5,038.5000	1,134.0000
1	Fallout 3	372,324.6235	7,511.8824	6,532.7412	979.1412
2	RuneScape	218,032.0000	3,424.0000	3,122.0000	302.0000
3	Fallout: New Vegas	217,451.6771	4,891.9148	4,419.1704	472.7444
4	Fallout 4	200,353.6503	5,172.6595	4,532.6288	640.0307
5	Banished	197,573.0000	7,015.0000	6,520.0000	495.0000
6	Fallout	166,122.8519	5,121.6296	4,535.0741	586.5556
7	Planet Zoo	146,643.5000	4,374.5000	3,865.5000	509.0000
8	Democracy 3	145,111.2500	3,071.7500	2,643.5000	428.2500
9	60 Parsecs!	134,724.0000	3,061.0000	2,835.0000	226.0000

Deployment II

- Recommendation System
  - Content based worked better
    - For youtubers, based on their top games rather than their entire library
    - Here the recommendations were done on the basis of views rather than “ratings”
  - Collaborative is difficult to judge





Road Ahead

- Steam skews the data
  - Platforms not available
  - Exclusives and the like could be draws for viewership
- Offload more complex parts
  - Streamlit cannot handle computational load
  - Scraping is very hacky, especially for larger youtubers
  - Basically streamlit isn't great for this use case and I made a mistake
- Larger samples might help collaborative filtering, but they also need success criteria