Where are the Cute Kittens and what happened to your attention spa...

P2 Deliverable, project proposal, and initial analyses for ADA - Fall 2022

Table of Contents

Abstract - Where are the Cute Kittens?

> *You sit comfortably on your couch and pick up your phone. You scroll through the latest TikTok trend, past that funny Reel about Gen-Z humor. You snug into a blanket, ready to dive into the new 20min episode by 3Blue1Brown on 'Why PI Is Everywhere'. Once on YouTube, though, you get sidetracked by a Short of Elon taking a puff with Joe Rogan, and you end up irritated by a Vox documentary on the Rise to Power of Vladimir Putin.* [1]

> You stop and think: *Where are the cute kittens?*

What happened to YouTube? How did we go from funny cat videos and epic fail montages to long-form podcasts, product reviews, and deconstructions of political rivalries with incredible production quality? How did YouTube not only survive but thrive in a world where short, mobile-first content became king?

In this project, we will travel through time, and explore YouTube's trends over the years, using the YouNiverse Dataset [2].

Research Question

Our aim is *not* to establish a causal relationship between particular features and engagement, nor to create a recipe for future channel success. Our focus is to present interesting trends that can be digested and appreciated regardless of how intertwined Social media as a whole have become.

- How did the length of videos change over the years?
- What are the most popular topics over the years and how are they distributed?
- Are there clear shifts in the behavior of the creators related to changes in the policy of the platform?
- How are particular real-life news and events reflected in the content?
- How did the frequency of uploads change over the years?
- How did the use of words in tags, titles, and descriptions change over time and what can we observe?
- Can we hypothesize how the rise of other Social Media and content platforms shaped YouTube? From a way to listen to free music to an alternative to mainstream media, from short funny videos to high production quality and solid monetization structure.

Additional Datasets

The main analysis will be performed on YouNiverse. Nonetheless, the focus of our project is to tie together an eagle-eye view of the internal changes in YouTube's content as well as some interactions with the external world, thus we will use data and results gathered from:

- Scientific papers
- Google Trends
- YouTube and Google public reports and official statements
- Independent reliable third parties like Socialblade and Channel Crawler [3]
- Reliable news channels for real-life world events

We don't plan on performing the analysis on raw data from these sources, but to obtain a coherent data story we will gather data from multiple sources and entrusted aggregators.

Methods

At this stage, we present partial results to show the feasibility of our project. That is, we run an initial analysis on the full dataset to make sure we are able to handle the data, but we don't have interactive versions or full-time series and results for each year. In particular, we want to optimize the handling of natural language in descriptions and titles. In P2 we show word clouds on tags for three different years, the evolution of the length of videos over time, and the average likes of videos in each category.

Key methods and aspects moving forward:

Data storage and computations

To work with this dataset we devised the following structure:

1. Creation of partial datasets with 1k, 10k, 100k, 1mln entries to devise

correct implementations on local machines

2. Use of Google Cloud Storage to run on a cluster with 600Gb of working memory and 90CPUs to handle the full dataset.
3. Vectorisation of natural language, optimization, parallelization of algorithms, and modularisation of code to make the notebook an optimized 'front-end' and run results on multiple cores.

Visualisation

Our goal is to make some visualizations interactive and informative. In particular, we plan the use of word clouds for topic popularity as well as graph views to show connections between descriptions and tags.

Proposed Timeline

We devised a tentative timeline to help us stay on track. Details are useful to our team and we keep them here in the official README since it's P2.

- ***[14/11 - 20/11]* - Week 9:** P2 - show the feasibility, clarity on research questions, initial exploration, and results.
- ***[21/11 - 04/12]* - Week 10-11:** Breadth: look for interesting trends, research monetization policies, and deeper analysis on the rise of long-format videos on YouTube alongside short-format videos on other platforms. Clean results on lengths of videos and time-series analysis. Refine NLP on descriptions, titles, and tags. Research optimization and parallelization of notebooks. Initial implementation of the graph and word cloud with vectorization.
-
    - ***[05/12 - 11/12]* - Week 12:** Outline of results, draft of data story. Finding signal over noise, and clearing up results narrowing down on what really seems interesting and surprising. NLP over titles and descriptions will need more revision. Length of videos and background on YouTube Policies and significant events should be clear and done.
- ***[12/12 - 18/12]* - Week 13:** Website mockup and design. Find captivating

angles to create an interesting story.

- ***[19/12 - 23/12]* - Week 14:** P3 - Clean up code, keep one main notebook, add useful comments, final touches on data story, redact final README with notes gathered during previous weeks.

Organization of work

Similar to the tentative timeline, we keep here a semi-detailed repartition of concerns. Useful for us even if we help each other in different aspects.

- ***[21/11 - 04/12]* - Week 10-11:**
  - **Mariia and Daria:** Research on big real-life events, and hypotheses on the effect of trends on YouTube. Research on NLP and methods to do sentiment analysis and exploring the idea of connected graphs to show correlations. Analysis of Frequency of uploads.
  - **Rouby:** Run clean notebook on the full dataset, further research on additional data that can be used, exploration of the contrast between emerging long-format videos on YouTube alongside short-term format around. Analysis of time series.
  - **Jacopo:** Research on YouTube monetization policies. Analysis of the length of videos. Trends at the top and the worst engagement. Research on animation and word clouds. Research on vectorization for word cloud and NLP.
  - **All:** Meeting, share notes on Google Docs, GitHub and WhatsApp. Individual setup of corresponding notebooks for each research and synchronization with the main, cleaner notebook.
- ***[05/12 - 11/12]* - Week 12:**
  - **Mariia and Daria:** Correct NLP implementation, analysis of tags, and descriptions to make top and worst topics emerge. A clear relationship between real-life events and YouTube trends.
  - **Rouby:** Clear results on the frequency of uploads, curation of the main notebook, and clear results from time series.
  - **Jacopo:** Clear results from the length of videos, animations, initial

    implementation of the website, and potential interactive elements like word cloud and graphs.
    - **All:** Draft of data story, eliminate uninteresting and insignificant results, tie up external resources, and ideas on how to make interesting results pop out. Don't fall in love with uninteresting yet complex results. Double-check results on NLP.
- ***[12/12 - 18/12]* - Week 13:**
    - **All:** Website mockup, design, a clear data story, and notebook run with clean code and comments. Draft of README evolved from notes + P2.
    - **Jacopo:** Website coding and deployment.
- ***[19/12 - 23/12]* - Week 14:**
    - **All:** Last cleanup of resources, data story, and README. Official deployment of the website. Correct details like grammar and check the story flow.

Questions to TA

- Suggestions on workflow and optimization, running Jupiter in parallel on multiple cores?
- Are the research questions clear?
- Suggestions on details in vectorization of words to better handle NLP?

Notebook

Explore Notebook provided with comments and compiled code.

Dependencies

Standard visualization libraries as used in the course. We will consider in the next steps the use of machine learning libraries like Scikit for clustering to reveal trends using tags and descriptions.

Authors

- @mariia
- @daria
- @rouby
- @jacopo

References

1. 3Blue1Brown - Joe Rogan - Elon Smoking Weed - Vox - Putin doc ↩

2. YouNiverse dataset - GitHub - Paper ↩

3. Channel Crawler - Social Blade ↩