Mini-Report: Exploring Filter Bubbles in YouTube Recommendations

Jader Silva

Introduction

This mini-project explores the dynamics of video retrieval and recommendation using three neutral seed queries: *healthy cooking*, *beginner guitar*, and *stretching exercises*. The purpose is to investigate how content clusters form around different seeds, to quantify diversity, and to assess overlap between themes. While simplified, the study illustrates core analytical skills in text processing, similarity analysis, and visualization of recommendation structures, all of which are transferable to larger-scale media ecosystem studies.

Method

Data collection was conducted through the YouTube Data API, retrieving ~150 videos per seed query. From this pool, a smaller set of representative "seed videos" was selected, and textual metadata (title + description) was preprocessed into a *text* column (lowercase, stripped punctuation).
A TF-IDF vectorization approach was applied to generate embeddings of video text. Using these embeddings:

1. Cosine similarity scores were computed to build edges between videos.
2. KMeans clustering was applied to partition videos into latent topics.
3. Metrics of diversity and overlap were computed at the seed level.
   Visualization was performed with *networkx* for graphs, and *matplotlib*/*seaborn* for bar charts and heatmaps.

Results

1. Entropy vs Step

Entropy curves show *stretching exercises* accumulating the highest diversity. *Healthy cooking* starts narrow at early steps but rises sharply by step 3, surpassing *beginner guitar*. This contrasts with the global diversity-by-seed metric, where *healthy cooking* remains minimal; the difference reflects that the step-wise curve captures the path of sampled recommendations, whereas the seed-level diversity summarizes the full distribution over clusters.

2. Jaccard Similarity Between Seeds

Video-level Jaccard similarity between seeds was **0**, confirming that no videos overlapped across queries. However, channel-level overlap was observed between *healthy cooking* and *stretching exercises* (e.g., recurring creators like Lilly Sabri and Massy Arias), suggesting cross-domain content strategies by certain channels.

3. Diversity per Seed

Normalized entropy scores reinforced the entropy curve findings:

- *Beginner guitar* = most balanced topical spread.
- *Stretching exercises* = moderate diversity.
- *Healthy cooking* = minimal diversity, with videos concentrating on similar phrasing and themes.

Limitations

This study is constrained by its scale: only 3 queries, ~450 videos total, and a reliance on textual metadata rather than full video transcripts or engagement signals. The Jaccard analysis, while informative, captures only surface-level overlap (IDs and channels), not thematic subtleties. The clustering method (KMeans on TF-IDF) provides a coarse partition, sufficient for a toy project but not for production-level recommendation audits.

Next Steps

Future iterations should:

- Expand seeds to politically sensitive or polarized queries, to test recommendation pathways in high-risk contexts.

- Incorporate temporal dynamics (video publication dates, trending status).
- Enrich the feature set with engagement metrics and full transcripts.
- Scale diversity and overlap metrics to larger graphs, enabling network analysis of recommendation bias.