YouTube History CLUSTERING

By Bailey Hall

Presentation Code: Drive

LIST OF CONTENTS

01 MY DATASET

02 MEANING FROM THE DATA

03 IR/DA TECHNIQUE

04 THE OUTCOME

05 THE FUTURE

MY DATASET:

Personal YouTube Watch History Data

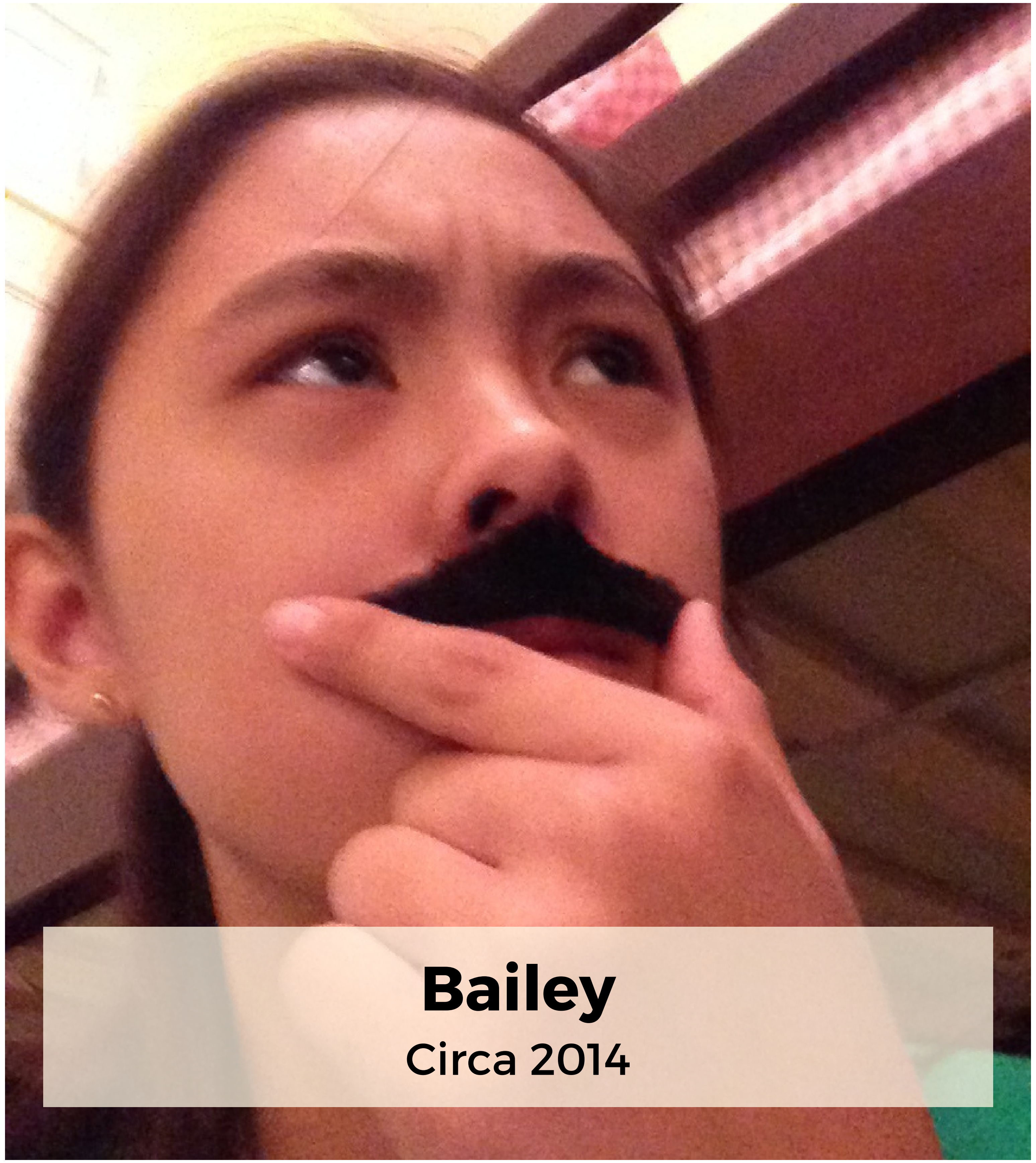
CONTAINS:

- **Title**
- Video URL
- **Channel Name**
- Channel URL
- Data and Time
- **Whether it is an ad**

*The information I use in this program is in bold

Why THIS DATA?

- The format of the data (.json) is not difficult to parse
- Was one of the few data sets with my personal data that had substantial amount of data
- I never had looked at my YouTube history before now
- What was 12 yr. old Bailey (when I started watching YouTube) watching vs. 17 yr. old Bailey vs. now?



Bailey

Circa 2014

MEANING FROM DATA

I wanted to categorize my data in some way to see the range of different types of content I'd consumed since childhood

I also wanted to see how many videos I watched per category to see what type of content I spent the most time consuming

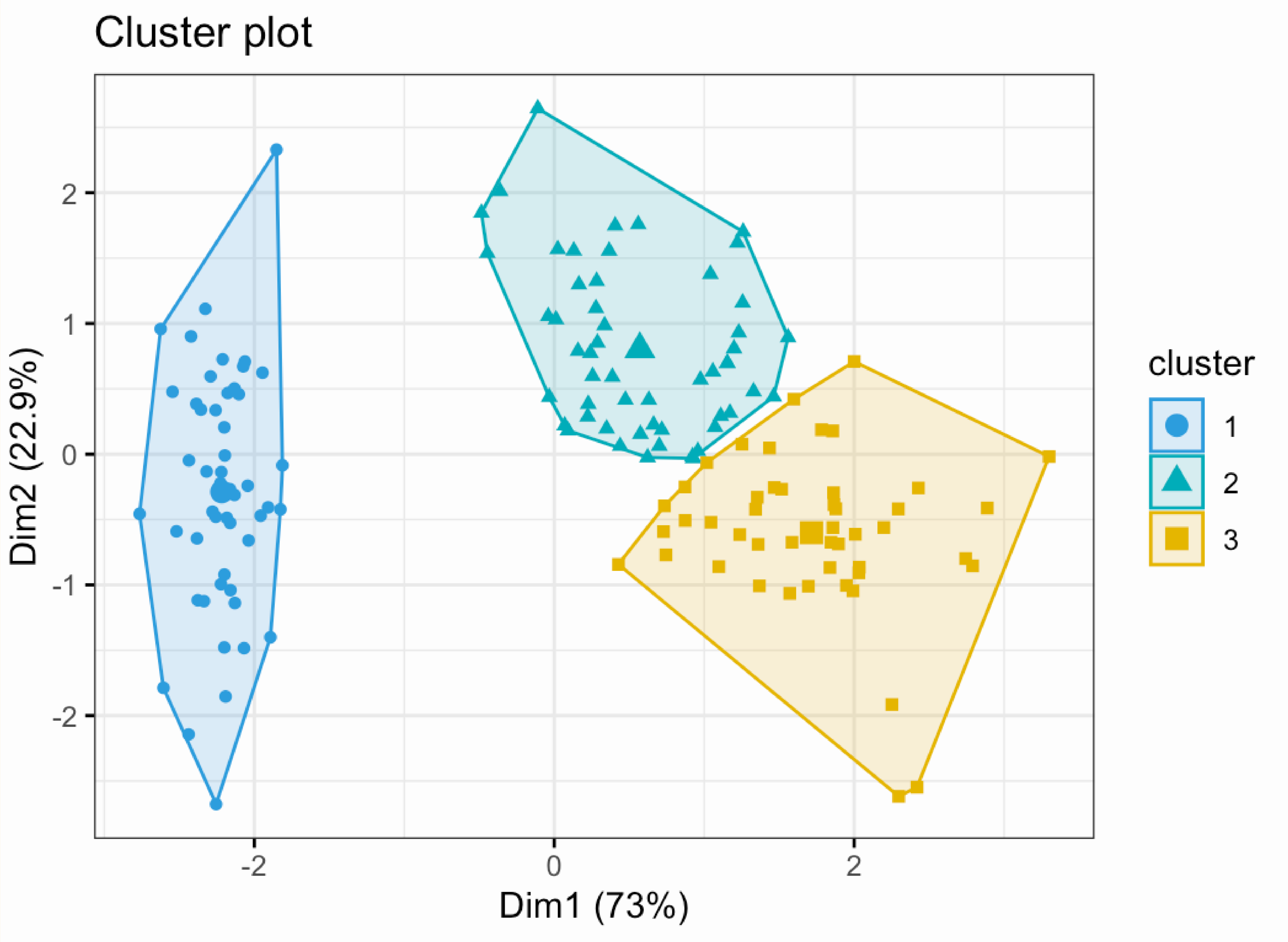
I don't spend much time thinking about the video content I consume, and I hoped to gain more insight in this area through this project

IR/DA TECHNIQUE

I chose to use clustering for my IR/DA technique

More specifically, the flat clustering algoritm, K-means.

I chose this because it is a concept we didn't cover in the course and it fits well with the meaning I set out to derive from the data



APPLICATION OF TECHNIQUE

Similar to the Vector Space Query assignment, each document is transformed into a TF-IDF vector

However, instead of comparing cosine similarities of the vectors, the Euclidean distance is calculated and compared between vectors

The vectors who have the smallest distance from each other are considered more similar and are clustered together.

$$|\mathbf{x}-\mathbf{y}| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

First I parsed the .json file

Then I utilized the sklearn library to vectorize the data on each video

Again using sklearn, ran the k-means algorithm over the vectors

Printed the specified number of clusters, how many videos were in each, and the top terms found in each cluster

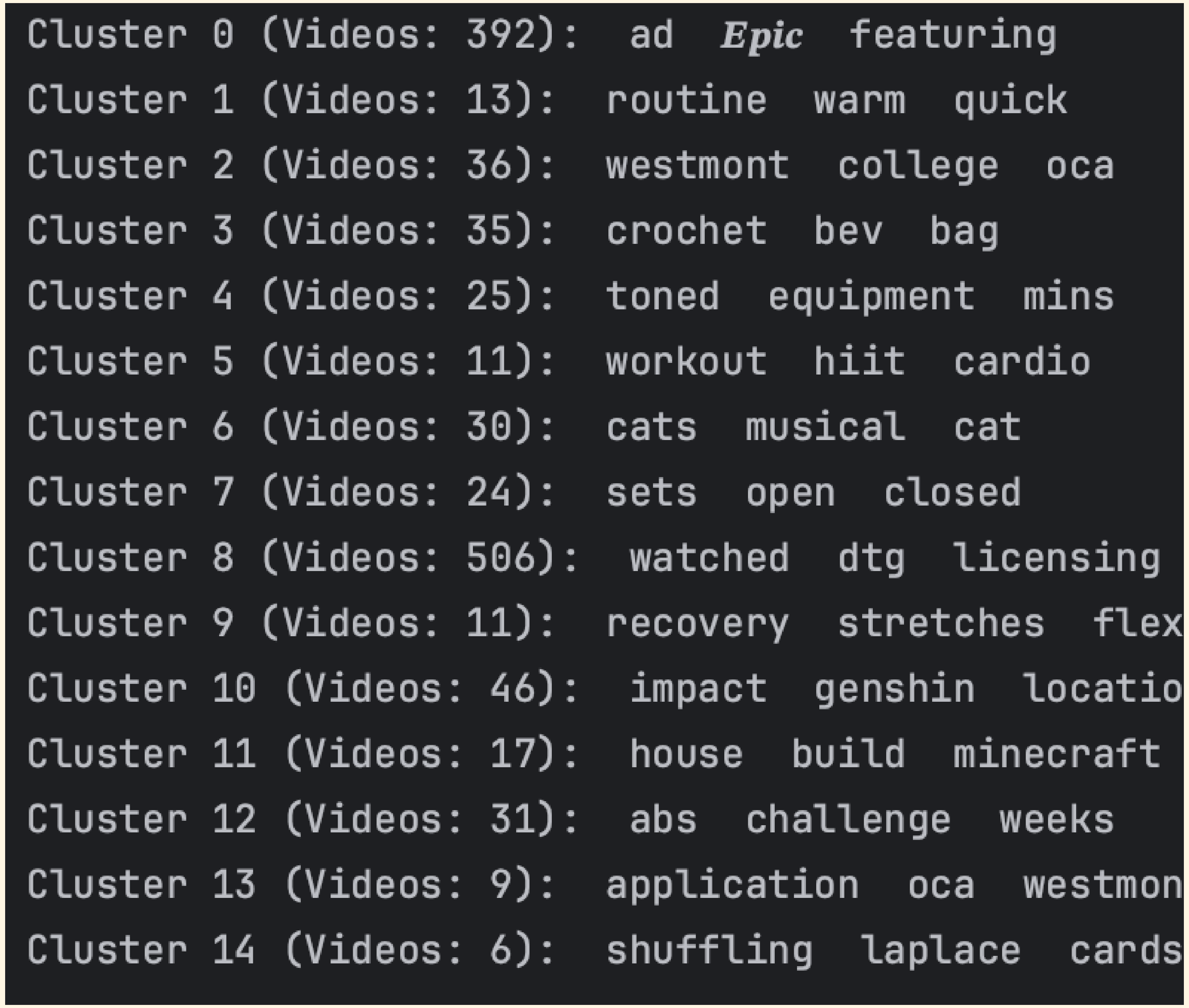
the OUTCOME

First of all: turns out my data only goes back into 2020, I must've deleted my previous watch history data

When reading this data, I can clearly see the different phases of interests I went through since 2020, and how much time I spent on each

A ton of what I watch are either 'ads' or in some 'dtg licensing' category (not sure what that is).

Note that results can look pretty different based on the number of clusters specified



```
Cluster 0 (Videos: 392):  ad  Epic  featuring
Cluster 1 (Videos: 13):  routine  warm  quick
Cluster 2 (Videos: 36):  westmont  college  oca
Cluster 3 (Videos: 35):  crochet  bev  bag
Cluster 4 (Videos: 25):  toned  equipment  mins
Cluster 5 (Videos: 11):  workout  hiit  cardio
Cluster 6 (Videos: 30):  cats  musical  cat
Cluster 7 (Videos: 24):  sets  open  closed
Cluster 8 (Videos: 506):  watched  dtg  licensing
Cluster 9 (Videos: 11):  recovery  stretches  flex
Cluster 10 (Videos: 46):  impact  genshin  locatio
Cluster 11 (Videos: 17):  house  build  minecraft
Cluster 12 (Videos: 31):  abs  challenge  weeks
Cluster 13 (Videos: 9):  application  oca  westmon
Cluster 14 (Videos: 6):  shuffling  laplace  cards
```

Program run with 15 clusters and 3 top terms

the FUTURE

This program specifically parses YouTube watch history data, so that is the only kind of data set it can handle

However, k-means clustering can be used to derive meaning from a wide range of data sets.

01

How much __ have I watched?

What amount of videos has one watched that are similar or related to a certain topic or keyword?

02

When did I watch the most?

Instead of by title, cluster by date watched to see at what point in time one has watched the most content, and what was the content about?

03

What kind of ads am I shown?

Right now the program just an ad an 'Ad' rather than its 'Title'. However, adjusting the program to filter out non-ads and consider ad titles could give some insight into what Google thinks you would want.

THANKS FOR LISTENING

Any questions??

My Info:

Bailey Hall

bahall@westmont.edu