COMP 4449: Summer 2024

Professor: Luke Sawle

*Student Name:* Evan Hollier

***Term Project Plan***

**Direction**: Address the questions in the following sections to create a project blueprint (or project plan for your midterm (This document is also applicable to the final project).

**Project Goals**

* What business goals are you trying to solve?

Measure audience overlap between YouTube channels with publicly available information. This will be done using the comments from a channel's videos.

* How will you measure the success of achieving the business goal?

Ability to input two or more channels and recieve percentage(s) of shared audience.

* What analytic or machine learning tasks will you be solving?

Data collection and preprocessing of YouTube comment data.

* How will you measure success for your analytic or machine learning solution? Hint: start by defining your baseline.

Baseline comparison with known audience overlap data from videos / channels with a small audience that will be easy to maunally verify.

**Significance of the Project**

* What is the expected impact of your project?
* Provide a statement that captures why your project is important with respect to the business goal(s).

Provide insights into YouTube ecosystem, which can help creators, advertisers, and other viewers understand the broad interests of different audiences. Given enough channels as input, I could see using this to recommend other channels.

**Methodology**

* What analysis method would you use to solve the problem?
* Statistical method? If you are using a statistical method, is it descriptive or inferential?

Descriptive statistics to summarize audience data. This is mostly me wanting to delve into data visualization, so I can't think of an actual statistical method I would need.

* Machine learning algorithm? Is it supervised or unsupervised machine learning?

I don't currently see a way to utilize machine learning with this data. If I wanted to do something with the content of comments, maybe there's an NLP angle, but that isn't my interest with this project.

* How does the selected method work to achieve the analysis goal?

There isn't really any statistics that I see needing done.

* What are the pros and cons of the method(s) selected compared to other methods for achieving the same task?

I can't think of another way to do this. The obvious alternative to measuring audience is by looking at subscribers, but that data is not readily available.

**Data Description**

* What does the data represent? Is it education data, health data, sensor data, etc.? Provide a general description, hosting site, and URL (if available).

Social media and user engagement data.

* Is the data structured or unstructured?

Unstructured

* What are the dimensions of the dataset?

My initial request for data from youtube's API gives me a list of the usernames for the comments on a specific video. So the size of this list depends on the video. But a typical video for a large channel probably has a few thousand comments. I will be combining the data for every video on a specific channel into one dataset, which will have duplicates removed. So a large channel will have tens or hundreds of thousands of usernames in the audience. These will be what I use to compare with each other.

* What data types are represented by features in the dataset.

Strings - usernames of users that posted comments

* Which variables will be used for analysis? If there are too many variables, try to use broader descriptions such as demographic variables instead of gender, ethnicity, age, etc.

Overlap of usernames between the audience for channels. Will mostly focus on percentages rather than a count, but both will be recorded.

**Data Preparation**

* What data preprocessing steps do you expect will be required?

Removing duplicates, merging data.

* How will you check and handle outliers and/or missingness?

One possible outlier that comes to mind is comments posted by bots. Will be hard to track without looking at the actual text content of their comments. These accounts will be present on any large channels that don't moderate their comments (which pretty much none of them do).

I can do a secondary pass of data collection for the content of comments for users that appear in too many channels to determine if they are bots. But initially collecting this data will make the data to large.

* What feature extraction or selection strategy would you use for dimensionality reduction, if any?

I will be removing duplicates. They are not needed for what I'm measuring.

**Data Exploration**

* What descriptive statistics will you generate for the different features represented in the data? Provide a rationale.

I'll only be looking at overlap. I can't think of any descriptive statistics I'll need.

* What types of data visualizations will be helpful in analyzing the dataset? Please provide the names of the graphs or charts and the appropriate variables for the graphs.

I know I want to generate tables of a list of channels' overlap. So it'd look like:  
 A B C

A -

B 30% -

C 80% 2% -

Only one side of the diagonal needs to be filled in

**Solution Development**

* What data product do you aim to achieve by the end of the analysis

A table or lists of tables of the overlap between the active audiences of different YouTube channels.

* Please outline the sequence of steps you expect to follow for selecting, building, and testing the algorithm(s) you intend to use to develop your solution. Try to be as detailed as possible.

For one channel, iterate through every video and collect unique usernames to build a dataset of that channel's audience. Do this for many channels (will pull from a list so it isn't one at a time). Once a collection of audiences is acquired, compare them to generate a table of overlap percentages.

* What computational resources will be required to ensure the feasibility of your modeling process.

Will be a lot of API calls, but the algorithm itself shouldn't be too computationally expensive.

* How will you use peer reviewed journal articles and other similar resources to guide your work?

I can do some research to see if there's any similar product that exists already.

* How will you ensure that your results are reproducible?

Results are expected to vary. New comments are posted every day, which will slightly shift the actual numbers of audience overlap.

**Conclusion and Recommendations**

* Who are the stakeholders or business users who will benefit from the solution you develop?

YouTube channel owners and advertisers. General social media analysts and researchers.

* What information will they want to know to assess the value of your outcomes in terms of their priorities?

The extent and significance of audience overlap between channels. Recommendations for collaboration and content strategy based on overlap findings.