**User Hack Clustering Analysis**

**Tanish Sanghvi, Asad Tanveer**

***Abstract*:** *This report/paper presents the results of a blueprint that aims to segment customers based on their social media activity. The process aims at identifying the ‘taste’ of users in order to segment them into distinguishable categories. The method that guided this process involved data extraction using the YouTube API, Natural Language Processing (NLP) for semantic understanding and K-Means for subsequent clustering. The proposed process is capable of generating solutions to questions such as ‘As a company can we divide our core audience into different groups based on certain shared characteristics? Can we then create context-based content that is specifically targeted towards the taste of these groups?’*  To answer these questions, one can use the above process and thus say that a given set of users is more likely to engage with content related to music while another is more likely to engage with content related to gaming.

***Keywords****: K-Means Clustering, User Segmentation, YouTube API, Elbow Curve, TF-IDF, PCA, NLP*

**1. Introduction**

Companies today are sitting on a treasure trove of data. Combine this with machine learning and the domain of analytics become enormous with applications such as sentiment analysis, market basket analysis, text mining and so on. This is more so true when it comes to Marketing. About 90% of the data collected today, by companies, is related to customer actions in marketing activities1. This can most likely be attributed to the era of personalization2, in which businesses are attempting to gather as much information about a user as possible in order to target them for revenue optimization, contextualized advertising, or enhanced engagement objectives3. A key challenge that marketing teams face in this sense is minimizing their ‘cost per acquisition’ with respect to their return on investment. One of the modern ways that companies tackle this is through Customer Segmentation. Customer segmentation is vital as it gives you a pathway to maximize your customer’s value to your business. The segmentation process simply deals with grouping customers and potential customers with similar characteristics into distinguishable segments. The major benefit of such a mechanism is that you can communicate to all of the individuals in that segment efficiently and with a sense of personal attention, without actually contacting each individual separately.

With the advent of data science, such form of segmentation has become pretty common. Clustering3 is a commonly utilized technology in industry today for these purposes since it provides a quantitative and visual way of analyzing audience groupings. A real world example of this can be seen in Target’s identity-driven offers which helped the company generate a 389% increase in the number of orders and a 413% increase in revenue, year over year4. The basis of this was identifying the identity of a group of customers and then advertising them offers relative to this identity2. In this project, we have tried to implement a similar workflow. Social media is at the heart of our lives now and the ability to create an internet-based presence, be it through engagements or advertising, is a goal that every company seeks. Thus by utilizing this readily available YouTube API data4, we aim to cluster a set of users belonging to a certain industry and define them with unique identities, in order to show the utility and insightfulness of customer segmentation.

**2. Methodology**

1. **Data Pre-Processing**

First and foremost, we needed to find videos (video ID’s in particular) belonging to a particular industry. This would put us in the position of being a company, from that industry, that is aiming to determine the preferences of people who watch these type of videos. We found a publicly available data set on [Kaggle](https://www.kaggle.com/harshitmakkar/youtube-videos-idtitledescription-and-category) that has 2607 videos along with their video ID’s and the category (food, science, manufacturing etc) that they belong to. We decided to choose ‘Travel’ as most of the videos (52% i.e 1355) belonged to this category. Now, the goal was to identify the people who have interacted with these videos

YouTube has 2 publicly available API’s (Reporting and Data) which provides a range of metrics – playlists belonging to a channel, number of subscribers, engagements on a video and so on. While the API provides the number of likes on a video, it does not give their user ID’s. However, it does provide the ID of the users who have commented on these videos. Therefore, in order to identify the people who have interacted with these videos, we decided to look at the commenters on a video and pull their ID’s.

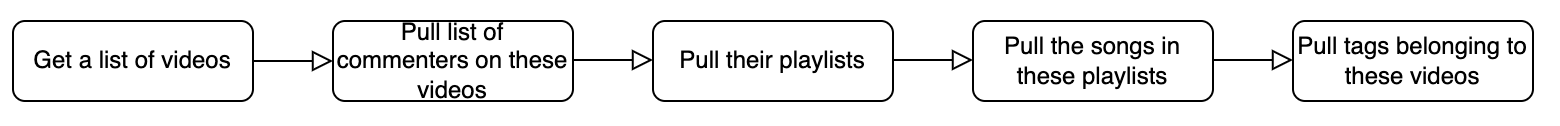
While the YouTube API is dynamic and easy to use (has nice tutorials at ….) it has a rate limit of 10,000 API calls per day. Therefore, we had to apply certain self-induced restrictions in order to prevent our data was being too large for the API to handle.

The first of these was to pull a maximum of 100 commenters per video. This gave us a total of 4,644 unique commenters. In order to determine the preferences of these users, we needed to extract their playlists (playlist ID’s in particular) as these would have the videos that these users like to view. Again, we used the API and restricted ourselves to a maximum of 20 playlists per user. This gave us a total of 17,297 playlists. This also reduced our number of users/commenters to 2,224 as the remaining did not have any playlists. As a final step, we used these playlist ID’s to extract the videos within them. Again, we restricted ourselves to a maximum of 50 videos per playlist thus leaving us with a total of ~2M videos.

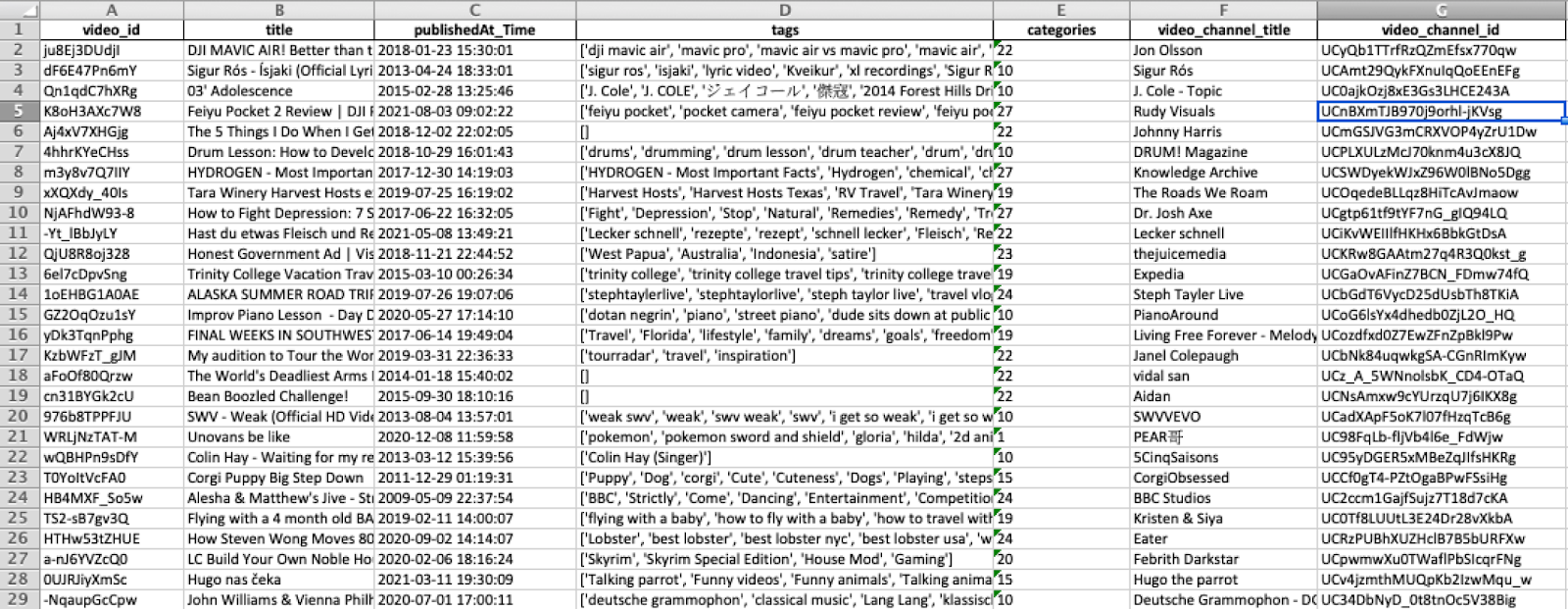
Now, in order to understand the type of these videos and thus segregate them, we needed some kind of description for these videos. A crucial metric in this case was ‘Tags’. As part of its Search Engine Optimization (SEO) system, YouTube requires channels to annotate their videos with ‘tags’ that describe what the video is about. This is essentially to dictate to the algorithm which videos to serve to the audience. So for example, a Yoga tutorial video might have tags such as ‘exercise’, mental health’, ‘meditation’ and so on. This is important to us, as using these ‘tags’ we can determine the range/type of videos that a user watches.

The YouTube API provides a mechanism for this as well wherein one can send a list of videos (25 max) and get their respective list of tags.

The below chart visualizes the workflow just mentioned –



The below screenshot is a sample of the data collected post all of the steps mentioned above:



As one can see, each video has a list of tags associated with it along with the playlist ID and the commenter ID that it belongs to.

1. **Data Post-Processing**

The end goal of the paper is to identify the ‘taste’ of its users and to ‘sensibly’ segment them into distinguishable categories. In order to identify their taste, we need our model to make sense of the semantics (in our case ‘tags) associated with these videos. As such the first step in modeling was to use Natural Language Processing (NLP) and in order to do so we needed to clean our tags.

A number of such ‘cleaning’ techniques were used, namely – apply lowercase, remove punctuations, remove non-alphabetic words, remove stop-words and remove words not in English. The intuition behind applying these cleaning techniques was to prepare the text for Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF returns a number/statistic for each word, reflecting how important the word is to a document in a collection of documents. This helped us in creating a sort of weighting of all the text contained in the tags.

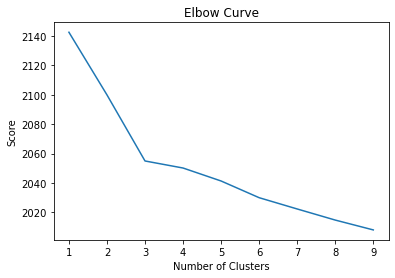
Considering we had ~2M videos, each with it’s own set of tags, we were looking at dealing with a very large set of features (~28K). This could lead to not just large space and memory requirements but also high noise sensitivity. Therefore we applied PCA as a next step in order to reduce the dimensionality of our dataset. We found ~1500 features to explain +90% of the variability in the data.

**3. Results**

Having performed the pre-processing and post-processing steps, we simply had to model our data with a clustering algorithm. We decided to use K-means as it is the most robust and most widely used clustering algorithm for unsupervised datasets.

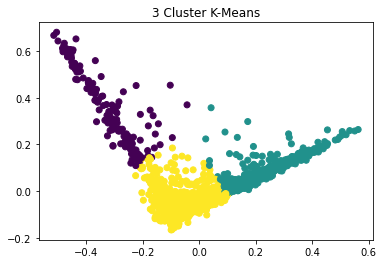
There are numerous ways to determine the number of clusters. We decided to use Elbow Curve as it is relatively easier to implement and better to visualize.

Below is the result of the Elbow Curve method applied on 1500 feature data set:



As one can see the graph has an elbow-like shape, wherein as we increase the number of clusters from 3, the model starts to give diminishing returns. Thus the cost of increasing k after 3 outweighs the loss in error.

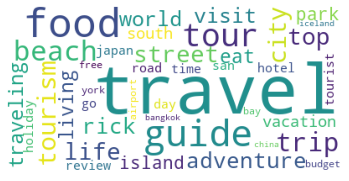
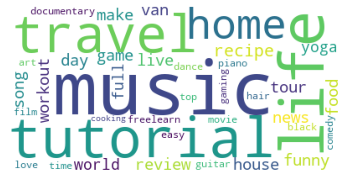
Having selected k=3 as the optimal number of clusters, we ran our K-means model and visualized the clustering of our 2,224 users. Below is a visualization of the K-Means clustering:



As one can see, our clusters are well-defined and distinguishable from each other. In other words, this portrays that our users do have a certain set of shared characteristics that distinguishes them from other sets of users.

We decided to make Word Clouds for each of our clusters as this would help us visualize the broad range of tags that distinguish each of these clusters. Following were the results:

**Cluster 1 Cluster 2 Cluster 3**

Following are some of the inferences we can make from the 3 word clouds

* The 1st group of users/cluster seem to have an inclination towards videos around Travel with tags such as adventure, beach, islands and so on.
* The 2nd group seems to be inclined towards music with tags such as remix, pop, dance and others.
* The 3rd cluster seems to be kind of ambiguous as it has music as well as travel with high weightage. However, one could see it being under the cloud of Tutorials with tags such as recipe, workout and yoga among others

As an addition, we could deep-dive into each cluster’s data and look for more insights. This could help give an added dimension to our knowledge of the cluster. Following are some insights that we found by looking at the data of each cluster of users.

Cluster 1:

* Most of the videos in the playlists were around ‘To-do lists’ and ‘Bucket list adventures’
* Touropia was the most popular channel followed by Ryan Shirley (10 and 9 videos respectively)

Cluster 2:

* Most of the videos in the playlists were from TrapMusicHDTV and TRAP MUSIC NOW’s channels (12 and 8 videos respectively)
* “Spider-man: Into..Soundtrack” and “Ariana Grande – Snow in California” were the videos added to most playlists (4 each)

Cluster 3:

* The most popular channels were ‘Annoying Orange’ followed by STEVEPUNCH 91TV (15 and 8 videos respectively)
* Majority of the videos in these channels were around entertainment such as challenges (drawing, word puzzle etc) and tutorials (cooking recipes, abs workout)

**4. Conclusion**

Leveraging the utility of YouTube API in providing readily available user data, we put forward a business-ready model that can help companies determine the ‘taste’ of its users and divide them into distinguishable categories. The model also provides an additional utility of providing cluster-specific insights. These can be utilized by sectors and businesses to make digital content-based judgments, produce context-based advertising, and a variety of other things that drive up engagements and revenue. As further scope, the project opens up scope for 1) the use of Hierarchical Clustering for alternative results, 2) application of Neural Networks for more optimized results, 3) the role of ethics in ‘hacking’ user data for segmentation purposes.

***References***

1. Rashid Kazmi, Ph.D. “Customer Segmentation.” *Medium*, The Startup, 14 Oct. 2020, https://medium.com/swlh/customer-segmentation-7ee1fb3f7ade.
2. Cheung, K.W., Kwok, J.T., Law, M.H. and Tsui, K.C., 2003. Mining customer product ratings for personalized marketing. *Decision Support Systems*, *35*(2), pp.231-243.
3. Wu, X., Yan, J., Liu, N., Yan, S., Chen, Y. and Chen, Z., 2009, June. Probabilistic latent semantic user segmentation for behavioral targeted advertising. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising* (pp. 10-17).
4. Tony Coray April 1, and Tony Coray. “How Top Brands Use Market Segmentation to Reach Customers.” *SheerID*, 25 June 2021, https://www.sheerid.com/blog/how-top-brands-use-market-segmentation-to-reach-customers/.
5. An, J., Kwak, H., Jung, S.G., Salminen, J. and Jansen, B.J., 2018. Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, *8*(1), pp.1-19.
6. Cross, G., 2008. Understanding your customer: segmentation techniques for gaining customer insight and predicting risk in the telecom industry.
7. Jansen, B.J., Jung, S.G., Salminen, J., An, J. and Kwak, H., 2018, October. Combining Behaviors and Demographics to Segment Online Audiences: Experiments with a YouTube Channel. In *International Conference on Internet Science* (pp. 141-153). Springer, Cham.