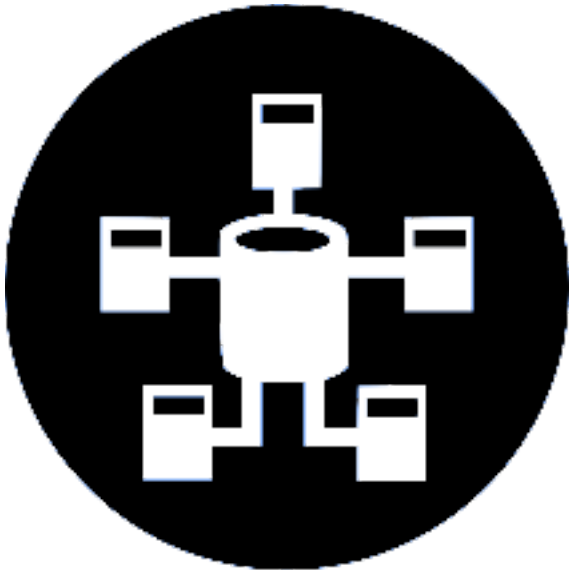

Attribute Oriented Induction

(Assignment-2)

Student1: Sarthak Singhal(20171091)

Student2: Sajal Asati(20171183)

Data set used: Trending Youtube Videos

Rationale for picking the data set

- To do any sort of analysis one should have some domain specific knowledge which helps a data analyst to analyse the data better.
- This is why we have chosen the dataset on Youtube videos as we are its active users and quite familiar with some of the criteria which tell about the popularity of the videos.
- Also, as this dataset was sufficiently large, we could get a vast range of data, hence generalization of attribute values could be done more accurately.
- One of the motivations was also to analyse what factors affect how popular a YouTube video will be.
- Also we expected to perform a simple sentiment analysis on the results, to notice any patterns if any in the people psychology about reacting to a video content.

The Process

- So in the beginning, our table contained 16 columns and about 50k rows. It had data about videos from various genres like Comedy, Education, Anime, Entertainment etc.
- To keep our analysis simpler and limit its size we picked a single genre - “Entertainment” to study the patterns in popularity of videos based on attributes like views, likes, comments and dislikes. This reduced the number of rows to roughly 13k.
- Since we were presenting data about how popular the video is, we removed some of the video related attributes such as video id, title, link, trending date etc. as these were unique for a particular video and had no generalization.
- Fields such as comments_dissabled, ratings_dissabled etc. were removed to further compress the data.
- After reducing the number of attributes, we started generalising attributes to reduce the domain size of each attribute.
- The concept hierarchy that we used is that we categorised all the attributes into ranges depending on number of records in a particular range for each attribute. For eg. we categorised views into three categories(<5lakh, 5lakh-5million, >5million). This hierarchy helps us to map a set of low-level concepts to higher-level concepts.
- Using this we could categorise all the views values as less, medium, high respectively.
- For eg. we had attributes like likes, views etc. so we created ranges for them and reduced their domain size to about 2 or 3.

- Then we aggregated all the records having the same values for all attributes and added an extra field of count to represent number of videos of each kind.
- Finally, after all these steps, the dataset was reduced to about 36 rows and 4 columns.

The Final Data

Category ID = 24 - Entertainment				
Views	**Likes**	**Dislikes**	**Comments**	**No. of Videos**
< 5 Lakh	< 10k	<500	< 1k	**5395**
< 5 Lakh	< 10k	<500	1k - 5k	**1063**
< 5 Lakh	< 10k	<500	>= 5k	**13**
< 5 Lakh	< 10k	>=500	< 1k	**201**
< 5 Lakh	< 10k	>=500	1k - 5k	**240**
< 5 Lakh	< 10k	>=500	>= 5k	**7**
< 5 Lakh	>= 10k	<500	< 1k	**167**
< 5 Lakh	>= 10k	<500	1k - 5k	**808**
< 5 Lakh	>= 10k	<500	>= 5k	**79**
< 5 Lakh	>= 10k	>=500	< 1k	**17**
< 5 Lakh	>= 10k	>=500	1k - 5k	**230**
< 5 Lakh	>= 10k	>=500	>= 5k	**92**
5 Lakh - 5M	< 10k	<500	< 1k	**325**
5 Lakh - 5M	< 10k	<500	1k - 5k	**221**
5 Lakh - 5M	< 10k	<500	>= 5k	**3**
5 Lakh - 5M	< 10k	>=500	< 1k	**209**
5 Lakh - 5M	< 10k	>=500	1k - 5k	**430**
5 Lakh - 5M	< 10k	>=500	>= 5k	**42**
5 Lakh - 5M	>= 10k	<500	< 1k	**123**
5 Lakh - 5M	>= 10k	<500	1k - 5k	**603**
5 Lakh - 5M	>= 10k	<500	>= 5k	**85**
5 Lakh - 5M	>= 10k	>=500	< 1k	**138**
5 Lakh - 5M	>= 10k	>=500	1k - 5k	**1124**
5 Lakh - 5M	>= 10k	>=500	>= 5k	**1437**
>= 5M	< 10k	<500	>= 5k	**1**
>= 5M	< 10k	>=500	< 1k	**1**
>= 5M	< 10k	>=500	1k - 5k	**2**
>= 5M	>= 10k	>=500	< 1k	**9**
>= 5M	>= 10k	>=500	1k - 5k	**44**
>= 5M	>= 10k	>=500	>= 5k	**342**
			TOTAL	**13451**

Some Conclusions Drawn

- Based on the ranges of likes, dislikes, comment numbers and views, we finally came up with the above table.
- Given a video with very less views and likes, people comment more on those videos which received lesser number of dislikes from people (Row 1 and Row 4).
- Videos with very high views (>=5M), and likes (>=10k) are also having high number of comments. They also have higher number of dislikes. Number of views decide here how many people are reacting to the videos through commenting, and liking/disliking, hence in a way also deciding how popular the video has become.
- Also we can see, the scale of viewership and liking/commenting are quite different. Most of the people (like about 90%) just watch videos, and don't usually react through ways like commenting or liking.
- Most of the videos receive about 5 Lakh views, as they make up about 50% of the video content. Similarly number of highly trending videos is quite less (4-5%).