Scraping YouTube to Collect Variables for Video Analysis

By Nandini Basu, Linyan Dai, Maxine Li

In partial fulfillment of the course BAX 422

Date of Submission: 3/15/2020

Table of Contents

Executive summary

Buzzfeed Tasty is an international media production channel that routinely produces high-quality videos that are primarily related to food and cooking. These videos are hosted on their website, Facebook and YouTube channels. They currently have over 1.7 million subscribers on YouTube. We wanted to help BuzzFeed Tasty evaluate the various factors that impact the number of views that a video gets. One of the primary sources of revenue for Buzzfeed Tasty is YouTube ads and sponsored brand partnerships. YouTube ads are the ads that other brands show on Buzzfeed Tasty's videos before the actual video starts. The sponsored brand partnerships are collaborations where other brands can feature their products in Buzzfeed Tasty's videos. Having more views enables Buzzfeed Tasty to not only get more ads on their platform, but also gives them the leverage to negotiate a better price for the ads and sponsorships.

Hence it would be important for them to understand how to increase video views without spending any more money on the actual video production. In order to further analyze the relationship between these factors and the video views, we had to conveniently capture and store these variables on file, preferably in a database. We decided to use web-scraping and the YouTube API key to get various variables related to a video such as the video's title, description, length, likes, dislikes, captions, comments and views. Once we had the variables, we did necessary transformations and eventually stored the data in MongoDB. We chose MongoDB because it allows for flexible scaling and we also didn't need a relational database. Additionally, we were storing captions and comments of variable length and MongoDB works well with such documents. With the variables now ready, we can do further data analysis to understand the relationship between the various video features and video views.

Introduction

Buzzfeed is a digital media company that has a worldwide following. They regularly post engaging online content such as news, blogs, lists, and videos. One of their key organizational goals is to create innovative or 'viral' content that their readers/viewers can engage with.[1] In July 2015, in an attempt to crack the Facebook video channel, Buzzfeed launched Tasty: a web-series on comfort food. In the past five years, the series has produced thousands of videos focused on food and food-related content.[2] These videos have different formats. Some are quick recipe tutorials without any voice overs, some are recipes with detailed voiceover instructions and the others include chefs sharing their favorite foods to eat or cook.

All of Buzzfeed Tasty's videos are produced by an experienced production team and are of high quality. Hence, Buzzfeed Tasty invests significant time, money and energy in producing these videos. These videos have indeed been well received by viewers and Buzzfeed Tasty currently has 1.7 Million subscribers on YouTube and their videos get hundreds of thousands of views. This viewership creates an opportunity for Buzzfeed Tasty to generate revenue by showing video ads on YouTube. Hence the organization is constantly trying to understand what it can do to increase viewership. We want to help Buzzfeed Tasty identify the factors, aside from the actual video production quality, that influence the number of views a video gets. In order to do so, we would have to collect various data points on its existing videos. The end goal would be to understand what the organization can do to organically[3] increase the number of views a video gets.

1*About BuzzFeed*, www.buzzfeed.com/about.

2Ting, Deanna, et al. "With Tasty, BuzzFeed Has a Multi-Revenue Stream Model." *Digiday*, 10 Dec. 2019, digiday.com/media/tasty-buzzfeed-multi-revenue-stream-model/.

3 An organic view is a view that is generated without the use of any sort of promotion.

Data characteristic

Before investigating the relation between the number of views on each Buzzfeed Tasty video, we had to identify the various features of the videos. The two kinds of features associated with the videos are business generated and user generated features. The business generated features include variables such as the video's title, description, length and captions, while the user generated features include the video views, likes, dislikes and comments. Even though these variables are available on Buzzfeed Tasty's YouTube page, they cannot be exported directly from the pages. Hence we decided to use a combination of web scraping and the YouTube API to store these features on file.

As for the data characteristics, these variables are divided into text and quantitative data. Most of the quantitative variables have right skewed distributions that need transformation for further analysis[Table 1-4], while the number of comments seems normally distributed with a mean of 6421[Table 5]. In addition, there is an obvious difference in total views between videos with captions and those without[Table 6]. Overall, all the quantitative variables are positively correlated[Table 7], reminding us to be considerate of the risk of multicollinearity among these variables.

Method and limitations

We decided to use web scraping to get the video's title, description, length, likes, dislikes and views. However it is difficult to simply scrape YouTube using the BeautifulSoup package as YouTube uses custom tags thus making it harder to find the regular HTML tags. Instead we used the Selenium package, which is an open-source web-based tool that can be used to automate web browser interaction.[4] It can also be useful for web scraping as it gives us the flexibility to extract the data and store it on file. We were able to extract the video links for 1300 videos. We chose only 1300 videos because we only wanted to consider the videos produced in 2019. This would allow us to control the quality of video production.

4 "Selenium." *PyPI*, pypi.org/project/selenium/.

We then sent a request to each link and saved each video's page on file. Once the page was stored on file, it was easier to extract the various video features. We also saved the main page where all the videos are listed as this was the only place we could extract the length of the videos. Once we had these features, we had to treat some variables to ensure they can be used for further analysis. Specifically, the date the video was posted on a was captured as a string value and had to be transformed into date-time variable. The length of the video was also stored in minutes and seconds and we had to convert this to seconds so it can be used as a continuous variable. Lastly, we used the YouTube API to get the captions of the videos but it would be important to note that not all the videos have captions. We were also able to use the YouTube Data API key to get the comments associated with each video. We decided to get only the top 100 comments because each API request was limited to a 100 results and because we only wanted to consider the comments which were most popular.

Once we had collected all the data, we decided to store it in MongoDB. This was an obvious choice since we were storing comments and captions and didn't need a relational database. MongoDB lets us store data without defining the field length, which can vary greatly for comments and captions. It is also easily scalable which would make it easy for us to add more data should we wish to do so in the future.

Recommendations for further analysis

We can now explore this stored data to better understand the factors that affect video views. The end goal would be to build a model that defines the relationship between the collected variables and the views that the video gets. We can explore how business generated features such as the title, description, length and captions can be tweaked so that videos get more views without having to invest more on video production quality. And we can also quantify how user generated features such as the likes, dislikes and comments influence the views that a particular video finally gets. We specifically want to look into the captions and the comments of each video so we can understand if there are certain sentiments or topics

that result in more video views. While Buzzfeed Tasty cannot directly control the number of likes or dislikes a video gets, understanding the sentiment or topics of the comments of popular videos can help them in many ways. They can seed content in the comments section to steer the conversation in a certain direction so that there is more engagement on the video which could possibly result in more views and virality.

Conclusion

Hence Buzzfeed Tasty can benefit greatly by better understanding the various features that influence the number of views a video gets. As mentioned before, having more video views enables Buzzfeed Tasty to earn more money from not only YouTube ads but also brand collaborations with other companies. For example, Heinz, a popular food processing company in the United States might want to advertise their new condiment in one of Buzzfeed Tasty's videos. Having more video views on average will enable Buzzfeed Tasty to charge a higher price for such a sponsorship deal and in turn enable them to be profitable in the long run.

Appendix

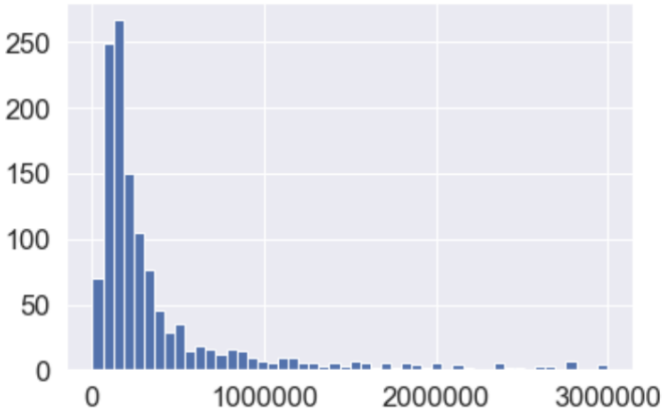

Table 1: distribution of views

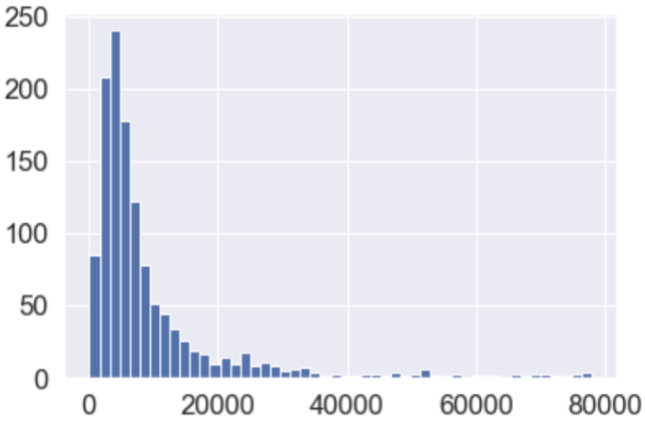

Table 2: distribution of likes

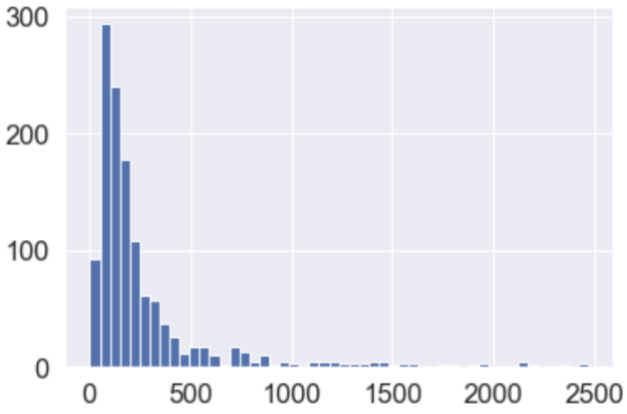

Table 3: distribution of dislikes

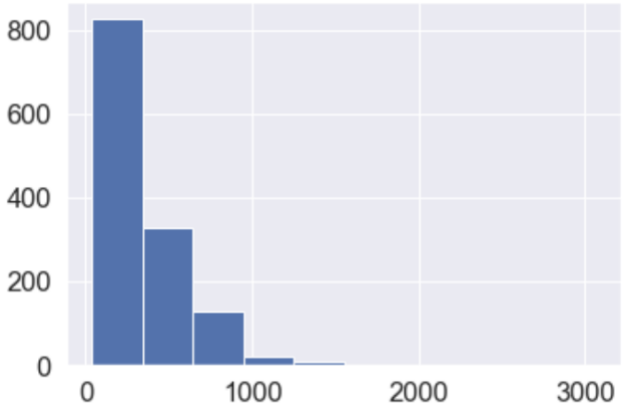

Table 4: distribution of length of videos

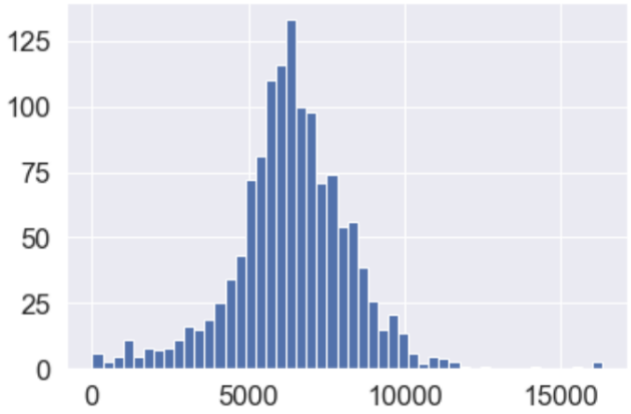

Table 5: distribution of comments

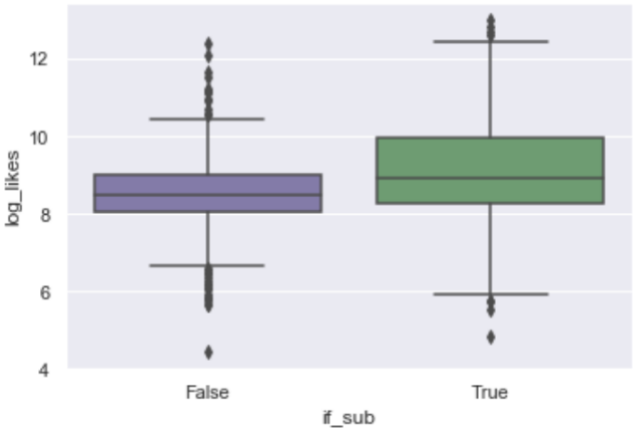

Table 6: views for videos with or without subtitle

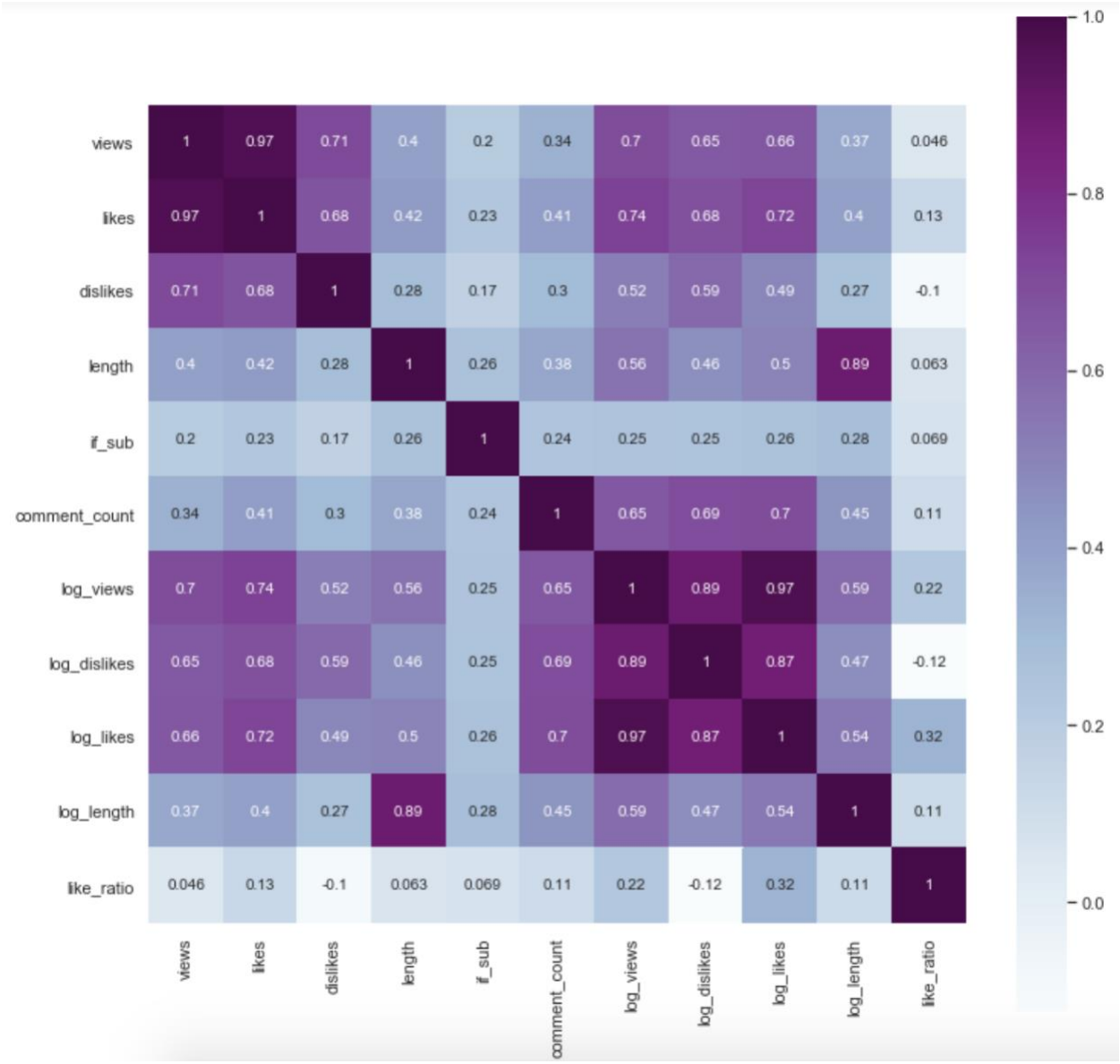

Table 7: correlation heat map

References

1. *About BuzzFeed*, www.buzzfeed.com/about.
2. Ting, Deanna, et al. “With Tasty, BuzzFeed Has a Multi-Revenue Stream Model.” *Digiday*, 10 Dec. 2019, digiday.com/media/tasty-buzzfeed-multi-revenue-stream-model/.
3. “Selenium.” *PyPI*, pypi.org/project/selenium/.