What are Various Popular Tech YouTubers Saying? A Topic Modeling Approach to Tech YouTube Transcripts

Vikas Gudhe
IPHS300 AI for the Humanities (Spring 2022) Prof Elkins and Chun, Kenyon College

{K}DH
IPHS 300: AI for the Humanities
Spring 2022
Professors Elkins & Chun

Research Question

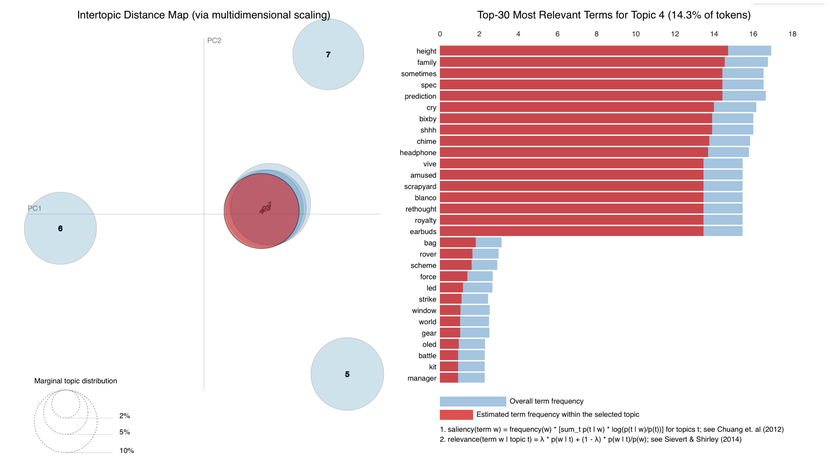
What can topic modeling tell us about the distribution of YouTube content concerning technology and what do content creators focus on when talking about the technology in particular?

Introduction

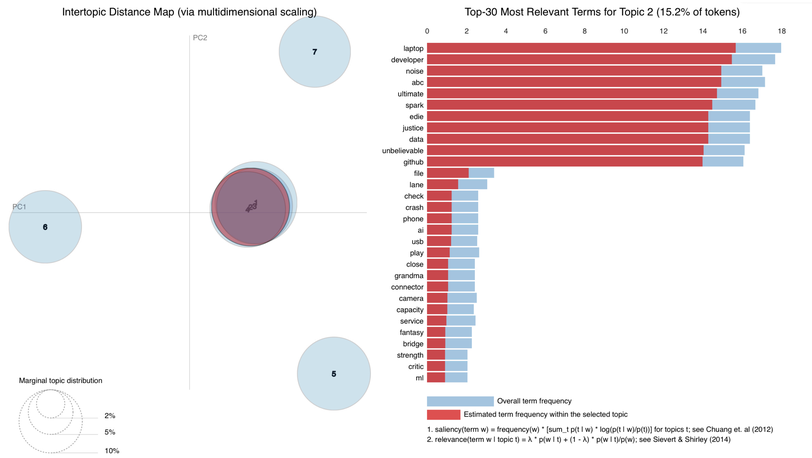
While tech reviews and educational content has largely been on TV and books respectively, increasingly people have been turning to YouTube to be filled in on the latest tech news, reviews of consumer technology, learning how to code, and much more. Unlike TV and other heavily produced content, YouTube allows more freedom for reviewers to express content and there are much larger amount of people expressing their opinions on consumer technology, talking about the latest tech news and even teaching. In fact, in February 2020, Statista found that there are over 500 hours of video content uploaded to YouTube every day. Given immense amount of content, it is up to the general populace to decide what to focus on and place importance on the content they are viewing. Based on this, I thought it would be important to take a look at the most popular content to understand where the interests of the viewers lie and what the content creators focus on providing.

In this project, I focused on 280 tech videos varying in content from general technology talk to comedy, news and informative/educational videos.

Visualized Models



Topic 4 visualized: Most likely pertaining to entertainment tech and gaming (Consumer Tech)



Topic 2 visualized: Most likely pertaining to portable computers and phones (Consumer Tech)

Data & Tools

For the purpose of the study, I chose four main areas to ensure a large corpus, as well as with enough variation to range from consumer technology aimed at the average audience to enthusiasts and even deeply technical videos concerning coding help.

The Categories Are:
Tech (General)
Tech, Comedy (Enthusiast)
Tech, Informative (Technical)
Tech, News (General)

The Dataset was taken from Kaggle in which the transcripts were relatively clean, but further measures were taken such as lemmatizing, tokenizing and removing stopwords in order to ensure a smooth Topic Modeling experience. The videos were taken from a large variety of YouTubers to ensure there was not overrepresentation of topics from one specific category or YouTuber – however all YouTube videos were from the most popular in each category to maintain relevance for the general public.

The main tools used analyze the text were Pandas for data management and clean up, NLTK for cleaning up the text within the transcripts and tokenizing, gensim for calculating bigrams, trigrams and training the LDA model used for final Topic Modeling. Finally, pyLDAvis was used for visualizing the Topic Modeling data for easier analysis. All of these are free to use in Python.

Methodology

1. First, download the Kaggle dataset “YouTuber’s saying things”.
2. Then convert the CSV file into a pandas DataFrame for easier cleanup to filter by “Tech, Tech,Comedy, Tech, Informative and Tech,News”
3. Given this data, combine all the Transcript values into one long string
4. Use NLTK to remove all “STOPWORDS” from the text to remove words that are not helpful in topic modeling
5. Utilize a python script to tokenize the reviews and generate text tokens based on the reviews, and clean the tokens of stopwords that will negatively impact the analysis, then lemmatize the tokens and join the lemmatized words together.
6. Using the gensim python module, generate the LDA model for analysis based on the number of topics and plot the topics as a bar graph.
7. Plot the models on graph with respect to changing the number of topics that go into generating the model using pyLDAvis
8. The dataset contains a combination of user-written transcripts, YouTuber-made scripts and auto generated subtitles from text-to-speech, so some results may not be entirely accurate, but given the limitations of data on human-written transcripts, we must use the available data to have a useful corpus for Topic Modeling

Analysis and Conclusion

Much of the project has been focused on cleaning up data and experimenting with the number of topics, as due to the stochastic nature of the model, the same code can generate various model-identified words topics and words within topics. Inherently, even within the four categories of technology-focused content, there are a large amount of topics covered, but consumer technology seemed to be the most common and prevelant.

Visualized above are topics 4 and 2 from the most coherent model, but even despite testing, this corpus did not prove to be the most helpful in answering the research question. Testing was done with topic numbers ranging from 5 to 25 and while an increased amount of topics generally proves to be more fine-grained, it generally did not seem to provide insight beyond the most common words. Topic 4 and 2 both seemed to focus on consumer technology, but interestingly they are varied in that they seem to be differentiated by portable computers and accessories respectively. Topic 4 in particular mentioned words like “oled”, “earbuds” and “vive” - which pertained to display technology (oled and vive- a VR headset) and accessories (earbuds). Topic 2 seemed to have differentiated by portable computers and smartphones as there is the obvious “laptop” and “phone”, but also talking about the “connector”, “camera”, “capacity”, which are descriptors and metrics reviewers use when talking about these products.

I believe there is still much work to do and a larger dataset yielding to a much larger corpus would be helpful. Furthermore, the closed captioning provided by YouTube’s text-to-speech is fairly good, but the ability to recognize specific product names may have been a limiting factor that ended up making the dataset appear smaller than it may be in reality. While a larger dataset would help, I believe utilizing different models like BERTopic and potentially even NMF models would provide a dimension that could help us understand the corpus better.

Ultimately, given the time constraints, there was limited analysis that could be done with the given data and single-model approach. There is a lot of experimentation that can be done in terms of model-approach, further exploration in topic-amounts and even scraping YouTube directly. This project highlights the difficulty of working with stochastic models such as these and proves the importance of human input and domain expertise to uncover more about the content at hand.

Acknowledgements

I would like to thank professor Chun and Elkins for the help guiding the project, the immense amount of code debugging and providing me with the knowledge to do this project.

Future Work

There is more work to be done, such as experimenting with models and scraping more content. But this provides a starting point for future work in more exhaustive topic modeling.