Going Viral

Olivia Carnes
Galvanize Data Science Immersive

SURGE
Empowering Publishers

Background

The only thing worse than being talked about is not being talked about. – Oscar Wilde

In the world of online publishing, understanding what makes content popular can be hugely beneficial for both consumers and content creators alike. When a publisher knows what's being shared and talked about, a more personalized experience can be tailored to the user based on current trends and user demographics.

From a publishing perspective, increased page views directly link an article's popularity with the ad revenue generated from a user clicking on that page. Publishing buzzworthy content is therefore acutely relevant to a publisher's bottom line.

Objectives

- Explore connections between popular articles published on a news and lifestyle site over the course of a few months during the summer of 2017
- Predict content popularity using page views as a target value and attributes such as article topic, sentiment analysis, imagery, and interactivity as features
- Compare article performance over time to trending topics on the same subject
- Create easily adapted prediction models for partner company Surge for use in online publishing analytics

Methods

- Scrape relevant content from all articles published within a certain timeframe on a news and lifestyle website.
- Clean text by removing uninformative words and stemming the remaining words to create a vocabulary list
- Rate importance of each word in vocab list to a given document based on term frequency and inverse-document frequency
- Determine latent topics and topic similarities using term frequency and Latent Dirichlet Allocation
- Add sentiment analysis to determine whether articles have an overall negative or positive sentiment, in addition to comparing whether an article employs extreme vs. neutral language, and subjective vs. objective language
- Evaluate importance of latent topics and sentiment to page views

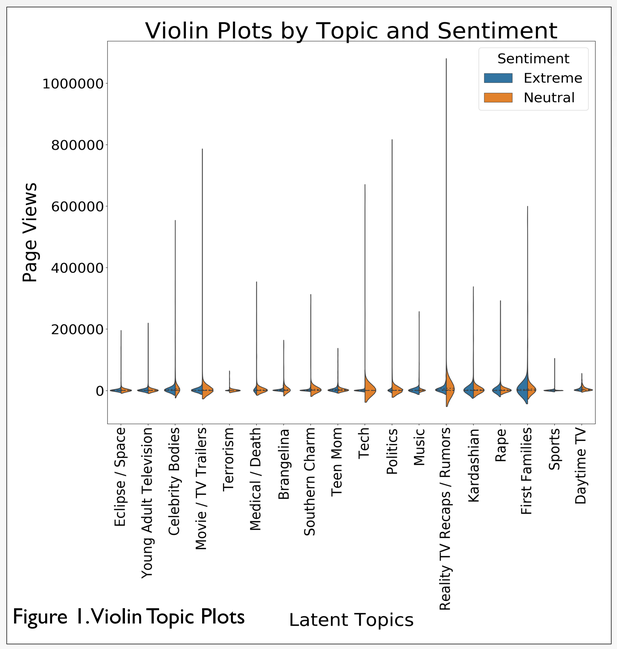


Figure 1. Violin Topic Plots

Results

Articles were ultimately split into the 18 topics seen in Fig. 2

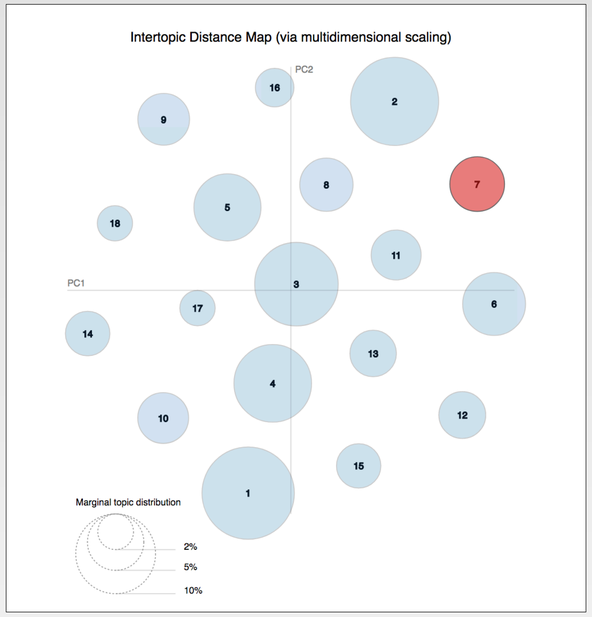


Figure 2. LDA of latent topics using pyLDAvis

Top 10 words for the highlighted topic 7 (sports) were: team, game, player, nba, trade, play, win, make, deal, free

Discussion

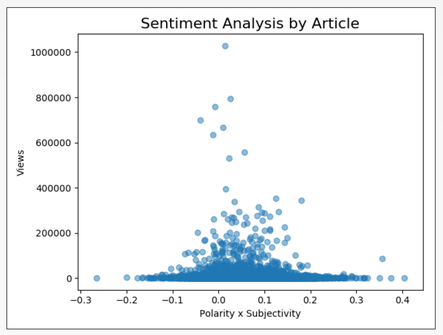
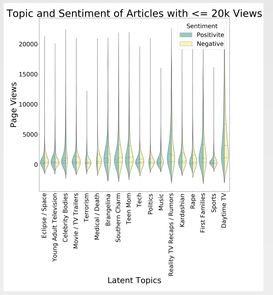


Figure 3. Sentiment Analysis by Article

Ultimately, popular articles seem to have relatively neutral language and subjectivity. Highly viewed articles also fell into average lengths and had low interactivity.

Conclusion

This particular site focuses mainly on pop culture and celebrity gossip, so the most popular articles are related to the lives of well-known celebrities like Beyonce and the Kardashians. In Figure 4 we can see that more positive articles dropout of the plots as we increase the threshold for minimum page views.



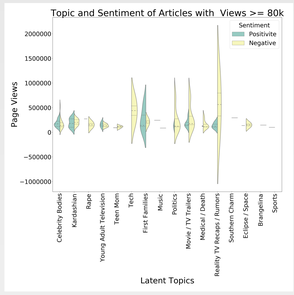


Figure 4a and b: Topic Plots for <= 20k views and Views >= 80k

Future Steps

- Analyzing popularity in terms of shares and reactions
- Adding trending topics from twitter
- Creating a web application where an article can be input for analysis and recommendations can be made on how to increase page views

Contact

olivia.a.carnes@gmail.com
github.com/ocarnes
linkedin.com/in/olivia-carnes