Capstone_DataStory

Jennifer Klein

3/31/2018

Introduction

Podcasts have emerged as the trendy new medium in media. But there's still so little we know about the audience listening to this nascent form of entertainment. We know how many times a podcast is downloaded, but we don't know audience demographics, how many people are skipping through ads, etc. That's all about to change with Apple's recent announcement that they are finally going to provide comprehensive analytics about listener behavior to podcast publishers. We can use data to seek out emerging trends in the genre.

Podcasts cultivate engaged listeners, have large audiences, and pose a legitimate challenge to the radio industry and other content mediums. And now, with podcasts slated to hit television and movie screens in the coming months, it's crucial to learn as much about Hollywood's hottest new source for material as possible.

According to Recode, the podcast industry's ad revenue will reach $220 million in 2017, up 85 percent from 2016's $119 million. The blog also discusses several interesting points regarding the medium's huge potential in both advertising revenue and growth rate:

- Up until now there has been comically little data about podcast consumption, especially compared to other digital media.
- This matters to podcast creators because they are unable to tell how the stuff they make performs — at best, they can usually only tell if someone has downloaded an episode or started to stream it.
- This also matters to podcast advertisers, who would like to know if people are listening to the ads they pay for. Right now, many of them are doing a crude end run around this data void by asking listeners to use a show-specific code when they visit a site after hearing an ad.
- Some podcast software has already provided some of this data. And the data that Apple is offering now is still fairly crude. But the majority of podcast consumption happens on Apple's software, and up until now it has been a black hole. So this is a big move for the industry, which generates a lot of attention (among media types, at least) but a very modest amount of money so far.

Despite the steady growth of podcast listening and spike in media attention over the past few years, the industry itself has trafficked in a relatively minuscule volume of cash money compared to its digital-media peers.

I would like to prove that certain topics are more appealing to the podcast listener and therefore publishers can tailor their content around these subjects to attract more listeners, garner more downloads and, ultimately, attract more advertisers and increase their revenue.

Some background on my dataset: it features 20 podcasts and 199 episodes. The shortest episode clocked in at a mere 15 minutes, while the longest was a whopping 6 hours. The most recent podcast episode was published August 31, 2017 (the day I pulled my data) and an episode of Dan. Carlin's Hardcore History from October 30, 2013 was still hanging in there with a 3 ranking. Unfortunately, I don't have the data to prove what sections of an episode a listener might skip over or where in the episode they might tune out.

Collecting My Data

When I first decided to investigate whether there was a correlation between podcast episode popularity and descriptions, I hoped to find a recent dataset on Kaggle or Data World that included that information. This proved difficult: the few datasets on podcasts I could find were either out of date or were simply lists of all

podcasts listed on iTunes. Knowing that the data I needed didn't exist, I set out to create the dataset myself from scratch.

This proved to be a fairly tedious task, but I knew that it would be a time saver when it came time for me to wrangle and clean my data. My initial thought was to use the top 100 podcasts overall on Apple podcasts (https://www.apple.com/itunes/podcasts/). On closer inspection, it became clear that the topics are so disparate – covering history, politics, pop culture, sports, etc. – that it would be hard to prove trends in keywords and popularity. I decided to stick to one subcategory – Society & Culture – to focus on for my dataset. I also decided that since each podcast can have dozens, if not hundreds, of episodes, I would scale my dataset down to the Top 50 podcasts in Society & Culture knowing that this would still yield hundreds of episodes for me to comb through.

I started collecting data in a CSV document focusing on collecting the podcast name, episode title, date, running time and, most importantly, episode description and popularity rating. I wanted to focus on the bare minimum knowing that cutting out the clutter would yield cleaner data. This yielded 199 rows in my dataset and since Apple's API is not the most user friendly, I scraped a different podcast hosting site, Stitcher (https://www.stitcher.com/), to obtain the data and used Apple Podcasts for their popularity rating.

Once my data was collected, the next step was to create a corpus. Since I was planning to do some text mining later in my analysis, creating the corpus was a crucial step.

Cleaning My Data

Once my data was collected, I set to work on cleaning it. For this, I used some basic data cleaning functions. I used:

- the tolower function to make my text uniform
- removed stopwords, numbers and punctuation
- used the stemDocument function

Once my data was cleaned, I could move forward with analysis.

Initial Findings

The majority of podcast episodes are only somewhat popular, with most being ranked at around 3. We can see that podcast episodes from the summer months appear with the highest density. This makes sense, as I pulled my data in the summer. There are definitely more podcast episodes pulled from the summer months of July and August, and their popularity ratings are pretty high - we can see a lot of 3s, 4s and 5s. There isn't really a correlation between episode length and popularity.

Going Forward

I initially wanted to use more text mining in my analysis and see if I could use term associations to see if there was a correlation between episode description and popularity. But now, I think I will use statistical analysis and create histograms, time-series plots and scatter plots to test my hypothesis. I will also create a term frequency model and word cloud to see the terms that pop up in episode descriptions again and again. And I might use Machine Learning to try to see if I can create a crude Episode Recommendation system.