Michael Winters
SF_DAT_15
Final Project Question and Data Set

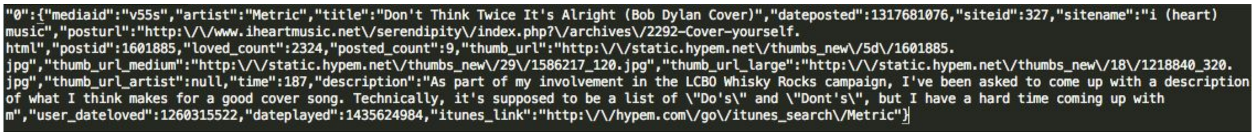
Question: Can I predict whether I'll like an unheard song on Hype Machine based on my listening and ratings history as well as the data about each song that are provided by the site? And can I pull and evaluate lists of unheard song IDs on a regular basis, thus creating an ongoing collection of songs I've never heard but am likely to enjoy?

Initially, I wanted to build a recommender for Vimeo, but I decided that I haven't generated enough viewing history to build a useful dataset, and I probably won't be able to build said useful data set during the span of this class.

What's Hype Machine? '*Hype Machine keeps track of what music bloggers write about. We've carefully handpicked a set of 871 music blogs and then present what they discuss for easy analysis, consumption and discovery. This way, your odds of stumbling into awesome music or awesome blogs are high.'*

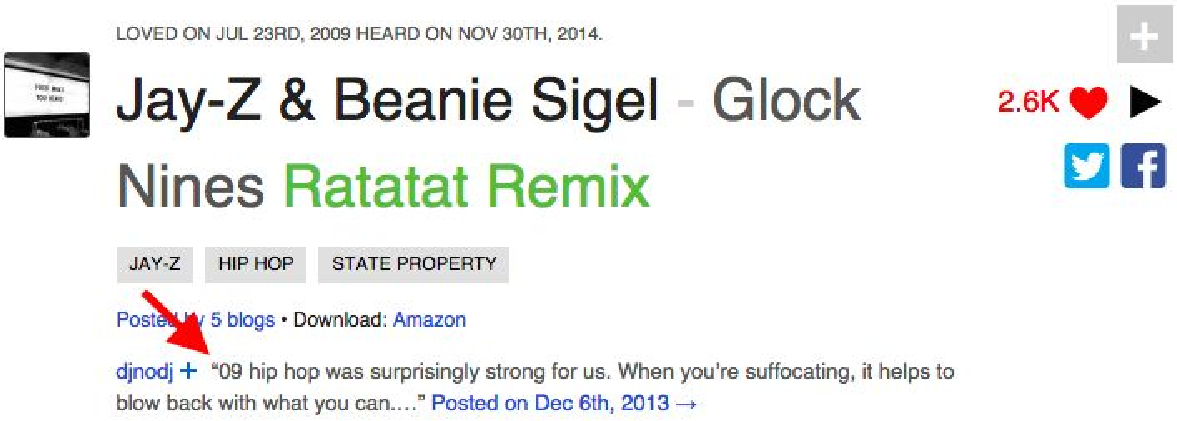
Datasets:

- ***Listening history:*** Because the folks over at Hype Machine are a bunch of saints, each user's listening history is available in JSON for free. The JSON URL format is straighforward (http://hypem.com/playlist/history/wints/json/1/data.js) for easy iteration through pages. Each 'history' entry includes, per song, some pretty exciting stuff:
    - Hype Machine song ID
    - Arist
    - Title
    - Name of blog whose post was featured on the site
    - Total times all users have 'loved' the song (loved song = Hype Machine's positive feedback mechanism)
    - Number of other blogs who posted the same song
    - Timestamp when I last played the song
    - Timestamp when I 'loved' the song (if I did in fact 'love' it--this field doesn't exist if I listed to the song but didn't 'love' it)



```
"0":{"mediaid":"v55s","artist":"Metric","title":"Don't Think Twice It's Alright (Bob Dylan Cover)","dateposted":1317681076,"siteid":327,"sitename":"i (heart)
music","posturl":"http:\/\/www.iheartmusic.net\/serendipity\/index.php?\/archives\/2292-Cover-yourself.
html","postid":1601885,"loved_count":2324,"posted_count":9,"thumb_url":"http:\/\/static.hypem.net\/thumbs_new\/5d\/1601885.
jpg","thumb_url_medium":"http:\/\/static.hypem.net\/thumbs_new\/29\/1586217_120.jpg","thumb_url_large":"http:\/\/static.hypem.net\/thumbs_new\/18\/1218840_320.
jpg","thumb_url_artist":null,"time":187,"description":"As part of my involvement in the LCBO Whisky Rocks campaign, I've been asked to come up with a description
of what I think makes for a good cover song. Technically, it's supposed to be a list of \"Do's\" and \"Dont's\", but I have a hard time coming up with
m","user_dateloved":1260315522,"dateplayed":1435624984,"itunes_link":"http:\/\/hypem.com\/go\/itunes_search\/Metric"}
```

- ***Song data:*** Each song ID (sample URL: http://hypem.com/track/rhs4) contains useful data, too, which will be scrapable using Beautiful Soup.
    - Tags (sample for song URL above: "tags":["jay-z","hip hop","state property"])
    - Song length
    - A snippet of text from the blog post that was featured on the site, example below:



LOVED ON JUL 23RD, 2009 HEARD ON NOV 30TH, 2014.

Jay-Z & Beanie Sigel - Glock Nines Ratatat Remix

2.6K

JAY-Z HIP HOP STATE PROPERTY

Posted by 5 blogs • Download: Amazon

djnodj + "09 hip hop was surprisingly strong for us. When you're suffocating, it helps to blow back with what you can...." Posted on Dec 6th, 2013 →

Why this project? And why Hype Machine? I'm very much interested in media recommendation engines because I constantly struggle to filter through the huge amounts of media in all formats (writing, podcasts, music, video) that are created every day so that I can find the few pieces I know I'll enjoy.

This is a particularly challenging problem for media from independent producers who don't have access to high-profile distribution channels--in other words, media that likely won't reach me via some channel's (like the NYT) push and that I have to seek out instead.

Hype Machine has solved part of this problem by aggregating the best and newest independent music from all over the internet. But Hype Machine:

a. does not offer any sort of personalized recommendations
b. built its core discovery engine around a 'Popular' list, which is generated based on the *community's* preferences rather than the *individual's* preferences. And therefore, there are lots of great songs buried in the site but difficult to discover due to the fact that they'll never make the 'Popular' list (largely, I believe, because the Hype Machine community's genre preferences are limited to dance / EDM / hip-hop / indie rock, in that order, while the site collects music from all genres). In addition, a song's tenure on the 'Popular' list is often limited to only a couple of days. So if I happen to miss a great song that makes the 'Popular' list, I'd need to expend significant time and energy to track it down.

I want to build myself a recommendation engine that will allow me to pull songs I'll love from the depth of Hype Machine so that I can enjoy the site's rich collection of music based on my own preferences rather than the community's preferences.

What are my concerns about the project?

- While I've 'loved' a significant number of songs (419 as of today), my engagement with the 'love' feature is inconsistent, so there are periods when I won't love songs simply because I'm working or distracted or not paying attention. So my listening history is not a true representation of songs I enjoy and songs I don't enjoy. Rather, it's a representation of:
    - songs I enjoy when I am in a mindset to 'love' a song on Hype Machine

    - songs I enjoy when I am not in a mindset to ‘love’ a song on Hype Machine
    - songs I don’t enjoy
- I tend to listen to Hype Machine in spurts, and I’ve been a listener since summer 2009. So over the course of the past 6 years, I’ll have 1-2 week streaks of high-volume listening followed by months of no listening. I am concerned this somewhat inconsistent distribution of ‘loves’ over time might impact my model.
- **so tldr:** I’m afraid my dataset isn’t great and won’t be able to generate solid recommendations

Why do I feel ok moving forward with the project even given these concerns / the fact that I might not be able to build a successful model with the data I have?

- I’ll still get to work through the entire process of creating a media recommender, and I am really stoked about that.
- It’s a really cool opportunity to think through how UX design enables better recommendations for users; if a media app is designed so that users can effortlessly provide feedback on the content they’re being served, data science teams are likely to receive richer datasets and will therefore be able to build superior recommendation models, thus creating a virtuous user experience cycle.
- So, as a part of my project, I might actually create mockups to show how Hype Machine could make it easier for users to provide feedback and therefore collect better data for recommendations. I’m a product manager, so I enjoy these types of design problems, too.