# What are They Even Saying???

An exploration of the linguistic styles of podcasts

### Plan and Hypothesis

I started with an ambitious but admittedly not revolutionary idea - train a model that could "read" podcast transcripts and learn genre, rating, topics, format, and year from various textual and non-textual features. Initially, I also thought that I would be able to clean the transcript text enough to parse out each host's speech, then use that data to train a model to assign a host-name label to a random string of text.

This is basically the story of how those plans collapsed . . . while providing some interesting data along the way.

#### Where did the data come from?

I collected transcripts from 24 podcasts, 5 of which fell victim to the re-reading of the website's copyright rules, so my total number of podcasts was 19.

I used transcripts from the podcast's official page, with one exception (NeoScum, whose transcripts were fan-made and in a Google drive with guest access).

#### Data overview:

- 1584 episodes

- 16,650,103 transcript tokens (as opposed to title tokens)







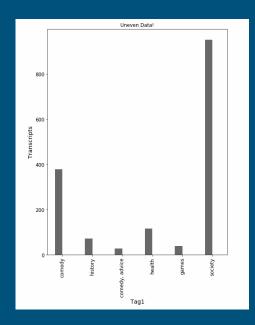


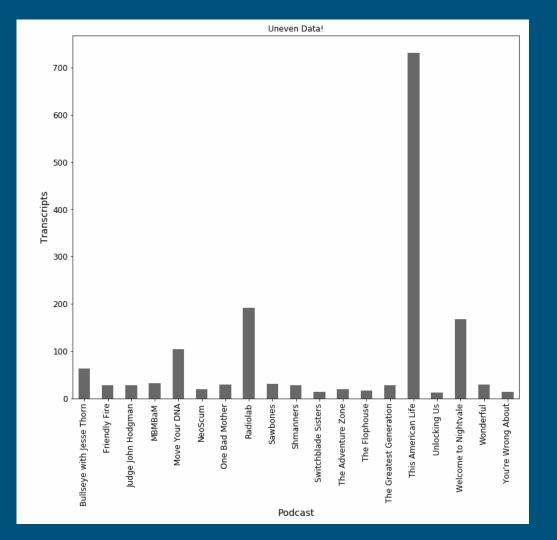
### Target features

```
pod feats = [['Welcome to Nightvale', 1, ['comedy', 'sci-fi'], 'scripted', 'fiction', 'news', 4.8],
                ['Move Your DNA', 2, ['health', 'fitness'], 'unscripted', 'nonfiction', 'chat', 4.8],
 2
                 ['You\'re Wrong About', 2, ['history', 'education'], 'unscripted', 'nonfiction', 'chat', 4.6],
                 ['Unlocking Us', 1.5, ['health', 'lifestyle'], 'unscripted', 'nonfiction', 'interview', 4.6],
                 ['Radiolab', 2, ['society', 'education'], 'unscripted', 'nonfiction', 'storytelling', 4.7].
                 ['This American Life', 1.5, ['society', 'history'], 'unscripted', 'nonfiction', 'storytelling', 4.6],
                 ['Bullseye with Jesse Thorn', 1.5, ['comedy', 'society'], 'unscripted', 'nonfiction', 'interview', 4.7],
                 ['One Bad Mother', 2.5, ['comedy', 'parenting'], 'unscripted', 'nonfiction', 'chat', 4.7],
                 ['Judge John Hodgman', 1.5, ['comedy, advice'], 'unscripted', 'nonfiction', 'chat', 4.8],
 9
                 ['The Flophouse', 3, ['comedy', 'movies'], 'unscripted', 'nonfiction', 'recap', 4.8],
10
                 ['Switchblade Sisters', 1.5, ['comedy', 'movies'], 'unscripted', 'nonfiction', 'chat', 4.9],
11
                 ['MBMBaM', 3, ['comedy', 'advice'], 'unscripted', 'nonfiction', 'chat', 4.9],
12
                 ['Sawbones', 2, ['history', 'medicine'], 'unscripted', 'nonfiction', 'storytelling'. 4.8].
13
                 ['Wonderful', 2, ['comedy', 'society'], 'unscripted', 'nonfiction', 'chat', 4.9],
14
                 ['The Greatest Generation', 2, ['comedy', 'TV'], 'unscripted', 'nonfiction', 'recap', 4.9],
15
                 ['Friendly Fire', 3, ['history', 'movies'], 'unscripted', 'nonfiction', 'recap', 4.6],
16
17
                 ['Shmanners', 2, ['society', 'advice'], 'unscripted', 'nonfiction', 'chat', 4.8],
                 ['The Adventure Zone', 4, ['games', 'RP'], 'unscripted', 'fiction', 'LARP', 4.9],
18
                 ['NeoScum', 5, ['games', 'RP'], 'unscripted', 'fiction', 'LARP', 4.9]]
19
20
21 # In case you're a cool person reading this and don't know, LARP is live action role playing.
```

### A disclaimer unbalanced data

Data isn't evenly distributed, so all findings come with a gigantic grain of salt. This American life had by far the most episodes available. Only 520 episodes had a year listed on the transcript page and 53 episodes did not list a title - they were all called "Final Draft" for some reason.





### Scraping and cleaning

```
['\n Smile My Ass\n ',
'\n January 29, 2021\n ',
"\n Jad: \r\n Wait, you're listening... \r\n \xa0 \r\n Speaker 2: \r\n Okay. \r\n \xa0 \r\n Jad: \r\n All right.
\r\n \xa0 \r\n Speaker 2: \r\n Okay. \r\n \xa0 \r\n Jad: \r\n All right. You are listening to radio lab radio lab. W. N Y. C.
all right. Latif, if you can rewind your mind back to a time when your life wasn't dominated by Allen Funt in Candid Camera.
How did this start? \r\n \xa0 \r\n Latif: \r\n So I first, unlike a lot of people, I did not grow up watching candid camera.
I had never heard of candid camera when I was a kid. \r\n \xa0 \r\n Jad: \r\n You never heard of candid camera? \r\n \xa0 \r
\n Robert: \r\n You've never heard of candid camera? \r\n \xa0 \r\n Latif: \r\n No. \r\n \xa0 \r\n Bobert: \r\n Have you heard of the Declaration of Independence, that ringing the bell? \r\n \xa0 \r\n Latif: \r
```

Smile My Ass

Jad: Wait, you're listening... Speaker 2: Okay. Jad: All right. Speaker 2: Okay. Jad: All right. You are listening to radio lab radio lab. W. N Y. r life wasn't dominated by Allen Funt in Candid Camera. How did this start? Latif: So I first, unlike a lot of people, I did not grow up watching You never heard of candid camera? Robert: You've never heard of candid camera? Latif: No. Jad: Wait. How... Robert: Have you heard of the Declarat hat? Robert: No, but it's sort of, he's up there with it. It's very noticeable. Latif: Okay, cool. Jad: He is actually, no BS, a founding father i I'm Robert Krulwich. Jad: This is radio lab... Robert: When you least expect it, you're addicted, you're the one today. Jad: Okay, just to set tha between show and life was really clear. Jad: Then along came a guy named Allen Funt who muddied that line in a way that was fascinating and would arks are on all of our butts. So check your tush and listen to this story from our producer Latif Nasser. Latif: So I first heard about candid cam

The unpredictable variation in website setup meant that I needed an individual scraper for each podcast or each podcast network. I would have liked to parse out host names and separate their speech, then analyze each host's style, but that was only possible for a select few podcasts. Radiolab, for instance, had four different formatting styles for speaker tags, and NeoScum's tags varied within each transcript.

# Machine Learning Part A: Regression



I went a little bit wild with extracting non-textual features. I'll briefly describe all of them now (briefly because there are 50 of them):

- Token count (int, transcript length)
- Token lengths (list of tuples: (token, length))
- Avgerage token length (float, mean of all alphabetic token lengths)
- TTR (float, type/token ratio measured against 300 characters)
- Avgerage kband (float, mean kband)
- Part of speech frequency (dictionary as {POS: % of entire document})
  - Noun
  - Verb
  - Adjective
  - Adverb
  - Interjection
  - Preposition
  - Conjunction
- Average sentence length (float, average sentence length over entire transcript)

- Part of speech length (dictionary as {POS: average POS length})
  - Noun
  - Verb
  - Adjective
  - Adverb
- Pronoun counts (dictionary as {pronoun: % of all pronoun occurrence that this pronoun makes up})
  - 0
  - You
  - She
  - He
  - O Thev
  - We
- Verb lemmas (dictionary of 20 most common verb lemmas as {lemma: % of all verbs that a verb comprises}) and their frequencies
  - Know
  - o Be
  - O Do
  - Mean
  - Make
  - O Go

- Entities (dictionary as {spacy's ent tag of token:
   % of ent occurrence over document length})
  - Organization
  - ⊃ **Art**
  - Geopolitical (countries, cities, etc)
  - O Cash
  - time\
  - product
- Opinioncount (float, occurrence of pronoun followed by optional auxiliary followed by lemma think or feel weighed against total verb occurrence)
- Prepositions per sent (float, average occurrence of prepositions per sentence)
- Donation appeal (int, count of "donate" occurring as a phrase root)
- Social count (int, count of how many times a social media platform is mentioned)
- And several more . . .

# Machine Learning Part B: Tfidf Vectorizer and Multinomial NB

stices besides my clock being. 4 seconds slow can we resolve here? [Jesse laughs quietly.] We can do our small part to resolve some small injus jesse Bullseye . Jesse Thorn. If a fan os: "My husband is obsessed with electric vehicles and free charging. We recently attempted to go ca coming . reminiscent of Daniel Klaus or Jit a motel, we spent the night in an empty hospital parking lot because there was a free charger for several covers for the New Yorker , incluer in order to save energy. We dream of building or buying a home soon, but while I look at school rtments. dates over video chat, Daniel Tihn You keep reading, I\'ll look it up. jesse "I seek an order that we use a gas vehicle on trips th **n it. he takes a little bit of a differen**ing. I would also like to use the dry cycle, or have him towel-dry all the dishes." john So an ICF book tours, trips to the doctor, even an thought it stood for "ice cream foams." That would have been delicious. jesse Yeah. But [laughs] ha ast Recommended if You Like . interviewede out of ice cream foam!" Ohhh, white chocolate flavor, please. No, it\'s of polystyrene foam with riter. [Music fades in.] So. get into it. ps the hot hot and the cool cool, jesse Got it. john verv. verv old reference, jesse Sure, john To t in 5 months from now. DANI: Not if you want to stay friends with me! I\'m ju now." You ever ne Hi Brian. How you doing? brian I am do<sup>b do ya</sup> rstand about other animals and how they use each other to get the group from pine Mount in I go as kind of this theme that just kept coming back up in my life. DANI: Mm.. KAT he McDLT in, I go re referring to [laughs] when you asked, ven revealing. But it\'ll probably come out in interviews that I\'ll be doing he to town to sell irst when all the plans changed, because listen to everything. Everything that I do on Movement Matters , I should prob/one\'s buying. [3 a little better than if we were standing should think so. I think this book\'s gonna be around for a long time. I think it\'s and and wife, whe conversation. I know felt less self-consc r. DANI: ...for an author and an educator. What? KATY: There\'s silver. The con is really your most personal and, I think he contents are pretty shiny too. KATY: There\'s some good puns. Did you catch as I prepared, but I was expecting some k ANI: But it was just, it\'s such a good book and to everybody listening, I cle t I depict within the book it because goo It\'s gonna stun you if you read it. DANI: You can ask my husband if he were he t and read it in the yard and stuff and he would watch me just close the kindle where I was like, I know so-and-so lives say, "Good book?" And I\'d just say, "It\'s heavy" There\'s a lot to think abo udience Q&A but probably would not get fr for putting it out there into the world. And for those of you listening to this ience might? adrian I was pretty sure tha ly think this is one of those books that will help you move through the world ake the book seem very casual and forthri t. DANI: Yeah. KATY: The community of the planet. DANI: And I don\'t feel like I felt like I could expose other people. u\'re very generous with your knowledge and stuff, but this one I just feel is eading that. But this one is just if you live in this universe, it\'s probably st because my nature is just love talking

s] onstage that I would expose them in a

uld read it. DANI: I do too. KATY: I think there should be some universal book

g me read it. KATY: Of course. DANI: An early read. And um.. KATY: And thanks

						Lo	gisti	ic Re	egre	ssio	n ad	ccur	acy:	98.	422	7%				
	The Adventure Zone -	11	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
nue label	You're Wrong About -	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	One Bad Mother -	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sawbones -	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Radiolab -	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	judge John Hodgman -	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
	NeoScum -	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
	The Greatest Generation -	0	0	0	0	0	0	0	37	0	1	0	0	0	0	0	0	0	0	0
	Welcome to Nightvale -	0	0	0	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0
	The Flophouse -		0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
	Switchblade Sisters -		0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
	Wonderful -		0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
	This American Life -		0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
	MBMBaM -		0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
	Unlocking Us -		0	0	0	0	0	0	0	0	0	0	0	0	0	146	0	0	0	0
	Bullseye with Jesse Thorn -		0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
	Move Your DNA -		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0
	Shmanners -		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0
	Friendly Fire -	- 1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
		The Adventure Zone	You're Wrong About	One Bad Mother	Sawbones	Radiolab	Judge John Hodgman	NeoScum	The Greatest Generation	Welcome to Nightvale	The Flophouse	Switchblade Sisters	Wonderful	This American Life	MBMBaM	Unlocking Us	Bullseye with Jesse Thorn	Move Your DNA	Shmanners	Friendly Fire
		ture	f buo	ad M	Sawb	Rad	Hod	Neo	ener	Nigh	Flop	de S	Wond	erica	MB	lockii	sse	Your	hmar	endl
		dven	e Wr	ne Bi			John		est G	oe to	Jhe J	chbla	_	Ame		5	ith Je	Move	O	Έ
		he A	You'n	0			adpr		Sreat	elcon		Swite		Ţ			ye w	_		
							_		The (	ž							allse			
		predicted label																		

# Test #1A: Predicting Podcast

125

- 100

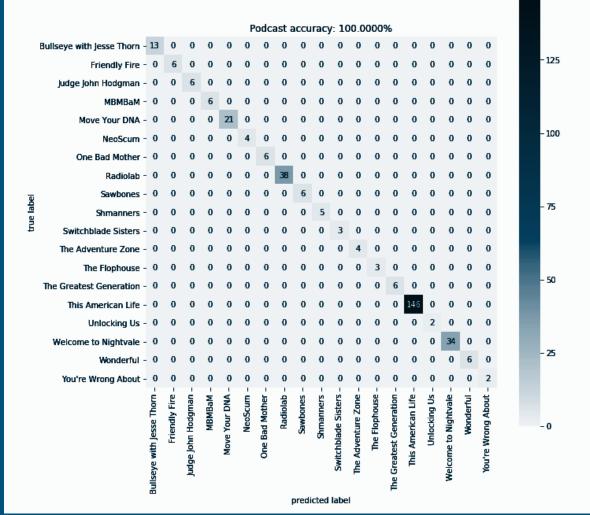
- 75

50

Logistic regression

## Test #1B: Predicting Podcast

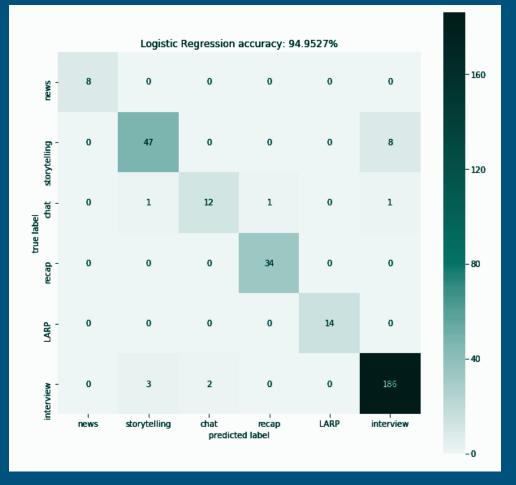
Tfidf and NB only 1,000 features!



# most features **Jnderstanding the** tive informat

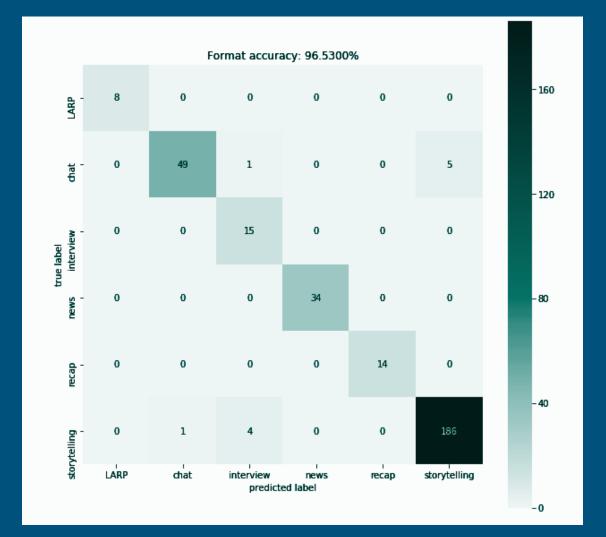
Friendly Fire: guy, laugh, music, really, know, right, clip, gonna, think, just, war, yeah, film, laughs, movie, like, john, ben, adam, host Judge John Hodgman: wanna, joel, think, mm, quietly, right, laughter, gonna, hm, don, uh, just, judge, know, laugh, yeah, like, laughs, jesse, john MBMBaM: good, got, wanna, laughing, right, think, oh, um, know, fucking, gonna, just, okay, uh, yeah, laughis, like, justin, griffin, travis Move Your DNA: don, lot, kind, way, things, stephanie, time, body, going, yeah, people, really, right, think, know, movement, just, like, dani, katy NeoScum: gonna, right, laughs, got, going, oh, guys, like, okay, just, tech, yeah, mm, dak, dr, es, tw, ct, bb, gr One Bad Mother: wanna, time, music, okay, job, good, host, think, doing, right, know, really, yeah, crosstalk, gonna, just, like, laughs, theresa, biz Radiolab: pat, annie, radiolab, right, matt, okay, people, speaker, clip, yeah, simon, know, krulwich, latif, just, molly, like, abumrad, robert, jad Sawbones: say, things, mean, really, medical, lot, gonna, right, yeah, think, okay, uh, people, just, um, know, laughs, like, justin, sydnee Shmanners: lot, hmm, thing, oh, say, people, yes, gonna, think, just, yeah, know, uh, laughs, right, um, okay, like, travis, teresa Switchblade Sisters: kelly, way, people, laughs, speaker, music, gonna, quote, really, yeah, movie, kind, think, uh, just, film, know, like, um, april The Adventure Zone: mean, ve, right, kind, oh, think, um, gonna, just, laughs, know, yeah, like, okay, uh, justin, clint, griffin, travis, fitzroy The Flophouse: justin, think, know, okay, people, laughter, uh, gonna, multiple, audience, just, laugh, yeah, movie, crosstalk, laughs, like, dan, stuart, elliott The Greatest Generation: scene, really, mm, make, just, gonna, think, know, right, episode, promo, music, uh, laughs, yeah, clip, like, ben, adam, host This American Life: act, way, did, say, right, got, ve, really, time, didn, think, said, going, don, know, people, just, glass, like, ira Unlocking Us: julie, way, book, right, time, black, love, said, want, yeah, white, really, say, going, know, think, just, people, like, bb Welcome to Nightvale: new, station, police, dana, old, secret, did, city, good, mayor, town, time, listeners, know, said, just, like, cecil, night, vale Wonderful: right, good, sort, thing, oh, gonna, kind, okay, lot, really, um, think, know, just, yeah, laughs, uh, like, rachel, griffin You're Wrong About: movie, okay, right, thing, time, kind, sort, really, going, don, think, yeah, know, people, just, michael, marshall, mike, sarah, like

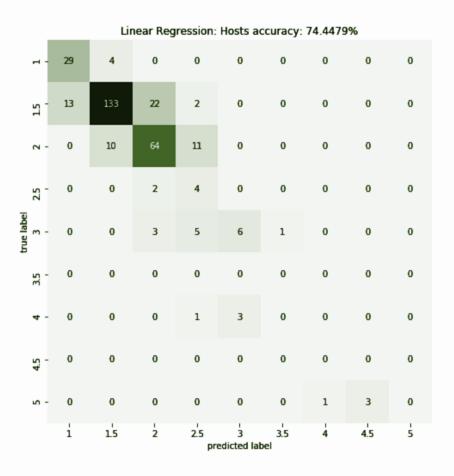
Bullseye with Jesse Thorn: did, promo, time, wanna, mean, gonna, people, chuckles, yeah, laughs, jordan, really, kind, think, fades, just, music, like, know, jesse



## Test #2A: Predicting format

## Test #2B: Predicting format





# Test #3A: Predicting hosts

- 125

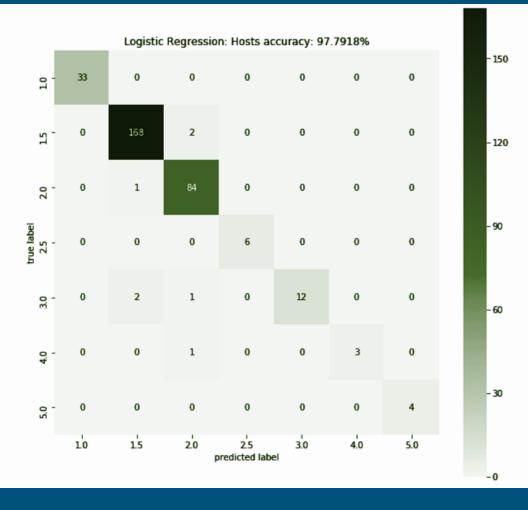
- 100

- 75

- 50

- 25

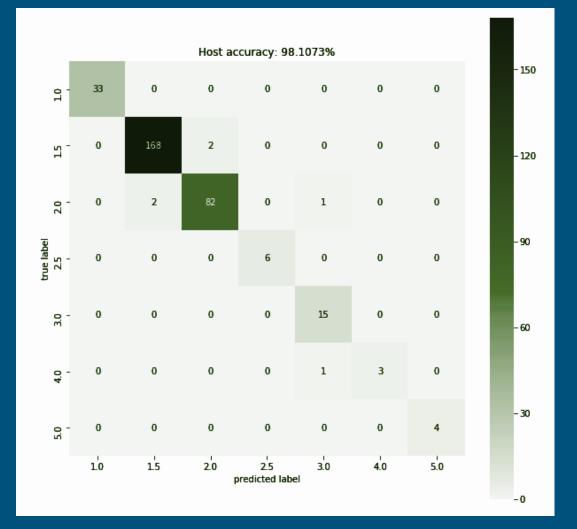
This one was a bit tricky, since the host numbers aren't actual values, but numerical representations, so they function as labels. Had to do some rounding of the predictions.

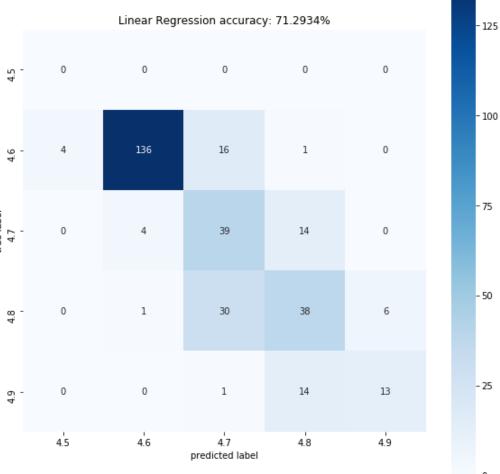


## Test #3A: Predicting hosts

Turns out logistic regression worked best on this, since the target values are technically labels.

## Test #3B: Predicting hosts





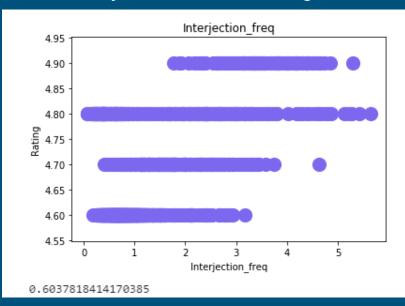
# Test #4A: Predicting rating

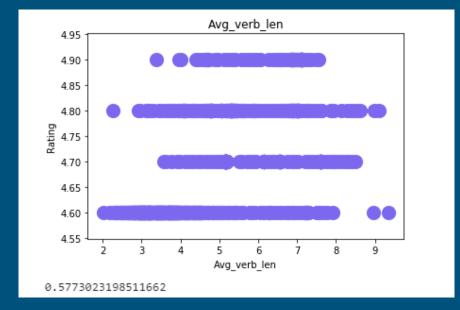
Ratings ranged from 4.6 to 4.9. This makes sense, since poorly-rated podcasts don't have transcripts. This tiny margin isn't ideal for training a model.

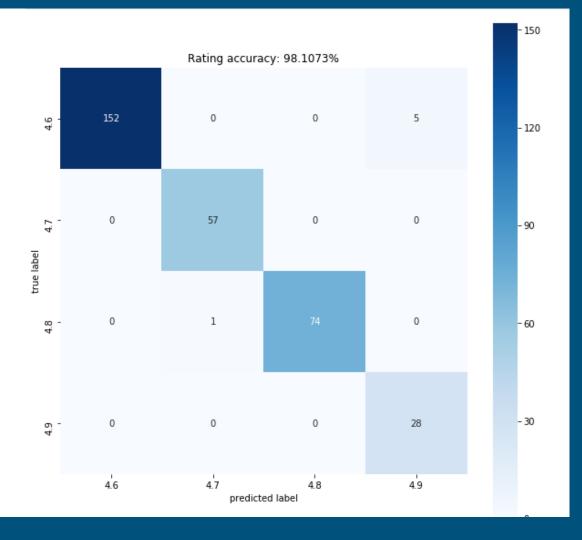
-0

# An aside . . . do any non-textual values correlate to rating?

Possibly? Here are the highest correlation values:



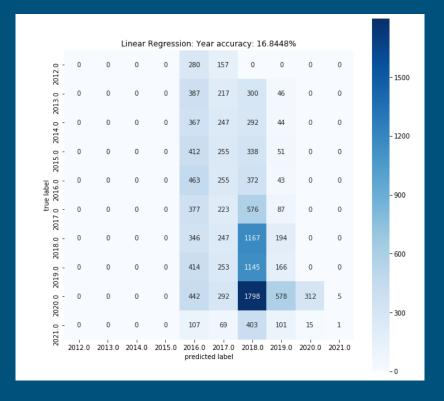




# Test #4B: Predicting rating

# Test #5A: Predicting year





\*only 520 of the transcript pages had scrapable year metadata. Extremely limited sample size! 205 of the 520 year-labeled transcripts were from 2020, so both models performed worse than if they would have guessed "2020" for all labels.

### Test #5B: Another failed attempt at predicting year

After running a gridsearch using a vectorizer, the best accuracy was still pretty abysmal at 0.39423. I will definitely say that, at least with this small of a data set, there is no way of predicting year from podcast.

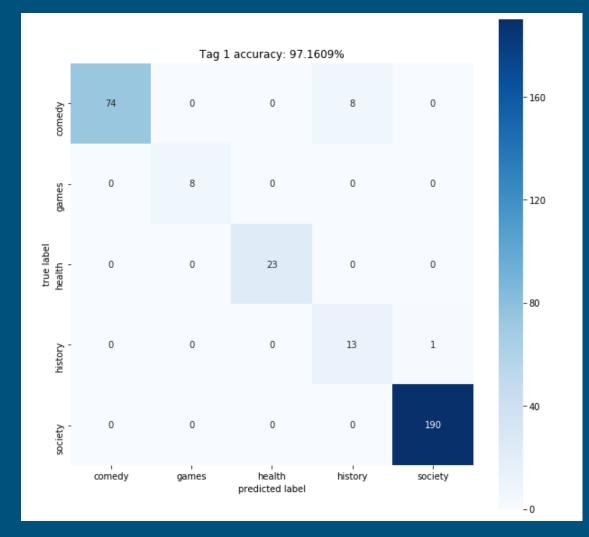
Maybe when more transcripts become available, a vectorizer model will be able to do this task with higher accuracy. It would be interesting to see the most informative features from year to year.

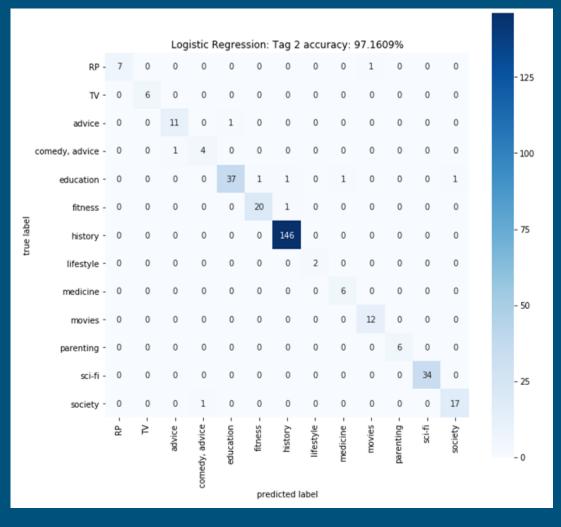




# Test #6A: Predicting Main Tag

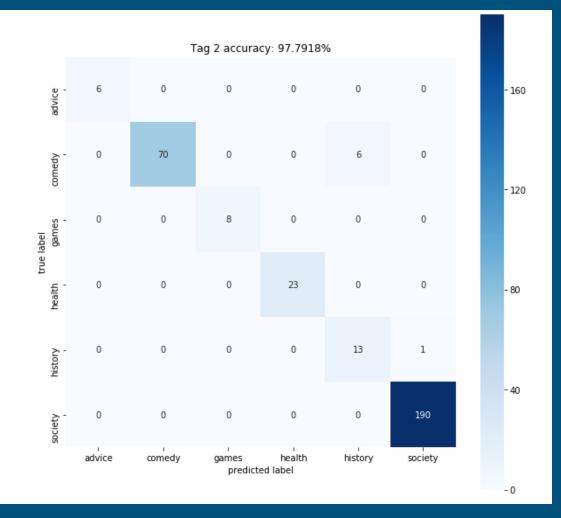
# Test #6B: Predicting Main Tag



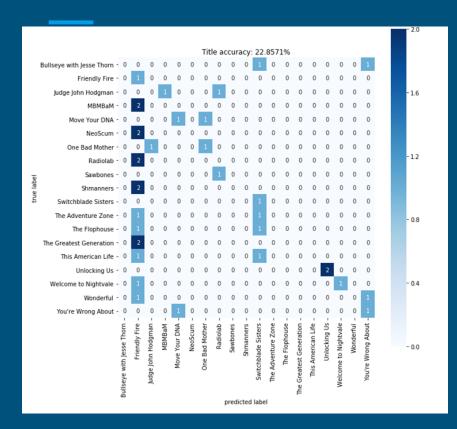


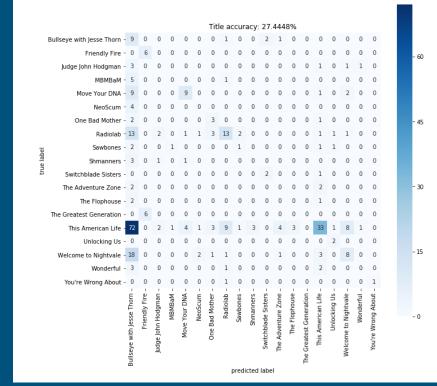
# Test #7A: Predicting Secondary Tag

# Test #7B: Predicting Main Tag



# Predicting podcast from vectorized title





<< even
representation
among podcasts
(228 total)</pre>

This American Life	731
Radiolab	192
Welcome to Nightvale	168
Move Your DNA	104
Bullseye with Jesse Thorn	63

### Using the classifier

#### Emma chapter 1 = This American Life

This makes sense, even though the classifier was probably just highly-weighted in favor of TAL. This American Life is about peoples' very personal stories, usually with a lot of emotional speech involved.

#### Leviathan Wakes (from The Expanse) = Welcome to Nightvale

This also tracks. Leviathan Wakes is a science fiction book, with lots of talk about space, spaceships, and the like. Welcome to Nightvale is also science-fictiony.

#### My Al Essay from Comp Ling = Move Your DNA

I don't quite get this one . . .

#### Real-world uses of this data

I personally think that podcasts are a very accurate representation of casual conversation. When I have to do another data science project, I'll find out how true that instinct is by comparing it to some real-world conversation in the form of corpora that we talked about in class.

If podcast data *does* reflect casual speech, transcripts would be very useful in training predictive text language models, speech synthesis, and second language acquisition.



### Thanks for listening!

Credits roll. [ending sax music

plays]

Make it so! Make make make-makemake-make make it so!

Kiss your dad square on the lips.

All right. Bye-bye.

But I will enjoy talking to you next time.

Thank you for listening. Please stay safe everybody.

And gonna do it for us, so join us again next week! No RSVP required.

Good night, listeners. Good night.

And we always appreciate you listening. Thank you for being with us. You especially. You, now, hearing this? You're our favorite.

And if there's more than one person in the room? Both of you are our favorites.

I mean, not like I'm your dad! be safe out there! And wash your hands!

And as always, don't drill a hole in your head!

Just remember: all great radio hosts have a signature sign off.