Can we predict a music review's numerical score from its text?

GA Data Science Lightning Talk

Clare McNeely

What's the data set like?

19K music reviews scraped from Pitchfork, dating from 1999 - 2017 (sourced from Kaggle).

Data is stored in a few tables that all have the common `reviewId`. Most importantly:

- **reviews table:** one row per reviewid; contains review attributes like artist, album title, genre, review author, author type (contributor vs. staffer) and pub date. Crucially, it also includes the "score" (1.0 - 10.0) of the review, and a flag that = 1 if the album in question was named "Best New Music"
- **content table:** one row per reviewid; contains reviewid and full text of the review

reviews.csv

reviewid	title	artist	url	score	best_new_music	au
22703	mezzanine	massive attack	http://pitchfork.com/reviews/albums/22703-mezz...	9.3	0	nat pa
22721	prelapsarian	krallice	http://pitchfork.com/reviews/albums/22721-prel...	7.9	0	zoe car
22659	all of them naturals	uranium club	http://pitchfork.com/reviews/albums/22659-all-...	7.3	0	da gli

content.csv

reviewid	content
22703	"Trip-hop" eventually became a '90s punchline,...
22721	Eight years, five albums, and two EPs in, the ...
22659	Minneapolis' Uranium Club seem to revel in bei...
22661	Kleenex began with a crash. It transpired one ...
22725	It is impossible to consider a given release b...

What questions could we ask of this data?

- Can we predict the numerical score of a review based on the text of the review?
- Can we predict whether or not a review labeled its subject “Best New Music” based on the text of the review?
- What words / phrases are most predictive of a good review? Of a bad review?
- What words / phrases are most associated with each genre?