
Team Lime
01 Aditya Chander
02 Ritika Khurana
03 Taylor Mahler
04 Yuchen Luo
GitHub
Spotify Podcast Recommender
Erdős Data Science Bootcamp
Mars Is a Cold Place
The 15th Planet
2:54
3:49

HOME

TABLE OF CONTENTS

THANKS!

Table of contents

Mars Is a Cold Place
The 15th Planet

2:54 3:49

Data gathering

- Podcast dataset provided by Spotify.
- We worked with around 40,000 English-language podcasts for this project, using the transcripts and topic categories.

Data cleaning

- We scraped the full transcripts from the provided .json files
- Some of the provided categories are highly granular, but the overall categorisation system (provided in the RSS metadata for the podcasts) is hierarchical and clean
- Mapping the granular subcategories from the original data to the top-level categories resulted in a reduction in the number of categories from 117 to 19

Modeling

We compared two encodings of the transcripts:

Tf-idf: for the 1000 most common words (excluding stopwords like “the” and “a”), we calculated the importance of these words to each transcript by multiplying:

- The number of occurrences of that word in the transcript
- The inverse of the number of transcripts in which that word appeared

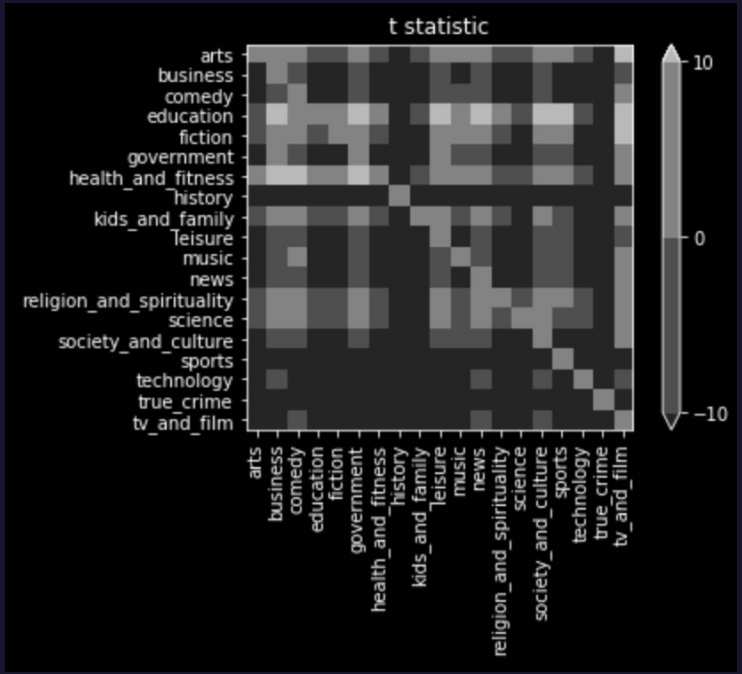
MiniLM-L6-v2: pre-trained transformer model computes 384-dimensional embeddings for each transcript

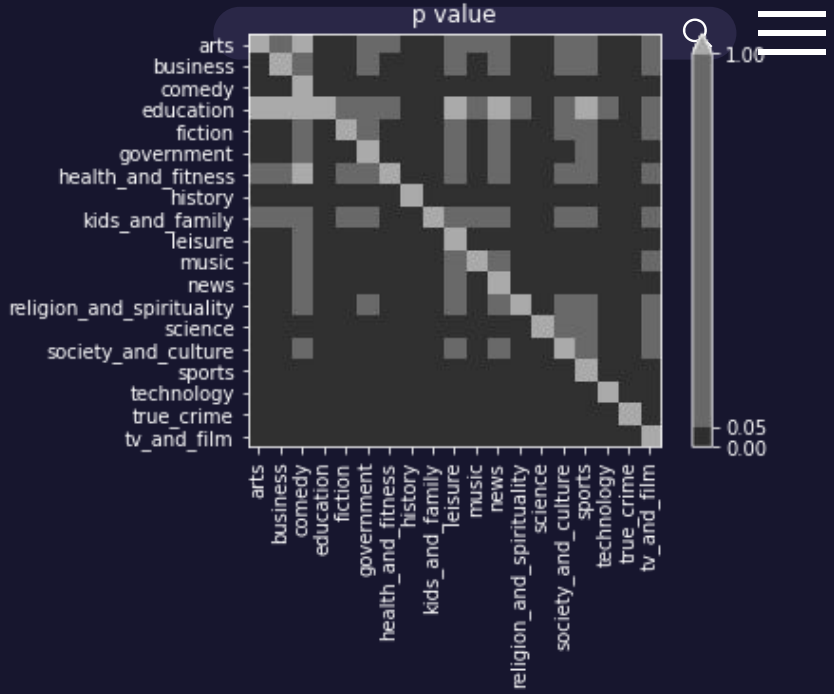
Modeling: evaluation

Are embeddings for episodes within a category more similar to each other than episodes from between categories?

Procedure:

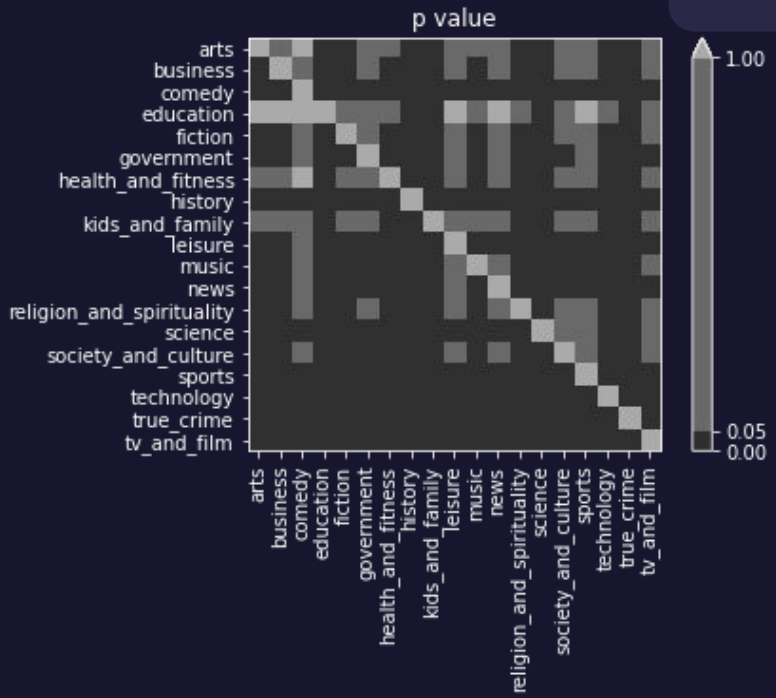
- Sample 50 random podcast episodes from each category
- For every pair of categories, compute the cosine similarity for every pair of episodes within a category and every pair of episodes between categories
- Examine within- vs. between-category cosine similarity distributions with a one-tailed *t*-test; see proportion of category pairs for which the result is significant





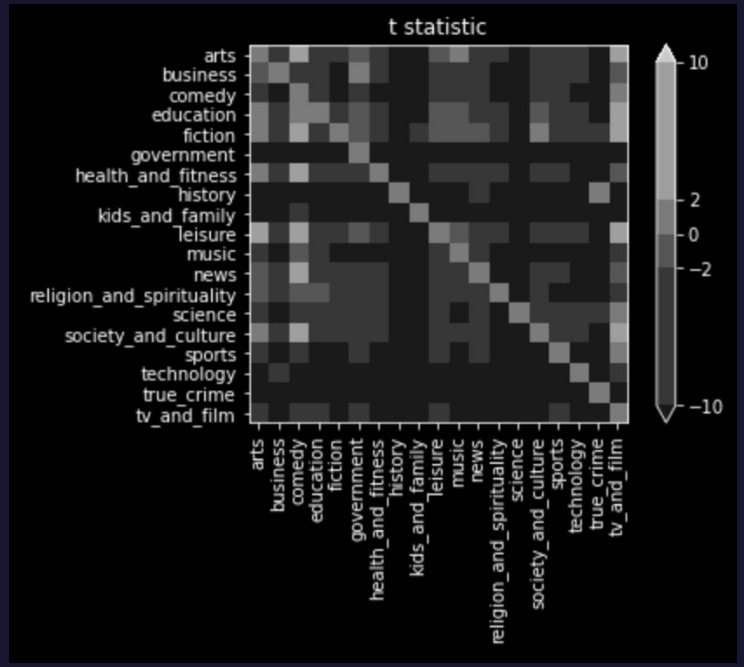
Modeling: Tf-idf 🥈

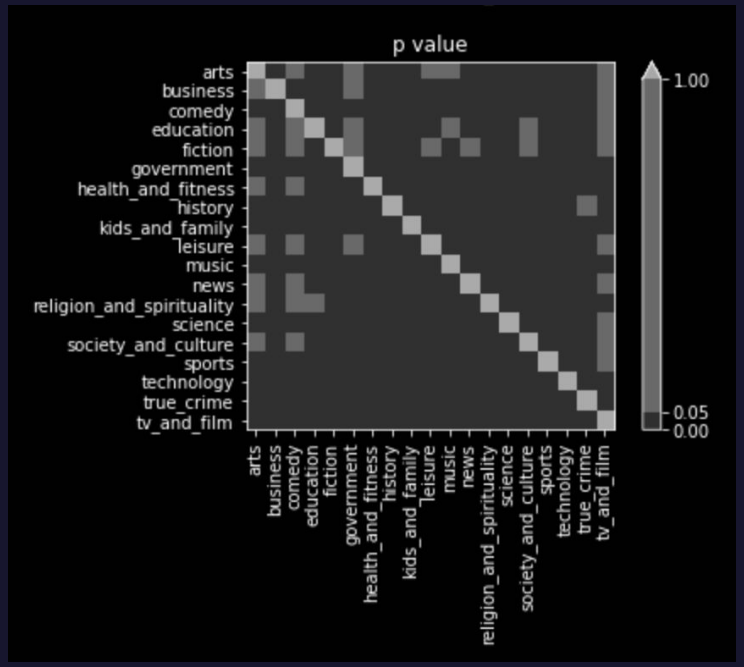
75.1% of ordered category pairs had lower between-category than within-category similarity scores



Modeling: Tf-idf 🥈

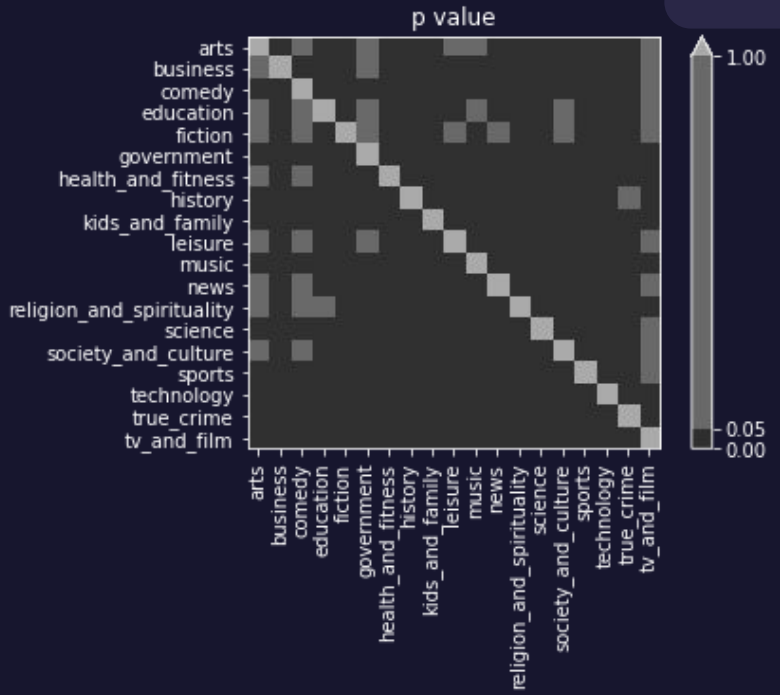
75.1% of ordered category pairs had lower between-category than within-category similarity scores





Modeling: MiniLM-L6-v2 🥇

88.3% of ordered category pairs had lower between-category than within-category similarity scores



Modeling: MiniLM-L6-v2

88.3% of ordered category pairs had lower between-category than within-category similarity scores

Results

Stream lit App!

Future Directions

- Take the app online
- Let the users tell us how good our recommendation engine is!
- Thank you!