



Podcast Clustering Engine at Spotify

Romauli Butarbutar
Suphanet Kotchum
Arushi Mishra
Edoarda Schoch
Christian Wiloejo

Introduction and Background Research

Spotify Podcast : “Digital audio series - Driving Growth Through Personalization”

What is the background project?

- + Spotify is known for its personalized playlist feature.
- + Spotify relies on machine learning to create recommendations based on users' engagement data.
- + Refer to HBS recommendations¹ : “using NLP to analyze podcast content to improve personalized playlist feature”

Why is this project interesting?

- + By implementing NLP techniques to analyze podcast data/metadata (show description, episode description, transcript), we can validate podcasters' descriptions of its content for classification.



Business Use Case

Where we are now

- + The significant growth of total monthly Spotify active users: **19%** year over year, to **381 million** in Q3 (up from 365 million)
- + Spotify Premium subscribers: **19%** to **172 million** in the quarter.
- + However no platforms provide recommendations for podcasts.
- + Existing recommendation tool: Spotify's Find the One online quiz (asks only a handful of questions, recommend the podcasts based on user responses)

Where we want to go

Build a classifier to help listeners discover new podcasts episodes or series

- + Using **Spotify Podcast Data Set**, we want to implement NLP techniques to provide a sophisticated recommendation lists by performing classification.



the Data and How it Was Accessed

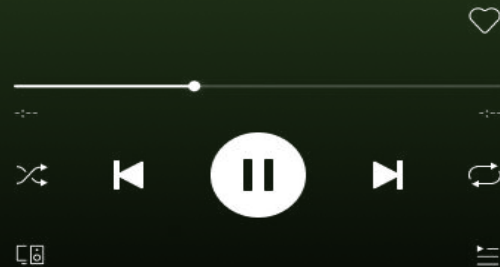
Spotify Podcast Data Set

- Around 100,000 podcast episodes containing show description, episode description, transcript, publisher, etc.
- Episodes are varied in nature

Procurement

- Podcast data readily available, to facilitate research on podcasts through the lens of speech and audio technology, natural language processing, information retrieval, and linguistics.
- Timeline end to end ~2 weeks

Data found at Podcast by Spotify Dataset [site](#)



Data Preview

show_uri	show_name	show_description	publisher	language	rss_link	episode_uri	episode_name	episode_description
spotify:show:2NYtxEZyYelR6RMKmjfPLB	Kream in your Koffee	A 20-something blunt female takes on the world...	Katie Houle	['en']	https://anchor.fm/s/11b84b68/podcast/rss	spotify:episode:000A9sRBYdVh66csG2qEdj	1: It's Christmas Time!	On the first ever episode of Kream in your Kof...
spotify:show:15iWCbU7QoO23EndPEO6aN	Morning Cup Of Murder	Ever wonder what murder took place on today in...	Morning Cup Of Murder	['en']	https://anchor.fm/s/b07181c/podcast/rss	spotify:episode:000HP8n3hNifglT2wSI2cA	The Goleta Postal Facility shootings-January ...	See something, say something. It's a mantra ma...

episode_uri	episode_name	episode_description	duration	show_filename_prefix	episode_filename_prefix	Unnamed: 0	episode	transcript
spotify:episode:000A9sRBYdVh66csG2qEdj	1: It's Christmas Time!	On the first ever episode of Kream in your Kof...	12.700133	show_2NYtxEZyYelR6RMKmjfPLB	000A9sRBYdVh66csG2qEdj	34866	000A9sRBYdVh66csG2qEdj	Hello. Hello. Hello everyone. This is Katie an...
spotify:episode:000HP8n3hNifglT2wSI2cA	The Goleta Postal Facility shootings-January ...	See something, say something. It's a mantra ma...	6.019383	show_15iWCbU7QoO23EndPEO6aN	000HP8n3hNifglT2wSI2cA	14162	000HP8n3hNifglT2wSI2cA	There were two more murders 15 miles away arri...

Design Choices and Rationale

- **Lemmatization**

- Filtering podcast descriptions for specific words
- Filtering for different forms of the same word

- **LDA**

- Understand the podcasts clusters

- **Word2Vec with LDA cluster**

- Similarity between words in LDA clusters and podcast show description

- **Google's Word2Vec with manual clusters**

- Similarity between words in the podcast show description and manually generated clusters

Method 1: Pre-processing through Lemmatization

Implementation

- + We used the “show_description” column in the dataset after removing the special characters and stop words
- + We then use bi-grams with lemmatization of the words to help filter out the podcasts by the meaning of words rather than the words itself while tokenizing them

Methodology

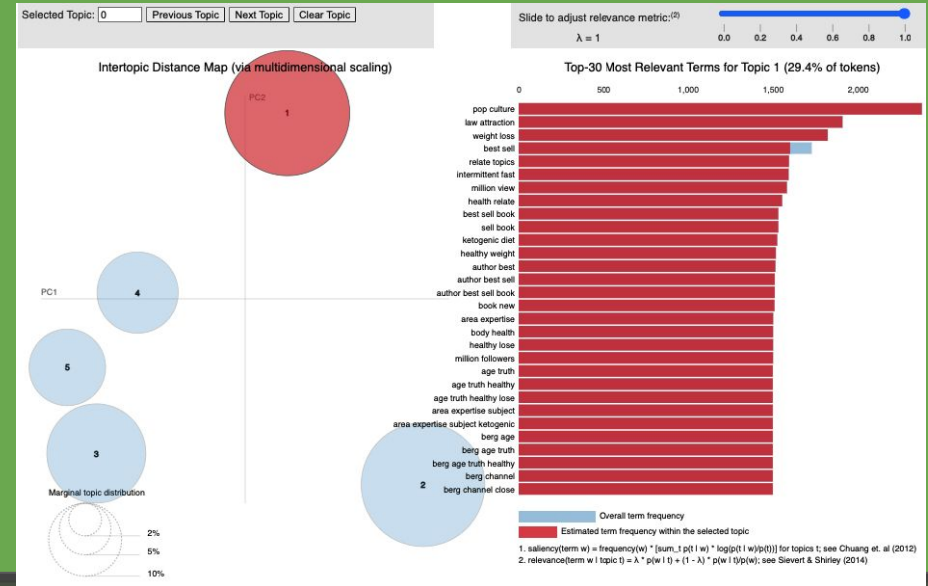
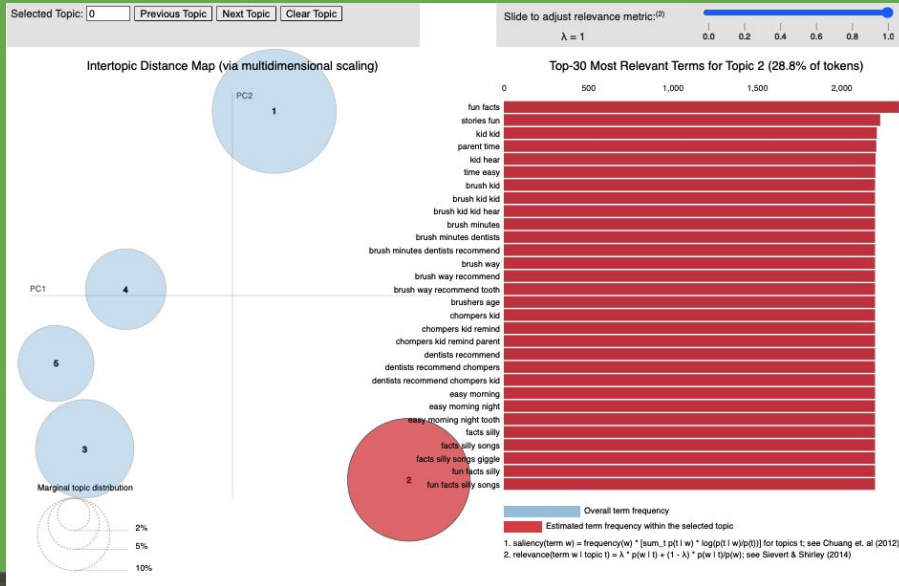
- + After splitting the “show descriptions” in words, removing the special characters as well as stop words, we will convert all the words in the lower case.
- + By using the the lemmatizer from the nltk package, we convert all the words to their verb form for capturing the essence of the word in different aspects it is used in.
- + Then we made the n-grams of 2 to 4 and converted them into vectors them for every “show_description”



Method 2: LDA Topic Modelling

Implementation

- Using “show_description” column, we built our scikit-learn LDA model and obtained 5 clusters describing different topics:



Method 2: LDA Topic Modelling (continued)

	topic_0	topic_1	topic_2	topic_3	topic_4
0	fun facts	social media	true crime	mental health	pop culture
1	stories fun	talk things	cover art	think optimal	law attraction
2	kid kid	personal development	audio experience	daily audioblog	weight loss
3	parent time	answer question	help revise	audioblog blogcast	best sell
4	kid hear	tip trick	core components	daily audioblog blogcast	relate topics
5	time easy	make sure	revision subject	best blog	intermittent fast
6	kid kid hear joke	best friends	complicate revision	health fitness	million view
7	silly songs	real life	help rock	share stories	health relate
8	chompers kid remind parent	parcast network	complicate revision subject	help people	sell book
9	chompers kid remind	sure subscribe	components help	read best	best sell book
10	chompers kid	cutler media	components help rock	start day	ketogenic diet
11	make tooth time	production cutler	exams series	change world	healthy weight
12	make tooth	production cutler media	core components help	discuss things	author best
13	silly songs giggle brush	peak performance	core components help rock	break date	author best sell book
14	make tooth time easy	fantasy football	rock exams	best content	author best sell
15	brushers age	coach teach	help rock exams	search best	book new
16	brush way recommend tooth	personal growth	components help rock exams	bother search best	area expertise
17	silly songs giggle	performance coach	serial killers	bother search	body health
18	songs giggle	audio book	students break	read think	healthy lose
19	kid remind parent time	free amazon	students break complicate revision	read think optimal	million followers

Methods 3a: Word2Vec with LDA clusters

Implementation

- + After getting the 5 clusters from topic modeling we joined the words in each cluster into a string.
- + Similarity scores are calculated for each of the topic keywords for each of the show descriptions which are then used to decide which topic that specific podcast falls into.

Methodology

- + After regex transformation of the topic tokens followed by removal of stop words, we create the word vectors for the search query use the word2vec function from the pyspark package.
- + We then defined our own similarity function to compare the vector for the search query with that of each descriptions' vectors.
- + The topic which gives the highest similarity score for a description is the label for that description.

Method 3b: Google's Word2Vec and Manual Generated Topics

Why?

Topic clusters from LDA topic modelling is almost random

Implementation

- + We manually built a bag of words for each genre: Comedy, Health and fitness, News, Politics, Pop culture, Religion, Sports and True crime.
- + Applied `calculate_similarity` function to every podcasts to every genre's joined bag of words
- + Word2Vec model: `GoogleNews-vectors-negative300.bin.gz`



Method 3b: Google's Word2Vec and Manually Generated Topics (continued)

	Comedy	Health and fitness	News	Politics	Pop culture	Religion	Sports	True crime
0	comedy	health	news	politics	pop culture	religion	sports	true crime
1	absurd	fitness	technical	arguments	pop	satanism	NFL	cold cases
2	funny	mental health	sports news	juicy	campus	bible	football	mystery
3	funnier	weight	weekly news	political	adultery	christianity	soccer	crime scene
4	effing	healthy	social news	journalist	hip-hop	religions	premier league	murder
5	fuckbois	healthy weight	market	investigative	film	faith	la liga	unsolved murders
6	idiots	ketogenic	stock		tv shows	spirituality	champions league	drama
7	ranting	diet	stock market		music	god	liga	victims
8	improvisers	body type			marvel	consciousness	league	victim
9	entertain	intermittent fasting			disney	third eye	boxing	
10	satire	weight loss			movie		fighters	
11	funniest	body health			controversial			
12		skincare			life			
13		makeup			stories			
14		beauty			DC			
15		treatment			movies			
16		self-help						
17		selfcare						
18		meditation						
19		mental health						
20		body image						
21		nutrition						
22		astrology						
23		self-development						
24		healing						



Evaluation and Results

Evaluation

- + Main evaluation metric: Model's Result vs. Manual Annotation
- + Reason: inability to A/B test real users and no labelled data
- + Based on manual tags we then were able to calculate the accuracy score of the models used.


Results

- + Based on the 2 ways of modelling used we can clearly see that the best performing model is manually generated clusters combined with Google's Word2Vec: 84% accuracy
- + The other model used had a lower accuracy of 28%
- + The results are all based on the use of manual tagging of first 50 podcast episodes

Demonstration: R Shiny App

Podcast Recommendations | **Podcast Cluster Analysis**

Select Podcast
Serial Killers

 Refresh Recommendations

Instructions: Select a podcast to view 10 podcast recommendations. Hit the 'Refresh Button' to view new recommendations!

	Show	Show URL	Description	Cluster
1	Morning Cup Of Murder	Click to Open Spotify!	Ever wonder what murder took place on today in true crime history If so sit back and grab a cup of coffee as you enjoy your daily dose of TC goodness Your host Korina Biemesderfer guides you through history with tales of murder abduction serial killers crimes of passion cults and more in this short form daily true crime podcast Support this podcast https://anchor.fm/morningcupofmurdersupport	Pop culture True crime Health and fitness
2	Unexplained Mysteries	Click to Open Spotify!	We dont know answers too many questions But in this podcast we dont take we dont know for an answer Every Thursday Unexplained Mysteries investigates the greatest mysteries of history and life on earth because the answer we dont know is always the scariest Unexplained Mysteries is part of the Parcast Network and is a Cutler Media production	Pop culture True crime Health and fitness
3	Con Artists	Click to Open Spotify!	You trust them with your life They seem like a friend Family even Anyone can fall victim to a con and many have What type of person intentionally tries to deceive manipulate and eventually destroy someone with their web of lies This Parcast Original peeks behind the masks of the most notorious Con Artists and explores how far someone will go in order to gain money power and respect New episodes are released every Wednesday	Pop culture True crime Health and fitness
4	The Outlander Podcast	Click to Open Spotify!	The Outlander After Show recaps reviews and discusses episodes of Starzs Outlander Show Summary After serving as a British Army nurse in World War II Claire Randall is enjoying a second honeymoon in Scotland with husband Frank an MI6 officer looking forward to a new career as an Oxford historian Suddenly Claire is transported to 1743 and into a mysterious world where her freedom and life are threatened To survive she marries Jamie Fraser a strapping Scots warrior with a	Pop culture True crime Health and fitness

Findings and Limitation

Findings

- + The best method was using manually generated genres' bags of words and Google's Word2Vec model
- + Podcasts tend to fall in multiple genres as opposed to one. Thus, we used the top 3 genres' similarity scores to tag each podcast.

Limitations

- + Lack of label made it challenging to evaluate our model performance
- + Very challenging to implement RoBERTa without having a labeled dataset
- + A/B testing real users could not be done as an evaluation in the limited timeframe
- + The original idea of using "transcript" column cannot be performed due to transcript length. Therefore, "show_description" column was used to perform the analysis.
- + The limitation of Spotify's Podcast API to obtain more metadata



Conclusion and Recommendation

Conclusion

- + Google's Word2Vec model still outperforms our custom-built Word2Vec model
- + Good first step, the project needs more work, as outlined in the limitations section
- + Spotify's podcast catalog is rather limited in nature

Recommendation

- + To perform analysis on the full transcript as opposed to just the show description. We needed to use the "show_description" due to local memory limitations.
- + As more observations are available, we presume building a customized Word2Vec would be better than using Google Word2Vec model
- + To perform other methods such as BERT and Word2Vec combined with K-means clustering