Entale

Podcast Recommendation

- The Challenge: Develop a podcast recommendation algorithm

User profile and/or recent episode they listened to → Recommender → Podcast episodes the user would enjoy

... but also expand their horizons!

- The Challenge: Develop a podcast recommendation algorithm

User profile and/or recent episode they listened to → Recommender → Podcast episodes the user would enjoy

... but also expand their horizons!

- The Data

Episodes:
- Title, show, episode description
- Transcription
- Named entities
- Category

Users:
- Episodes listened (small overlap with episode data) (1 year)
- Month of each listen
- Some show subscriptions (5 years)

Outline

- High level overview of approaches
- Different ways of representing and comparing podcasts
    - Networks
    - Topic models
    - Pre-trained embeddings
- The Rabbit-hole recommender
- Evaluation and Limitations
- Future Work

Approach Overview

Approach 1: “Distance” based recommenders

Embed episodes into a graph or vector space

This lets us compare and cluster them mathematically

Approach 1: “Distance” based recommenders

Different embeddings encode different notions similarity between episodes → Embed episodes into a graph or vector space ← This lets us compare and cluster them mathematically

Approach 1: "Distance" based recommenders

Embed episodes into a graph or vector space

Consider a user's most recent listen

Recommend another episode that is "nearby"

Approach 1: “Distance” based recommenders

Embed episodes into a graph or vector space

Consider a user’s most recent listen

Represent the *user* in the same “space” - e.g. average their listen history

Recommend another episode that is “nearby”

Approach 2: Modern “Collaborative Filtering”

User listens as indicators of user preferences

Representations of both users and podcasts which jointly explain observed behaviour

Use these to make recommendations about new podcasts for the users

Approach 2: Modern "Collaborative Filtering"

User listens as indicators of user preferences → Representations of both users and podcasts which jointly explain observed behaviour → Use these to make recommendations about new podcasts for the users

Techniques in the pipeline:

Matrix factorisation

Convolutional matrix factorisation

Neural collaborative filtering

Approach 3: The Rabbit Hole

Identify **specific entities** discussed in a podcast

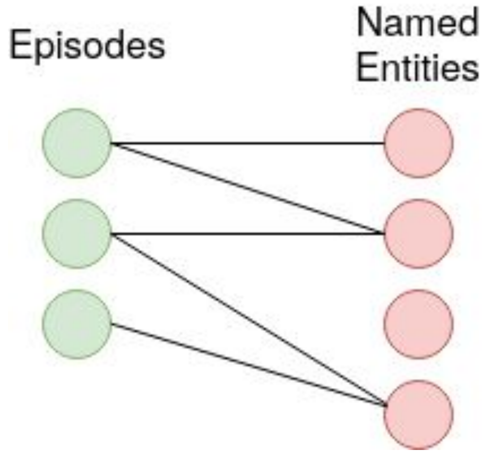
Identify “candidates” for recommendation which focus in one of these entities

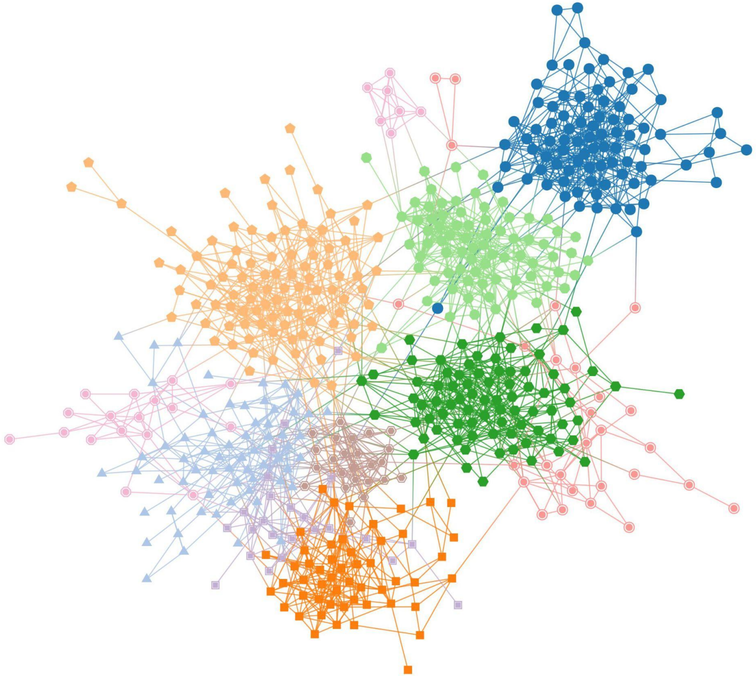
Sort these candidates **according “similarity” with user preferences**

Episode Representation and Similarity Modelling

Network-based recommendation

One step neighbour → Community structure





Network-based recommendation

Hierarchical Stochastic Block Models (hSBMs)

- **nonparametric**: infer the number of topics

- **hierarchical** clustering (topics): clustering different resolution -> ***serendipity***

- recommendations based on community membership, equivalently, **topic distribution**


Wimbledon GitHub Cloud US Open Andy Murray
El Paso Mercedes Tick-Tock Father's Day Lexus Calgary Ontario
Europe American Germany Italy Earth China Christian London France
Man University Burger King Chad Stitch Northern Ireland Harry Potter Chanel Boulder
Medicines West Wing Harlem Sawbones
Netsuite U2 Fitbit The Bachelor Escape Pod Day Love Escape Creative
Kate Baker Oslo Taipei radio Dot CrossFit
Saturn Venus Moon NASA Mars Sun Jupiter Mercury Galaxy iHeartRadio
Federal Reserve Vogue White Central Bank S&P 500 House Warren Buffett
Play Batman Ottawa GameStop Seas Kyoto Manila
Beverly Hills County Kickstarter Orange County Kenya Bravo Andy Cohen Bali Porsche
New York City Washington California New York Christmas Canada
Game of Thrones Disney Oscar Anchor Fox NBC
San Diego World War Two Nazis Maryland Muslim World War II
Jim Cramer iPad Pro Robin Hood CNBC IOS AT&T Tim Cook Apple TV MacBook Pro
Greenland Agatha Christie Kanye West World Health Organization Arctic Antarctica Albuquerque Kim Kardashian Amsterdam
Comedy Central Iowa George Floyd Trevor Noah CBS New Hampshire Minneapolis Central Mississippi
Himalaya Theater York City Travis Vengroff Ear Hustle Hudson Institute
Financial Times Tanzania American Academy of Pediatrics Ripple Krypton Andy Earl Parkland
Charles Dickens Columbus, Ohio Nikes Global Newsweek Mel Gibson ASMR Mansion
South 9/11 Russia City United Kingdom Ukraine Iran
Vikings Patriots Rams Ravens Giants Eagles Saints Tom Brady Cowboys
White House Congress Democrats Senate Donald Trump Republicans
Jurassic Park Spider-Man Avengers Rogue Hulu
Portland, Oregon instagram Alcove Marshall Frozen google Exeter Harvard Medical School
Jim Clemente Ben Affleck Nancy A Grace David Fincher Croatia Eden Sundance 19 Wonder Woman
Sky Toyota Ireland Russian Central Bank of Ireland
New Yorker Nobel Prize Island WNYC Studios Austria Caribbean Network Cuba
Albert Einstein org Box Neil deGrasse Tyson
Indiana Jones Malcolm Gladwell Farmland Valley Art Channel Tim Harford Adobe Prague Queen
Christmas Eve Santa Claus East Park Max Cutler Ron Cutler Charles Manson
Tulsa Justin Bieber Tom Cruise Community Joe Rogan skillshare Illuminati
Molly Bloom The Matrix Academy Retreat Paradigm music
Han Solo Buddhism FaceTime X Factor Luke Skywalker Canon George Lucas School
WrestleMania Raw SWAT Vince McMahon WWE Fallout NXT Smackdown
Old Testament Heaven Thomas Aquinas atlas obscura SiriusXM New Testament
Black Friday Best Buy Peloton Black Oracle Getty Images
Affordable Care Act Breitbart John Roberts Paypal Department of Labor Department of Health and Human Services
Premier League Manchester United Liverpool Society Arsenal
Nia Kent French Revolution North Africa Crusades University College London

Using Topic Models for recommendations

- **Input**: a collection of documents (i.e. episode named-entities, episode transcriptions, episode descriptions)
- **Models**: Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) implemented with **tomotopy**
- **Output**: Some topics that are given by a distribution of the words based on how likely they occur in the topic

Using Topic Models for recommendations

- **Example output of a topic model:**
  - **Topic A:** Trump (2.9%), President (1.5%), Election (1.2%), Biden (1.1%), ..
  - **Topic B:** Football (1.1%), Season (0.7%), Player (07%), Manchester United (0.6%), ….
  - …
- **Understanding of the topics requires human interpretation:**
  - **Topic A:** “US Politics”
  - **Topic B:** “European Football”
  - …
- **We can also obtain the topic distribution for each episode:**
  - Episode 1: Topic A (10%), Topic B (0.5%), ….
  - ...

Using Topic Models for recommendations

Example recommendation:

0. Choose an embedding and what episode metadata to use

- (LDA with k=20 topics on episode descriptions)

1. What has the user listened to?

- **Barca Blanugranes**: Chatting about redemption for Coutinho, Griezmann...
- **Between the Links**: Jon Jones Vacates, Future of 205, Frankie Edgar vs. ...
- **MMA Fighting**: Luke Rockhold explains why he's returning to fighting...
- **Barca Blanugranes**: A chat about Ronald Koeman's time at Everton...

Using Topic Models for recommendations

Example recommendation:

2. Combine the episode embeddings

- Used an exponential weighted average (and normalise)
  - Newer episodes get a larger weight

3. Compute the similarity between the combined episode embedding to other episodes

- Used the Wasserstein distance

Using Topic Models for recommendations

Example recommendation:

1. **The Ornstein & Chapman Podcast**: Chelsea sack Lampard - The Inside Story (WD = 0.002966)
2. **Move the Sticks**: Draft Scenarios for Teams Picking 3-5… (WD = 0.002969)
3. **The Ringer NBA Show**: The Trade Deadline Extravaganza… (WD = 0.002987)
4. **Why Always Us? A show about Manchester City**: Take The Shot! (WD = 0.003089)
5. **The Odd Couple**: Bucs Taking a Big Risk with Antonio Brown (WD = 0.003094)

Using Topic Models for recommendations

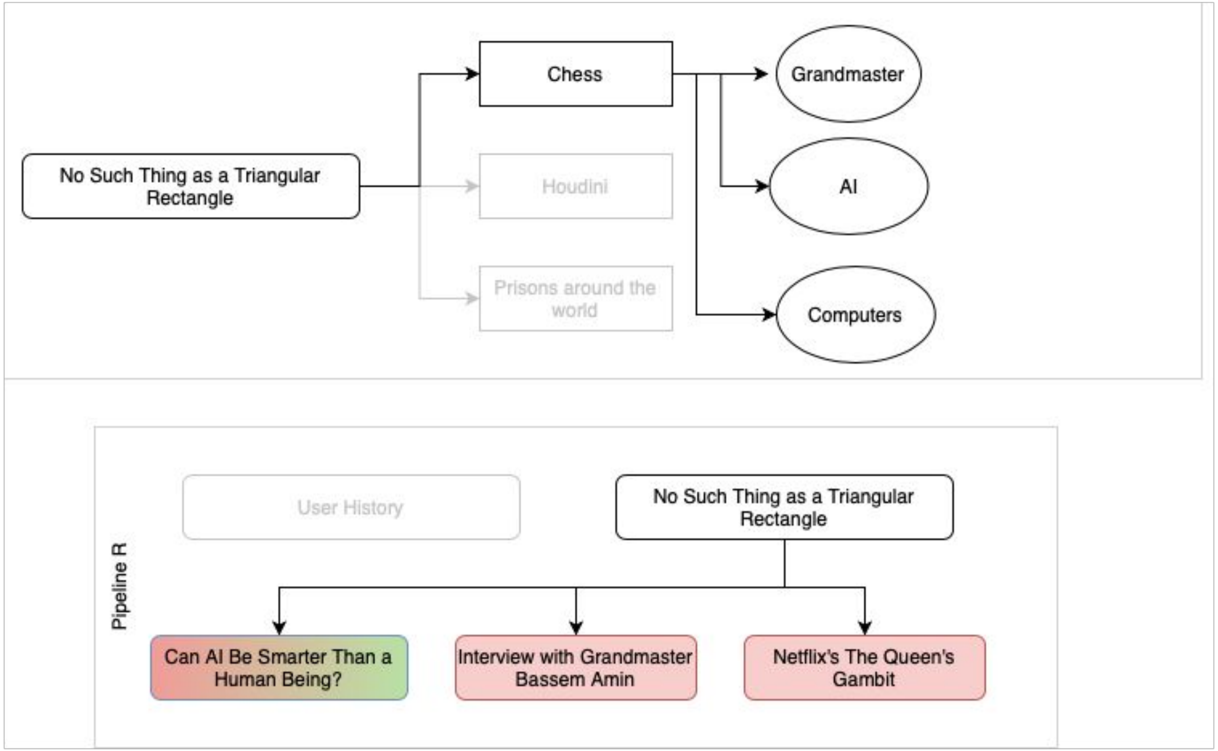
Flexible approach:

- Options to use different embeddings
- Various ways to combine the user listened episodes
- Various ways to compare similarity
- Many possible extensions
  - Clustering on the episode embedding
  - Dimension reduction on the episode embeddings

Going Down a Rabbit Hole

- Recommend podcasts which pick up on something discussed in what you were just listening to

Going Down a Rabbit Hole



Limitations

- **Time** is an important missing feature for inferring user intention/preference (duration of listen + date and time) - ***a listener might want to listen to different genre depending on the time of the day***
- **Transcripts** for all the episodes
- **Disambiguation** of **named entities** vs. lack of information on **frequency** in transcript
- **Niche** areas **themselves result in a spiral**, need a way to **get people out** of that

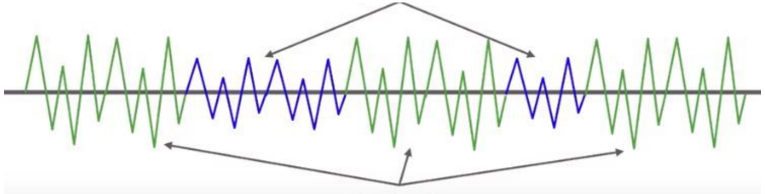
Future Work (Directions)

Future Work: Other Features

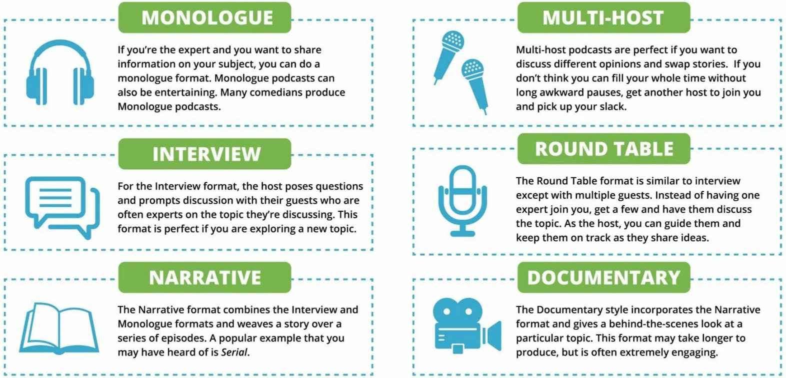
Sentiment

Tone

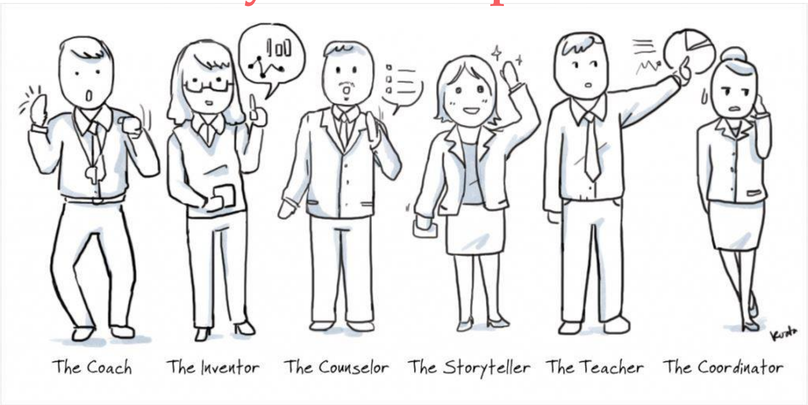
Speaker Diarization



Format



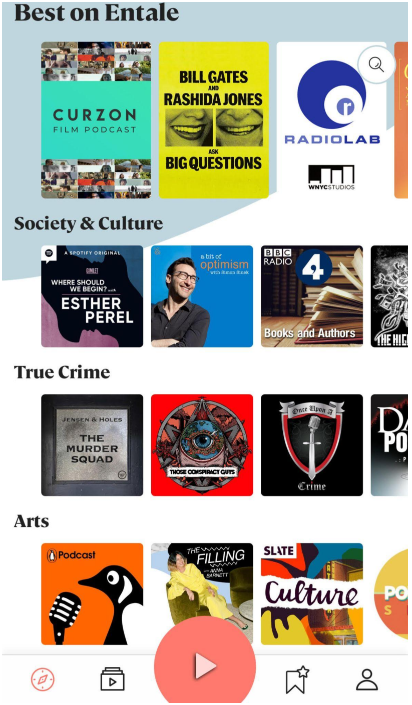
Style of the Speaker



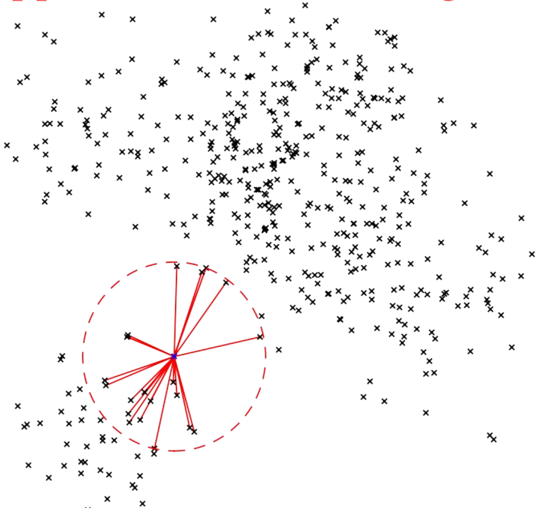
Future Work (Directions)

Ranking in a category

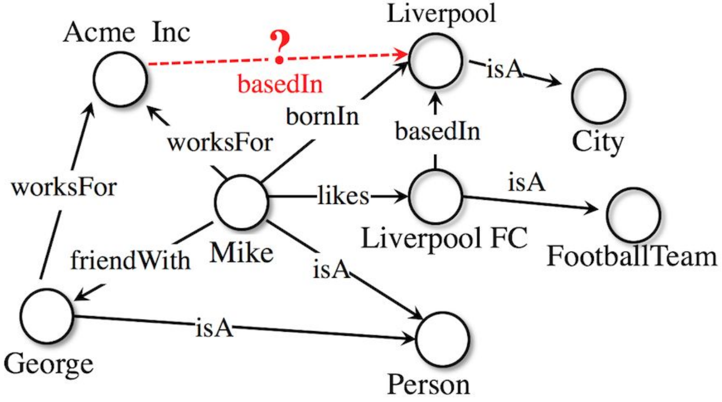
Category Ranking



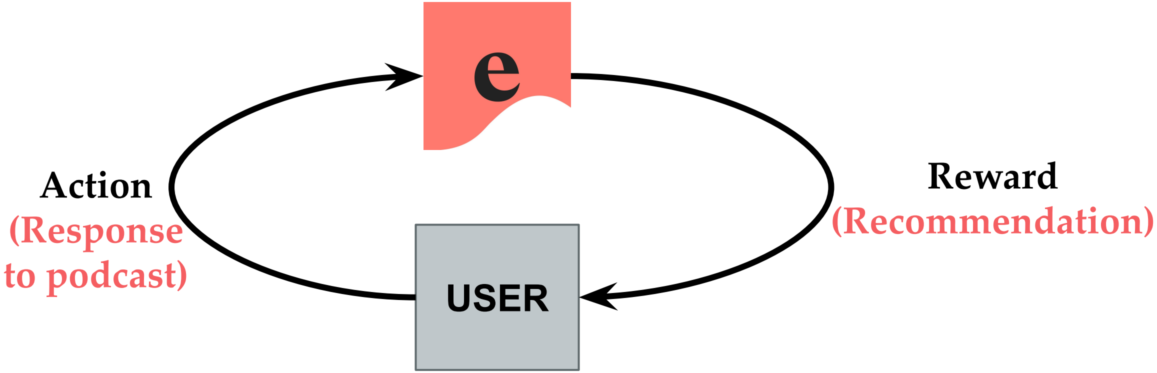
Approximate Nearest Neighbors



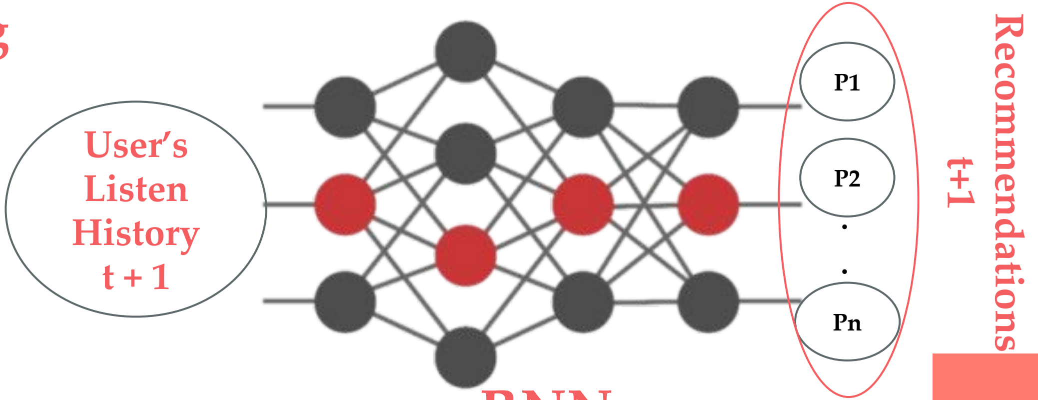
Knowledge Graph -- Embedding



Future Work (Directions)



Reinforcement Learning



RNN

Future Work: Evaluation

Strategy 1: *Comparing different Recommendation Systems*

Held out episode E_t → **Score** for E_t from different Recommender systems → **Score** for ***random episodes*** from different Recommender systems → **Compare the statistics**

Strategy 2: *Evaluating a Recommendation system*

At time ***t:*** episodes till ***t-1*** → **Score** for E_t from the Recommender system → Check the scores of E_t for all ***t***

Thanks for listening!

The Entale DSG Team:

- Ryan Chan
- Jayesh Choudhari
- John Fitzgerald
- Jev Gamper
- Erfan Loghmani
- Jamie McGowan
- Vanessa Pope
- Ilan Price
- Kirstin Roster
- Lizhi Zhang

Special thanks to Olek and the Entale team for their consistent support throughout the challenge

References

- Speaker presenter style -- https://www.facebook.com/melaniewoodspeakingstyles/posts/2370882959673070/
- Tone: https://blog.inkforall.com/tone-words
- Sentiment: https://medium.com/@sam_hames/four-reasons-sentiment-analysis-is-misinterpreted-4d9bb59b41b9
- Format: https://www.singlegrain.com/podcast/podcast-trends-2021/
- Ranking: https://www.topbots.com/netflix-movie-recommender-system-rework/
- ANN: https://sdjournal.org/annoy-approximate-nearest-neighbors-in-cpython/
- Knowledge Graph: https://docs.ampligraph.org/en/1.0.3/
- RNN: https://www.edureka.co/blog/recurrent-neural-networks/