Measuring the Similarity of Song Artists using Topic Modelling

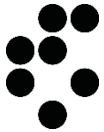
Erik Calcina, Erik Novak

Jožef Stefan Institute

Ljubljana, Slovenia

Introduction

- Finding suitable songs or artists from a large selection of songs
- An aspect to consider can be the song topic interpreted as
  - An emotion
  - An event
  - A message
- Topic modeling-based approach for measuring the similarity of the music artists based only on their song lyrics



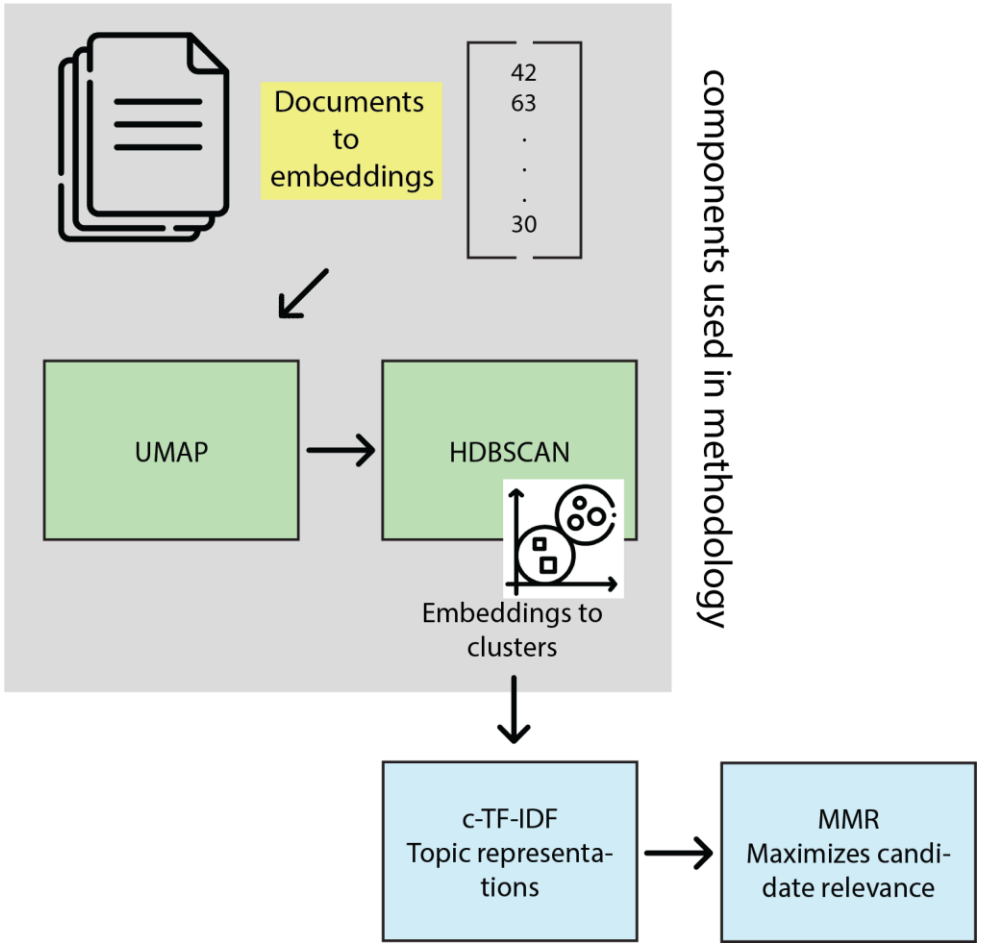
Outline

- Methodology
  - BERTopic
  - Measuring Artists' Similarity
- Dataset
- Results
- Discussion
- Conclusion

Methodology BERTopic

- Language model creates vectors from lyrics
- Reducing dimensionality with UMAP
- Topic clusters are created using HDBSCAN

- BERTopic also creates topic word desciptions which is not used in our artists similarity measuring

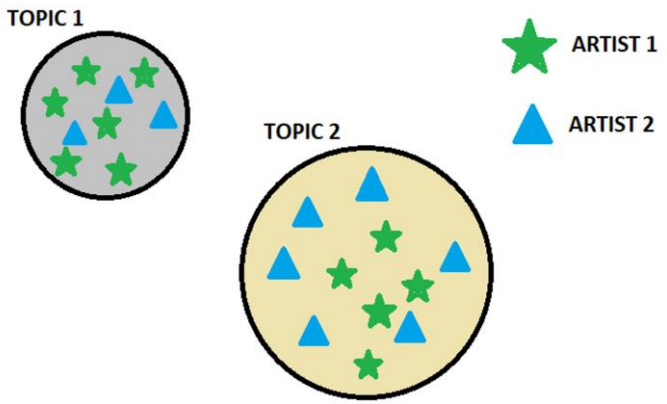


Measuring Artists' Similarity

- For each topic we count the songs that corresponds to a particular artist

- To ensure that songs were not assigned to a cluster by coincidence we set a threshold to 5 songs, otherwise we remove the artist from the topic

- Finally, for each pair of artists we count the number of common topics



Dataset

- Each song consists of its **name**, **release year**, **artist**, **genre** and **lyrics**
- Original dataset consists of 218,210

Artist	genre	# songs	avg. length
black-sabbath	Rock	160	184
bon-jovi	Rock	320	266
dio	Rock	127	203
aerosmith	Rock	208	226
ac-dc	Rock	171	193
coldplay	Rock	138	174
50-cent	Hip-Hop	318	502
2pac	Hip-Hop	259	648
eminem	Hip-Hop	369	640
black-eyed-peas	Hip-Hop	119	463
celine-dion	Pop	182	230
britney-spears	Pop	225	313
frank-sinatra	Jazz	356	133
ella-fitzgerald	Jazz	503	156
Together	-	3,455	319

Generated Topics
Results

- Experiment generated 215 topics
- 107 have at least one artists with more than 5 songs
- Artists with a larger number of songs are spread over several topic clusters than those with less songs

Artist	topics	#avg. songs
black-sabbath	6	5
bon-jovi	10	6
dio	4	7
aerosmith	9	6
ac-dc	7	5
coldplay	2	5
50-cent	17	9
2pac	13	9
eminem	18	9
black-eyed-peas	3	12
celine-dion	8	6
britney-spears	12	6
frank-sinatra	16	8
ella-fitzgerald	28	8

Artists‘ Similarity (absolute count)

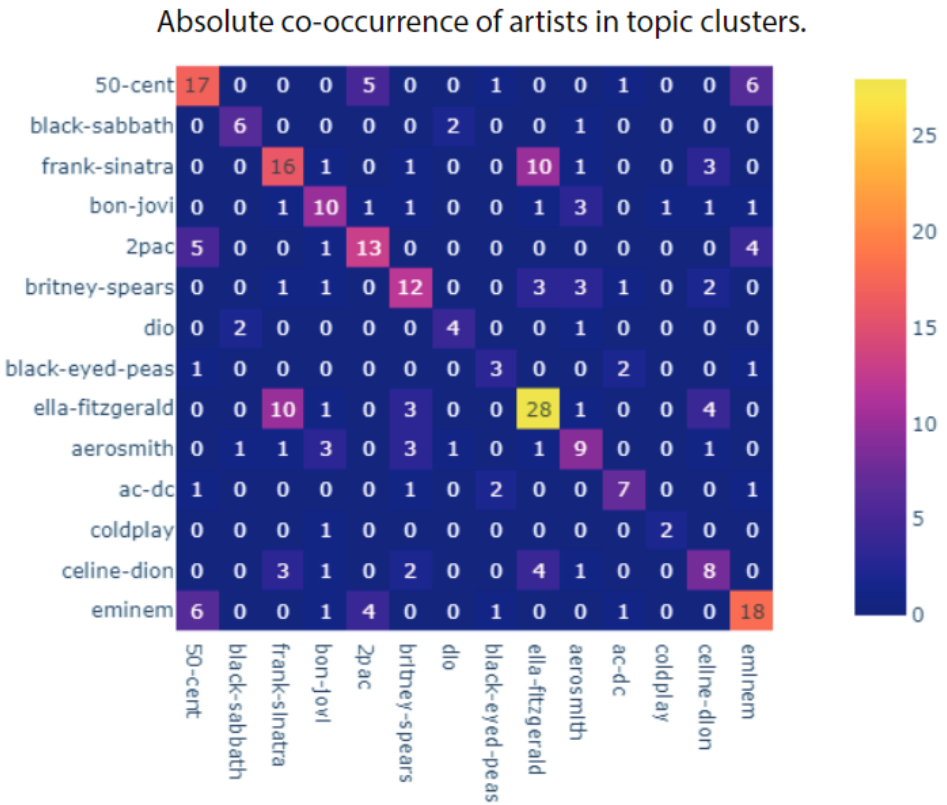
Results

50-cent	17	0	0	0	5	0	0	1	0	0	1	0	0	6

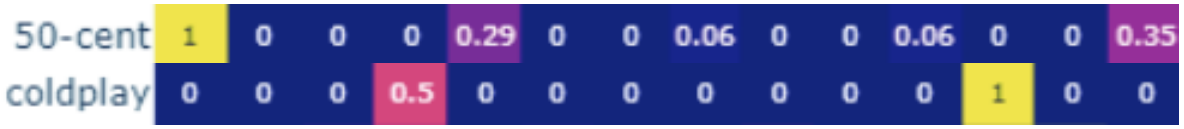
Rows depict the number of common topics with other artists

Example: 50-cent (17 topics)

- 5 with 2pac
- 1 with black-eyed-peas
- 1 with ac-dc
- 6 with eminem



Artists‘ Similarity (relative count)
Results

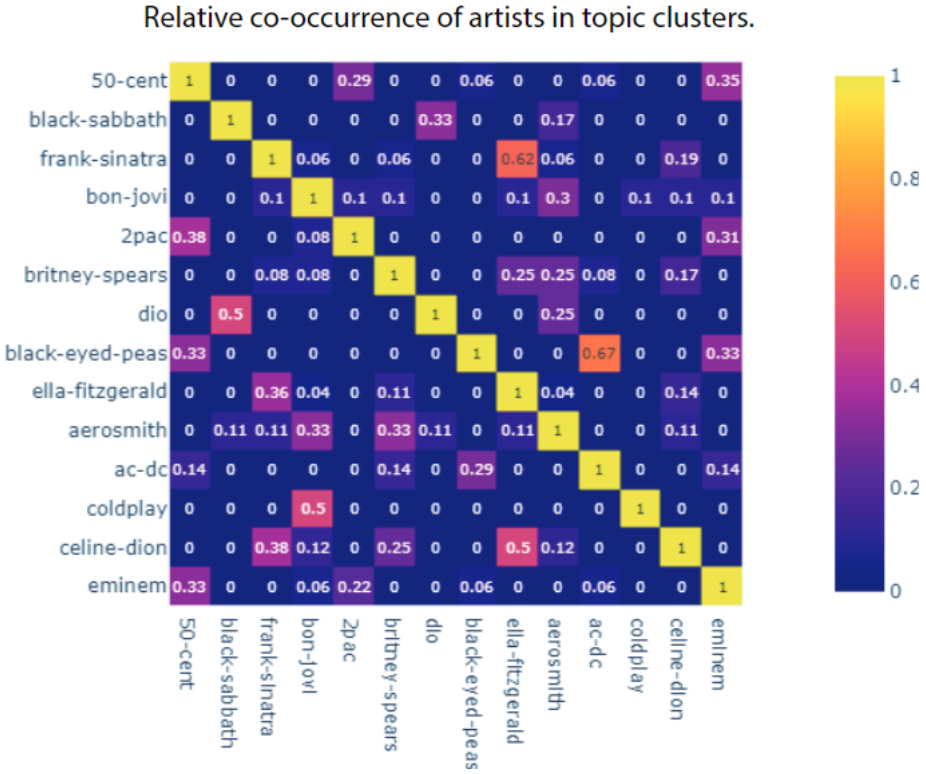


Calculated using the following equation:

$$\text{sim}(a, b) = \frac{|A \cap B|}{|A|}$$

where A is the set of topics of artist a, and B is the set of topics of artist b

Artists with smaller number of topics can result in higher similarity with other artists



Discussion

- Language Model Limitations
  - Language model max. sequence length is 384 tokens
  - However, it might capture the majority because of the song's repeated text

- Clustering Algorithm Selection
  - HDBSCAN can label songs which do not fall into any topic clusters as outliers
  - Downside is when the majority of songs are labeld as outliers

Conclusion

- We present a way to measure similarity between artists using topic modeling
- We clustered lyrics and compared artists based on generated topic clusters
- The results have shown that the approach finds similar artists.

Future Work

1. Apply the methodology on a larger dataset
2. Use all of the cluster information (including topic word decription)
3. Reduce the lyrics length by filtering the repeated chorus to take into account the language model‘s input limit