Lyric/Genre Analysis

Comanici Mario, Salhofer Eileen, Sterner Thomas, **Group04**

https://github.com/thenextmz/AdvancedInformationRetrieval_WS22

Motivation/Research Question

- Genre defined by literary technique, tone, **content**
- How much information you can get from text regarding genre
- Does analysing text lead to good genre classification?
- Our dataset has no audio == we rely purely on text
- Goal: classify genre by text analysis and ML
    - + General genre analysis by text metrics eg. Wordcloud, TF, n-grams…

Data + Methods

- *Kaggle dataset: Lyrics from 79 Genres, 4239 Artists, 379.893 Songs
- Original data:
  - Split into artists and song csv
  - Preprocessing step 1: unify in one file + break down size (runtime)
- Preprocessing step 2:
  - stopword removal
  - punctuation removal
  - number removal
  - lemmatization
  - tokenization
  - remove “non-relevant” words (chorus, verse, …)
- Analysing general word distribution per genres (#songs, avg. #word, avg. #uniquewords)

*https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres?select=lyrics-data.csv

Data + Methods

- Analysing Term frequencies and n_grams
  - Top terms per genre
  - n_grams one directional -> “We will rock you” => [(we, will)(will,rock)(rock,you)]
    - See word sequences and repetitions
    - most common ones per genre
    - percentage of “unique” two-word phrases -> (la, la)
    - percentage of phrase overlap in genres
- Visualization of terms as word-cloud
- Basic sentiment analysis + visualisation as plot
- Classification of lyrics
  - unsupervised
    - k-means clustering
      - td-idf model
  - supervised
    - support vector classifier
      - w2v model

Results - [Romantico, Rap, Gospel/Religioso, Country, Heavy Metal, Hardcore]

General Features

Genre	#Songs	#Avg.Words	#Avg.Unique	Unique Words
Romantico	16516	115	59	51%
Rap	16820	284	166	58%
Gospel/Religioso	5888	90	51	56%
Country	9749	103	60	58%
Heavy Metal	19562	101	60	60%
Hardcore	5002	94	54	57%

Results - [Romantico, Rap, Gospel/Religioso, Country, Heavy Metal, Hardcore]

Top 4 Terms per genre

Genre	#1	#2	#3	#4
Romantico	love	know	like	oh
Rap	like	ni**a	get	got
Gospel/Religioso	love	know	oh	like
Country	love	bill	monroe	know
Heavy Metal	time	know	one	see
Hardcore	know	like	never	time

Results - [Romantico, Rap, Gospel/Religioso, Country, Heavy Metal, Hardcore]

Top 4 word sequences

Genre	#1	#2	#3	#4
Romantico	oh oh	la la	love love	let go
Rap	yeah yeah	oh oh	feel like	ain’t got
Gospel/Religioso	oh oh	let go	holy holy	love love
Country	bill monroe	i’ve got	oh oh	let go
Heavy Metal	let go	oh oh	yeah yeah	feel like
Hardcore	oh oh	let go	i’ve got	feel like

Results - [Romantico, Rap, Gospel/Religioso, Country, Heavy Metal, Hardcore]

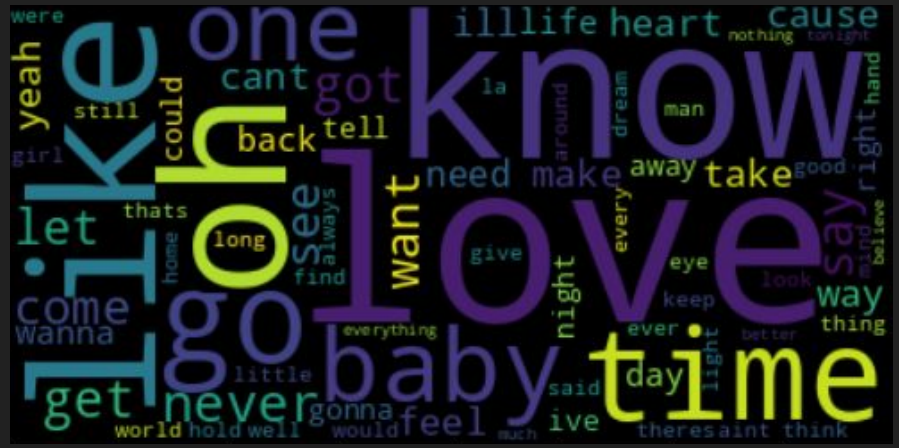
N-Gram - Outstanding Genre Overlap

	Romantico	Rap	Gospel/Religioso	Country	Heavy Metal	Hardcore
Romantico		9%	11%	16%	13%	11%
Rap	9%		4%	7%	8%	5%
Gospel/Religioso	11%	4%		12%	9%	11%
Country	16%	7%	12%		11%	11%
Heavy Metal	13%	8%	9%	11%		9%
Hardcore	11%	5%	11%	11%	9%	

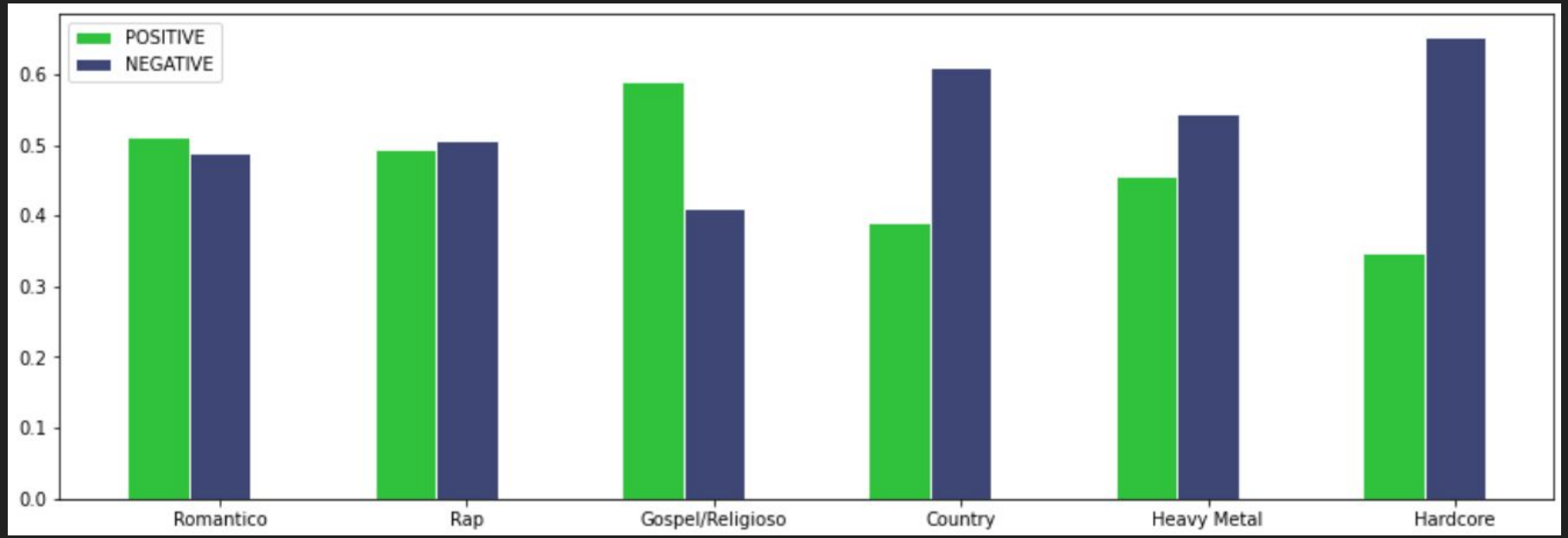
Word Clouds

Country

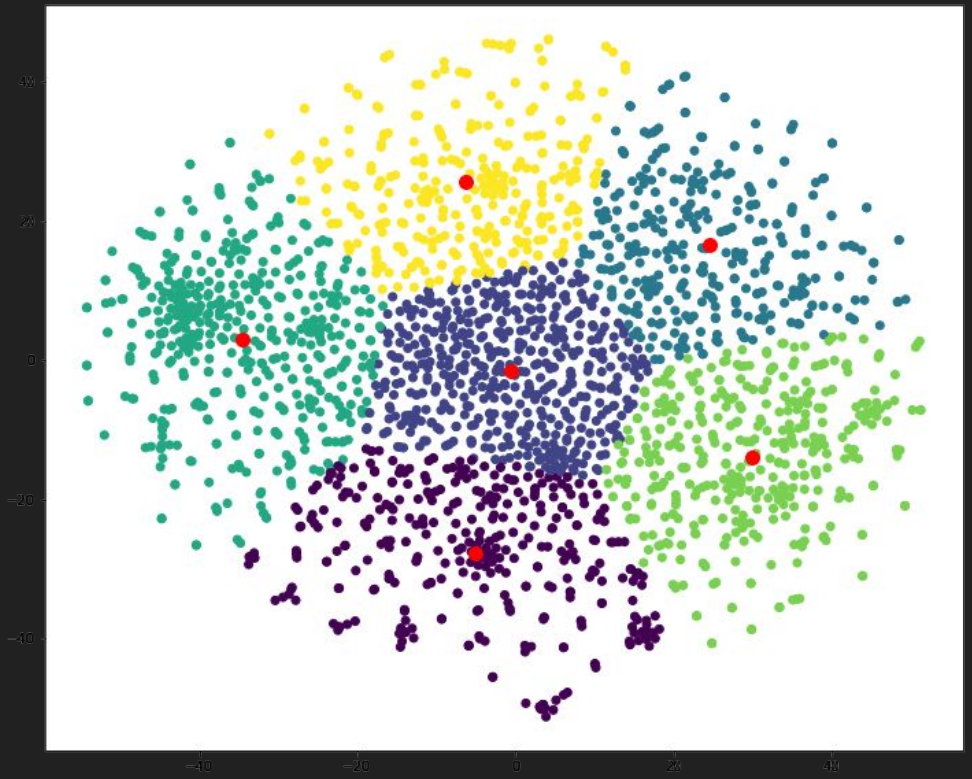
Romantico



Sentiment Analysis



k-means clustering



Classification of lyrics

accuracy	k-means clustering	super vector classifier
Romantico	29.25%	51.76%
Gospel/Religioso	40.75%	51.28%
Rap	75.50%	83.72%
Hardcore	21.25%	50.00%
Heavy Metal	28.75%	35.80%
Country	26.75%	61.84%
Average	37.04%	56.04%

Conclusion

- Genre analysis on text not best way
- Prediction not optimal as many genres overlap in text/word frequency
- More distinct data from beginning would also help
- For better classification -> use audio + text analysis
- Rap stands out the most, best results (unique choice of words)