732A92 - Text Mining

Lyrical cleverness across musical genres

Individual Project - João Patrício (joapa307)

Introduction

In music composition, the diversity composition style and arrangement lead to the categorization of musical pieces into different genres. Songs in the same genre typically share similar musicality or arrangement conventions [1]. Genres are in fact a well established classification model in mainstream music culture. It is not uncommon to frequent music listeners to express their individual preferences in the form of what genre they like best. But what does genre tells us about the songwriting? Is it the case that genres can be associated to more complex storytelling?

Informally, text cleverness refers to the degree of intricacy some text has. There has been scarce research on this topic and therefore not a lot of conventions have been established as to what criteria should one look for to calculate such attribute. However, two distinct approaches have been used for such end, the first relies on simple measures like word difficulty and sentence length, the second one is bound to textual cohesion or in other words, the relatedness of distinct components of the text [2][3].

In the following sections an empirical approach is described using text mining techniques to measure lyrical complexity metrics. Subsequently, an analysis of their variation across mainstream music genres is conducted, which will then be evaluated using a rank correlation coefficient.

Method

All the data used in this project was fetched from the Musixmatch API[1]. Musixmatch is a platform that contains millions of music lyrics as well as associated metadata, allowing to filter songs according to criteria like, genre, language, and being or not instrumental. Furthermore, Musixmatch's partnership with large music streaming services also allows it to assign a popularity rating to songs. This functionality is helpful to make sure that a representative sample of songs is chosen. The code for API communication and data handling was written in Python, with heavy usage of the requests[2] and NLTK[3] libraries.

A total of 12 genres (Blues, Classical, Country, Eletronic, Jazz, Latin, Pop, R&B/Soul, Hip-Hop/Rap, Alternative, Rock and Reggae) were considered in this experiment. The 2000 most popular song lyrics (according to Musicmatch's metrics) for each of them was used.

Preprocessing

The used API provides song lyrics as raw text. Typically they are formatted such that each sung verse is separated by a newline. Punctuation marks of any kind are non existent, apart from the occasional comma, often used to demark small pauses in the singing. Another observed artifact in the lyrics are pitch/duration marks. When the singer extends a syllable or changes pitch in a word, these are reflected in the text as repeating a vowel several times or putting between hyphens a subsection of the word (see last line - table 1).

Raw Text	RegExp Word tokenizer	White Space tokenizer
I know it's hard sometimes. Yeah, I think about the end just way too much But it's fun to fantasize On my enemies who wouldn't wish who I was But it's fun to fantasize Oh, oh, oh, oh Oh, oh, oh, oh I'm fallin' so I'm taking my time on my ri-i-i-ide	'i', 'know', 'it', 's', 'hard', 'sometimes', 'yeah', 'i', 'think', 'about', 'the', 'end', 'just', 'way', 'too', 'much', 'but', 'it', 's', 'fun', 'to', 'fantasize', 'on', 'my', 'enemies', 'who', 'wouldn', 't', 'wish', 'who', 'i', 'was', 'but', 'it', 's', 'fun', 'to', 'fantasize', 'oh', 'oh', 'oh', 'oh', 'oh', 'oh', 'oh', 'i', 'm', 'fallin', 'so', 'i', 'm', 'taking', 'my', 'time', 'on', 'my', 'ri', 'i', 'i', 'ide'	'i', 'know', 'it', "'s", 'hard', 'sometimes', '.', 'yeah', ',', 'i', 'think', 'about', 'the', 'end', 'just', 'way', 'too', 'much', 'but', 'it', "'s", 'fun', 'to', 'fantasize', 'on', 'my', 'enemies', 'who', 'would', "n't", 'wish', 'who', 'i', 'was', 'but', 'it', "'s", 'fun', 'to', 'fantasize', '.', 'oh', ',', 'oh', ',', 'oh', ',', 'oh', 'oh', ',', 'oh', ',', 'oh', ',', 'oh', '.', 'i', "'m", 'fallin', "'", 'so', 'i', "'m", 'taking', 'my', 'time', 'on', 'my', 'ri-i-i-ide'

Table 1 - A segment of the lyrics from *Ride*, a song by "Twenty One Pilots", and corresponding possible tokenizations.

In order to facilitate metric calculations the raw text was preprocessed into a more convenient format. Preprocessing consisted in normalization and tokenization. After tokenization an additional token clean up was performed to remove single punctuation marks or uninformative characters that could affect the final calculations.

For tokenizing, two approaches were tested. One was NLTK's regular expression tokenizer, which splits tokens by any non-alphanumeric character found. The second, is a

[1] http://developer.musixmatch.com
[2] http://docs.python-requests.org/en/master/
[3] http://www.nltk.org/

whitespace tokenizer, which, despite the naming, uses both white space and the apostrophe mark to split tokens. Table 1 contains an illustrative example of applying either tokenization to raw lyrics. Using the regular expression approach we get a much ‘cleaner’ result when compared to the “white space” tokenizer, which has several gibberish tokens consisting solely of punctuation mark. However, it’s more interesting to analyse how both of these tokenizers handle lyrical artifacts as is the case of the last word of the raw lyrics. “Ri-i-i-ide” is one of the aforementioned cases where the extended pronunciation time of the word while singing is represented by repeated vowels in the word. For the regular expression tokenizer, “ri-i-i-ide” is split into 4 different tokens: 'ri', 'i', 'i', 'ide'. An event that by itself is innocuous, but repeated several times may lead to very different results in measures of vocabulary size and average word length. The white space tokenizer instead generates the single token ‘ri-i-i-ide’ which even though can still bear some significance in the final calculations, it will prove less impactful, which is the reason why this strategy was prefered and further preprocessing was performed to remove the unwanted tokens.

As a final note, even though it’s usage for the final results was considered, performing stemming was ultimately deemed non contributive as it displayed similar but more diluted results for cross-genre variation.

Calculating ‘cleverness’

As suggested in the introduction section, two distinct approaches are used in the literature to measure textual complexity. A first is bound to simple statistical measures. The second, relates to textual cohesion and complexity. A concrete description of the criteria used in each follows.

Simple Statistical attributes

While parsing the collected data for all genres, three variables were measured for each:

- The Vocabulary size - How many different words, in total the songs of a particular genre employ.
- Average word length - The mean of the number of letters of each word has in all songs of a genre.
- Vocabulary diversity - A measure given by the vocabulary size divided by the total amount of words used.

The reasoning behind the choice of these three was their potential meaning in the given context. A vocabulary size shows how rich the lexic of a genre is, the average word length may provides a hint on the complexity of the vocabulary and the vocabulary diversity rewards non-repetition.

For each genre the three values were calculated and then normalised and averaged together to form a final ‘complexity’ measure.

Lyrical cohesion

While informative, the previously described metrics can be considered a naïve approach to complexity measure. They fail to capture subtleties of text construction on a deeper level. There’s been a number of measures shown to have good correlation rate to cohesion and coherence levels of the text [2]. However, implementing these on regular text proved significantly different than in a lyrical context. Songs lyrics do not obey the same rigid

composition that is normally observed. That is to say, it is not always the case that the syntactic structure rules are followed. Additionally, because no punctuation is present in the lyrics, there is no practical sentence demarcation existent in the raw lyrics. This makes the construction of a syntax tree or using flesh-kinkaid scoring variants [5], virtually impossible. However, not all metrics require a regular syntax. [4] proposes a number of concrete measures that aim to measure cohesion that don't need analyse syntatic trees:

- Lexical overlap - How often repeated lemmas occur in subsequent verses.
- Givenness - The proportion of single occurrence lemmas.
- Type–token ratio - The variation of parts of speech types divided by the total number of words.

Results

Simple Statistical attributes

As mentioned in the method section, three different components were used to measure the complexity of lyrics across genres. Below a plot for the vocabulary size of the tested genres:

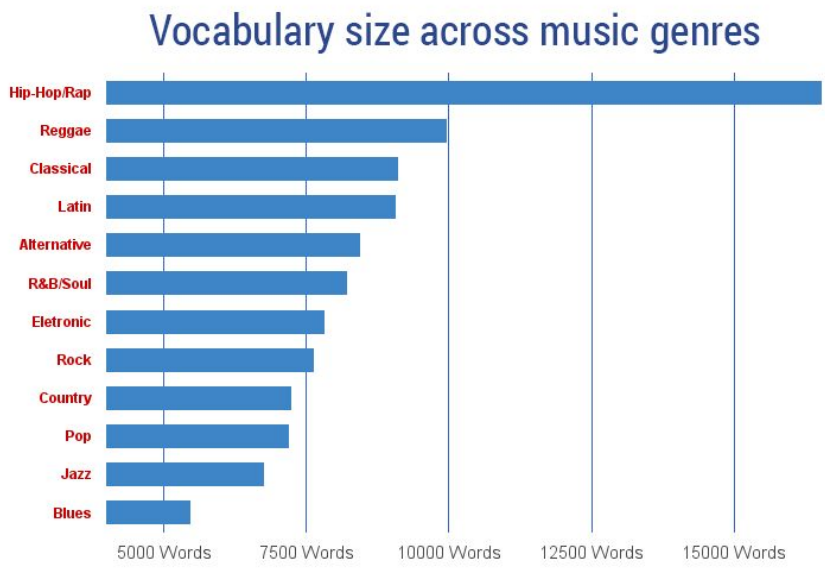


Figure 1: Vocabulary size across music genres

As shown the hip-hop/rap genre has an overwhelming predominance in terms of vocabulary size. However interesting these results may be, it's good to bear in mind that it may simply be the case that some genres have a larger overall lyric text size which may result in a wider vocabulary. In fact, by dividing the total amount of words, by the vocabulary size, we obtain the "inverse vocabulary diversity" measure. Informally, this can be described as the occurrence mean number of words between which a 'new' word is used.

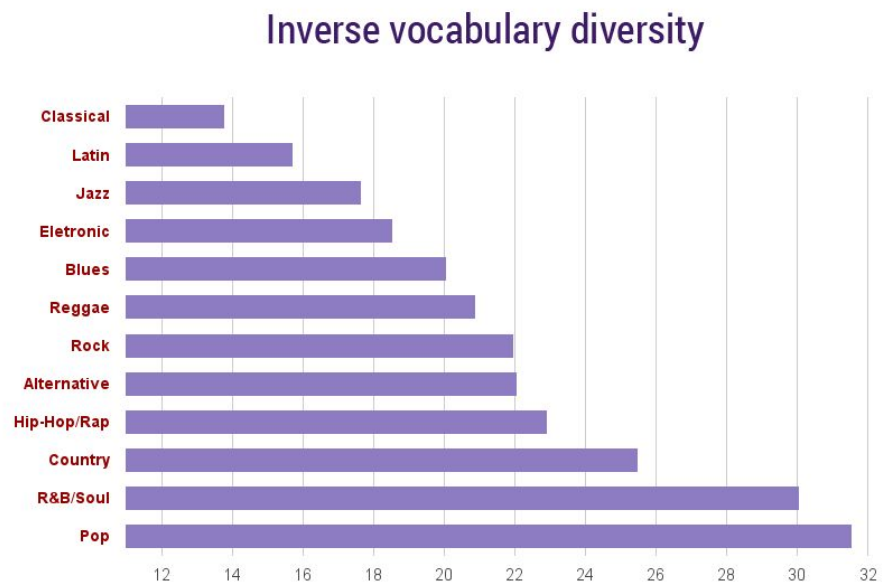
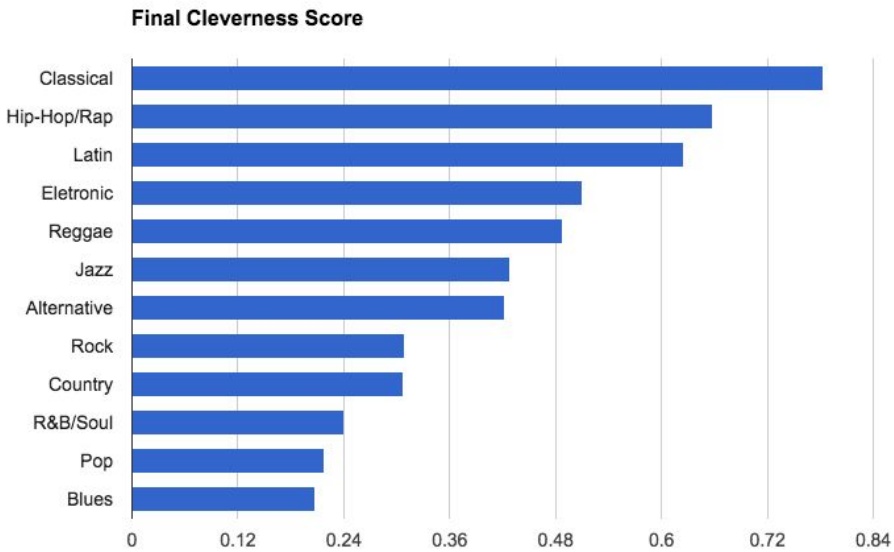


Figure 2: Inverse vocabulary diversity

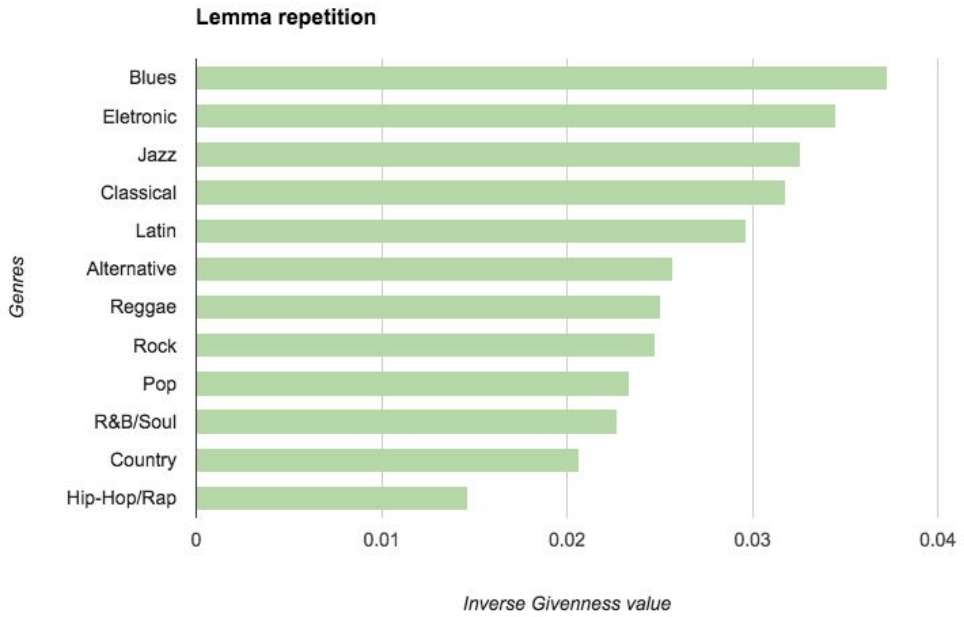
Essentially, figure 2 tells us how much a genre resorts to word repetition, where larger values represent larger repetition rates.

Experimentally, it was found that vocabulary size had a large correlation value with the average word length. In order to balance these variables weights, the final cleverness was measured by giving a 50% weight to Vocabulary diversity and 25% to both vocabulary size and average word length. Below the final calculated ‘cleverness’ of each genre.

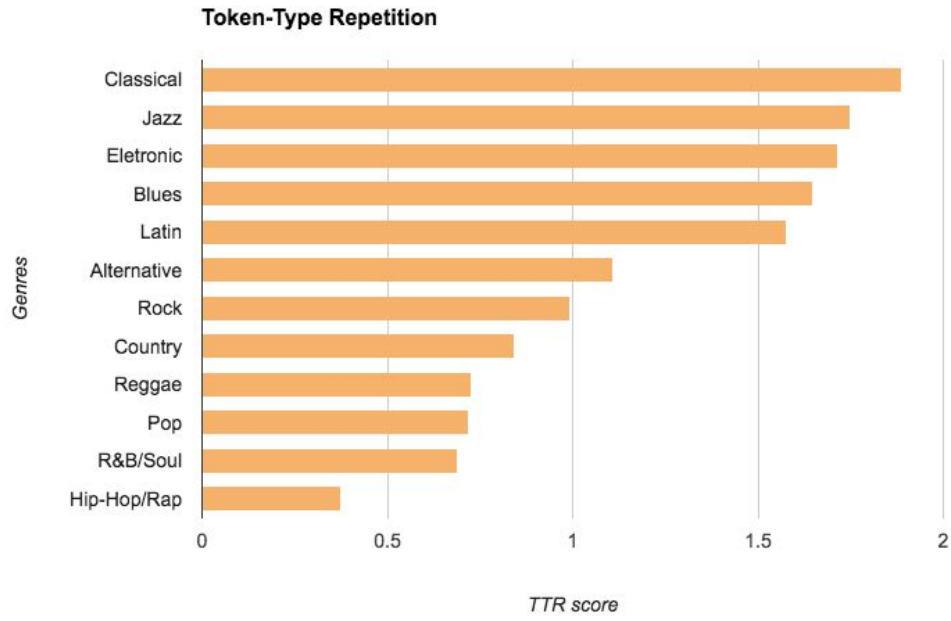


Lyrical cohesion

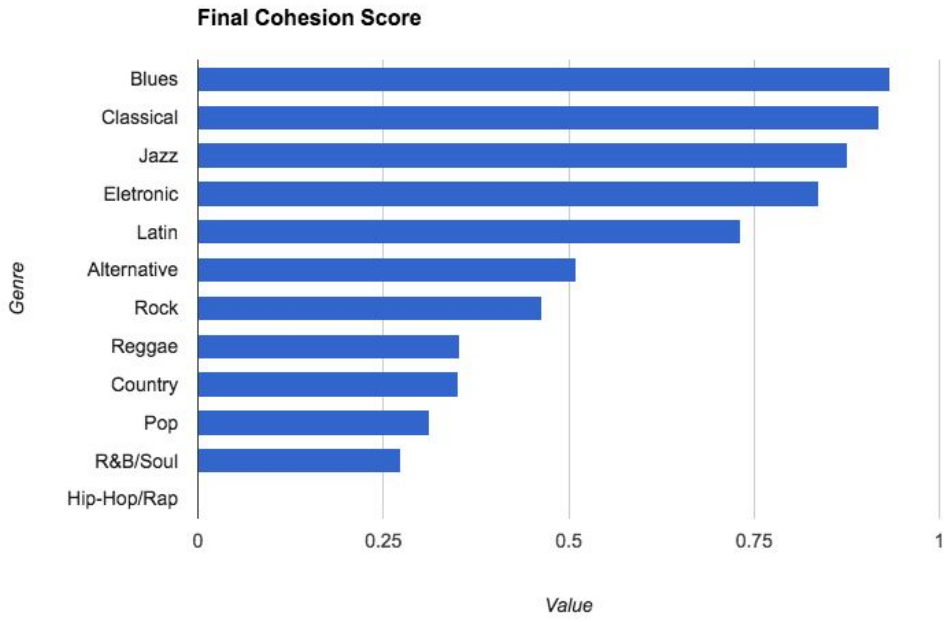
For the experiments made on a syntactic level, the measures evidenced quite a different trend. Below a plot showing the relation between givenness, the lemma repetition score and genres:



The values were inverted so that higher score represents higher complexity. As we can see, there has been quite a shuffling of the genres when compared to the simple statistical analysis. The most notable of mention is perhaps Hip-Hop/Rap falling into last place in the classification hinting that perhaps its lyrical complexity fails to hold when a more in-depth analysis is performed. This is further evidenced in the TTR score chart:



Using the three assessed metrics in this section, the final score was determined.



Evaluation

The aim of this project was to determine if a correlation can be found between a song's lyrics complexity score and its genre. In this context, correlation can be calculated as coefficient that measures the extent to which variation in one variable is associated with variation in another (not making any assumptions on how the variables are related e.g. linearly). When selecting a

correlation coefficient one has to be careful about the type of data being analysed. In this scenario the continuous variable ‘complexity’ is being correlated to the categorical variable ‘genre’. This excludes a number of popular correlation coefficients like Spearman’s rank or Kendall’s rank because of their assumptions of having numerical data when performing calculations. Indeed, there’s a limited number of coefficient rank’s that are appropriate for the scenario at hand, and even less stable python libraries that implement such methods. One possibility is using the point biserial correlation [7]. It contemplates scenarios where the relation between continuous and categorical variables is being verified. However, it has the downside of accepting only two different category types. A second option is to discretize the continuous variable by ‘binning’ it into range groups.

Using the discretized data approach, the correlation between individual song’s ‘complexity’ and their genre was calculated for the entire dataset. The used coefficient ranks were Spearman’s and Kendall. Both have gotten values very close to zero, evidencing a non-existing correlation. This result can be understood by analysis of the data. Even though there is clear demarcation in the score of some genres, when considering all 12 of them the pool of data becomes quite diluted and it would only be surprising if such clear correlation was found considering that would mean the mainstream genres could be easily ranked in terms of lyrical cleverness.

In order to reduce the pool of genres while keeping the dataset size, a study in of music genealogy [8] was used to ‘merge’ groups deemed similar. Below the correlation coefficient for different considerations on number of genres.

Correlation coefficient	Simple Statistical Attributes		Lyrical Cohesion	
#genres used	Spearman	Kendall	Spearman	Kendall
12 genres	0.045	0.036	0.016	0.014
6 genres	0.231	0.163	0.215	0.158
4 genres	0.24	0.19	0.207	0.164

By analysis of the table, we can see that even though the correlation results are still quite low, they seem to improve as similar genres are merged. This is an interesting observation considering that the same trend was not observed when merging groups arbitrarily, which motivates the given argument of diluted scores and also hints that genres of similar musical origin may be more likely to have a closer score.

All-in-all these correlation values are not sufficient to identify a correlation between genres and verse complexity. A motivation for this observation is given above. However, from the results section it is clear that are quite noticeable disparities between some genres. To further investigate just how significant they are, the point biserial correlation coefficient can be used when only two genres are being compared. For instance, using the simple statistical measures, the correlation coefficient using the data set of Hip-Hop/Rap songs and Blues songs determines a correlation coefficient of $\simeq 69.7\%$. The same genres’ correlation coefficient using the lyrical cohesion metrics scores $\simeq 64.8\%$. This represents a positive correlation which means that even though doing such thing as predicting a song genre by its lyrics ‘cleverness’ alone might prove an

impossible task, estimating the likelihood of it belonging to a genre or another might be a lot more feasible.

Discussion

Despite the correlation evaluation having failed to hold when the entire dataset is considered, the found results regarding how lyrical complexity plays out across music genres has in itself room for some interesting analysis. Firstly and possibly more discussion prone are the discrepancies between the complexity results of the simple statistical measures and the lyrical cohesion. In the latter, Hip-hop/Rap scored second while Blues scored last. Their position then almost reversed having Blues placing first and Hip-hop Rap getting last place. To understand this occurrence an analysis of the lyrics of each of them may help. In the Hip-hop genre, slang and mispronunciation are a dominant theme. This causes a large heightening of the genre's vocabulary size but it also leads to a poor performance in token-type ratio. Blues on the other hand, while placing well in terms of repetivity, it seems to be very themed, which causes a very low score in vocabulary size. However, it's proper syntax, and low repetition are highly valued in the used cohesion metrics which boosts his score. Classical music has placed well with both metric types. This is a consequence of its recurs to old english (heightening vocabulary size), the prevalence of sound syntactic structure and low repetition.

When it comes to evaluating a musical piece, lyrics are often a component that plays an important role in the process. However, songwriting is a space where artistic freedom bends and breaks the rules of textual composition. Therefore lyrical complexity does not hold the exact principles of measure as regular text pieces. Most studies of textual complexity measures recur to human evaluators to provide a gold standard of what constitutes a rich piece of text and then derive metrics based on what approximates the most such standard. Although such study type would fall outside the scope of this project, the research conducted in it can be seen as a naïve approach to lyrical complexity measure and attempt at understanding the underlying characteristics of songwriting across mainstream genres.

Bibliography

[1] Moore, A. F. (2001). Categorical Conventions in Music Discourse: Style and Genre. *Music {&} Letters*, *82*(3), 432-442.

[2] Nelson, J., Perfetti, C., Liben, D., Liben, M.: Measures of text difficulty. Technical Report, Gates Foundation (2011)

[3] Sheehan, K. M., Kostin, I., & Futagi, Y. (2009). When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty? *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, (2006), 1978-1983.

[4] Crossley, S.A., Kyle, K. & McNamara, D.S. Behav Res (2016) The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion, Behavior Research Methods, 48, 1227-1237.

[5] Ease, F. R. (2009). Flesch–Kincaid readability test. Reading 70, 8-10.

[6] Kendall, M. (1938). "A New Measure of Rank Correlation". Biometrika. 30 (1–2): 81–89

[7] Gupta, S. D. (1960). Point biserial correlation coefficient and its generalization. *Psychometrika*, *25*(4), 393-408. Springer-Verlag.

[8] Crauwels, Kwinten. January 20, 2017. The genealogy and History of Popular Music Genres. Retrieved from https://musicmap.info