From Words to Movement: Predicting Danceability Using Lyrical Repetition Metrics

Abstract

This project investigates whether lyrical repetition helps explain why some songs are more danceable than others. To address this question, we merged Billboard Hot 100 data (2000–2023) with Spotify audio features and constructed a "repetition ratio" that measures the proportion of repeated lyrical content after removing filler words.

Our analysis revealed a distinctly nonlinear relationship between repetition and danceability. A quadratic regression model fitted the data substantially better than linear and transformation-based specifications. Incorporating the repetition ratio and its quadratic term into a full model containing Spotify's standard musical features (energy, tempo, valence, etc.) further improved predictive performance. An Extra Sums of Squares F-test ($p < 0.001$) confirmed that adding lyrical repetition provides statistically significant explanatory power beyond using traditional audio features alone.

Overall, our results show that repetition in lyrics is a meaningful predictor of danceability. Even after accounting for core musical characteristics, songs with moderate – but not excessive – lyrical repetition tend to be more danceable. These findings highlight the value of integrating lyrical structure when considering listener engagement.

Introduction

As music consumption shifts toward streaming platforms, listeners have shown a growing preference for songs that are more danceable (Interiano et al., 2018). This preference reflects the importance of danceability, which captures how strongly a song encourages physical movement (Wu, 2025). Prior work shows that songs with higher danceability tend to receive more streams and are more frequently included in playlists, reflecting its role in shaping listener engagement (Wu, 2025).

Spotify's audio feature dataset quantifies song's production features like energy, loudness, speechiness, valence, tempo, etc. However, while these acoustic features are well-studied, the structural features of lyrics, particularly the repetition of words in a song has received far less analysis. Repetition is a fundamental tool in song writing since songs with more repetitive lyrics are easier for listeners to process and gain wider adoption in the market (Nunes & Nalsesia, 2014). These qualities may naturally correlate with the perception of danceability of songs. Despite the intuitive importance, we have not found reports on the relationship between lyrical repetition and danceability. Our research aims to fill this gap by examining whether lyrical repetition is related with danceability in popular music.

Methods

a. Data Collection and Cleaning

Our analysis required a comprehensive dataset linking musical audio features with lyrical content for popular songs. The core variable, danceability, is a metric provided by Spotify that describes how suitable a track is for dancing based on a combination of musical elements. Although Spotify does not disclose the exact algorithm, the score ranges from 0 (not danceable) to 1 (highly danceable). We constructed our analytic dataset by combining two primary data sources from Kaggle: (1) a compiled dataset of Billboard charting songs from 2000–2023 containing lyrics and audio features, and (2) a Spotify Top Songs dataset containing complete audio features from Spotify Web API. This section details the full data processing pipeline focused on merging these sources to create a clean and complete dataset for our analysis.

The Billboard dataset includes 24 years of Hot 100 chart information. Our initial exploration revealed two primary data quality issues: duplicate entries for songs that charted across multiple years or featured multiple artists, and a significant proportion of missing values for key audio features like danceability. The Spotify dataset contained multiple entries for the same track sourced from different distributors, leading to near-duplicate records.

The initial step involved de-duplication. The Billboard data was consolidated to ensure each unique song appeared only once. For the Spotify data, the multiple entries for a given track name were resolved by retaining the first instance, under the assumption that any minor variations in audio features between different distributors would be negligible for our large-scale analysis.

Then, we merged the Spotify audio features onto the Billboard dataset using song titles and artist names. Songs in the Billboard dataset that lacked audio features were updated with values from their Spotify match if one was found; if no match was found, these songs were deleted from the analytic set. For songs that already had audio features in the Billboard data, their values were replaced with the Spotify data upon a successful match to ensure consistency, while unmatched songs with existing features were retained.

b. Exploratory Analysis

Because we wanted to measure how repetitive each song's lyrics are, we defined our own metric: "repetition ratio." To compute it, we first broke the lyrics into individual words and counted the total word count for each song. After removing filler expressions (a, the etc.), we counted how many times each remaining word appears and summed the counts of words that occur more than once. Finally, we divided this number by the song's total word count to obtain the repetition ratio.

During the initial exploration stage, we fitted the scatterplot (Figure 6) to explore the relationship between repetition ratio and danceability, which suggests a weak position relationship.

c. Analytic Methods and Results

First, we fitted a linear regression model between danceability and repetition ratio to explore their relationship. The result showed a positive linear association between the predictor and the response:

$$mean(danceability) = 0.614 + 0.253 * repetitionratio$$

However, with an R^2 of 0.0382, this simple model accounted for only 3.82% of the variation in danceability. Because the scatterplot between danceability and repetition ratio (Figure 1) displayed a pattern that is more curved than linear, and because the residual plot of this single-variable model (Figure 2) showed violations of the constant-variance assumption of linear regression, we explored both model transformation and model expansion to capture the relationship more accurately. We considered three transformation methods: Box-Cox transformation with λ=1.5 (Figure 3), log transformation, and logit transformation. We also examined an expanded model that included a quadratic term to capture the curved pattern shown in the scatterplot. The four residual plots of the resulting models (Figure 4) indicated that the quadratic model performed best: its residuals displayed the most random scatter and the most stable variance. As shown in the table below, the quadratic model accounted for the greatest variation in danceability (highest R^2) and demonstrated the strongest model fitness without being too complex (lowest BIC score). In contrast, the transformation-based models did not meaningfully improve the fit and often produced more complicated patterns. When we fitted a blue quadratic curve onto the scatterplot (Figure 5), it clearly matched the observed pattern more closely than the simple linear regression line (Figure 6). Therefore, we concluded that the following quadratic model best captured the relationship between danceability and repetition ratio:

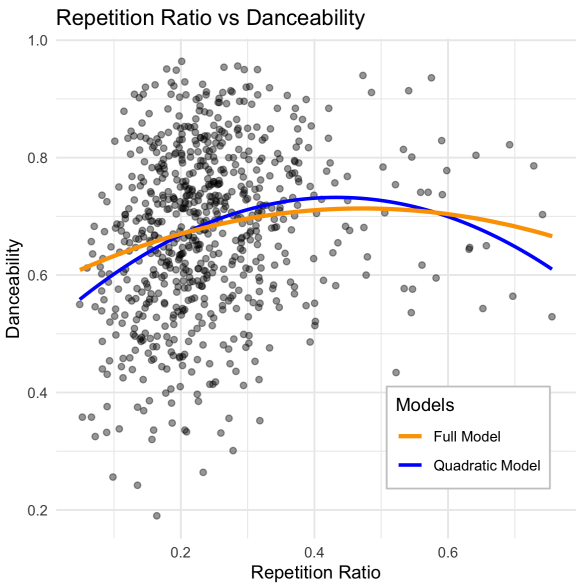


Figure 5

$$mean(danceability) = 0.512 + 1.019 * repetitionratio - 1.176 * repetitionratio^2$$

	Single-Var Model	Box-Cox Model	Logged Model	Logit Model	Quadratic Model
R^2	0.038	0.037	0.040	0.034	0.070
Adjusted R^2	0.037	0.035	0.039	0.032	0.068
BIC	-767.479	-503.782	-58.273	1558.784	-784.885

Table 1

After identifying this quadratic relationship, we also examined whether incorporating repetition ratio into the best base model for predicting danceability would improve overall model performance. Although the exact formula for Spotify's danceability score is unknown, we knew that it was computed from several musical elements, such as tempo, energy, and acousticness. Repetition ratio was not officially included among these features, but we sought to determine whether it could nonetheless enhance predictive accuracy.

To address this, we first constructed a base model using all available musical features in our merged dataset: energy, loudness, speechiness, valence, tempo, acousticness, instrumentalness, and liveness. The resulting base model is as follows:

$$mean(danceability) = 0.857 - 0.308 * energy + 0.007 * loudness + 0.289 * speechiness + 0.279 * valence - 0.0007 * tempo - 0.161 * acousticness - 0.015 * instrumentalness - 0.066 * liveness$$

Then, we construct a full model by adding repetition ratio and its quadratic term to the base model. Our full model is as follows:

$$mean(danceability) = 0.759 - 0.301 * energy + 0.007 * loudness + 0.269 * speechiness + 0.265 * valence - 0.0006 * tempo - 0.147 * acousticness - 0.026 * instrumentalness - 0.070 * liveness + 0.553 * repetitionratio - 0.586 * repetitionratio^2$$

Next, we performed an Extra Sums of Squares F-test to determine whether adding repetition ratio and its quadratic term significantly improved the model. The null hypothesis stated that these additional predictors provided no improvement. With a p-value below 0.001, the ANOVA result provides strong evidence that at least one of the two repetition-ratio predictors contributes meaningfully to improving model fit. The table of R^2, adjusted R^2, and BIC indicated that the full model explains more variation in danceability and achieved a better fit without increased complexity. According to the Variable Importance Plot (Figure 7), the repetition ratio and its quadratic term are the 6th- and 7th-most important predictors in the model, exceeding traditional audio features like loudness, liveness, and instrumentalness. Thus, we concluded that incorporating the quadratic relationship involving repetition ratio yielded a model with superior predictive ability. To visualize this superiority, we fitted an orange curve of the full model onto the scatterplot between danceability and repetition ratio, which captures the pattern better than the blue curve of the simple quadratic model (Figure 5).

	Base Model	Full Model
R^2	0.322	0.346
Adjusted R^2	0.314	0.337
BIC	-968.843	-981.472

Table 2

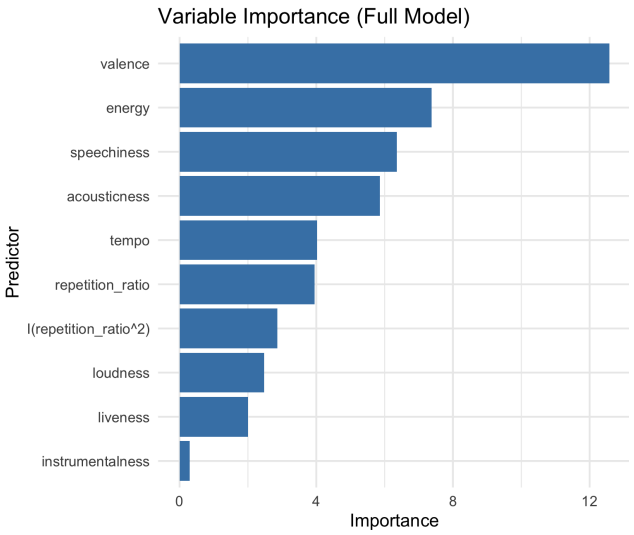


Figure 7

Discussion and Conclusion

Our analysis showed that while repetition ratio has a weak linear association with danceability, the relationship is essentially nonlinear. A quadratic model captures this curvature better than then simple linear model and improves the model fitness at a statistically significant level. This makes intuitive sense because moderate lyrical repetition increases a song's danceability by making it easier to follow. However, beyond a certain point, too much repetition makes the song monotonous and thus less engaging to dance with. We then looked at whether repetition ratio adds explanatory value beyond Spotify's standard musical features. The results showed that including both the repetition ratio and its quadratic term improves prediction accuracy over the base model by having higher adjusted R^2 and lower BIC score. Besides, the p-value of repetition ratio and its quadratic term are both significant in the Extra Sums of Squares F-test. Therefore, our results suggest that lyrical repetition meaningfully contributes to danceability, even after accounting for other musical features. This emphasizes the importance of lyrical structure as an overlooked but important aspect in illustrating what makes a song more "danceable."

One limitation of our analysis is that the base model does not include several musical elements, such as key, mode, and time signature, because these features were unavailable in our merged dataset. In addition, the repetition ratio was invented for this paper, which introduces subjectivity and may not perfectly capture a song's true lyrical repetition. Moreover, the repetition ratio metric only removed filler expressions in English but not Spanish, which makes repetition ratios incomparable between languages.

The future work could be to include a dummy variable for the Spanish songs in the regression model to test its significance. A spanish version of the filler words can also be created and added to the calculation of the repetition ratio.

Generative AI statement

The use of generative AI in this paper followed the course guideline where GenAI is only allowed to assist coding but not writing the paper.

References

[1] Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, *5*(5), 171274. https://doi.org/10.1098/rsos.171274

[2] Wu, W. (2025). Predicting danceability and song ratings using Deep Learning and auditory features. *PeerJ Computer Science*, *11*. https://doi.org/10.7717/peerj-cs.3009

[3] Nunes, J., & Valsesia, F. (2014). The power of repetition - repetitive lyrics in a song increase processing fluency and drives market success. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2938838

Appendix

Figure 1: Scatterplot between Danceability and Repetition Ratio

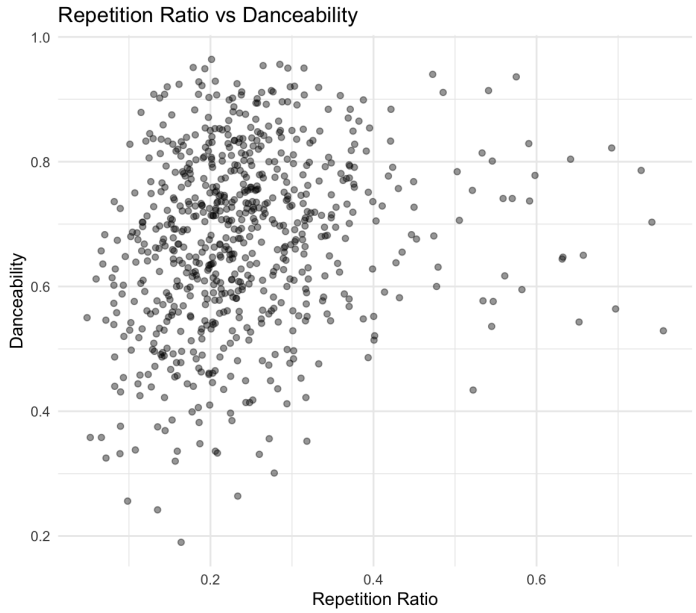


Figure 2: Residual vs. Fitted Values Plot of the Model with only one predictor (Repetition Ratio)

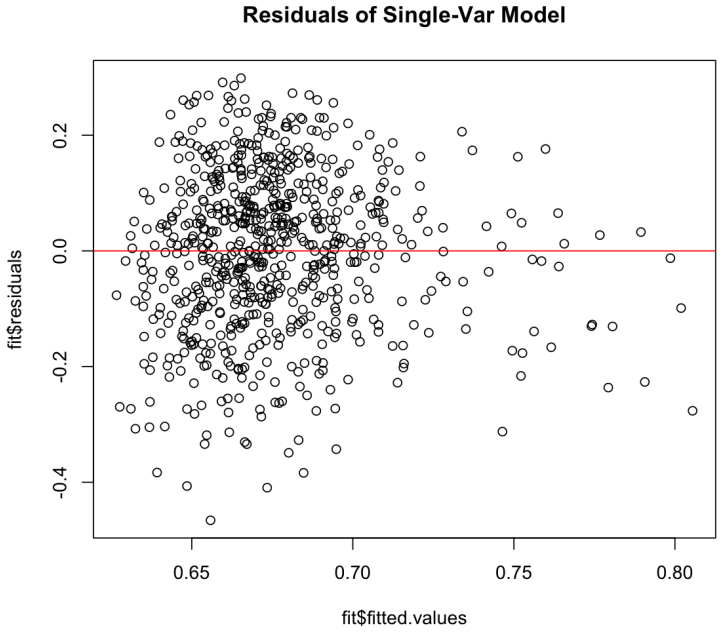


Figure 3: Log-Likelihood Plot to identify the λ for Box-Cox Transformation (λ=1.5)

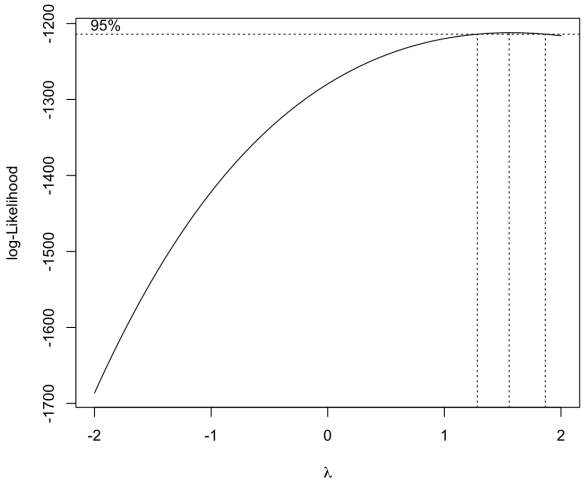


Figure 4: Residual Plots for Box-Cox Model, Log Model, Logit Model, and Quadratic Model

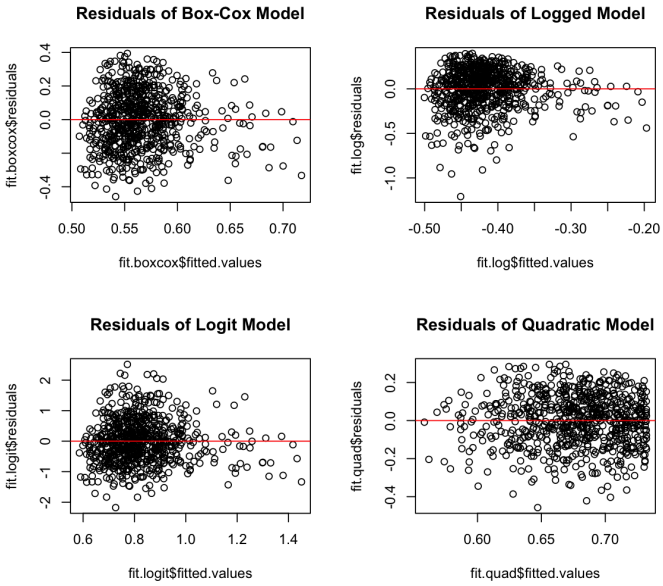


Figure 5: Fitted Quadratic Regression Between Danceability and Repetition Ratio vs. Fitted Full Model

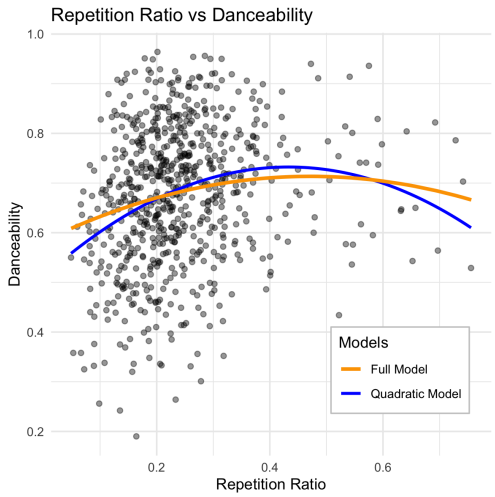


Figure 6: Fitted Linear Regression Between Danceability and Repetition Ratio

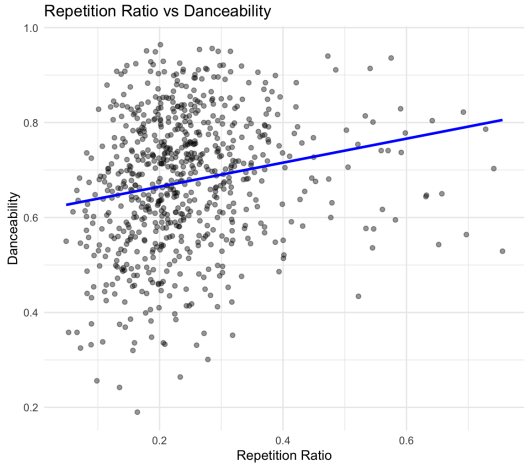
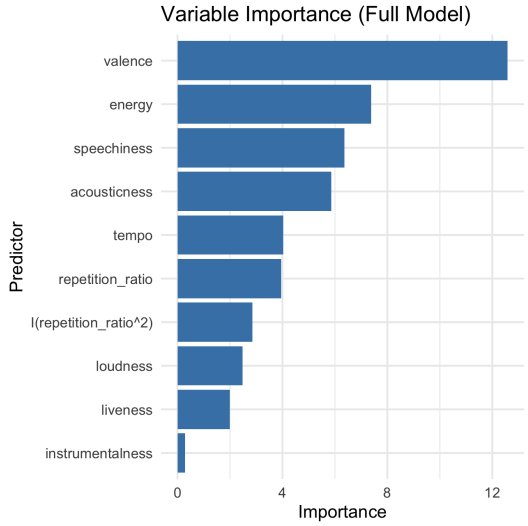


Figure 7: Variable Importance Plot of the Full Model with all predictors.



Examples: Comparison between songs with high and low Repetition Ratios and their Danceability.

	Lyrics	Repetition Ratio	Danceability
Blue (Da Ba Dee)	Yo, listen up, here's the story About a little guy that lives in a blue world And all day and all night and everything he sees is just blue Like him, inside and outside Blue his house with a blue little window And a blue Corvette and everything is blue for him And himself and everybody around 'Cause he ain't got nobody to listen (To listen, to listen, to listen) I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di I have a blue house with a blue window Blue is the color of all that I wear Blue are the streets and all the trees are too I have a girlfriend and she is so blue Blue are the people here that walk around Blue like my Corvette, it's standing outside Blue are the words I say and what I think Blue are the feelings that live inside me You might also like I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di I have a blue house with a blue window Blue is the color of all that I wear Blue are the streets and all the trees are too I have a girlfriend and she is so blue Blue are the people here that walk around Blue like my Corvette, it's standing outside Blue are the words I say and what I think Blue are the feelings that live inside me I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Inside and outside Blue his house with a blue little window And a blue Corvette and everything is blue for him And himself and everybody around 'Cause he ain't got nobody to listen (To listen) I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di I'm blue, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di Da ba dee da ba di, da ba dee da ba di	0.69	0.82

	Lyrics	Repetition Ratio	Danceability
The Good Stuff	Well, me and my lady had our first big fight So I drove around 'til I saw the neon lights Of a corner bar, and it just seemed right so I pulled up Not a soul around but the old bar keep Down at the end and looking half asleep But he walked up, and said, "What'll it be?" I said, "The good stuff" He didn't reach around for the whiskey He didn't pour me a beer His blue eyes kind of went misty He said "You can't find that here" 'Cause it's the first long kiss on a second date Mama's all worried when you get home late And dropping the ring in the spaghetti plate 'Cause your hands are shaking so much And it's the way that she looks with the rice in her hair Eating burnt suppers the whole first year And asking for seconds to keep her from tearing up Yeah, man, that's the good stuff He grabbed a carton of milk and he poured a glass And I smiled and said, "I'll have some of that" We sat there and talked as an hour passed, like old friends I saw a black and white picture and it caught my stare It was a pretty girl with bouffant hair He said, "That's my Bonnie Taken about a year after we were wed" You might also like He said, "Spent five years in the bottle When the cancer took her from me But I've been sober three years now 'Cause the one thing stronger than the whiskey Was the sight of her holding my baby girl The way she adored that string of pearls I gave her the day that our youngest boy, Earl Married his high school love And it's a new T-shirt saying, "I'm a Grandpa" Being right there as our time got small And holding her hand, when the Good Lord called her up Yeah, man, that's the good stuff" He said, "When you get home, she'll start to cry When she says, "I'm sorry," say, "So am I" And look into those eyes, so deep in love And drink it up 'Cause that's the good stuff That's the good stuff Embed	0.06	0.61