Análise de correlação semântica entre músicas originais RaussTuna

Semantic correlation analysis between original RaussTuna songs

Daniela Filipa Pereira Fontinha[1[a52076]], Gabriel Ribeiro Carneiro[2[a60447]], Gonçalo Martins Pereira[3[a48168]], Tiago Filipe Santos Guedes[4[a61250]], Vinícius Nascimento Silva[5, 6[a62860]]

[1] Instituto Politécnico de Bragança, Portugal, a52076@alunos.ipb.pt
[2] Instituto Politécnico de Bragança, Portugal, a60447@alunos.ipb.pt
[3] Instituto Politécnico de Bragança, Portugal, a48167@alunos.ipb.pt
[4] Instituto Politécnico de Bragança, Portugal, a61250@alunos.ipb.pt
[5] Instituto Politécnico de Bragança, Portugal, a62860@alunos.ipb.pt
[6] Centro Federal de Educação Tecnológica de Minas Gerais Campus Nova Gameleira, Brasil, vnszero@gmail.com

Resumo

Este estudo propõe uma análise de correlação semântica entre as canções originais da *RaussTuna - Tuna Mista de Bragança* cuja produção cultural apresenta uma identidade única. A análise procura identificar padrões de significado entre as letras das músicas, apesar das variações no estilo musical, nos temas e nos compositores. Para tal, o estudo emprega técnicas de *Natural Language Processing* (NLP) e modelos de *embeddings* de palavras, como o *Word2Vec*, para converter letras de canções em representações vetoriais numéricas. Estes vetores captam semelhanças semânticas entre as letras, permitindo uma comparação objetiva do seu conteúdo temático. A semelhança entre as músicas é quantificada através da *Cosine Similarity*, uma medida matemática de proximidade textual. Como cada música é constituída por múltiplas palavras, o *embedding* da canção final é obtido pela agregação dos *embeddings* das palavras que a compõem em um único vetor representativo. Para melhor compreender as relações entre as músicas, é aplicado um processo de redução de dimensionalidade em duas fases: a *Principal Component Analysis* (PCA) é utilizada primeiro para reduzir o ruído e enfatizar características essenciais, seguida pela *t-Distributed Stochastic Neighbor Embedding* (t-SNE), que projeta os dados em um espaço bidimensional para preservar as vizinhanças semânticas. Esta abordagem permite uma representação gráfica das distâncias semânticas entre as músicas. Este estudo analisa 32 de 35 canções originais; instrumentais são excluídos devido à ausência de letras. Os resultados devem oferecer compreensões relevantes sobre a estrutura e criatividade presentes na produção musical do grupo, ao valorizar a singularidade artística e cultural de cada original. Espera-se que a análise contribua para uma compreensão mais profunda da produção musical da tuna e para a preservação das raízes culturais da região em um contexto globalizado.

Palavras-Chave: *Processamento de linguagem natural, análise musical, identidade cultural.*

Abstract

This study proposes a semantic correlation analysis between the original songs of *RaussTuna - Tuna Mista de Bragança* whose cultural production presents a unique identity. The analysis seeks to identify patterns of meaning between the lyrics of the songs, despite the variations in musical style, themes and composers. To this end, the study employs Natural Language Processing (NLP) techniques and word embedding models, such as Word2Vec, to convert song lyrics into numerical vector representations. These vectors capture semantic similarities between the lyrics, allowing an objective comparison of their thematic content. The similarity between the songs is quantified through Cosine Similarity, a mathematical measure of textual proximity. Since each song is made up of multiple words, the embedding of the final song is obtained by aggregating the embeddings of the words that compose it into a single representative vector. To better understand the relationships between the songs, a two-stage dimensionality reduction process is applied: Principal Component Analysis (PCA) is used first to reduce noise and emphasize essential features, followed by t-Distributed Stochastic Neighbor Embedding (t-SNE), which projects the data into a two-dimensional space to preserve semantic neighborhoods. This approach allows a graphical representation of the semantic distances between the songs. This study analyzes 32 of 35 original songs; instrumentals are excluded due to the absence of lyrics. The results should provide relevant insights into the structure and creativity present in the group's musical production, while valuing the artistic and cultural uniqueness of each original. It is expected that the analysis will contribute to a deeper understanding of the tuna's musical production and to the preservation of the region's cultural roots in a globalized context.

Keywords: *Natural language processing, music analysis, cultural identity.*

Referências

Chen, S., Moore, J. L., Turnbull, D., & Joachims, T. (2012). Playlist prediction via metric embedding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 714–722. https://doi.org/10.1145/2339530.2339643

Sergl, M. J. (2013). Identidades sonoras na ditadura militar brasileira (1964-1985). *Journal Lumen et Virtus, Brasil, IV (8)*, 124–151.

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Computer Science*, *117*, 256–265. https://doi.org/https://doi.org/10.1016/j.procs.2017.10.117