Análise - Page Rank Musical

O relatório do *Page Rank Musical* é dividido em seções e a análise será também disposta dessa forma. Cada parte refere-se a uma seção do relatório, com exceção da Parte 3 deste relatório, que une a terceira e quarta parte do relatório original.

# Parte 1 - Entender e implementar Page Rank em matriz de adjacência igual ao vídeo

Nessa parte analisarei o código do arquivo ***algo\_inicial.js*** do repositório <https://github.com/joaoapel/page-rank>.

Conceitualmente, definimos uma matriz de adjacências A, onde A[i][j] = 1 se há uma ligação do nó j ao nó i, e A[i][j] = 0, caso contrário.

Para transformar essa matriz de adjacências em uma matriz de transição válida, normalizamos as colunas, dividindo cada valor da coluna pela soma de todos os elementos da coluna, ou seja:

P[i][j] = A[i][j] / sum(A[:,j])

Para evitar que os nós da rede que a matriz representa estejam desconectados, aplicamos um processo de dampening ou ***amortecimento***. Isso garante que, com iterações suficientes, não ficaremos presos em partes do grafo que estão desconectadas do resto ou que entraremos em loops de páginas que se referem apenas a elas mesmas. Definimos uma matriz Q, onde esse processo é aplicado - basicamente estamos adicionando um pouco de chance de pular aos nós que não estão conectados a cada nó - e retirando um pouco de chance de pular aos nós que estão de fato conectados, para garantir que a probabilidade total de cada coluna continuará sendo *aproximadamente* 100%.

Na fórmula a seguir

α é o ***fator de amortecimento***, um número pequeno,

N é o ***número de nós***.

Q[i][j] = (1 - α) \* P[i][j] + α / N

Definimos um vetor de probabilidade inicial

v = [1/N, 1/N, ..., 1/N]

Agora que temos todos os componentes, podemos iniciar o processo para buscar o *ranking* de nós, que será um ***v*** estável (cuja alteração será mínima para cada iteração conseguinte). O processo é definido por

Após um número suficiente de iterações, obteremos um vetor estável, onde

Na álgebra linear, chama-se esse vetor ***v*** de autovetor de , com um autovalor de 1.

Conclusão da análise da parte 1 - ✅

Apesar do código implementado ter executado apenas uma iteração do processo, a teoria garante que após algumas iterações o vetor obtido seria o que buscamos - um vetor com um número correspondente à importância de cada artista em cada posição.

# Parte 2 - Buscar maneiras de parsear as letras das músicas

# O objetivo do *Page Rank Musical* é determinar a relevância de artistas/bandas no contexto de músicas de outros artistas. Utilizando o algoritmo de Page Rank, definimos os artistas//bandas como nós de um grafo e suas referências a outros artistas em suas músicas como as arestas que os conectam. Para isso, precisamos de uma forma eficiente de extrair as referências à artistas nas letras de músicas. Isso é uma tarefa difícil, já que muitos artistas escolhem nomes artísticos de objetos, lugares ou até mesmo referências a outros artistas.

O artigo original do *Page Rank Musical* lida com essa dificuldade utilizando a IA de linguagem natural (LN) da Google, que classifica as palavras de um texto em categorias, como na imagem abaixo.



Nesse exemplo, podemos perceber alguns erros. Como a caracterização de Cidade Negra, banda de reggae, como Organização e baby (Baby Consuelo), cantora da banda Novos Baianos, também como Organização.

O website da IA de LN da Google não especifica seu funcionamento, mas é possível assumir que ela utiliza alguma arquitetura moderna de modelo de linguagem e compara a vetorização contextual / ***contextual embedding*** de cada palavra de uma das últimas camadas de sua rede neural a exemplos conhecidos e classificados de cada categoria definida (evento, organização, pessoa, etc.). Logo, uma possível melhoria na classificação para as categorias que estamos interessados (artistas, bandas) seria classificar cada palavra das letras das músicas de acordo com apenas essas categorias e comparar a vetorização de cada palavra aos casos-base relevantes. Essa tentativa fugiria do escopo dessa análise e requer acesso a algum modelo de linguagem sendo executado local ou remotamente.

Conclusão da análise da parte 2 - ✅

A dificuldade de receber os *nomes* dos artistas foi relativamente bem solucionada pelo uso da API. Nota-se que no código em *page\_rank.js,* a busca por nomes está sendo feita apenas para nomes de pessoas, ignorando bandas e músicas. Isso aumentaria a complexidade da tarefa, já que teríamos que relacionar todas as músicas aos artistas que a compuseram, uma tarefa extremamente complicada, visto que há muitas músicas com o mesmo nome.

Na função parseEntidades, há um erro na função filter, ao invés de filtrar as entidades que foram caracterizadas como “Person”, a função está atribuindo “Person” a todas as entidades. Porém, já que posteriormente no código considera-se apenas as entidades que estão na lista de artistas com músicas compostas (com a finalidade de formar uma matriz N x N), isso não deve ocasionar erros.

const listaFiltrada = await lista.filter((entidade) => {

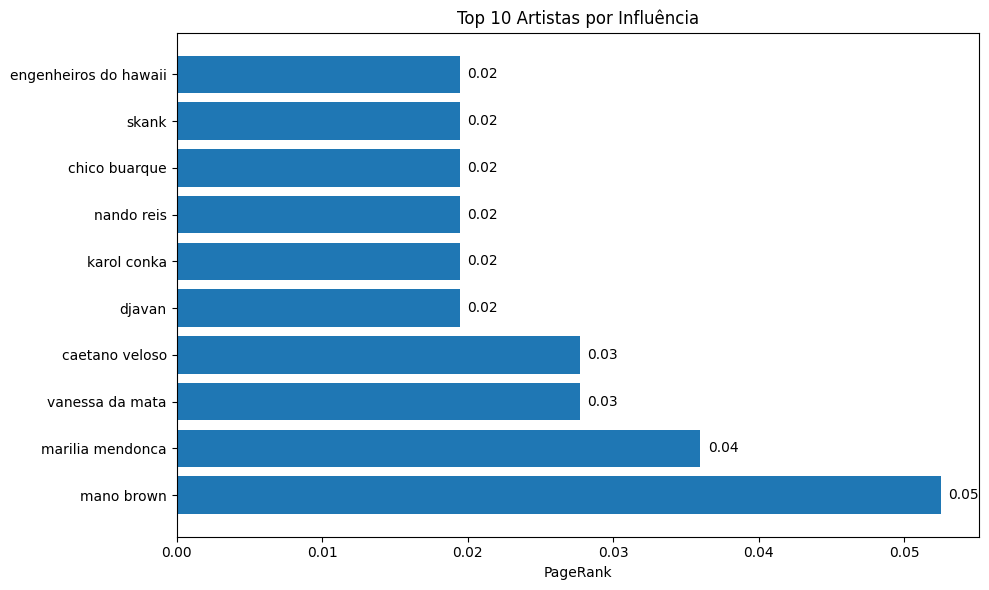
return entidade.type = 'PERSON'

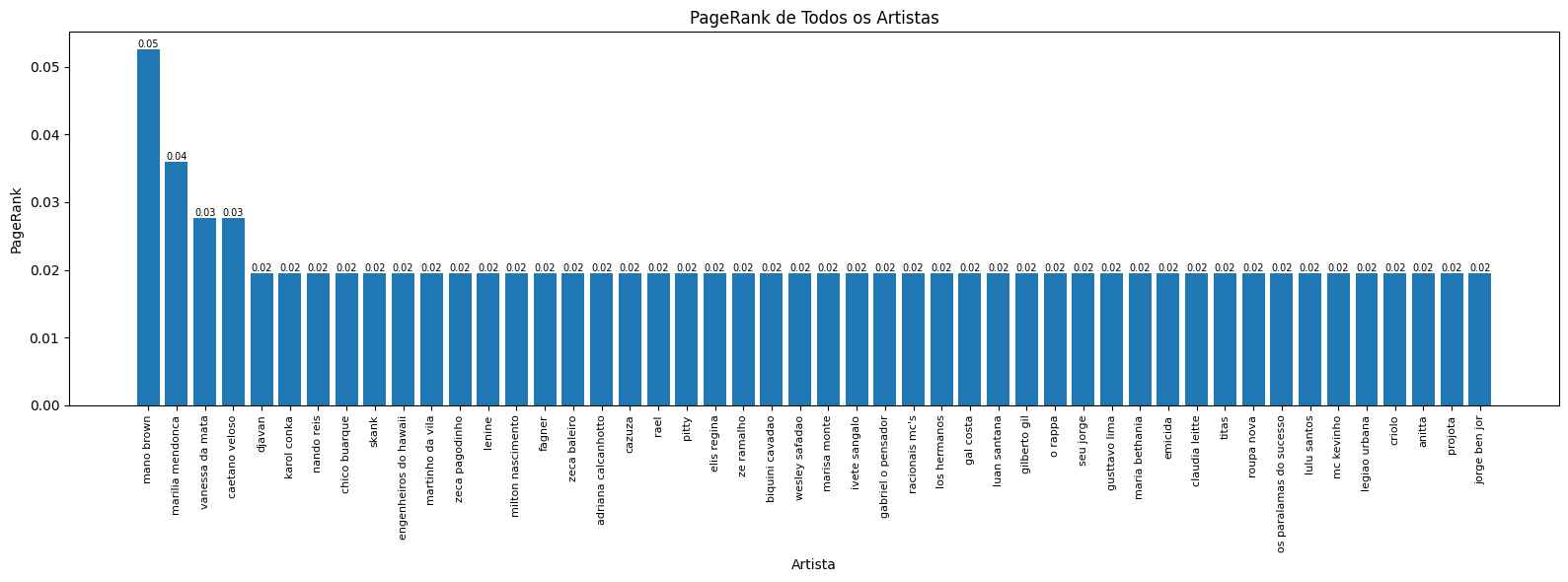
}).map(user => user.name);

# Parte 3 - Implementar Matriz Adjacência, Matriz de Transição e aplicar o algoritmo

Não foi registrado nenhum erro na implementação do algoritmo. Porém é necessário notar que a escolha das músicas teve impacto considerável nos resultados, foram músicas intencionalmente escolhidas devido ao número alto de menções a outros artistas.

Devido às limitações de uso gratuito da API da Google e com o intuito de teste, re-implementei o algoritmo em notebook Python e no lugar da API de LP executei o modelo de reconhecimento de entidade nomeada *lfcc/bert-portuguese-ner*, disponível no site [huggingface.co](http://huggingface.co), localmente. Para aumentar a base de dados utilizei a API do [last.fm](http://last.fm) para obter o título de até 10 músicas de 60 artistas proeminentes de língua portuguesa e a API lyrics OVH para obter as letras dessas músicas. Uma seleção menos enviesada de músicas e artistas acabou criando um grafo muito esparso, após a aplicação do algoritmo foi obtida uma distribuição no vetor de importância quase uniforme entre todos os artistas. As imagens abaixo mostram isso.





Diversas técnicas poderiam ser empregadas numa tentativa de melhorar os resultados: Poderia haver um esforço maior para limpar as letras das músicas e garantir que os nomes dos artistas sempre correspondam a quem se referem, assim como seria possível não depender totalmente de modelos de linguagem e buscar os nomes dos artistas diretamente nas letras de outros artistas. Porém os resultados nessa etapa me levam a crer que isso não deve ser feito agora.

Observação : Deve-se notar que o desempenho do modelo de classificação disponível gratuitamente parece ser um pouco pior do que da Google, o que era de se esperar.

Conclusão

O algoritmo *Page Rank* aplicado com a finalidade de determinar a importância de diferentes artistas teve uma prova de conceito no relatório porém há considerações importantes. Raramente em músicas citam-se artistas diretamente, normalmente referências são mais sutis, contendo partes de letras, nomes de músicas e crucialmente para essa tentativa - temas musicais. Sendo a música fundamentalmente um meio não exclusivamente textual seria muito difícil determinar a influência de artistas em outras músicas utilizando apenas técnicas de Processamento de Linguagem Natural *sobre apenas as letras de músicas*. Os grafos obtidos utilizando essa técnica geralmente são muito esparsos (os grafos obtidos para a implementação do PageRank para websites também são muito esparsos) e até agora levaram a resultados insatisfatórios para esse caso de uso.   
 Para aumentar o grau médio do grafo, outras referências poderiam ser utilizadas. Caso tentássemos buscar o nome de músicas em letras de outras músicas, por exemplo, teríamos que enfrentar um problema maior ao parsear esses dados, visto que músicos constantemente escolhem nomes de músicas que já existem ou até mesmo escolhem o nome de suas bandas inspirando-se em músicas.

Apesar dessas considerações, caso conseguirmos utilizar uma base de dados em que artistas enunciam abertamente suas inspirações talvez seria possível aplicar o mesmo algoritmo e obter resultados melhores. Na próxima seção deste relatório este parágrafo será melhor explicado.

Parte 4 - Alternativa : Utilizar bases de dados alternativas

Previamente neste relatório, buscamos referências a outros artistas no corpo das letras de suas músicas. Apesar dos resultados obtidos não terem sido satisfatórios, talvez seria possível utilizar as mesmas técnicas de processamento de linguagem natural sobre um corpo de entrevistas, redações ou gêneros do tipo, já que esses meios geralmente tendem a tratar diretamente dos assuntos de referências e influências musicais, precisamente o assunto que o *Page Rank Musical* utiliza para ranquear os artistas*.*

Algumas bases possíveis e observações

Entrevistas em revistas - Entrevistas comumente tratam de assunto de influência de outros artistas, porém contém o viés do entrevistador, que pode influenciar o artista entrevistado a mencionar artistas que ele não mencionaria. Além disso, não há, até onde sei, uma base de entrevistas de artistas brasileiros ou lusófonos. Haveria trabalho considerável de web-scraping para obter essas entrevistas e associar aos artistas entrevistados.

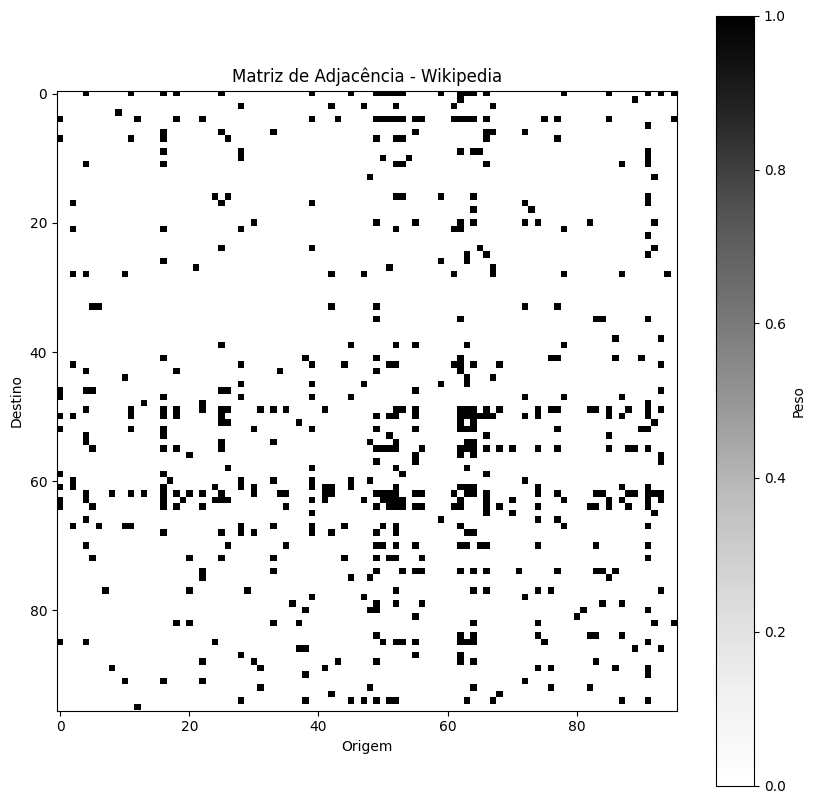
Podcasts - Assim como em entrevistas, o assunto que buscamos é comumente tratado, porém são um meio recente e artistas mais antigos possivelmente não seriam contemplados. Além disso, haveria uma etapa a mais, a de transcrição dos podcasts. Isso provavelmente não seria um problema, visto que a API de transcrição do Youtube é excelente.

Wikipédia - Das bases contempladas até agora, a Wikipedia é o único meio em que os dados estariam concentrados em uma página só. Porém, também é o único meio em que o artista não toma nenhuma decisão sobre o conteúdo disposto. O editor da página pode ou não mencionar artistas relacionados de acordo com seu viés pessoal. Por enquanto, essa será a base alternativa escolhida.

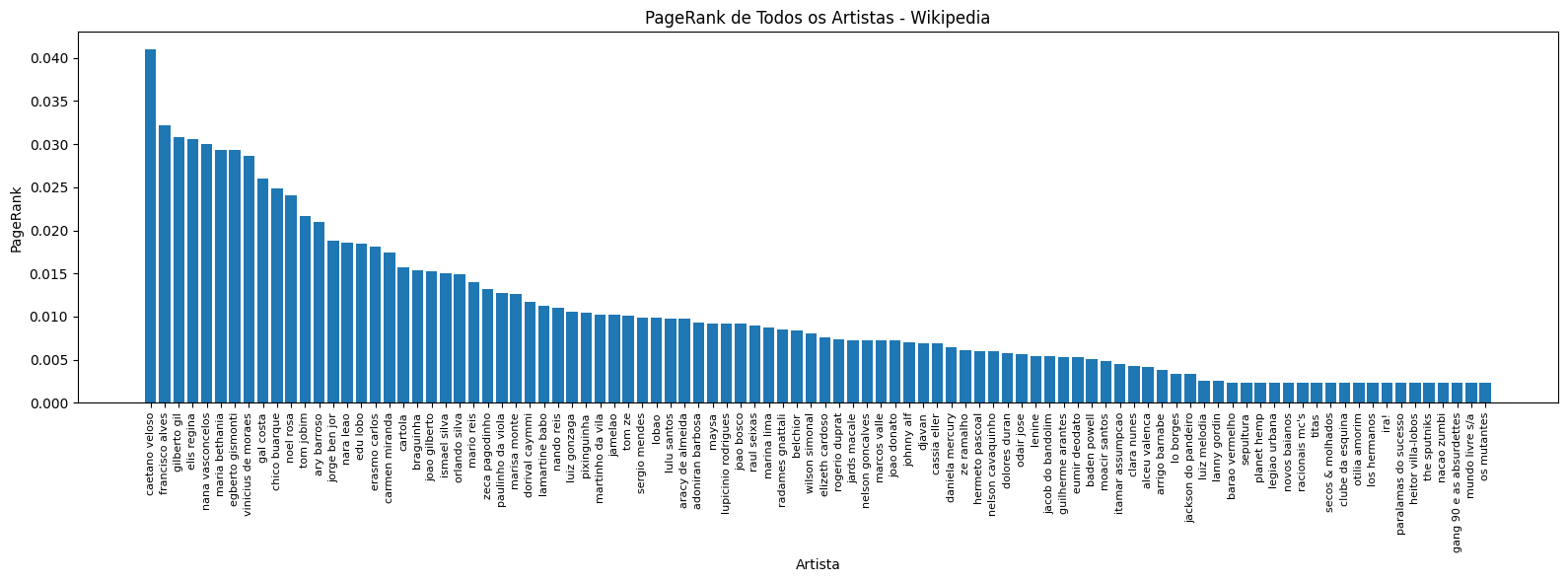
Parte 4.1 - Wikipedia

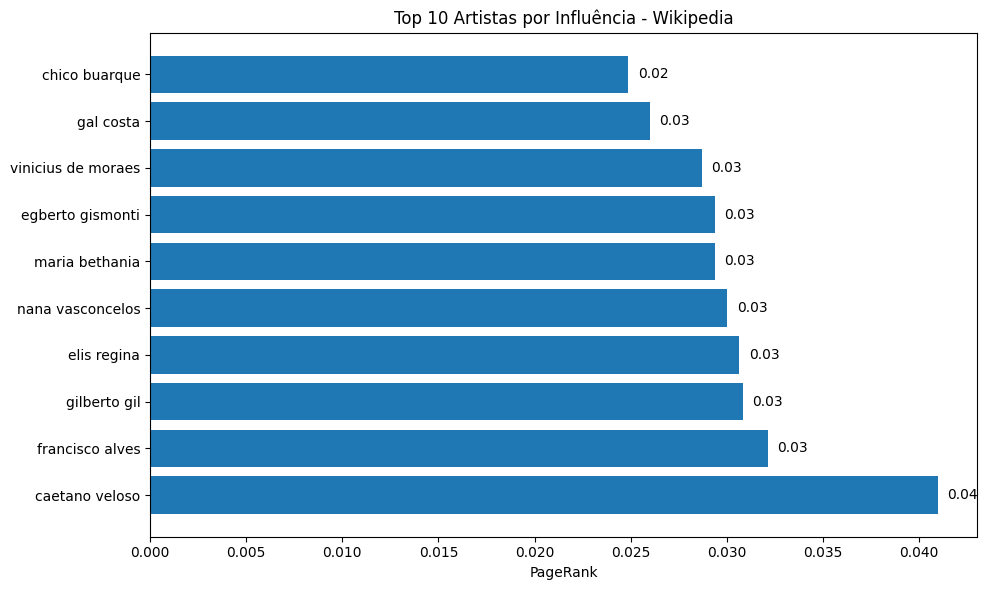
Para a implementação utilizando a Wikipédia, os artistas foram selecionados por uma lista da revista Rolling Stones Brasil- Os 100 maiores artistas da música brasileira. O conteúdo das páginas foi obtido utilizando a API da Wikipédia. Após essa etapa, o procedimento realizado foi idêntico aos descritos nas partes 1,2 e 3 deste relatório. Para fins de comparação, os resultados obtidos utilizando as letras das músicas desse mesmo grupo de artistas estarão disponíveis ao final desta seção.

Nota-se na representação da matriz de adjacência abaixo, que foi obtida a partir do corpo dos textos da Wikipédia, que a densidade do grafo obtido é muito maior, devido ao número expressivo de referências a outros artistas na página de cada um dos artistas.

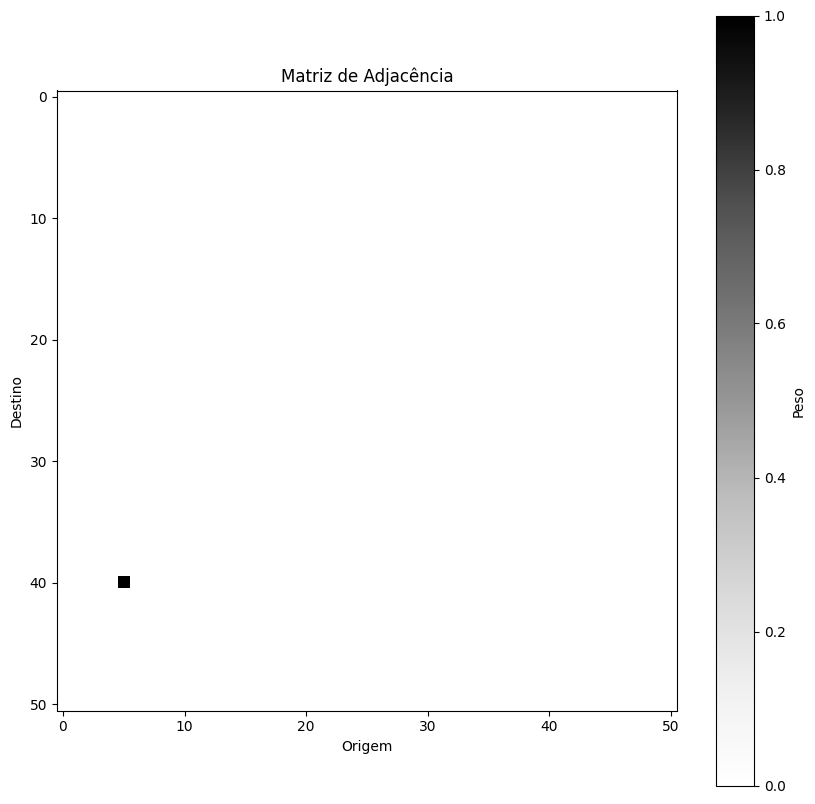


Também percebe-se uma distribuição menos uniforme no resultado final do algoritmo Page Rank, resultado direto do maior número de referências. Abaixo estão as distribuições do ranking da totalidade dos artistas contemplados e os 10 primeiros do ranking, respectivamente.

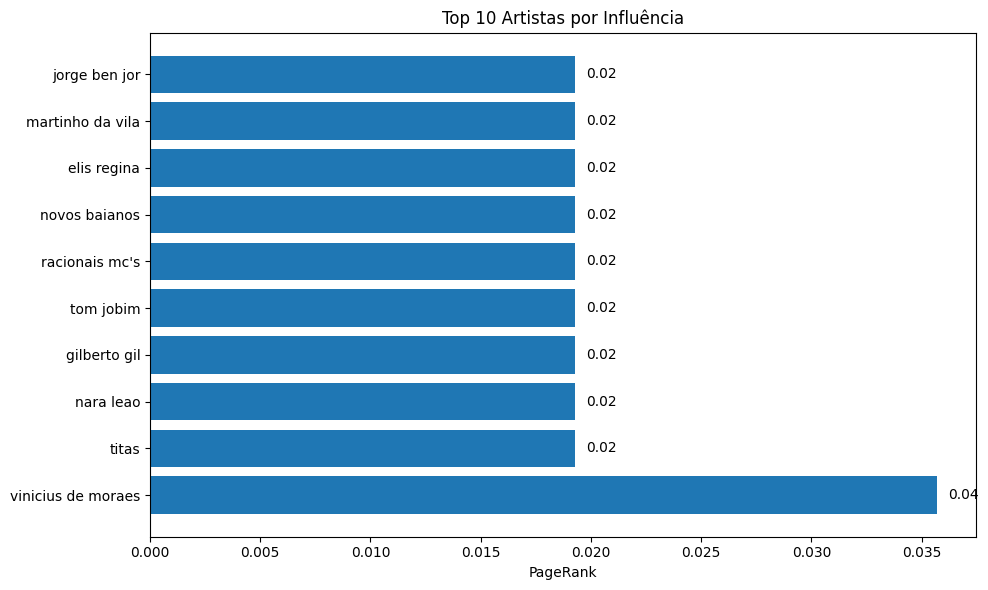




Matriz de Adjacência e Resultado do Page Rank para os 100 artistas mais influentes segundo a Rolling Stones utilizando as letras das músicas.



Percebemos que nas letras das músicas retornadas pela API há apenas uma referência captada entre os artistas presentes na lista.



Vinícius de Moraes foi o único artista que se destacou dos demais, portanto foi o destinatário da única referência.

O código utilizado para a construção deste relatório está disponível em

<https://github.com/arthur273/page-rank-musical>

Peço desculpas pelos resquícios de execução no notebook; optei por não executá-lo novamente, pois a execução anterior levou cerca de 60 minutos.