UNIVERSIDADE FEDERAL DO PARANÁ

MARCELO BATISTA DE CARVALHO

APLICAÇÃO DE TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS: UM ESTUDO NO GRAMMY SONG OF THE YEAR DATABASE

CURITIBA

2017

MARCELO BATISTA DE CARVALHO

APLICAÇÃO DE TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS: UM ESTUDO NO GRAMMY SONG OF THE YEAR DATABASE

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção de grau de Bacharel no curso de Gestão da Informação, Departamento de Ciência e Gestão da Informação do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Prof.ª Dr.ª Denise Fukumi Tsunoda

CURITIBA

2017

TERMO DE APROVAÇÃO

MARCELO BATISTA DE CARVALHO

APLICAÇÃO DE TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS: UM ESTUDO NO GRAMMY AWARD DATABASE

Trabalho apresentado como requisito parcial à obtenção do grau de bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, pela seguinte banca examinadora:

_____

Prof.ª Dr.ª Denise Fukumi Tsunoda

Orientadora - Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Dr. Cícero Aparecido Bezerra

Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. André José Ribeiro Guimarães

Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

Curitiba, 06 de dezembro de 2017

AGRADECIMENTOS

A meus pais, Cleonice e Valmor, que me deram segurança para concluir esta etapa de minha vida.

À Karen, por me suportar em momentos difíceis.

A meus colegas de ensino médio, especialmente à Elaine e ao Renato, pelo incentivo para que eu continuasse meus estudos.

Aos colegas de graduação, por enfrentarem ao meu lado os desafios da vida acadêmica.

A todos os professores do curso de Gestão da Informação, em especial a Dra. Patrícia Zeni Marchiori e Dra. Denise Fukumi Tsunoda, minha orientadora, pelo apoio e incentivo durante minha graduação.

“Your time is limited, so don’t waste it living someone else’s life. Don't be trapped by dogma - which is living with the results of other people's thinking. Don't let the noise of others' opinions drown out your own inner voice. And, most important, have the courage to follow your heart and intuition. They somehow already know what you truly want to become. Everything else is secondary.”

Steve Jobs

RESUMO

Estudo da aplicação de métodos de descoberta de conhecimento em dados de músicas indicadas ao prêmio Canção do Ano do GRAMMY Awards. Disponibilizou-se a base com dados de 300 canções participantes das 59 edições da premiação para acesso pela internet. Seguiram-se as etapas do processo de KDD, extraindo-se quatro diferentes configurações de tabelas atributo-valor no pré-processamento das letras das canções e aplicando-se os algoritmos J48 e SMO para a classificação dos registros conforme indicados e vencedores do prêmio. As taxas de precisão das classificações figuraram entre 67,46% e 80,14%. Considerando-se apenas as canções vencedoras, as melhores taxas de precisão foram de 18,96% com o J48 e 25,86% com o SMO. Os resultados foram satisfatórios do ponto de vista da Descoberta de Conhecimento em Bases de Dados, mas não são suficientes para a predição de canções vencedoras. Trabalhos futuros podem considerar outras variáveis, como popularidade das canções, ou ainda outras categorias do GRAMMY, como Melhor Canção de Rock ou Melhor Canção de Pop, por exemplo.

Palavras-chave: GRAMMY. Gestão da Informação. Descoberta de Conhecimento em Bases de Dados. Mineração de Dados. Mineração de Textos.

ABSTRACT

Study on application of knowledge discovery methods on data about GRAMMY's Song of the Year award nominees. The database with data about 300 songs participating on the 59 editions was published on internet. KDD process' stages were followed, extracting four different settings of attribute-value tables from songs' lyrics preprocessing and applying J48 and SMO algorithms for entries classification according to nominees and winners. Classifications' precision rate figured between 67.46% and 80.14%. Considering only winner songs, best precision rate was 18.96% with J48 and 25.86% with SMO. Results were satisfactory from the KDD point of view, but they are not enough to predicting winner songs. Future works may consider other variables, such as songs' popularity, or yet other categories from GRAMMY, as Best Rock Song or Best Pop Song, for instance.

Keywords: GRAMMY. Information Management. Knowledge Discovery on Databases. Data Mining. Text Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

HTML	–	Hypertext Markup Language
KDD	–	Knowledge Discovery in Databases
KDT	–	Knowledge Discovery in Texts
MINE	–	Maximum Information-based Nonparametric Exploration
RIAA	–	Recording Industry Association of America
SGBD	–	Sistema de Gerenciamento de Banco de Dados
SVM	–	Support Vector Machine
SMO	–	Sequential Minimal Optimization
TF	–	Term Frequency
TF-IDF	–	Term Frequency – Inverse Document Frequency
TF-Linear	–	Term Frequency – Linear
UFPR	–	Universidade Federal do Paraná
WoS	–	Web of Science

SUMÁRIO

1 INTRODUÇÃO

GRAMMY Award é um prêmio concedido anualmente pela Recording Academy, uma organização norte-americana de profissionais da indústria fonográfica, como forma de reconhecimento de excelência técnica e artística de músicos, compositores, intérpretes, engenheiros e outros profissionais do ramo musical, sem considerar desempenho em vendas ou posições em outras tabelas (THE RECORDING ACADEMY, 2017a).

Os membros da Academia elegem os indicados dentre os inscritos e, depois, os vencedores dentre os indicados por meio de votação. Na edição de número 59 (cinquenta e nove), relativa ao ano de 2016, foram premiadas 84 (oitenta e quatro) categorias, divididas em 31 (trinta e um) gêneros (THE RECORDING ACADEMY, 2017c).

Informações sobre os indicados e os vencedores podem ser recuperadas pelo sítio oficial da premiação[1], porém, não de forma estruturada. Além disso, poucos estudos foram realizados utilizando dados sobre o prêmio e seus indicados e vencedores.

Uma das possibilidades de análise desse conjunto de dados é a Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD), apresentada como uma alternativa para ampliar as capacidades humanas de análise e lidar com a quantidade de informação coletada (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Nesse cenário, este trabalho visa a análise de dados sobre o GRAMMY por meio de técnicas de descoberta de conhecimento em bases de dados e, para tal, a criação e publicação de uma base de dados sobre a categoria Canção do Ano. Diante disso, apresenta-se o problema de pesquisa a seguir.

1.1 PROBLEMA DE PESQUISA

Potencialmente, a utilização de técnicas de KDD numa base de dados pode revelar padrões que descrevam os dados, indicando relações previamente desconhecidas entre variáveis, ou que prevejam valores futuros. Numa base de dados

[1] O sítio do GRAMMY está disponível em: <http://www.grammy.com>.

relativos ao GRAMMY, há a possibilidade de verificar relações entre vencedores de dada categoria, identificar características que os diferenciam dos indicados não premiados e predizer, com base numa lista de indicados, qual seria o vencedor de uma premiação futura, por exemplo.

A construção de uma base de dados estruturados deve ser considerada para a aplicação do KDD neste contexto. Em relação à base, dois fatores podem ser considerados essenciais para o sucesso do estudo: as características consideradas (quais variáveis seriam relevantes para a realização das análises?) e a qualidade dos dados cadastrados (de qual fonte os dados serão recuperados?).

Diante da diversidade de técnicas de KDD existentes, outro fator importante é a escolha de heurísticas adequadas para a descoberta de conhecimentos relevantes acerca dos dados.

Neste contexto, a questão de pesquisa deste trabalho é: “A aplicação de métodos de Descoberta de Conhecimento seria adequada para a recuperação de conhecimento implícito em base de dados sobre o GRAMMY Award?”.

1.2 OBJETIVOS

Como forma de responder à questão de pesquisa apresentada anteriormente, o objetivo geral deste projeto é aplicar métodos de descoberta de conhecimento em dados de músicas indicadas aos prêmios GRAMMY.

Para alcançar tal objetivo, os seguintes objetivos específicos foram definidos:

a) construir uma base de dados sobre as canções indicadas e distribuí-la livremente pela web;
b) recuperar na literatura os métodos mais utilizados em contextos similares;
c) estudar e escolher as heurísticas a serem aplicadas à base de dados construída;
d) avaliar, por meio de comparações, os experimentos realizados.

1.3 JUSTIFICATIVA

A seguir, apresenta-se a justificativa em três níveis: para a área de Mineração de Dados, para o curso de Gestão da Informação e para o autor do trabalho.

1.3.1 Para a área de Mineração de Dados

As áreas de aplicação da Descoberta de Conhecimento em Base de Dados e da Mineração de Dados são variadas. Podem ser citadas, por exemplo, análise de crédito, análise de concorrentes, descoberta de relações entre produtos, classificação de consumidores, previsão de vendas, inferência de necessidades, descoberta de causas, entre outras (CARVALHO, 2005; SILBERSCHATZ; KORTH; SUDARSHAN, 2006).

Por outro lado, o uso em dados relativos ao GRAMMY Awards é ainda escasso. Uma pesquisa realizada na plataforma Web of Science (WoS)[2] pelos termos *grammy* e *data* (no campo *title*) retornou apenas um resultado: um trabalho dos autores Peacock e Hu (2013).

O estudo inclui a análise de dados dos prêmios GRAMMY, Emmy e Academy (ou Oscar). Quanto ao GRAMMY, os autores utilizaram dados sobre o prêmio de canção do ano num período de dez anos, incluindo posição em tabelas Billboard, números de certificação da Recording Industry Association of America (RIAA) e dados da Amazon. Aplicaram regressão logística, regressão linear e *Maximum Information-based Nonparametric Exploration* (MINE). O estudo indicou que a quantidade de vendas de uma canção é relevante, mas não prediz o sucesso no GRAMMY Awards. Os dados também indicaram uma relação complexa entre a duração da música e seu desempenho no GRAMMY (PEACOCK; HU, 2013).

Ao ampliar o escopo da pesquisa na WoS, retirando-se o termo *data*, a plataforma retornou 38 (trinta e oito) resultados. Com a pesquisa no Google Acadêmico[3] por *grammy data* (no título dos registros), recuperou-se apenas o mesmo artigo de Peacock e Hu (2013) mencionado anteriormente. De forma similar, retirando-se o termo *data*, 63 (sessenta e três) itens são retornados. A mesma busca realizada na plataforma SciELO[4], pelo termo *grammy* nos títulos dos registros, resultou em registro algum.

---

[2] Pesquisa realizada em 02 de maio de 2017 no endereço <http://apps-webofknowledge.ez22.periodicos.capes.gov.br/>.
[3] Pesquisa realizada em 02 de maio de 2017 no endereço <http://scholar.google.com.br>.
[4] Pesquisa realizada em 02 de maio de 2017 no endereço <http://search.scielo.org>.

A quantidade de estudos recuperados com as buscas realizadas na WoS, no Google Acadêmico e no SciELO ressalta a existente demanda por novas pesquisas com o tema.

1.3.2 Para o curso de Gestão da Informação da UFPR

No âmbito do curso de Gestão da Informação da Universidade Federal do Paraná (UFPR), esta pesquisa justifica-se por estar aderente ao curso ao basear-se em conhecimentos de disciplinas como Introdução à Teoria da Informação, Banco de Dados e Mineração de Dados. Além disso, o estudo fundamenta-se em conceitos que complementam os aprendidos no curso, como o de Mineração de Textos.

Ao pesquisar-se por Trabalhos de Conclusão no sítio do curso de Gestão da Informação da UFPR[5], nenhum dos documentos retornados aborda a descoberta de conhecimento em bases de dados relacionadas à música ou às premiações como o GRAMMY. Desta forma, o presente estudo contribuirá, potencialmente, para o avanço do conhecimento produzido no ambiente do Curso sobre tais temas.

1.3.3 Para o autor

O tema do presente estudo foi escolhido pelo autor devido a seu interesse na Descoberta de Conhecimento em Bases de Dados e Mineração de Dados e seu apreço pela Música. Assim, este trabalho se configurou numa oportunidade de estudar as duas áreas e aprimorar os conhecimentos sobre ambas.

1.4 DELIMITAÇÃO DA PESQUISA

O Grammy Song of the Year Database consiste numa base de dados sobre as canções indicadas ao prêmio de canção do ano do GRAMMY nas edições 1 a 59. Estes dados incluem nome dos compositores; título, letras e idioma das canções; nome do intérprete; ano e resultado da premiação. Não foram coletados dados referentes às outras premiações do GRAMMY.

[5] Pesquisa realizada em 28 de maio de 2017 no endereço <http://www.decigi.ufpr.br>.

1.5 ESTRUTURA DO DOCUMENTO

Este documento está organizado como segue. A revisão de literatura é apresentada na Seção 2, no qual são descritos os temas que embasam o estudo: Dado, Informação e Conhecimento; Banco de Dados; Descoberta de Conhecimento em Base de Dados; Mineração de Dados; Mineração de Textos; Pré-processamento de Textos e GRAMMY Awards. Na seção 3 apresenta-se o encaminhamento metodológico, por meio de sua caracterização e materiais e procedimentos metodológicos adotados. Na seção 4 são apresentados os resultados finais e análises e, na seção 5, são feitas as considerações finais.

2 REVISÃO DE LITERATURA

Esta seção apresenta a fundamentação teórica da pesquisa, abordando os GRAMMY Awards; os conceitos de dado, informação e conhecimento; bancos de dados; Descoberta de Conhecimento em Bases de Dados; Mineração de Dados e Mineração de Textos.

2.1 DADO, INFORMAÇÃO E CONHECIMENTO

Para a melhor compreensão do leitor, diferencia-se aqui *dado* de *informação* e *conhecimento*, já que, embora aparentemente similares, são conceitos que não devem ser confundidos (DAVENPORT; PRUSAK, 1998a). De Sordi (2001) apresenta as características diferenciais entre os termos, conforme QUADRO 1.

QUADRO 1 – CARACTERÍSTICAS DIFERENCIAIS ENTRE DADOS, INFORMAÇÃO E CONHECIMENTO

Características	**Dado**	**Informação**	**Conhecimento**
Estruturação, captura e transferência	Fácil	Difícil	Extremamente difícil
Principal requisito para sua geração	Observação	Interpretação consensual	Análise e reflexão
Natureza	Explícita	Predominantemente explícita	Predominantemente tácita
Percepção de valor no contexto administrativo	Baixa	Média	Grande
Foco	Operação	Controle e gerenciamento	Inovação e liderança
Abordagens administrativas que os promovem	Execução de transações de negócios, processamento de dados	Gerenciamento de sistemas de informação	Gestão do conhecimento e aprendizagem organizacional
Tecnologias que os promovem	Sistemas de processamento de dados e transações via Internet	Sistemas de informação gerenciais, sistemas analíticos, sistemas de suporte à decisão e sistemas de informações executivas	Mineração de dados, mineração de textos, sistemas de processamento de linguagem natural, sistemas especialistas e sistemas de inteligência artificial

FONTE: Adaptado de DE SORDI (2001).

Dados são simples registros de acontecimentos. De acordo com Davenport e Prusak (1998a, p. 2), são conjuntos de “[...] fatos distintos e objetivos, relativos a eventos”. Para os autores, num contexto organizacional, eles estão relacionados a registros de transações. Davenport e Prusak (1998b, p. 19) definem dados como “[...] observações sobre o estado do mundo”. Assim, eles seriam “[...] registros ou fatos em sua forma primária [...]” (BEAL, 2008, p. 12).

Por serem simples registros, os dados, por si só, não possuem significado ou valor, a menos que lhes seja atribuído por meio de contextualização, categorização, cálculo, correção ou condensação, por exemplo. Ao acrescentar-lhes significado ou valor, os dados tornam-se informação (DAVENPORT; PRUSAK, 1998a).

Desta forma, informação são dados organizados ou combinados de forma significativa (BEAL, 2008, p. 12). Neste processo, as pessoas possuem um papel importante. Para Davenport e Prusak (1998a), o receptor de uma mensagem decide se ela constitui informação ou não. Drucker (1988 apud DAVENPORT; PRUSAK, 1998b, p. 19) define informação como dados dotados de relevância e propósito, características que lhes são atribuídas por pessoas.

O conhecimento tem origem ao agregarmos valor às informações, seja por comparações, conexões, conversações ou verificação de consequências, por exemplo (BEAL, 2008; DAVENPORT; PRUSAK, 1998a). De acordo com Davenport e Prusak (1998a), o conhecimento “[...] é uma mistura fluída de experiência condensada, valores, informação contextual e *insight* experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações”. O ser humano é ainda mais importante para a geração de conhecimento, já que é ele quem dá à informação contexto, significado e interpretação (DAVENPORT; PRUSAK, 1998b).

2.2 BANCO DE DADOS

Conforme mencionado anteriormente, um dos objetivos específicos deste trabalho é a elaboração de um banco de dados (chamado de GRAMMY Song of the Year Database). Um banco de dados pode ser entendido como uma coleção de dados que descrevem as atividades de uma ou mais organizações relacionadas (RAMAKRISHNAN; GEHRKE, 2000) ou, num sentido mais amplo, como um conjunto

de dados integrados que atendem a um conjunto de sistemas ou a uma comunidade de usuários (HEUSER, 1998).

Bancos de dados são normalmente gerenciados por meio de Sistemas de Gerenciamento de Banco de Dados (SGBDs). Esses sistemas reúnem funções que são comuns a diversas aplicações de forma a facilitar o desenvolvimento destas. Segundo Silberschatz, Korth e Sudarshan (2006, p. 1), um SGBD "[...] é uma coleção de dados inter-relacionados e um conjunto de programas para acessar esses dados". Heuser (1998, p. 4) define um SGBD como um "[...] software que incorpora as funções de definição, recuperação e alteração de dados em um banco de dados".

2.3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

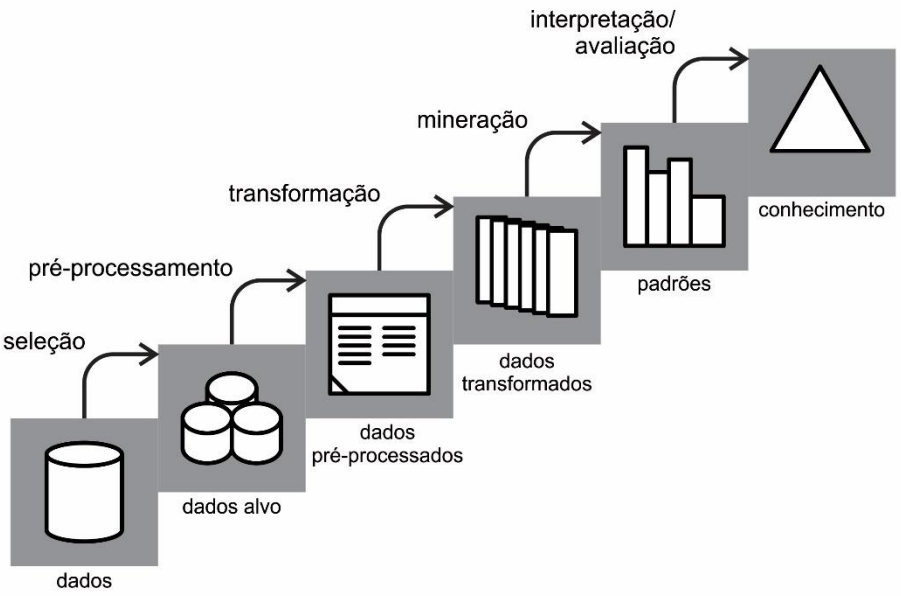
O objetivo geral deste trabalho engloba a aplicação de métodos de Descoberta de Conhecimento em Bases de Dados (ou *Knowledge Discovery in Databases* - KDD) no GRAMMY Song of the Year Database. Assim, abordam-se a seguir os conceitos básicos sobre o KDD.

O termo Descoberta de Conhecimento em Bases de Dados refere-se ao processo de busca e extração de conhecimento de bases de dados. Fayyad, Piatetsky-Shapiro e Smith (1996 apud GOLDSCHIMIDT; PASSOS, 2005) definem o KDD como "[...] um processo, de várias etapas, não trivial, interativo e iterativo, para a identificação de padrões compreensíveis, válidos, novos, e potencialmente úteis a partir de grande conjunto de dados". Para Frawley, Piatetsky-Shapiro e Matheus (1992, p. 58, tradução nossa), "a descoberta de conhecimento é a extração não-trivial de informações implícitas, previamente desconhecidas e potencialmente úteis de dados"[6]. As informações extraídas são utilizadas "[...] para aumentar os ganhos, reduzir os custos ou melhorar o desempenho do negócio [...]", por exemplo, (THOMÉ, s.d., p. 11).

De acordo com Fayyad, Piatetsky-Shapiro e Smith (1996), o KDD possui os seguintes passos (FIGURA 1):

[6] "Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data." (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992)

FIGURA 1 - VISÃO GERAL DOS PASSOS QUE COMPÕEM O PROCESSO DE KDD



FONTE: Adaptado de FAYYAD, PIATETSKY-SHAPIRO E SMYTH (1996).

a) *seleção*: é a definição do objetivo do processo do ponto de vista do usuário. Envolve a compreensão do domínio da aplicação e de conhecimentos prévios relevantes;
b) *dados alvo*: a criação de um conjunto de dados alvo inclui a seleção de um conjunto de dados ou o foco num subconjunto de variáveis ou amostra de dados, no qual a descoberta será realizada;
c) *pré-processamento*: pode incluir remoção de ruído, reunião de informações necessárias para a modelagem ou correção de ruído e decisão de estratégias para lidar com dados ausentes;
d) *decisão de método de mineração de dados*: escolha de um método específico (sumarização, classificação, regressão, agrupamento etc.) que condiga com os objetivos do processo;
e) *análise e modelo exploratório e seleção de hipótese*: escolha do algoritmo de mineração de dados e do método de seleção a ser usado na busca de padrões. Inclui a decisão de modelos e parâmetros apropriados;
f) *mineração de dados*: é a procura de padrões de interesse numa forma representacional específica ou num conjunto dessas representações. Inclui

regras ou árvores de classificação, regressão e agrupamento. Esse passo depende significativamente do bom desempenho dos anteriores;

g) *interpretação/avaliação*: interpretação dos padrões minerados, com possível retorno a algum passo anterior para interação adicional. Pode envolver a visualização dos padrões e modelos extraídos ou a visualização dos dados tendo em conta os modelos extraídos;

h) *ação sobre o conhecimento descoberto*: é o uso direto do conhecimento, incorporando-o em outro sistema para ação futura ou documentando-o e relatando para os interessados. Envolve a verificação e resolução de potenciais conflitos com conhecimentos anteriores.

Destas, a Mineração de Dados é considerada a principal etapa. Nela é de fato realizada a busca por padrões relevantes, conforme abordado a seguir.

2.4 MINERAÇÃO DE DADOS

Para Silberschatz, Korth e Sudarshan (2006, p. 496), o termo *mineração de dados* “[...] refere-se, em geral, ao processo de analisar grandes bancos de dados de forma semiautomática para encontrar padrões úteis”. Segundo AMARAL (2001), “[m]ineração de dados é o processo de busca de relacionamentos e padrões globais existentes nas bases de dados”.

Na Mineração de Dados definem-se algoritmos e técnicas a serem aplicados no problema em questão. Redes neurais, algoritmos genéticos, modelos estatísticos e probabilísticos e árvores de decisão são exemplos de técnicas utilizadas (GOLDSCHIMIDT; PASSOS, 2005; AMARAL, 2001).

Os dois principais objetivos da Mineração de Dados são a predição e a descrição. O primeiro envolve o uso de variáveis ou campos do banco de dados para predizer valores desconhecidos ou futuros de outras variáveis. O segundo trata da descoberta de padrões que descrevam os dados e sejam interpretáveis por humanos (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Fayyad, Piatetsky-Shapiro e Smith (1996) listam os seguintes métodos de Mineração de Dados:

a) *classificação:* descoberta de uma função que mapeie (classifique) um item de dados em um conjunto de classes pré-definidas;

b) *regressão:* descoberta de uma função que mapeie um item de dados em uma variável de predição de valor real;
c) *agrupamento (clusterização):* identificação de um conjunto finito de categorias (clusters) que descrevam os dados;
d) *sumarização:* busca de uma descrição compacta para um subconjunto de dados;
e) *modelagem de dependência:* busca de um modelo que descreva as dependências significativas entre as variáveis;
f) *detecção de mudança e desvio*: descoberta das mudanças mais significativas nos dados a partir de valores normativos ou previamente medidos.

Goldschimidt e Passos (2005) acrescentam dois itens a essa lista: a descoberta de associação, que consiste na "[...] busca de itens que frequentemente ocorram de forma simultânea em transações do banco de dados"; e a descoberta de sequências, uma extensão da anterior, "[...] em que são buscados itens frequentes considerando-se várias transações ocorridas ao longo de um período".

A Mineração de Dados pode ser usada em análises de crédito, análise de concorrentes, descoberta de produtos que são comprados em conjunto, descobertas de causas etc. (SILBERSCHATZ; KORTH; SUDARSHAN, 2006). Carvalho (2005, p. 3) indica o uso para "[...] descobrir relações entre produtos, classificar consumidores, prever vendas, localizar áreas geográficas potencialmente lucrativas para novas filiais, inferir necessidades, entre outras".

Carvalho (2016) indica métodos e algoritmos de Mineração de Dados mais encontrados em trabalhos sobre Mineração de Textos disponíveis na Web of Science. Entre eles estão os métodos de classificação, associação e agrupamento; e os algoritmos Bayes, Regressão, Máquinas de Vetores de Suporte (*Support Vector Machines – SVM*), K-médias, Markov, Árvores de Decisão e C4.5, entre outros.

A seguir, são tratadas duas abordagens para a classificação na Mineração de Dados: Árvores de Decisão e SVM.

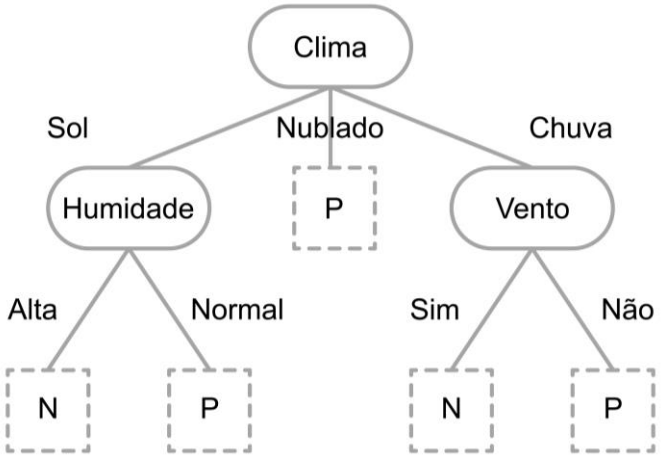
2.4.1 Árvores de decisão

Alguns sistemas de classificação recebem um conjunto de casos pré-classificados, no formato de vetores de atributos, e fazem um mapeamento dos

valores dos atributos em relação às classes. Um exemplo desses sistemas é o C4.5 que gera classificações em árvores de decisão (QUINLAN, 1996).

Uma árvore de decisão pode ser entendida como um procedimento de classificação para determinado conjunto de dados. É um diagrama composto basicamente de nós e folhas. Cada nó indica uma decisão ou teste aplicado a uma instância de dados. Cada possibilidade de decisão é indicada abaixo de um nó como linhas que o ligam a outros nós ou a folhas. O procedimento de classificação inicia, assim, pelo primeiro nó, segue os testes indicados até encontrar uma folha. As folhas indicam as classes para as instâncias de dados testadas (QUINLAN; RIVEST, 1989). Um exemplo de árvore de decisão é ilustrado na FIGURA 2.

FIGURA 2 – EXEMPLO DE ÁRVORE DE DECISÃO



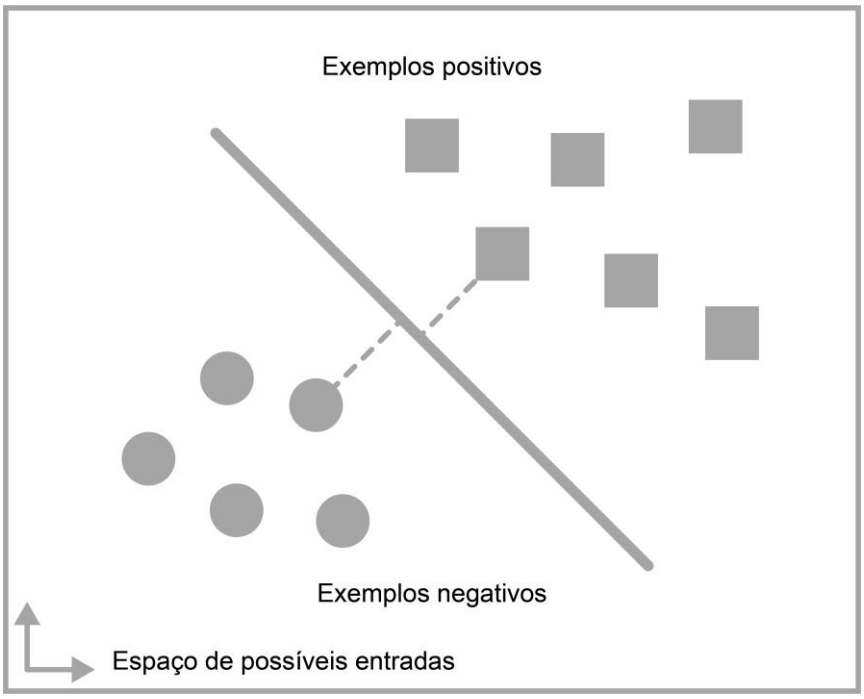
FONTE: adaptado de Quinlan e Rivest (1989).

O J48 é uma implementação do algoritmo C4.5 para a geração de árvores de decisão binárias. Maiores detalhes sobre o algoritmo podem ser encontrados no trabalho de Patil e Sherekar (2013).

2.4.2 Máquinas de Vetores de Suporte

Máquinas de Vetores de Suporte (ou SVMs) são algoritmos para classificação que utilizam um hiperplano para separar exemplos positivos de exemplos negativos, maximizando a distância entre o hiperplano e os exemplos mais próximos (FIGURA 3). SVMs podem gerar classificadores lineares ou não-lineares (PLATT, 1998).

FIGURA 3 – REPRESENTAÇÃO DO HIPERPLANO UTILIZADO EM SVM



FONTE: adaptado de Platt (1998).

Um algoritmo melhorado para treinar SVMs, chamado de Otimização Mínima Sequencial (*Sequential Minimal Optmization* – SMO), foi apresentada por Platt (1998).

2.5 MINERAÇÃO DE TEXTOS

A Mineração de Textos, também chamada de Mineração de Dados em Textos ou Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts* - KDT), é uma variante da Mineração de Dados. É o processo de extração de padrões ou conhecimento não-trivial, inesperado, útil e de interesse de documentos de texto não-estruturado (TAN, 1999). O termo *mineração de textos* é usado, em geral, para indicar um "sistema que analisa grande quantidade de texto em linguagem natural e detecta padrões de uso léxico ou linguístico na tentativa de extrair informações provavelmente úteis (ainda que apenas provavelmente corretas)"[7] (SEBASTIANI, 2002 apud WITTEN, 2004, tradução nossa).

[7] "The phrase 'text mining' is generaly used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract problably useful (although only probably correct) information." (SEBASTIANI, 2002 apud WITTEN, 2004)

Diferentemente da Mineração de Dados, a Mineração de Textos suporta conjuntos de dados não-estruturados e semiestruturados, como correios eletrônicos, arquivos de Linguagem de Marcação de Hipertexto (HTML) e documentos de texto em geral (VIJAYARANI; MUTHULAKSHMI, 2013).

Witten (2004) compara a Mineração de Textos e a Mineração de Dados em três aspectos, de acordo com os resultados gerados por ambas as metodologias: implicitude, desconhecimento prévio e utilidade potencial. Para o autor, a Mineração de Dados é caracterizada por extrair informações implícitas, previamente desconhecidas e potencialmente úteis de dados.

Enquanto na Mineração de Dados os dados de entrada possuem informações implícitas, na Mineração de Textos, as informações já estão explícitas nos textos de entrada. Além disso, as informações já são conhecidas previamente para um ser humano, ainda que de maneira relativamente lenta ao se comparar com a velocidade de computadores. Os desafios aqui estão relacionados à adequação dos textos ao processamento automático por máquinas (WITTEN, 2004).

Quanto à utilidade das informações extraídas, Witten (2004) difere as informações capazes de prover uma base para a tomada automática de decisões das informações voltadas para a compreensão humana. Na Mineração de Textos, diferentemente da Mineração de Dados, ambas as características são de difícil identificação.

Oliveira et al. (2004) indicam usos da Mineração de Textos em estudos econômicos, descobrindo associações entre países e organizações ou realizando previsões sobre tecnologias, e na Internet, revelando associações, autoridades e eixos (*hubs*) de alguma área, por exemplo.

Matsubara, Martins e Monard (2003) indicam quatro fases essenciais para a Mineração de Textos, independentemente da tarefa específica a ser executada, são elas:

a) *coleta de documentos*: recuperação de documentos relevantes para a extração de conhecimento a ser realizada;
b) *pré-processamento:* transformação dos documentos para formatos adequados à aplicação de algoritmos de extração de conhecimento;
c) *extração de conhecimento:* descoberta de padrões úteis e desconhecidos nos documentos;

d) *avaliação e interpretação dos resultados:* verificação do alcance do objetivo e da necessidade de refazer alguma etapa.

Algumas tarefas de Mineração de Texto são apresentadas a seguir, divididas entre as voltadas a textos puros e as voltadas a textos semiestruturados.

2.5.1 Mineração de texto puro

Aqui são indicadas maneiras para minerar um texto puro em linguagem natural. Witten (2004) classifica as tarefas de mineração de texto puro em três categorias, de acordo com seu objetivo: extração de informações para uso humano, avaliação da similaridade de documentos e extração de informações estruturadas.

A primeira categoria é a situação em que o resultado da mineração é voltado para o consumo humano. Como as informações extraídas não podem basear tomadas de decisão automáticas, estas tarefas estão nos limites do que é considerado Mineração de Texto (WITTEN, 2004). São elas:

a) *sumarização de texto*: a representação condensada de um texto, voltada para consumo humano. Pode resultar em extratos, resumos (indicativos ou informativos) ou resenhas, por exemplo (WITTEN, 2004). De acordo com Gong e Liu (2001), enquanto motores de busca atuam como filtros, retornando um conjunto de documentos aparentemente relevante para o usuário, sumarizadores de texto atuam como detectores de informação, produzindo resumos que possibilitam a verificação rápida dos documentos encontrados;
b) *recuperação de documentos*: é a tarefa de identificar e retornar os documentos 25 mais relevantes de um corpus, de acordo com a pesquisa de um usuário (WITTEN, 2004). Pode ser baseada em metadados dos documentos ou no texto completo;
c) *recuperação de informações*: para Turtle e Croft (1989), é a seleção de objetos de uma coleção (tais como livros, museália ou correio eletrônico, por exemplo) que podem ser de interesse do usuário, definição similar à da tarefa anterior. Porém, Witten (2004) indica que na recuperação de informações os resultados de uma busca são tratados para condensar ou extrair a informação exata procurada pelo usuário.

A avaliação da similaridade de documentos, segunda categoria, envolve o agrupamento de itens em classes pré-estabelecidas ou não, problemas comuns também na Mineração de Dados. As tarefas incluem (WITTEN, 2004):

a) *categorização de textos*: é a tarefa de atribuir documentos à uma ou mais categorias pré-definidas de acordo com seu conteúdo (WITTEN, 2004; JOACHIMS, 1998);
b) *agrupamento de documentos*: é a procura por grupos de documentos que são similares em algum aspecto. Porém, os grupos não são pré-estabelecidos como na categorização de texto. De acordo com Zamir e Etzioni (1998), esta técnica é utilizada para melhorar o desempenho de ferramentas de busca, pré-agrupando todo o corpus ou aplicando a técnica após a recuperação;
c) *identificação de idioma*: reconhecimento da língua utilizada no texto com base nas sequências de letras encontradas;
d) *atribuição de autoria*: categorização de textos voltada para a descoberta de autores de um documento;
e) *identificação de frases-chave*: atribuição de palavras-chave ou frases-chave para representar um documento.

A extração de informação estruturada compreende tarefas que buscam informações como endereços, números de telefone, sumários, resumos, listas de referências e etc. podemos citar (WITTEN, 2004):

a) *extração de entidades*: reconhecimento de construções que representam objetos do mundo, como nomes de pessoas, endereços, datas, valores monetários etc.;
b) *extração de informações*: a tarefa de completar modelos, ou fichas, a partir de textos em linguagem natural;
c) *aprendizagem de regras a partir de textos*: um passo além da extração de informações, é a identificação de regras que representam o conteúdo dos textos.

2.5.2 Mineração de texto semiestruturado

São abordadas a seguir as tarefas para minerar textos semiestruturados, presentes na web. São elas: indução de *wrappers*, agrupamento de documentos com links, e determinação 26 de autoridade de documentos web.

A primeira tarefa refere-se à inferência automática de *wrappers* para a recuperação de informações em páginas da web. *Wrappers* são procedimentos de extração de conteúdo de um determinado recurso. É um tipo de mineração de textos em que a entrada deve seguir um formato determinado para que a informação seja extraída de maneira algorítmica. Elaborar manualmente os procedimentos para extração pode se tornar uma tarefa demorada, com difícil escalabilidade e com suscetibilidade a erros. Assim, a elaboração automática de *wrappers* é uma alternativa que visa minimizar tais problemas (KUSHMERICK; WELD; DOORENBOS, 1997; WITTEN, 2004).

O agrupamento de documentos com *links* é a procura de grupos nos documentos de acordo com links para outros documentos, além de seu conteúdo. Para este agrupamento, considera-se um grafo em que os nós são documentos web e os elos são *hyperlinks* entre eles. A quantidade de elos que separam um documento de outro, a existência de documentos ancestrais e/ou descendentes em comum, além do próprio conteúdo dos documentos analisados, são fatores que podem ser utilizados no agrupamento (WITTEN, 2004; AGRAWAL; BATRA, 2013).

Outra tarefa baseada na estrutura de *links* é a determinação de autoridade de documentos web. Os documentos são pontuados de acordo com o número de outros documentos que o citam. A pontuação dos documentos citantes também pode ser considerada, assim como os *links* presentes no documento analisado. Esta tarefa é utilizada em motores de busca na ordenação dos resultados das pesquisas (WITTEN, 2004; AGRAWAL; BATRA, 2013).

2.6 PRÉ-PROCESSAMENTO DE TEXTOS

Conforme mencionado anteriormente, o pré-processamento de textos é considerado uma das etapas do processo de Mineração de Textos. Neste contexto, o pré-processamento consiste na transformação de dados em documentos textuais em um formato estruturado, como uma tabela atributo-valor (conforme TABELA 1), para

que então sejam aplicados algoritmos de aprendizado de máquina (MATSUBARA; MARTINS; MONARD, 2003; SOARES; PRATI; MONARD, 2008).

TABELA 1 – TABELA ATRIBUTO VALOR PARA REPRESENTAÇÃO DE DOCUMENTOS

	t_1	t_2	t_3	...	t_M	C
d_1	a_{11}	a_{12}	a_{13}	...	a_{1M}	c_1
d_2	a_{21}	a_{22}	a_{23}	...	a_{2M}	c_2
d_3	a_{31}	a_{32}	a_{33}	...	a_{3M}	c_3
...	...	...	...	...	...	...
d_N	a_{N1}	a_{N2}	a_{N3}	...	a_{NM}	c_N

FONTE: adaptado de Santos, Pradi e Monard (2008).

Uma das abordagens possíveis para esse processo é o uso de *bag-of-words*, na qual se representa cada documento como um vetor de palavras presentes no documento. O conjunto dos vetores pode ser representado em tabela chamada atributo-valor. Essas tabelas são compostas de uma linha para cada vetor *d* e uma coluna para cada palavra (ou atributo) *t*, sendo uma interseção de linha e coluna a_{12} um valor para a palavra t_2 no documento d_1, por exemplo. Na última coluna da tabela pode encontrar-se a classe *C* do documento (MATSUBARA; MARTINS; MONARD, 2003; SANTOS; PRATI; MONARD, 2008).

2.6.1 Escolha de atributos

Alguns métodos podem ser utilizados para a redução da quantidade de atributos no momento da transformação dos documentos em dados estruturados, de maneira a buscar maior eficácia na posterior extração de conhecimentos. São abordados a seguir a *tokenização*, o *stemming*, a remoção de palavras vazias (*stopwords*), os cortes baseados em frequência e os n-gramas.

A *tokenização* consiste na divisão do documento por palavras, também chamadas de *tokens*. Para a realização deste processo de forma automática, é necessária a remoção anterior de caracteres indesejados, como sinais de pontuação (SOARES; PRATI; MONARD, 2008).

O *stemming* é um processo de redução das palavras a seus radicais. Assim, diferentes variações de uma palavra são representadas por um radical comum, também chamado de *stem*. O processo de *stemming* é composto pela remoção de prefixos e de sufixos (SOARES; PRATI; MONARD, 2008).

Algumas palavras presentes no documento podem ser consideradas irrelevantes por serem muito comuns ao idioma em questão, como artigos, pronomes, preposições, conjunções etc. A elas dá-se o nome de palavras vazias (*stopwords*). Tais palavras podem ser removidas do processamento do texto, de maneira a otimizar os resultados obtidos.

Outra forma de redução que pode ser utilizada é o corte de atributos de acordo com sua frequência. Luhn (1958 apud SOARES; PRATI; MONARD, 2008) propôs o uso de dois pontos de corte para excluir atributos irrelevantes por serem muito comuns ou muito raros. Assim, os atributos mais relevantes estariam entre os pontos de corte e seriam nem tão comuns e nem tão raros. Ainda assim, a determinação dos pontos de corte envolve certa arbitrariedade.

Ainda, os n-gramas são agrupamentos de palavras consecutivas para melhorar a relevância dos atributos e o poder de predição. O *n* indica a quantidade de palavras unidas para a criação de um atributo. Geralmente, muitos atributos gerados contêm palavras unidas simplesmente pelo acaso, porém os atributos mais frequentes podem ser bastante relevantes para o aprendizado (SOARES; PRATI; MONARD, 2008).

2.6.2 Valores dos atributos

Dentre as possibilidades de atribuição valores aos atributos dos vetores podem ser utilizados valores booleanos, term frequency (tf), term frequency linear (tf-linear) ou term frequency – inverse document frequency (tf-idf), por exemplo (MATSUBARA; MARTINS; MONARD, 2003; SANTOS; PRATI; MONARD, 2008).

Para valores booleanos atribui-se valores diferentes para a presença e a ausência da palavra no documento, geralmente 1 e 0 ou *verdadeiro* e *falso*, respectivamente.

Com o *term frequency* (*tf*), contabiliza-se as aparições da palavra no documento (frequência absoluta), conforme equação (1).

$$a_{ij} = freq(t_j, d_i) \quad (1)$$

De forma similar ao anterior, o método *term frequency linear* (*tf-linear*) contabiliza a frequência das palavras, porém aquelas que aparecem na maioria dos documentos têm um peso de representação menor. O fator de ponderação utilizado é linear, conforme equação (2).

$$a_{ij} = freq(t_j, d_i) \times (1 - \frac{d(t_j)}{N}) \quad (2)$$

Outra medida para ponderação do peso de representação das palavras é o *term frequency – inverse document frequency* (*tf-idf*). Neste caso, o fator de ponderação é inversamente proporcional ao logaritmo do número de documentos em que a palavra aparece no número total de documentos, como indica a equação (3).

$$a_{ij} = freq(t_j, d_i) \times (log \frac{N}{d(t_j)}) \quad (3)$$

Em conjunto com essas medidas podem ser utilizadas suavizações e normalizações. Suavizações são normalmente empregadas quando uma palavra aparece em todos os documentos, de forma que os fatores de ponderação *idf* e *linear* se tornariam nulos, o que inutilizaria o termo na tabela atributo-valor. A suavização altera os valores de ponderação temporariamente para evitar esses valores nulos. Normalizações são utilizadas para solucionar problemas relacionados a diferenças de tamanho entre os documentos, o que resulta em diferenças também nas frequências dos termos. As normalizações podem ser realizadas por linhas (documentos) ou por colunas (palavras), de forma linear ou quadrática (MATSUBARA; MARTINS; MONARD, 2003; SANTOS; PRATI; MONARD, 2008).

2.7 GRAMMY AWARDS

GRAMMY Awards são prêmios relativos à indústria fonográfica apresentados anualmente pela Recording Academy, concedidos pelos membros votantes da

Academia como reconhecimento de excelência nas artes e ciências fonográficas (THE RECORDING ACADEMY, 2017a).

O número de categorias já chegou a ser maior que cem, mas se mantém próximo aos oitenta nas últimas edições da premiação (LOS ANGELES TIMES, 2015). A edição de número 59 (cinquenta e nove), relativa a 2016, premiou 84 (oitenta e quatro) categorias, divididas em 31 (trinta e um) grandes grupos. Gravação do ano, álbum do ano, canção do ano e artista revelação compõem as categorias no principal grupo, as categorias gerais. Os outros grupos são chamados de grupos de gênero e incluem as categorias melhor gravação, melhor desempenho e melhor álbum para gêneros como country, gospel, pop, rap e rock, por exemplo (THE RECORDING ACADEMY, 2017c). Uma relação das categorias da edição 59 (cinquenta e nove) é apresentada no Apêndice A.

2.6.1 Recording Academy

A Recording Academy (Academia Fonográfica, em tradução livre) é uma organização estadunidense de músicos, compositores, produtores, engenheiros e profissionais fonográficos, fundada em 1957. De acordo com THE RECORDING ACADEMY (2017a, não paginado, tradução nossa), é o “principal canal para honrar realizações nas artes fonográficas e apoiar a comunidade musical”.

Reconhecida principalmente pelo GRAMMY Awards, a Academia também busca a melhoria da condição cultural e qualidade de vida de músicos (THE RECORDING ACADEMY, 2017a), sendo responsável por “inovações em desenvolvimento profissional, enriquecimento cultural, advocacia, educação e programas de serviços humanos” (THE RECORDING ACADEMY, 2017b, não paginado, tradução nossa).

A Academia é gerenciada por um grupo de profissionais experientes na área da música. Eles supervisionam doze capítulos regionais em cidades dos Estados Unidos. Cada capítulo é responsável pelo desenvolvimento de suas comunidades musicais locais (THE RECORDING ACADEMY, 2017b).

Fazem parte da Recording Academy (THE RECORDING ACADEMY, 2017b):

a) *Grammy Foundation*: organização que visa contribuir para o entendimento, apreciação e avanço da contribuição da música gravada para a cultura norte-americana;

b) *Music Cares*: uma rede de assistência a pessoas ligadas à música e em situação de necessidade;
c) *Advocacy & Government Relations*: escritório que busca representar artistas, compositores e outros profissionais da música em assuntos políticos norte-americanos;
d) *The Latin Recording Academy*: organização responsável pelo GRAMMY Latino, premiação equivalente ao GRAMMY original com enfoque em música latina;
e) *GRAMMY Museum*: museu, situado na cidade de Los Angeles, que oferece atrações relacionadas à música.

Sendo organizadora do GRAMMY Awards, a Recording Academy gerencia o processo de votação nos indicados e vencedores da premiação, conforme abordado a seguir.

2.6.2 O processo de votação

O processo de votação do GRAMMY inicia com membros da Recording Academy e gravadoras submetendo inscrições, que são verificadas quanto à elegibilidade e à categorização. Então, os membros votantes participam do processo de escolha primeiramente dos cinco finalistas e depois do vencedor de cada prêmio. Cada etapa do processo é detalhada a seguir (THE RECORDING ACADEMY, 2017a):

a) *submissão*: os membros da Academia e as companhias gravadoras inscrevem gravações e vídeos musicais, publicados durante o ano de elegibilidade, que julgam merecedores de reconhecimento pelo GRAMMY;
b) *verificação*: especialistas em vários domínios se certificam de que as inscrições estão de acordo com qualificações e estão categorizadas adequadamente (como rock, country, gospel etc.). Nesta etapa, não são feitos julgamentos artísticos ou técnicos;
c) *indicação*: a primeira rodada de cédulas é enviada para os membros votantes. Os membros podem votar em até quinze categorias de gênero mais as quatro categorias gerais (gravação do ano, álbum do ano, canção do ano e artista revelação) de acordo com suas áreas de especialidade. As células são tabuladas por uma empresa independente. Em algumas

categorias especializadas, os indicados são decididos por um comitê nacional;

d) *votação final*: a última rodada de cédulas é enviada aos membros votantes. Os indicados pelo comitê especial são incluídos nestas cédulas. Os membros, mais uma vez, votam em quinze categorias de gênero mais as quatro categorias gerais. Os resultados são novamente tabulados por uma empresa independente;

e) *resultados*: os nomes dos vencedores dos prêmios são mantidos em segredo até a cerimônia de premiação, quando os resultados são entregues em envelopes selados pela empresa responsável pela tabulação. Os vencedores são revelados durante a transmissão da cerimônia.

Os temas apresentados são o aporte teórico da pesquisa. Com base neles, previram-se os procedimentos metodológicos abordados a seguir.

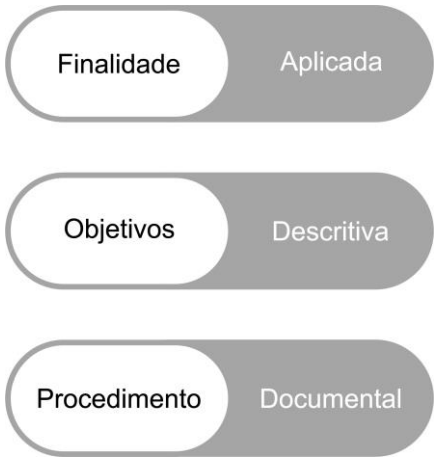
3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção aborda a caracterização da pesquisa, os materiais e métodos empregados, os custos estimados e o cronograma de atividades da pesquisa.

3.1 CARACTERIZAÇÃO DA PESQUISA

Caracteriza-se a pesquisa de acordo com sua finalidade, objetivos e procedimentos, conforme Gil (2002) e Gil (2008). Esta caracterização é indicada na FIGURA 4.

FIGURA 4 - CARACTERIZAÇÃO DA PESQUISA



FONTE: o autor (2017).

Quanto à finalidade, trata-se de uma pesquisa aplicada, uma vez que seu interesse é a utilização do conhecimento. Segundo Gil (2008), ainda que dependam das descobertas de pesquisas puras e se enriqueçam com elas, pesquisas aplicadas estão menos voltadas para o "[...] desenvolvimento de teorias de valor universal que para a aplicação imediata numa realidade circunstancial" (GIL, 2008, p. 27).

Considerando-se os objetivos apresentados, esta é uma pesquisa descritiva, devido à "[...] descrição das características de determinada população ou fenômeno ou, então, o estabelecimento de relações entre variáveis" (GIL, 2002, p. 42). A exposição das características se dará com a reunião de dados e a criação de uma base e a busca por relações entre as variáveis coletadas, com a aplicação de métodos de mineração de dados.

Já quanto aos procedimentos, esta pesquisa é documental por utilizar dados disponíveis principalmente no sítio do GRAMMY e na ferramenta Genius[8]. Para Gil (2008), a pesquisa documental diferencia-se por valer-se de "materiais que não receberam ainda um tratamento analítico, ou que ainda podem ser reelaborados de acordo com os objetivos da pesquisa" (GIL, 2008, p. 51).

3.2 MATERIAIS E MÉTODOS

A metodologia utilizada neste estudo baseou-se no processo do KDD apresentado anteriormente, conforme a seguir.

As etapas de *seleção* e *dados-alvo* incluiu a construção da base de dados GRAMMY Song of the Year Database. A base é composta de dados sobre o prêmio de canção do ano nas edições 1 a 59 do GRAMMY, a saber: nome dos compositores; título, idioma e letras das canções; nome do intérprete; ano e resultado da premiação.

Tais dados foram obtidos no sítio do evento e reunidos em planilha eletrônica com a ferramenta Google Sheets[9], com exceção das letras das músicas, obtidas no sítio Genius e gravadas em arquivos de texto separados em pastas de acordo com o ano da premiação.

No *pré-processamento*, *redução* e *projeção de dados* utilizou-se a ferramenta PreTexT2[10], gerando tabelas de atributo-valor com os termos utilizados nas canções. O PreTexT2 realiza primeiramente uma limpeza dos documentos, removendo pontuações, símbolos e palavras vazias (*stopwords*). Então são gerados n-gramas de acordo com as configurações estabelecidas. Por fim, os n-gramas são utilizados na criação da tabela de atributo-valor.

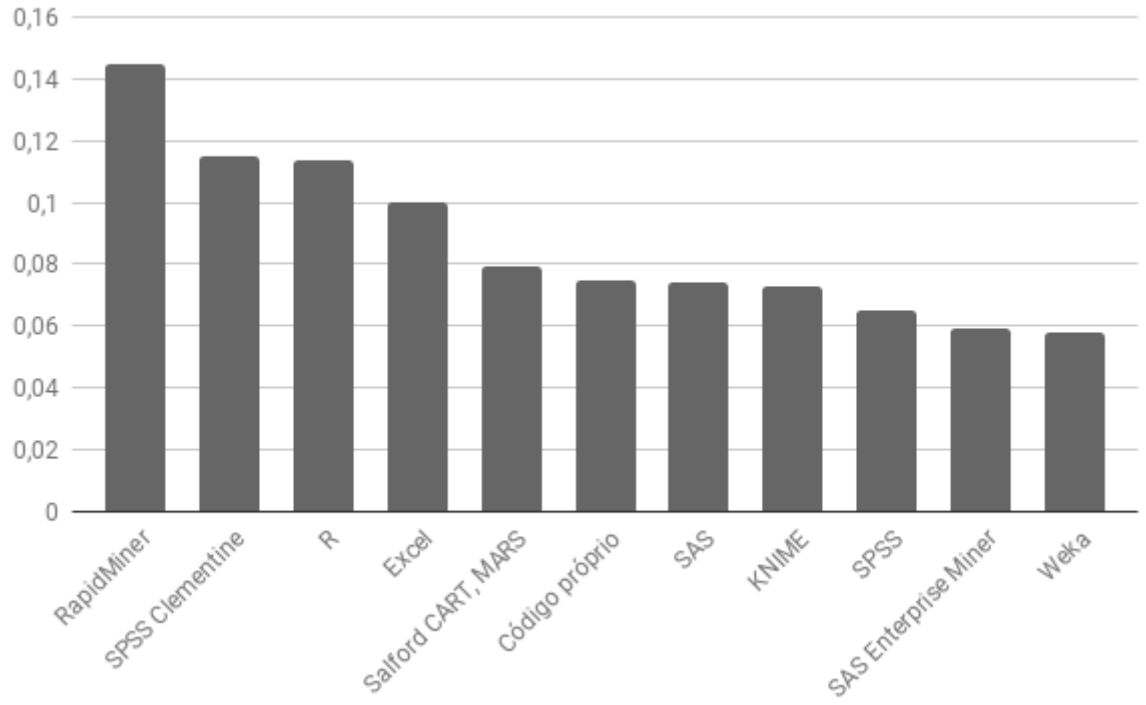
Na etapa de *decisão de método de mineração de dados* decidiu-se aplicar métodos de classificação na tabela atributo-valor gerada anteriormente, buscando identificar padrões para classificar corretamente uma canção como vencedora ou não do prêmio. Nas etapas seguintes, *análise e modelo exploratório* e *seleção de hipótese*, foram escolhidos os algoritmos J48 e SMO.

---

[8] Genius é um sítio para recuperação de letras de músicas, disponível em: <https://genius.com>.
[9] Google Sheets, ou Planilhas Google, é uma ferramenta para criação e edição de planilhas eletrônicas, disponível em: <https://www.google.com/sheets/about/>.
[10] PreTexT2 é uma ferramenta de pré-processamento de textos implementada pelo Laboratório de Inteligência Computacional (LABIC) da Universidade de São Paulo (USP). Disponível em <http://sites.labic.icmc.usp.br/pretext2/index.html>.

Nesta fase, como parâmetro de escolha de *softwares* a serem utilizados para a aplicação das heurísticas, verificou-se pesquisa realizada por Wang, Wang e Gu (2011), na qual se avaliaram o fator de impacto de 38 (trinta e oito) ferramentas de mineração de dados utilizadas em projetos. As onze ferramentas com melhores resultados são indicadas na FIGURA 5.

FIGURA 5 – FATOR DE IMPACTO DE SOFTWARES DE MINERAÇÃO DE DADOS



FONTE: Adaptado de WANG, WANG E GU (2011).

Devido ao uso anterior durante as disciplinas do curso de Gestão da Informação, a ferramenta Weka[11] foi escolhida para aplicação das heurísticas de mineração de dados desta etapa da pesquisa.

Diante das características do problema em questão, avaliou-se que o método de classificação seria o mais adequado para aplicação na base de dados. Considerando-se os algoritmos de classificação elencados por Carvalho (2016) quais estão disponíveis no Weka, optou-se pelo uso do J48 e o SMO.

[11] Weka é uma ferramenta que disponibiliza uma coleção de algoritmos para tarefas de mineração de dados, disponível em: <http://www.cs.waikato.ac.nz/ml/weka>.

A última etapa realizada neste trabalho foi a interpretação e avaliação dos resultados obtidos. Nesta etapa, os resultados foram analisados e comparados, sendo eles abordados na próxima seção.

4 RESULTADOS E ANÁLISES

O primeiro objetivo estabelecido para este trabalho engloba a construção de uma base de dados relacionada ao GRAMMY e sua disponibilização na Internet. Como resultado, a GRAMMY Song of the Year Database foi alimentada com registros das 300 canções indicadas ao prêmio de canção do ano entre os anos de 1959 e 2017.

A base de dados foi submetida para inclusão na coleção do sítio UCI Machine Learning Repository[12]. Após aprovação pela equipe administradora do repositório, a base poderá ser livremente descarregada e utilizada para estudos posteriores.

4.1 ANÁLISE DA BASE DE DADOS

Com 450 compositores com obras indicadas ao prêmio de Canção do Ano, o total de indicações foi 589 considerando que uma canção pode possuir mais de um compositor e um compositor pode ser indicado mais de uma vez, com canções diferentes. A TABELA 2 lista os onze compositores com maior número de indicações. Lionel Richie e Paul McCartney dividem a primeira posição na tabela, com seis nomeações cada (1,02% do total de indicações).

[12] A UCI Machine Learning Repository é uma coleção de bases de dados utilizados para análise empírica de algoritmos de Aprendizado de Máquina. Disponível em: <https://archive.ics.uci.edu/ml/index.php>.

TABELA 2 – COMPOSITORES COM MAIOR NÚMERO DE INDICAÇÕES AO GRAMMY DE CANÇÃO DO ANO

COMPOSITOR	INDICAÇÕES		
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA RELATIVA ACUMULADA (%)
Lionel Richie	6	1,02	1,02
Paul McCartney	6	1,02	2,04
Burt Bacharach	5	0,85	2,89
John Lennon	5	0,85	3,74
Alan Bergman	4	0,68	4,42
Billy Joel	4	0,68	5,10
Jimmy Van Heusen	4	0,68	5,78
Max Martin	4	0,68	6,46
Sammy Cahn	4	0,68	7,14
Sting	4	0,68	7,82
Will Jennings	4	0,68	8,50
Outros	539	91,50	100,00
Total	589	100,00	100,00

FONTE: o autor (2017).

A TABELA 3 lista os compositores com maior número de canções vencedoras no GRAMMY de Canção do Ano. Nove nomes foram premiados duas vezes cada nesta categoria e outros noventa e sete receberam uma premiação cada.

TABELA 3 – COMPOSITORES COM MAIOR NÚMERO DE PREMIAÇÕES NO GRAMMY DE CANÇÃO DO ANO

COMPOSITOR	INDICAÇÕES		
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA RELATIVA ACUMULADA (%)
Adam Clayton	2	1,74	1,74
Adele Adkins	2	1,74	3,48
David Evans	2	1,74	5,22
Henry Mancini	2	1,74	6,96
James Horner	2	1,74	8,70
Johnny Mercer	2	1,74	10,44
Larry Mullen Jr.	2	1,74	12,18
Paul Hewson	2	1,74	13,92
Will Jennings	2	1,74	15,66
Outros	97	84,34	100,00
Total	115	100,00	100,00

FONTE: o autor (2017).

As 300 canções indicadas possuem 238 intérpretes listados, num total de 326 indicações (cada canção pode possuir mais de um intérprete e cada intérprete pode ser indicado mais de uma vez com canções diferentes), conforme TABELA 4. Frank Sinatra é o intérprete com maior número de registros, somando sete canções indicadas (2,15% do total de indicações).

TABELA 4 – INTÉRPRETES COM MAIOR NÚMERO DE CANÇÕES INDICADAS AO GRAMMY DE CANÇÃO DO ANO

COMPOSITOR	INDICAÇÕES		
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA RELATIVA ACUMULADA (%)
Frank Sinatra	7	2,15	2,15
Barbra Streisand	5	1,53	3,68
The Beatles	5	1,53	5,21
Tony Bennett	5	1,53	6,74
Billy Joel	4	1,23	7,97
Glen Campbell	4	1,23	9,20
U2	4	1,23	10,43
Vários Artistas	4	1,23	11,66
Outros	288	88,34	100,00
Total	326	100,00	100,00

FONTE: o autor (2017).

Considerando-se apenas os vencedores, sete intérpretes tiveram duas músicas premiadas como Canção do Ano cada e outros 52 tiveram uma música premiada cada, conforme TABELA 5.

TABELA 5 – INTÉRPRETES COM MAIOR NÚMERO DE CANÇÕES VENCEDORAS DO GRAMMY DE CANÇÃO DO ANO

COMPOSITOR	INDICAÇÕES		
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA RELATIVA ACUMULADA (%)
Adele	2	3,03	3,03
Barbra Streisand	2	3,03	6,06
Bette Midler	2	3,03	9,09
Eric Clapton	2	3,03	12,12
Henry Mancini	2	3,03	15,15
Roberta Flack	2	3,03	18,18
U2	2	3,03	21,21
Outros	52	78,79	100,00
Total	66	100,00	100,00

FONTE: o autor (2017).

4.2 MINERAÇÃO DE DADOS

Para a etapa de mineração de dados, elegeram-se os algoritmos J48 e SMO para aplicação na base de dados pré-processada, referendados pela literatura pesquisada. Os testes foram feitos em 4 versões da tabela atributo-valor. A primeira foi construída com valores booleanos (0 ou 1) para os atributos e as demais com frequência TF. Nas duas primeiras versões não se realizaram cortes de atributos mais frequentes e menos frequentes. Na terceira e na quarta, aplicaram-se cortes de 0,5 desvio padrão e 1 desvio padrão, respectivamente.

Dentre as versões estudadas, a base com valores em TF e corte de 0,5 desvio padrão foi a com maior precisão de classificação no J48, obtendo 217 das 300 canções classificadas corretamente (74,31%). Por outro lado, a versão com valores em TF e sem cortes classificou corretamente 11 canções entre as 58 premiadas (18,96%), maior valor obtido com o J48 (TABELA 6).

TABELA 6 – PRECISÃO DAS CLASSIFICAÇÕES COM O J48 EM DIFERENTES CONFIGURAÇÕES DA TABELA ATRIBUTO-VALOR

VERSÃO DA TABELA	CANÇÕES CLASSIFICADAS CORRETAMENTE		CANÇÕES PREMIADAS CLASSIFICADAS CORRETAMENTE	
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)
Valores booleanos	197	67,46	9	15,51
Valores em TF sem corte	205	70,20	11	18,96
Valores em TF com corte de 0,5 desvio padrão	217	74,31	5	8,62
Valores em TF com corte de 1 desvio padrão	204	69,83	9	15,51

FONTE: o autor (2017).

Aplicando o SMO, os resultados foram similares. A base com valores em TF e corte de 0,5 desvio padrão também obteve a maior precisão (234 das 300 canções classificadas corretamente - 74,31%). Com este algoritmo a segunda versão, com valores em TF e sem cortes, também se obteve o maior número de canções premiadas classificadas corretamente, 15 canções de 58 premiadas (25,86%) (TABELA 7).

TABELA 7 – PRECISÃO DAS CLASSIFICAÇÕES COM O SMO EM DIFERENTES CONFIGURAÇÕES DA TABELA ATRIBUTO-VALOR

VERSÃO DA TABELA	CANÇÕES CLASSIFICADAS CORRETAMENTE		CANÇÕES PREMIADAS CLASSIFICADAS CORRETAMENTE	
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)
Valores booleanos	224	76,71	8	13,79
Valores em TF sem corte	224	76,71	15	25,86
Valores em TF com corte de 0,5 desvio padrão	234	80,14	0	0
Valores em TF com corte de 1 desvio padrão	225	77,05	8	13,79

FONTE: o autor (2017).

A TABELA 8 compara os valores obtidos com os dois algoritmos estudados apenas na tabela atributo-valor com valores em TF e sem cortes, versão que obteve melhores resultados entre as canções premiadas.

TABELA 8 – COMPARAÇÃO DOS RESULTADOS DOS ALGORITMOS J48 E SMO APLICADOS NA TABELA ATRIBUTO-VALOR COM VALORES EM TF E SEM CORTES

ALGORITMO	CANÇÕES CLASSIFICADAS CORRETAMENTE		CANÇÕES PREMIADAS CLASSIFICADAS CORRETAMENTE	
	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA ABSOLUTA	FREQUÊNCIA RELATIVA (%)
J48	205	70,20	11	18,96
SMO	224	76,71	15	25,86

FONTE: o autor (2017).

Conforme a TABELA 8 indica, usando-se o algoritmo J48, 11 canções vencedoras foram classificadas corretamente (18,96%). Assim, 47 canções

vencedoras (81,04%) receberam classificação equivocada pelo algoritmo. Dentre as canções indicadas, 40 (17,09%) foram classificadas como vencedoras e 194 (82,91%) como apenas indicadas. A TABELA 9 apresenta a matriz de confusão com tais resultados.

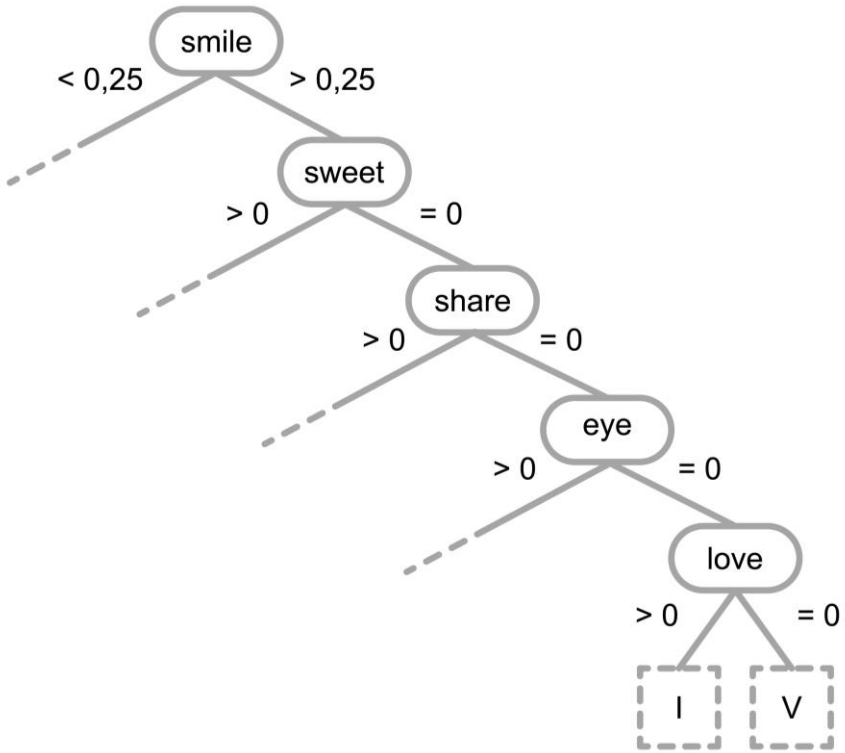
TABELA 9 – MATRIZ DE CONFUSÃO DO ALGORITMO J48 APLICADO NA TABELA ATRIBUTO-VALOR COM VALORES EM TF E SEM CORTES

CLASSE	CLASSIFICADAS COMO		TOTAL
	VENCEDORAS	INDICADAS	
Vencedoras	11	47	58
Indicadas	40	194	234
Total	51	241	292

FONTE: o autor (2017).

A classificação do J48 é baseada em árvores de decisão, conforme abordado anteriormente. Neste caso, a árvore gerada possui 38 folhas. Uma seção da árvore é ilustrada na FIGURA 6, indicando que seriam vencedoras canções cujas letras possuam a palavra *smile* com peso maior que 0,25 e não possuam as palavras *sweet*, *share*, *eye* ou *love*.

FIGURA 6 – SEÇÃO DA ÁRVORE DE DECISÃO RESULTANTE DO J48 APLICADO NA TABELA ATRIBUTO-VALOR COM VALORES EM TF E SEM CORTES



FONTE: o autor (2017).

As árvores de decisão geradas pelo J48 são inteiramente apresentadas no Apêndice B.

Ainda conforme a TABELA 8, 15 canções vencedoras foram classificadas corretamente (25,86%) com o SMO. Assim, 43 canções vencedoras (74,14%) foram classificadas equivocadamente. Nas canções indicadas, o algoritmo classificou 25 (10,68%) como vencedoras e 209 (89,32%) corretamente como apenas indicadas. A TABELA 10 aponta tais resultados com a matriz de confusão.

Sob o ponto de vista da Descoberta de Conhecimento em Bases de Dados, os resultados obtidos com as classificações foram satisfatórios, considerando-se os atributos presentes na base de dados. Porém, analisando-se apenas as classificações das canções premiadas, os resultados são baixos demais para permitir o uso deste método na predição de novos vencedores.

TABELA 10 – MATRIZ DE CONFUSÃO DO ALGORITMO SMO APLICADO NA TABELA ATRIBUTO-VALOR COM VALORES EM TF E SEM CORTES

CLASSE	CLASSIFICADAS COMO		TOTAL
	VENCEDORAS	INDICADAS	
Vencedoras	15	43	58
Indicadas	25	209	234
Total	40	252	292

FONTE: o autor (2017).

Outros fatores, não registrados na GRAMMY Song of the Year Database, podem ser relevantes para a premiação de uma canção na categoria Canção do Ano. A base de dados não considera, por exemplo, a popularidade das canções, o contexto histórico e cultural ou características como ritmo, tonalidade ou estilo musical, por exemplo.

5 CONSIDERAÇÕES FINAIS

Métodos de Descoberta de Conhecimento em Bases de Dados são aplicáveis, a princípio, em bases de dados de diferentes domínios do conhecimento humano visando encontrar, potencialmente, conhecimento novo, relevante e útil a partir dos dados de entrada. A escolha de uma base de dados relacionados ao GRAMMY Awards se mostrou uma oportunidade de verificar uma possibilidade de aplicação ainda pouco explorada do KDD.

Além disso, a pesquisa possibilitou ao autor o estudo de assuntos não abordados nas disciplinas da graduação em Gestão da Informação, como o uso de ferramentas de Mineração de Textos e o contexto do GRAMMY Awards, revelando outras interdisciplinaridades possíveis. Também foi possível o reforço de temas já estudados, como Bancos de Dados e Mineração de Dados.

A seguir, apresentam-se as considerações finais quanto aos objetivos propostos para este trabalho e, na sequência, apresentam-se possibilidades para trabalhos futuros.

5.1 ALCANCE DOS OBJETIVOS

O objetivo geral proposto para este trabalho foi a aplicação de métodos de descoberta de conhecimento em dados de músicas indicadas aos prêmios GRAMMY. O alcance desse objetivo pode ser confirmado ao verificar-se o cumprimento dos objetivos específicos, conforme a seguir.

O primeiro objetivo específico estabelecido para este trabalho foi a criação de uma base de dados sobre uma das premiações do GRAMMY Awards e sua disponibilização para estudos futuros. Os dados foram obtidos do sítio da premiação e da ferramenta Genius. A base de dados foi disponibilizada na plataforma UCI Machine Learning Repository.

Através de busca na literatura (segundo objetivo específico), identificaram-se métodos usados em contextos de mineração de texto (como Bayes e K-médias, por exemplo), o que influenciou na escolha do J48 e do SMO para aplicação na base de dados estudada.

Também se realizaram o estudo e a decisão de métodos para aplicação na base dados, presentes no terceiro objetivo específico. Verificou-se o funcionamento

das ferramentas e dos algoritmos escolhidos para decidir como se configuraria a metodologia desta pesquisa. Assim, adaptou-se o processo de descoberta de conhecimentos para o contexto do estudo, utilizando métodos de mineração de textos na etapa de pré-processamento dos dados. Conforme mencionado anteriormente, para a mineração de dados escolheram-se, com base na literatura, duas heurísticas disponíveis na ferramenta Weka, utilizada no estudo.

A avaliação, por meio de comparações, dos diferentes estudos realizados com os algoritmos de mineração de dados (quarto objetivo específico) foi abordada na seção anterior. Encontraram-se padrões nas letras estudadas por meio dos algoritmos, ainda que não sejam suficientes para a correta classificação de uma canção indicada como vencedora do prêmio ou não.

Desta forma, o objetivo geral do trabalho foi atingido. Responde-se afirmativamente à questão de pesquisa “A aplicação de métodos de Descoberta de Conhecimento seria adequada para a recuperação de conhecimento implícito em base de dados sobre o GRAMMY Award?”. Ainda que, no escopo deste estudo, os resultados obtidos não fossem suficientes para prever canções vencedoras, a aplicação dos métodos se mostrou adequada, do ponto de vista da descoberta de conhecimento implícito.

5.2 TRABALHOS FUTUROS

Trabalhos futuros podem empregar outros algoritmos de classificação, além do J48 e SMO utilizados aqui, ou outros métodos de mineração de dados (inclusive utilizando-se de inteligência artificial) e mineração de textos na base criada neste trabalho, uma vez que esta estará disponível livremente na internet. Também podem ser empregados outras abordagens de pré-processamento de textos nas letras das canções.

Os dados coletados para a base de dados incluem nome dos compositores; título, letras e idioma das canções; nome do intérprete; ano e resultado da premiação. Outros dados podem ser considerados em estudos futuros, como vendas de álbuns ou estilo musical, por exemplo.

Ainda, outras premiações do GRAMMY (tais como Melhor Canção de Rock ou Melhor Canção de Pop, por exemplo) podem ser estudadas com técnicas similares.

REFERÊNCIAS

AGRAWAL, R.; BATRA, M. A detailed study on text mining techniques. **International Journal of Soft Computing and Engineering**, v. 2, n. 6, p. 118–121, 2013. Disponível em: <https://goo.gl/A1o1o8>. Acesso em:18 jun. 2017.

AMARAL, F. C. N. do. **Data mining**: técnicas e aplicações para o marketing direto. São Paulo: Berkeley Brasil, 2001.

BEAL, A. **Gestão estratégica da informação**: como transformar a informação e a tecnologia da informação em fatores de crescimento e de alto desempenho nas organizações. São Paulo: Atlas, 2008.

CARVALHO, L. A. V. de. **Datamining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Ciência Moderna, 2005.

CARVALHO, M. B. de. **Análise de Dados de Artigos Recuperados da Web of Science (WoS)**. Curitiba: Universidade Federal do Paraná, 2016. Relatório de Iniciação Científica.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento empresarial**: como as organizações gerenciam seu capital intelectual. 2. ed. Rio de Janeiro: Campus, 1998a.

DAVENPORT, T. H.; PRUSAK, L. **Ecologia da Informação**: por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998b.

DE SORDI, J. O. **Administração da Informação**: fundamentos e práticas para uma nova gestão do conhecimento. São Paulo: Saraiva, 2001.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. From datamining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, Fall 1996. Disponível em: <https://goo.gl/Jn7KB5>. Acesso em: 25 fev. 2016.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. **AI Magazine**, v. 13, n. 3, p. 57–70, Fall 1992. Disponível em: <https://goo.gl/Bt1fY2>. Acesso em: 21 jan. 2016.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GOLDSCHIMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

GONG, Y.; LIU, X. Generic text summarization using relevance measure and latent semantic analysis. In: SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT

IN INFORMATION RETRIEVAL, 24, 2001, [Tokyo]. **Proceedings…** [Tokyo]: ACM, 2001. p. 19–25. Disponível em: <https://goo.gl/RXjKJ9>. Acesso em: 18 jun. 2017.

HEUSER, C. A. **Projeto de banco de dados**. 4. ed. [Porto Alegre]: Sagra Luzzatto, 1998. Disponível em: <https://goo.gl/a4jiEn>. Acesso em: 05 fev.2016.
JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. Machine learning: In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 10, 1998, [Chemnitz, Germany]. **Proceedings…** [Dortmund]: Springer, 1998. p. 137–142, 1998. Disponível em: <https://goo.gl/DyHkd1>. Acesso em 18 jun. 2017.

KUSHMERICK, N.; WELD, D. S.; DOORENBOS, R. **Wrapper induction for information extraction.** 1997. Tese (Doutorado) – University of Washington, Washington, 1997. Disponível em: <https://goo.gl/wsLjUR>. Acesso em: 18 jun. 2017.

LOS ANGELES TIMES. Grammys history and winners through the years. **Los Angeles Times**, Jan. 2015. Disponível em: <http://timelines.latimes.com/grammy-awards/>. Acesso em: 28 maio 2017.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. **PreTexT**: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. São Carlos: ICMC, 2003. Disponível em: <http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_209.pdf>. Acesso em: 23 nov. 2017.

OLIVEIRA, J. P. M. de et al. Applying text mining on electronic messages for competitive inteligence. In: INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE AND WEB TECHNOLOGIES - EC-WEB, 5, 2004, Zaragoza. **Proceedings…** Zaragoza: Spain, 2004. Disponível em: <https://goo.gl/Dq5ru8>. Acesso em: 27 jan. 2016.

PATIL, T. R.; SHEREKAR, S. S. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. **International Journal of Computer Science and** Applications, v. 6, n. 2, Apr. 2013, p. 256-261. Disponível em: <http://keddiyan.com/files/AHCI/week2/9.pdf>. Acesso em: 23 nov. 2017.

PEACOCK, D. E.; HU, G. Analyzing Grammy, Emmy, and Academy Awards data using regression and maximum information coefficient. In: INTERNATIONAL CONFERENCE ON ADVANCED APPLIED INFORMATICS, 2013, [Matsue]. **Proceedings…** [Matsue]: IEEE, 2013. p. 74–79. Disponível em: <https://goo.gl/jzCJZa>. Acesso em: 18 jun. 2017.

PLATT, J. C. **Sequential Minimal Optimization**: a fast algorithm for training Support Vector Machines. 1998. Disponível em: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf>. Acesso em: 23 nov. 2017.

QUINLAN, J. R. Improved Use of Continuous Attributes in C4.5. **Journal of Artificial Intelligence Research**, v. 4, 1996, p. 77-90. Disponível em: <http://www.jair.org/media/279/live-279-1538-jair.pdf>. Acesso em: 23 nov. 2017.

QUINLAN, J. R.; RIVEST, R. L. Inferring decision trees using the minimum description lenght principle. **Information and Computation**, v. 80, n. 3, March 1989, p. 227-248. Disponível em: <https://ac.els-cdn.com/0890540189900102/1-s2.0-0890540189900102-main.pdf?_tid=5c007dfc-ce53-11e7-85fa-00000aacb362&acdnat=1511224488_966554fb417b5c9abffb486c63c1dc0a>. Acesso em: 23 nov. 2017.

RAMAKRISHNAN, R.; GEHRKE, J. **Database management systems**. 2000. Disponível em: <https://goo.gl/wPY8Ap>. Acesso em: 15 fev. 2016.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados**. Rio de Janeiro: Elsevier, 2006.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. **PreTexT**: a reestruturação da ferramenta de pré-processamento de textos. São Carlos: ICMC, 2008. Disponível em: <http://sites.labic.icmc.usp.br/pretext2/pretext/RT-Pretext-Draft.pdf>. Acesso em: 23 nov. 2017.

TAN, A.-H. Text mining: The state of the art and the challenges. In: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999, [Beijing]. **Proceedings…** [Beijing]: PAKDD, 1999. Disponível em: <https://goo.gl/wm86HB>. Acesso em: 21 jan. 2016.

THE RECORDING ACADEMY. **GRAMMY.org**: the official site for the recording academy. 2017a. Disponível em: <http://www.grammy.org>. Acesso em: 28 maio 2017.

THE RECORDING ACADEMY. **GRAMMYPro**. 2017b. Disponível em: <https://www.grammypro.com>. Acesso em:18 jun. 2017.

THE RECORDING ACADEMY. **The GRAMMYs**. 2017c. Disponível em: <http://www.grammy.com>. Acesso em: 28 maio 2017.

THOMÉ, A. C. G. **Redes neurais**: uma ferramenta para KDD e data mining. Disponível em: <https://goo.gl/b7U4SU>. Acesso em: 26 jan. 2016.

TURTLE, H.; CROFT, W. B. Inference networks for document retrieval. In: SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 12, 1989, [Massachusetts]. **Proceedings…** [Massachusetts]: ACM, 1989. p. 1–24. Disponível em: <https://goo.gl/ef2bSB>. Acesso em 18 jun. 2017.

VIJAYARANI, S.; MUTHULAKSHMI, M. Comparative analysis of Bayes and Lazy classification algorithms. **International Journal of Advanced Research in Computer and Communication Engineering**, v. 2, n. 8, Aug. 2013. Disponível em: <https://goo.gl/3fRj8W>. Acesso em: 25 fev. 2016.

WANG, Y.; WANG, H.; GU, Z.-G. A survey of data mining softwares used for real projects. In: INTERNATIONAL WORKSHOP ON OPEN-SOURCE SOFTWARE FOR SCIENTIfiC COMPUTATION – OSSC, [s.n.], 2011, [Beijing]. **Proceedings…**

[Beijing]: IEEE, 2011. p. 94–97. Disponível em: <https://goo.gl/HNp5KR>. Acesso em: 18 jun. 2017.

WITTEN, I. H. **Text mining**. 2004. Disponível em: <https://goo.gl/2VAfxd>. Acesso em: 03 jun. 2017.

ZAMIR, O.; ETZIONI, O. Web document clustering: A feasibility demonstration. In: SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21, 1998, [Melbourne] . **Proceedings…** [Melbourne]: ACM, 1998. p. 46–54. Disponível em: <https://goo.gl/5ZChQo>. Acesso em 18 jun. 2017.

APÊNDICE A – CATEGORIAS PREMIADAS NO 59º GRAMMY AWARDS

A seguir são listadas as categorias premiadas na 59ª edição do GRAMMY, divididas de acordo com os respectivos gêneros:

Geral

a) Gravação do ano
b) Álbum do ano
c) Canção do ano
d) Artista revelação

Pop

a) Melhor desempenho solo de pop
b) Melhor desempenho em dupla ou grupo de pop
c) Melhor álbum vocal de pop

Pop tradicional

a) Melhor álbum vocal de pop tradicional

Dance e música eletrônica

a) Melhor gravação de dance
b) Melhor álbum de dance ou música eletrônica

Música instrumental contemporânea

a) Melhor álbum instrumental contemporâneo

Rock

a) Melhor desempenho de rock
b) Melhor desempenho de metal
c) Melhor canção de rock
d) Melhor álbum de rock

Música alternativa

a) Melhor álbum de música alternativa

Rhythm and blues (R&B)

a) Melhor desempenho de R&B
b) Melhor desempenho de R&B tradicional
c) Melhor canção de R&B
d) Melhor álbum urbano contemporâneo
e) Melhor álbum de R&B

Rap

a) Melhor desempenho de rap
b) Melhor desempenho de rap ou sung
c) Melhor canção de rap
d) Melhor álbum de rap

Country

a) Melhor desempenho solo de country
b) Melhor desempenho em dupla ou grupo de country
c) Melhor canção de country
d) Melhor álbum de country

New age

a) Melhor álbum de new age

Jazz

a) Melhor solo improvisado de jazz
b) Melhor álbum vocal de jazz
c) Melhor álbum instrumental de jazz
d) Melhor álbum de jazz em conjunto
e) Melhor álbum de jazz latino

Gospel e música cristã contemporânea

a) Melhor desempenho ou canção gospel
b) Melhor desempenho ou canção de música cristã contemporânea
c) Melhor álbum de gospel
d) Melhor álbum de música cristã contemporânea
e) Melhor álbum de gospel de raiz

Música latina

a) Melhor álbum latino de pop
b) Melhor álbum latino de rock, urbano ou alternativo
c) Melhor álbum de música regional mexicana (incluindo tejano)
d) Melhor álbum de música tropical latina

Música americana de raiz

a) Melhor desempenho de raiz americana
b) Melhor canção de raiz americana
c) Melhor álbum de americana
d) Melhor álbum de bluegrass

e) Melhor álbum de blues tradicional
f) Melhor álbum de blues contemporâneo
g) Melhor álbum de folk
h) Melhor álbum de música regional de raiz

Reggae

a) Melhor álbum de reggae

Música mundial

a) Melhor álbum de música mundial

Música para crianças

a) Melhor álbum para crianças

Declamação

a) Melhor álbum de declamação (incluindo poesia, áudio livros e narrativas)

Comédia

a) Melhor álbum de comédia

Teatro musical

a) Melhor álbum de teatro musical

Música para mídia visual

a) Melhor compilação de trilha sonora para mídia visual
b) Melhor trilha sonora original para mídia visual
c) Melhor canção escrita para mídia visual

Composição e arranjo

a) Melhor composição instrumental
b) Melhor arranjo, instrumental ou à capela
c) Melhor arranjo, instrumentos e vocais

Pacote

a) Melhor pacote de gravação
b) Melhor disco em edição especial limitada

Encarte

a) Melhor encarte

Histórico

a) Melhor álbum histórico

Produção não-clássica

a) Melhor engenharia em álbum não-clássico
b) Produtor do ano de música não-clássica
c) Melhor gravação remixada

Som Surround

a) Melhor álbum em som surround

Produção clássica

a) Melhor engenharia em álbum clássico
b) Produtor do ano de música clássica

Clássica

a) Melhor desempenho de orquestra
b) Melhor gravação de ópera
c) Melhor desempenho de coral
d) Melhor desempenho de música de câmara e grupo pequeno
e) Melhor solo instrumental de música clássica
f) Melhor álbum vocal solo de música clássica
g) Melhor compêndio de música clássica
h) Melhor composição de música clássica contemporânea

Clipe e vídeo

a) Melhor clipe
b) Melhor vídeo musical

APÊNDICE B – ÁRVORES DE DECISÃO GERADAS NOS ESTUDOS

Abaixo são apresentadas as árvores de decisão obtidas com a aplicação do J48 nas quatro versões da tabela atributo-valor estudadas durante a pesquisa (com valores booleanos, com valores TF e sem cortes, com valores TF e corte de 0,5 desvio padrão e com valores TF e corte de 1 desvio padrão).

VALORES BOOLEANOS

```
car <= 0
| lai <= 0
| | think <= 0
| | | beat <= 0
| | | | step <= 0
| | | | | game <= 0
| | | | | | hour <= 0
| | | | | | | gotta <= 0
| | | | | | | | reason <= 0
| | | | | | | | | long <= 0
| | | | | | | | | | lover <= 0
| | | | | | | | | | | hang <= 0
| | | | | | | | | | | | chanc <= 0
| | | | | | | | | | | | | open <= 0
| | | | | | | | | | | | | | ya <= 0
| | | | | | | | | | | | | | | girl <= 0
| | | | | | | | | | | | | | | | stranger <= 0
| | | | | | | | | | | | | | | | | door <= 0
| | | | | | | | | | | | | | | | | | mile <= 0
| | | | | | | | | | | | | | | | | | | chang <= 0
| | | | | | | | | | | | | | | | | | | | empti <= 0: nom (59.0/2.0)
| | | | | | | | | | | | | | | | | | | | empti > 0
| | | | | | | | | | | | | | | | | | | | | ll <= 0: nom (3.0)
| | | | | | | | | | | | | | | | | | | | | ll > 0: win (2.0)
| | | | | | | | | | | | | | | | | | | | chang > 0
| | | | | | | | | | | | | | | | | | | | | scare <= 0: win (4.0/1.0)
| | | | | | | | | | | | | | | | | | | | | scare > 0: nom (2.0)
| | | | | | | | | | | | | | | | | | | | mile > 0: win (4.0/1.0)
| | | | | | | | | | | | | | | | | | | door > 0
| | | | | | | | | | | | | | | | | | | | walk <= 0: win (5.0/1.0)
| | | | | | | | | | | | | | | | | | | | walk > 0: nom (3.0)
| | | | | | | | | | | | | | | | | | | stranger > 0
| | | | | | | | | | | | | | | | | | | | letter <= 0: nom (3.0/1.0)
| | | | | | | | | | | | | | | | | | | | letter > 0: win (2.0)
| | | | | | | | | | | | | | | | | | | girl > 0: nom (15.0)
| | | | | | | | | | | | | | | | | | ya > 0
| | | | | | | | | | | | | | | | | | | gui <= 0: win (4.0)
| | | | | | | | | | | | | | | | | | | gui > 0: nom (2.0)
| | | | | | | | | | | | | | | | | | open > 0
| | | | | | | | | | | | | | | | | | | lean <= 0: win (4.0)
| | | | | | | | | | | | | | | | | | | lean > 0: nom (2.0)
| | | | | | | | | | | | | | | | | | chanc > 0
| | | | | | | | | | | | | | | | | | | don <= 0: win (7.0/1.0)
| | | | | | | | | | | | | | | | | | | don > 0: nom (3.0)
| | | | | | | | | | | | | | | | | | hang > 0: nom (7.0)
```

```
| | | | | | | | | | lover > 0: nom (10.0)
| | | | | | | | | long > 0: nom (23.0)
| | | | | | | | reason > 0
| | | | | | | | | world <= 0
| | | | | | | | | | love <= 0: win (4.0/1.0)
| | | | | | | | | | love > 0: nom (6.0)
| | | | | | | | | world > 0: win (3.0)
| | | | | | | gotta > 0: nom (10.0)
| | | | | | hour > 0: nom (11.0)
| | | | | game > 0: nom (14.0)
| | | | step > 0
| | | | | walk <= 0
| | | | | | melt <= 0: nom (8.0/1.0)
| | | | | | melt > 0: win (2.0)
| | | | | walk > 0: win (3.0)
| | | beat > 0: nom (18.0)
| | think > 0
| | | till <= 0
| | | | wish <= 0
| | | | | time <= 0
| | | | | | don <= 0: win (4.0/1.0)
| | | | | | don > 0: nom (4.0)
| | | | | time > 0: nom (6.0)
| | | | wish > 0: win (2.0)
| | | till > 0: win (3.0)
| lai > 0
| | won <= 0
| | | troubl <= 0: nom (9.0/1.0)
| | | troubl > 0: win (2.0)
| | won > 0: win (4.0)
car > 0: nom (15.0)
```

VALORES EM TF SEM CORTE

```
smile <= 0.25
| wake <= 0
| | pass <= 0
| | | lie <= 0
| | | | car <= 0
| | | | | think <= 0
| | | | | | rememb <= 0
| | | | | | | beat <= 0
| | | | | | | | tomorrow <= 0
| | | | | | | | | sea <= 0
| | | | | | | | | | step <= 0
| | | | | | | | | | | insid <= 0
| | | | | | | | | | | | lai <= 0
| | | | | | | | | | | | | sit <= 0
| | | | | | | | | | | | | | send <= 0
| | | | | | | | | | | | | | | knew <= 0
| | | | | | | | | | | | | | | | long <= 0
| | | | | | | | | | | | | | | | | ya <= 0
| | | | | | | | | | | | | | | | | | moon <= 0
| | | | | | | | | | | | | | | | | | | hurt <= 0
| | | | | | | | | | | | | | | | | | | | word <= 0: nom (72.0/8.0)
| | | | | | | | | | | | | | | | | | | | word > 0
| | | | | | | | | | | | | | | | | | | | | turn <= 0: win (3.0)
| | | | | | | | | | | | | | | | | | | | | turn > 0: nom (3.0)
| | | | | | | | | | | | | | | | | | | | hurt > 0
```

```
| | | | | | | | | | | | | | | | | | | door <= 0: nom (2.0)
| | | | | | | | | | | | | | | | | | | door > 0: win (2.0)
| | | | | | | | | | | | | | | | | | moon > 0
| | | | | | | | | | | | | | | | | | | light <= 0: win (3.0)
| | | | | | | | | | | | | | | | | | | light > 0: nom (3.0)
| | | | | | | | | | | | | | | | | | ya > 0
| | | | | | | | | | | | | | | | | | yeah <= 0: win (3.0)
| | | | | | | | | | | | | | | | | | yeah > 0: nom (2.0)
| | | | | | | | | | | | | | | | | | long > 0: nom (14.0)
| | | | | | | | | | | | | | | | | knew > 0
| | | | | | | | | | | | | | | | | | knew <= 0.5: win (4.0)
| | | | | | | | | | | | | | | | | | knew > 0.5: nom (2.0)
| | | | | | | | | | | | | | | | | send > 0
| | | | | | | | | | | | | | | | | | girl <= 0: win (4.0)
| | | | | | | | | | | | | | | | | | girl > 0: nom (2.0)
| | | | | | | | | | | | | | | | | sit > 0: nom (7.0)
| | | | | | | | | | | | | | | | | lai > 0
| | | | | | | | | | | | | | | | | | bright <= 0: win (5.0)
| | | | | | | | | | | | | | | | | | bright > 0: nom (2.0)
| | | | | | | | | | | | | | | | | insid > 0: nom (18.0)
| | | | | | | | | | | | | | | | | step > 0
| | | | | | | | | | | | | | | | | | move <= 0: win (6.0/1.0)
| | | | | | | | | | | | | | | | | | move > 0: nom (3.0)
| | | | | | | | | | | | | | | | | sea > 0
| | | | | | | | | | | | | | | | | | fine <= 0: win (6.0/1.0)
| | | | | | | | | | | | | | | | | | fine > 0: nom (2.0)
| | | | | | | | | | | | | | | | | tomorrow > 0: nom (9.0)
| | | | | | | | | | | | | | | | | beat > 0: nom (12.0)
| | | | | | | | | | | | | | | | rememb > 0: nom (16.0)
| | | | | | | | | | | | | | | think > 0
| | | | | | | | | | | | | | | | | strong <= 0
| | | | | | | | | | | | | | | | | | like <= 0.5
| | | | | | | | | | | | | | | | | | | version <= 0: win (8.0/1.0)
| | | | | | | | | | | | | | | | | | | version > 0: nom (3.0/1.0)
| | | | | | | | | | | | | | | | | | like > 0.5: nom (2.0)
| | | | | | | | | | | | | | | | | strong > 0: nom (4.0)
| | | | | | | | | | | | | | | | car > 0: nom (13.0)
| | | | | | | | | | | | | | | | lie > 0: nom (13.0)
| | | | | | | | | | | | | | | pass > 0: nom (14.0)
| | | | | | | | | | | | | | wake > 0: nom (15.0)
smile > 0.25
| sweet <= 0
| | share <= 0
| | | eye <= 0
| | | | love <= 0: win (4.0)
| | | | love > 0: nom (2.0)
| | | eye > 0: win (4.0)
| | share > 0: nom (2.0)
| sweet > 0: nom (3.0)
```

VALORES EM TF COM CORTE DE 0,5 DESVIO PADRÃO

```
smile <= 0.25
| reason <= 0: nom (257.0/42.0)
| reason > 0
| | gave <= 0
| | | doesn <= 0: nom (16.0/4.0)
| | | doesn > 0: win (2.0)
| | gave > 0: win (2.0)
smile > 0.25
```

```
| smile <= 0.5
| | head <= 0
| | | grammy-year <= 1970: win (2.0)
| | | grammy-year > 1970: nom (8.0/1.0)
| | head > 0: win (2.0)
| smile > 0.5: win (3.0)
```

VALORES EM TF COM CORTE DE 1 DESVIO PADRÃO

```
smile <= 0.25: nom (277.0/50.0)
smile > 0.25
| sweet <= 0
| | share <= 0: win (10.0/2.0)
| | share > 0: nom (2.0)
| sweet > 0: nom (3.0)
```