Sistema de indicação literária utilizando Naive Bayes
Literary Indication System (LIS)

Cryslene C. de Oliveira[1], Emilda S. Oliveira[1], Natália F. Araújo[1]

[1]Faculdade de Engenharia de Computação – Universidade Federal do Pará (UFPA)
Tucuruí – PA – Brasil

{cryslenecl, emildaoliv, taiaraujo20}@gmail.com

Resumo. *O presente artigo faz uma abordagem sobre os conceitos relacionados à mineração de dados e suas aplicações no projeto de indicação de livros utilizando o algoritmo Naive Bayes, mostrando o processo de execução, os métodos utilizados e o resultado de todas essas aplicações. Visando um usuário final, foi idealizada uma plataforma computacional que irá conter livros de diversos gêneros, o usuário irá determinar uma certa pontuação em cada categoria de livros que, através de um dataset e sua mineração dos dados, será mostrado indicações de livros relacionados de acordo com essa avalição e que esse usuário pode ler posteriormente. Visando a incentivar na continuação do processo de leitura pelo usuário.*

Abstract. *The present article is an approach on concepts related to data mining and its applications in the design of a Naive Bayes algorithm application window, showing the execution process, the methods used and the results of all the applications. Aiming at one end user, a computational platform was designed that will show the groups of several genres, as will be the case of a series in each category of books, through a set of data and a data mining, the indications of related books according to this rating and which this user can read later. Aiming at the continuation of the verification process by the user.*

1. Introdução

Falar de leitura é bastante desafiador, pois se trata de uma ação que envolve o ser humano com a busca de conhecimento, interatividade ou lazer. No Brasil esse paradigma vem sendo discutido por muitas áreas, em busca de chaves que associem essa necessidade de se promover a leitura através de projetos existentes ou que possam surgir no decorrer do tempo, associações com o governo e através de políticas públicas como, por exemplo, o Instituto Pró-Livro (IPL) que vem mostrar dados significativos para análise do processo de leitura.

> "O Instituto Pró-Livro é uma organização social civil de interesse público – uma Oscip – criada por três das principais entidades do livro no Brasil: Câmara Brasileira do Livro (CBL), Sindicato Nacional de Editores de Livros (Snel) e Associação Brasileira de Editores de Livros (Abrelivros). [...]. Seu Objetivo principal é viabilizar ações para ajudar a fomentar a leitura e o livro no Brasil. [...]." (AMORIM, 2008).

O livro "*Retratos da leitura no Brasil*", do organizador Geleno Amorim, mostra muitas análises estatísticas em relação ao processo de leitura no Brasil, dentre elas foi realizada uma pesquisa feita pelo IPL, onde 95,6 milhões (55%) da população estudada são leitores autodeclarados e leram pelo menos um livro durante os três meses que antecederam a pesquisa em 2007.

Segundo Amorim (2008), essas informações dadas pelo IPL foram muito significativas, já que implicaram de forma considerável em indicadores de leitura. Com isso, surgiu a ideia de uma ferramenta que auxiliasse leitores, como público alvo, a escolher qual seria o próximo possível livro que ele poderia ler, de acordo com os livros que já foram lidos.

A base de dados contém mais de 56 mil livros e 9 gêneros diferentes, onde esses gêneros foram definidos baseados na pesquisa feita pelo IPL dos livros mais importantes na vida dos leitores, segundo mostra a tabela.

Tabela 1. Livros mais importantes na vida dos leitores – Pesquisa IPL – Adaptado de "*Retratos da leitura no Brasil" de AMORIM (2008)*

Livros mais importantes na vida dos leitores*			
1)	Bíblia	16)	O Menino Maluquinho
2)	O Sítio do Pica-pau Amarelo**	17)	A Escrava Isaura
3)	Chapeuzinho Vermelho	18)	Romeu e Julieta
4)	Harry Potter	19)	Poliana
5)	O Pequeno Príncipe	20)	Gabriel Cravo e Canela
6)	Os Três Porquinhos	21)	Pinóquio
7)	Dom Casmurro	22)	Bom Dia Espírito Santo
8)	A Branca de Neve	23)	A Moreninha
9)	Violetas na Janela	24)	Primo Basílio
10)	O Alquimista	25)	Peter Pan
11)	Cinderela	26)	Vidas Secas
12)	Código da Vinci	27)	Carandiru
13)	Iracema	28)	O Segredo
14)	Capitães de Areia	29)	A Ilha Perdida
15)	Ninguém é de Ninguém	30)	Meu Pé de Laranja Lima
* Resposta espontânea e com uma única opção. ** Embora não conste da bibliografia brasileira, é uma referência à obra de Monteiro de Lobato.			
59% dos leitores (56,2 milhões) souberam citar o livro mais marcante.			
O número de citações da Bíblia é 10 vezes maior que a do 2º colocado.			
2 em cada 3 entrevistados (contando os não-leitores) não souberam dizer ou não informaram um livro marcante.			

A pergunta de proposta ao banco de dados é: “Segundo o perfil do usuário, definido através de uma análise feita a partir da avaliação atribuída por ele aos livros já lidos, qual será o próximo livro que vai ser indicado ao usuário e que ele vai decidir avaliar com a máxima pontuação?”

Todo o processo para se obter a resposta será descrito neste artigo, mostrando a análise e implementação de algoritmo e tratamento de dados para indicar ao leitor uma próxima leitura de acordo com o perfil que será mostrado pelo usuário, mas que será determinado pela execução do projeto.

2. Metodologia

Com uma base de dados adquirida da Amazon, que contém a listagem de livros do *site*, pelo Akshay Bhatia, surgiu um projeto que fez parte do Spikeway Technologies AI Lab (SAIL), onde coletou a base de dados de livros da Amazon e através do algoritmo de Regressão Logística, NaiveBayes Multinominal, Multi Layer Perceptron e XGBoost fez a classificação dos livros por gêneros criando uma nova base de dados, que continha 207.591 livros e 32 gêneros diferentes, utilizando o conceito de Deep Learning e Long Short Term Memory (LSTM), disponibilizando esse material no seu GitHub em 2018.

Em uma adaptação para esse projeto, foi realizada a análise, filtragem e tratamento desse *dataset*, que passou a conter 56.244 livros com 9 gêneros diferentes, sendo esse o *dataset* principal com várias informações, dentre elas será destacado o nome do livro, o autor, a categoria, a identificação da categoria, bem como a identificação do livro, através do código definido pela ISBN (*International Standard Book Number*), como seu próprio *site* explica, “é um sistema internacional padronizado que identifica numericamente os livros segundo o título, o autor, o país, a editora, individualizando-os inclusive por edição. Utilizado também para identificar software, seu sistema numérico é convertido em código de barras, o que elimina barreiras linguísticas e facilita a circulação e comercialização das obras”.

A partir desse processo, tem-se outra base de dados que contém toda a classificação inicial de cada uma das categorias de acordo com seu respectivo livro, sendo gerada a partir de uma pesquisa inicial com alguns leitores, que fazem parte de clubes de livros, onde eles determinaram uma pontuação para cada gênero.

Em um terceiro *dataset*, criado para servir como base de exemplo, será definido um perfil literário do usuário, contendo uma identificação, uma pontuação dada pelo usuário a cada gênero, uma pontuação dada pelo usuário a cada gênero em relação aos livros lidos por ele e uma avaliação de 1 a 5 estrelas em relação ao livro lido.

Foi definido o algoritmo Naive Bayes para fazer a etapa de classificação e indicação dos livros de acordo com a pontuação definida para cada categoria pelo leitor.

> “O aprendizado bayesiano é do tipo supervisionado, já que são fornecidas ao algoritmo de AM as instâncias juntamente com seus rótulos (ou seja, as classes). Seguindo o paradigma estatístico, o algoritmo faz uso de fórmulas estatísticas e cálculo de probabilidades para realizar a classificação (Mitchell, 1997)”. (PARDO; NUNES, 2002).

2.1. Tratamento dos dados - Dataset (Dados, Dados-alvos, Dados Processados, Dados Transformados)

Usando a Base de dados da Amazon que continha, inicialmente, 32 gêneros e 207.572 livros, esse conjunto de dados encontrava-se no arquivo, com extensão CSV, com a seguintes informações: "ISBN"; "Filename"; "Image URL"; "Title"; "Author"; "Category ID"; "Category" de cada livro e necessitou ser feita a conversão para XML, já no arquivo convertido foi necessária a eliminação das colunas: "Filename"; "Image URL" ; "Category", que são dados que não influenciam na análise da base de dados.

Após a fase da limpeza de dados, já na parte da integração houve-se a necessidade da inclusão de colunas com subcategorias com um sistema de pontuação para que o algoritmo possa classificar e fazer a predição. A planilha de 12 colunas e 71 linhas (incluindo o cabeçalho), resultante da limpeza e integração de dados, contém 9 gêneros e 70 livros selecionados da base de dados da Amazon manipulados 1 a 1, onde os livros foram categorizados pelo gêneros literários e seus subgêneros, em seguida salvo em formato CSV, pois, o restante do pré-processamento será realizado pelo algoritmo que consiste em: eliminação de ruídos, identificação, remoção de desvios, normalização e agregação, para uma qualidade maior dos dados.

A organização dos dados que é fundamental para obter resultados satisfatórios na Mineração de Dados foi descrita por atributos nos quais são os gêneros e subgêneros. Os 9 gêneros dos livros selecionados são:

Tabela 2. Lista de Categorias dos Livros e Seu ID de Identificação – Adaptado de https://github.com/akshaybhatia10/Book-Genre-Classification

ID_CATEGORY	NOME_DA_CATEGORIA
4	Livros infantis
5	Quadrinhos e Romances gráficos
13	Humor e Entretenimento
15	Literatura e Ficção
17	Mistério e Suspense
21	Religião e Espiritualidade
22	Romance
24	Ficção Científica e Fantasia
27	Adolescente e Jovem Adulto

2.2. Mineração dos dados (Padrões – Nayve Bayes)

A partir de todo o tratamento realizado na base de dados da Amazon, foi gerado um *dataset* contendo 70 livros diferentes, com a pontuação de cada gênero aplicada através um seguinte cenário: em uma escala de 0 a 10, foi determinada uma pontuação que demonstrasse a intensidade em que cada gênero se encaixaria de acordo com o sentido do livro e tendo como prioridade máxima, sendo a pontuação de valor 10, a categoria definida pelo *dataset* da Amazon.

Iniciando o processo de implementação desses dados, foi importado para a plataforma *Python* o *dataset* com os 70 livros, transformando-o em uma *array* utilizando a biblioteca *numpy*, e a base de dados criada para o perfil do usuário. Com isso, se obtém um *array* sendo uma lista muito mais maleável do que uma lista normalmente é, já que se pode fazer várias funções que em uma lista comum não é possível.

A partir disso, é feito um tratamento mais aprofundado, gerando uma *array* de dados contendo todas as características contidas no *dataset.* Nesses dados irá ter exatamente o *dataset* que foi criado, só que dentro do *Python* e em uma *array*, definindo as variáveis, sendo o ISBN, nome do livro e autor serão caracteres e a pontuação das categorias será um valor inteiro.

Posteriormente é criado duas *arrays*, uma de x com o perfil do usuário seguido do perfil de um determinado livro e de y com a avaliação atribuída ao livro pelo leitor.

array → x (perfil do usuário, perfil do livro)

array → y (avaliação do livro feita pelo usuário)

Na lista de x o campo terá várias informações, enquanto em y o campo terá apenas uma, que seria a avaliação em formato de variação de 1 a 5 estrelas. Com isso, foi criada uma base de treino com 10 livros, ou seja, um usuário leu 10 livros e deu uma pontuação para as categorias desse livro, sendo armazenada na base de dados, depois disso é utilizado todo o restante do *dataset* para treinar de acordo com os 10 livros.

A partir do perfil do usuário e o perfil do livro é feito uma predição, de acordo com essa estrutura e utilizando o algoritmo do Naive Bayes, através da função *GaussianNB* que é própria do *Python*. Essa predição irá gerar um arquivo temporário, que se assemelha a execução do processo da base de treino, mas ao invés de ser utilizado 10 livros, passa a ser utilizado os outros 60 livros.

Com essa pontuação definida, passa a ser analisado quais os livros tiveram 5 estrelas como resultado da predição, selecionando os cinco livros, que não estão na lista que o usuário já leu e que estão no *dataset* da Amazon, que tiveram a melhor avaliação pelo Naive Bayes e, por fim, através do identificador desse livro, mostra para o usuário essas indicações. Com a resposta da predição, será mostrado o nome do livro para o usuário, tendo um resultado da avaliação.

3. Resultados

A filtragem dos dados para a recomendação e indicação de livros ocorre através de análises dos dados obtidos no dataset em união com o traço do perfil de cada usuário em forma de pontuação das características dos livros.

Foram feitos alguns testes com outros algoritmos de mineração, como por exemplo o KNN, que faz uma classificação analisando vizinhos mais próximos, porém o algoritmo de Naive Bayes se mostrou mais eficiente em relação ao outros.

Assim os resultados obtidos através do algoritmo de Naive Bayes foram cinco livros que mais se aproximavam dos livros já pontuados pelo usuário na nossa base de dados.

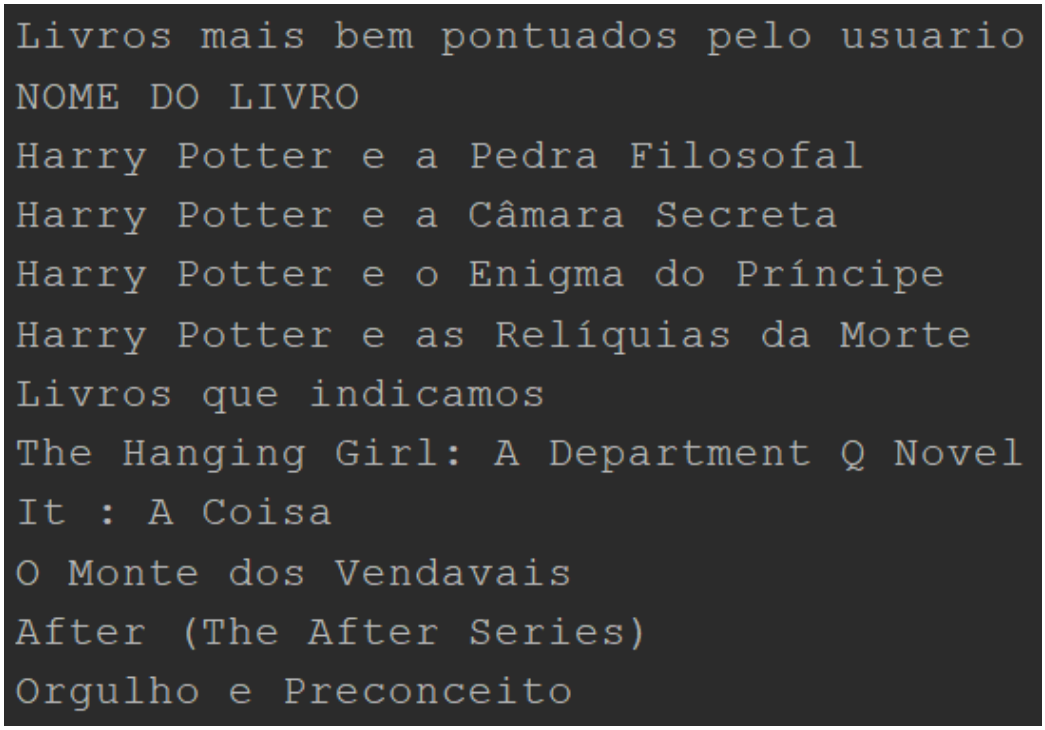

Figura 1. Perfil do usuário e indicação de livros

Com o Python foi gerado a matriz de confusão para demonstrar o quão preciso é o algoritmo utilizado na sua indicação. Confirmando a acurácia de 66,6% como mostrado na figura 2.

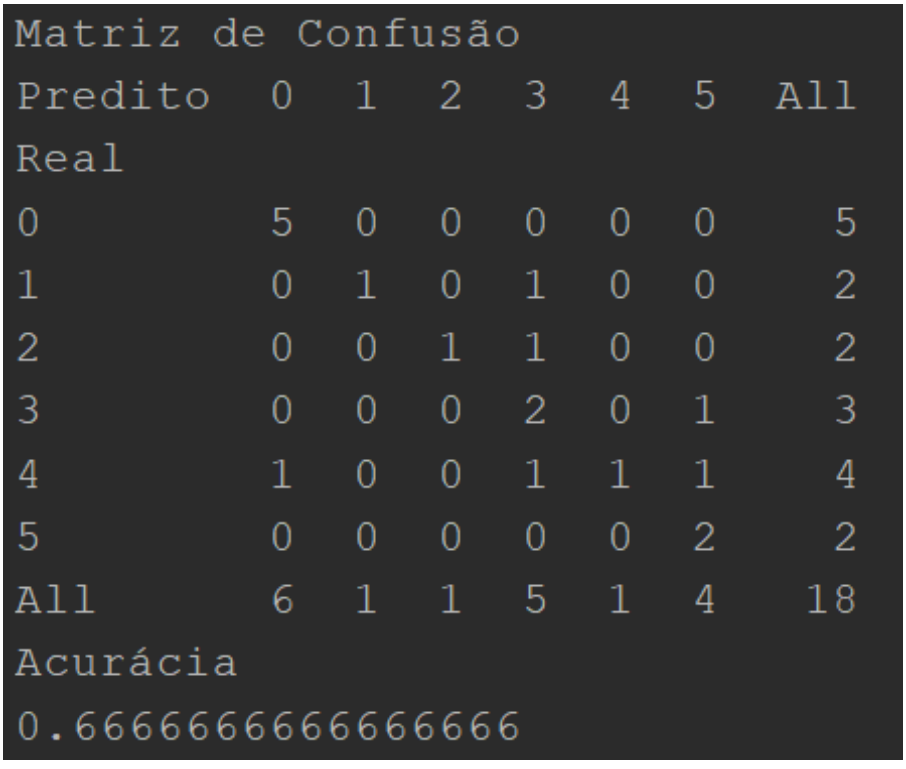

Figura 2. Matriz de Confusão e Acurácia

4. Considerações Finais

A área de mineração de dados vem ganhando cada vez mais espaço na área da tecnologia, através de pesquisas e estudos para evolução desse campo. Atualmente estamos vendo isso na prática, através do crescimento da Inteligência Artificial e dos benefícios que esse contexto vem trazer, mostrando que está se tornando real e com o tempo será uma estratégia trivial a ser seguida.

Foram encontrados muitos desafios a serem quebrados durante a construção desse projeto, desde o tratamento da base de dados, já que continha um enorme número de informações, até a resposta final ao usuário com uma precisão eficiente na indicação do livro, pois é uma área ainda bastante complexa por envolver conhecimentos como essenciais, como processo de mineração de dados, base de dados ou *dataset*, lógica de programação, linguagem *Python*, entre outros que quando implementadas em conjunto passa a resolver questões de forma eficaz no processo.

Além disso, mesmo que se tenha conseguido o resultado da pergunta feita ao banco de dados, sobre qual livro seria indicado ao usuário de acordo com seu perfil e a pontuação dos gêneros, outras implementações podem ser feitas posteriormente, como uma ferramenta que seria apresentada em formato de *site*, onde o usuário iria atribuir uma pontuação as categorias relacionadas a cada livro que já tenha lido, criando uma lista, e determinaria uma pontuação também para o livro indicado pelo *site*, que vai variar de uma a cinco estrelas e para execução dessa plataforma poderiam ser utilizados *Python* 3.7, django 2.2.3, HTML5 e CSS3.

Por fim, essas aplicações e execuções realizadas em todo o processo desse projeto foram de uma gratificante significância, utilizando os conhecimentos adquiridos na vida acadêmica e aplicando tudo o que foi estudado nas aulas de Mineração de Dados, residindo a importância de tais discussões nos meios acadêmicos.

5. Referencias

Amorim, Galeno. (2008) “Retratos da Leitura no Brasil”. In: Instituto Pró-Livro, Imprensa Oficial do Estado de São Paulo. Governo do Estado de São Paulo.

Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6, 4, Article 13 (December 2015), 19 pages.DOI: http://dx.doi.org/10.1145/2843948.

Pardo, Thiago A. S.; Nunes, Maria G. V. (2002) “Aprendizado Bayesiano Aplicado ao Processamento de Línguas Naturais”. Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, São Carlos, SP, Brasil.

Pinheiro, Marcel. SlideShare. (2013) “Construindo Sistemas de Recomendação com Python”. https://pt.slideshare.net/marcelcaraciolo/construindo-sistemas-de-recomendao-com-python.

Bhatia, Akshay. (2018) “Book-Genre-Classification”. https://github.com/akshaybhatia10/Book-Genre-Classification.