SISTEMA DE RECOMENDAÇÃO
Utilizando conceitos de Ciência de Dados e Aprendizagem de Máquina

Robson J. Reis[1], **Talita G. Santos**[2], **Luiz C. Gz**[2], **Abraão B. Brandão** [2] , **Gabriel A. Teixeira** [2]

[1]Licenciatura em Computação – Universidade do Estado do Amazonas (UEA)

[2]Engenharia de Computação – Universidade do Estado do Amazonas (UEA)

rjrsj.lic17@uea.edu.br, tsg.eng17@uea.edu.br, glomyerjunior@hotmail.com

abraaobritof10@gmail.com, gaflt.eng17@uea.edu.br

Escola Superior de Tecnologia - EST
Endereço: Av. Darcy Vargas, 1200 - Parque Dez,
Manaus - AM, 69050-020

Resumo. *Este artigo tem como finalidade a aplicação dos conceitos de Ciência de Dados no desenvolvimento de um sistema de recomendação. Partimos das definições básicas de Ciência de Dados com o objetivo de obter embasamento teórico sobre análise de dados e, posteriormente, adotando conceitos de Aprendizagem de Máquina dispusemos da teoria necessária para a construção do sistema de recomendação. Para a concepção do sistema foi proposta a utilização do dataset MovieLens e tendo como algoritmo-base a filtragem colaborativa.*

1. Ciência de Dados - Importância, Avanços e Aplicação no Mercado

Ciência de Dados em definição torna-se um termo difícil de expor, pois o mesmo é de multidisciplinar e sua aplicação está integrada em muitas áreas da ciência. Em princípio podemos adicionar alguns termos para melhor entendimento: Os "dados" são como pequenas partes de informações que podem ser manipulados de forma ordenada para gerar uma informação, sendo que dados aleatórios não têm nenhum sentido já que não montam uma estrutura coesa e coerente. Como visto informação é um conjunto de dados ordenados dos quais através destas informações é possível fazer análises que consistem em estudar de forma detalhada e minuciosa esses conjuntos informacionais para uma tomada de decisão baseadas nas mesmas consistindo assim na resolução de alguma problemática ou até mesmo o aprimoramento de algum existencial processo.

Introduzindo esse conceito geral podemos dizer de forma grosseira que Ciência de Dados consiste em coletar dados e usá-los para melhorar a informatização de um processo como um todo, sendo que sua importância vai além do mercado, de forma aplicativado hoje basicamente tudo em relação a busca e consulta.

O avanço tecnológico nesta área é gigantesca e com uma gama de dados que é notório, um exemplo disso é o salto que foi dado na medicina usando dados de pacientes fornecido por hospitais é possível fazer um análise de dados para, por exemplo, prever os futuros problemas em um certo paciente como a infecção generalizada e isso apenas com comparação de sua situação atual com dados antigos que já foram avaliados.

A Ciência de Dados pode ser aplicada em várias situações, pois hoje grandes empresas usam isso para aumentar suas vendas, por exemplo, os mercados têm grandes problemas relacionados em que lugar posicionar os objetos a serem vendidos. Uma solução para essa problemática é posicionar as compras com combinações diferentes, assim, pode-se usar as que mais combinaram e alavancaram as vendas. Foi dessa forma que algumas empresas de cereais aprenderam onde por suas embalagens porque foi feito um estudo neste cenário que mostra que cereais vendem mais quando posicionados nas prateleiras mais baixas por causa das crianças que podem ver os cereais e alguns desses com personagens fazendo as empresas venderem mais ou terem uma tendência maior de vendas e tudo por uma certa interação com os pequeninos.
[1] [3] [2]

2. Aprendizagem de Máquina

O aprendizado de máquina é um campo da computação que se baseia no reconhecimento de padrões e em inteligência artificial. Por meio de entrada de dados e de algoritmos de aprendizagem o computador "aprende com seus erros" e começa a fazer previsões cada vez mais precisas de um determinado fato, como a probabilidade de chuva dadas as atuais condições climáticas.

A generalização é um conceito fundamental do aprendizado de máquina: um computador bem treinado tem um poder maior de generalização e de percepção de dados aparentemente ocultos. Uma aplicação prática seria o uso de aprendizado de máquina em supermercados, associando tipos de clientes a diferentes tipos de produtos. Por exemplo, vários clientes que compram produtos A ou B geralmente levam também um produto do tipo C, tendo isso em mente, poderia-se adotar uma estratégia de marketing oferecendo os dois produtos AB ou AC com um determinado desconto.

É um campo fundamentalmente ligado à estatística computacional. Há várias modalidades de aprendizado, mas apenas dois tipos fundamentais: o aprendizado supervisionado e o aprendizado não-supervisionado. Os algoritmos supervisionados possuem uma espécie de professor, que diz qual o tipo de comportamento é esperado do programa. Os não-supervisionados são baseados em observações e descobertas.
[9]

3. Filtragem Colaborativa

A Filtragem Colaborativa é um dos algoritmos mais utilizados na criação de um sistema de recomendação, pois utiliza o histórico de objetos ou informações de interesse do usuário para sugerir novos itens, produtos, ou conteúdo de uma forma geral. É um algoritmo vastamente utilizado por lojas online e sites de conteúdo geral. O algoritmo é baseado em usuários e seus históricos de interesses, que podem ser filmes assistidos, músicas ouvidas, livros lidos entre outros, e criar perfis semelhantes para usuários com mesmas preferências ou estilos. Pode ser baseada na recomendação pela semelhança entre usuários ou pela semelhança entre itens.
[4]

4. DataSet: MovieLens

O dataset MovieLens é disponibilizado pelo GroupLens, um laboratório da Universidade de Minnesota especializado em sistemas de recomendação e comunidades on-

line. Como dito no seu próprio site: “Nós avançamos a teoria e a prática de computação social através da construção e compreensão de sistemas usado por pessoas reais”, em tradução livre.

O grupo tem uma longa história com a pesquisa em sistemas de recomendação, que começou com um artigo na USENET sobre o desenvolvimento de filtragem colaborativa automática. Nos dias de hoje a pesquisa continua, com o grupo gerenciando vários serviços relacionados à recomendação.

O banco de dados que utilizamos consiste em cem mil avaliações, entre 1 e 5, que foram coletadas durante um período de sete meses, a partir de setembro 1997. O projeto foi algo bastante grande para a sua época, sendo composto de 1682 filmes e 943 usuários. Os dados ainda sofreram uma leve filtragem, desconsiderando usuários que avaliaram menos de vinte filmes para dar uma acurácia maior às informações compiladas. Para o aprendizado de máquina os campos de ocupação e gênero de uma pessoa estão disponíveis.
[8]

5. K-Nearest Neighbors - KNN (K-Vizinhos Próximos)

KNN é o algoritmo de aprendizado de máquina mais simples, utilizado para problemas de classificação (aprendizagem supervisionada). Sempre que o modelo é alimentado com dados de teste, ele encontra a distância, geralmente utilizando a Distância Euclidiana, daquele ponto com todos os outros pontos nos dados de treinamento. Em seguida, ele encontra os membros k mais próximos para esse ponto. Isso auxilia na classificação dos dados em um grupo específico.

O valor de k é o número de vizinhos mais próximos a considerar ao classificar o novo ponto de dados. A escolha do valor de k é muito importante e o valor correto para k ajudará a melhorar a precisão do modelo. Se o valor de k for mantido alto, poderá resultar na redução da variância, mas poderá levar a uma situação em que os padrões pequenos nos dados são omitidos.
[7] [5] [10] [6]

6. Produzindo um Sistema de Recomendação com Python

A proposta da disciplina de Ciência de Dados era o desenvolvimento de um sistema de recomendação que usasse o banco de dados MovieLens para recomendar filmes. Utilizando um método de Aprendizagem de Máquina com conceitos de Probabilidade e Estatística para pressupor um resultado aceitável. Foi acolhido como auxílio técnico o curso da Udemy - Produzindo um Sistema de Recomendação: https://www.udemy.com/inteligencia-artificial-sistemas-de-recomendacao-em-python/learn/v4/ onde, por meio deste, demonstrou-se, matematicamente e por algoritmo, como é produzido as funções e a utilização dos métodos matemáticos como a Distância Euclidiana para demonstrar a Similaridade de Usuários (S.U.) e Itens (S.I.), onde ambos tem um tratamento diferente com os dados. O algoritmo foi produzido em Python 3 e os testes, assim como apresentação do código pela IDE Jupyter Notebook.

O sistema prático dispõe de um banco de dados com seis usuários e sete filmes, para ter uma representação mais lúdica. Não foi utilizado o banco de dados MovieLens

no sistema, pois o mesmo não contém nomes de usuários, ao invés disto, usa-se um ID numérico por questões de privacidade do usuários que fizeram parte do projeto MovieLens. Com base nisso para melhor demonstração adotamos um banco de dados fictício menor para melhor funcionalidade.

7. Conclusão

O resultado deste projeto foi um sistema de recomendação feito em arquivo executável de Python (Prompt de Comando/cmd) com as seguintes funcionalidades: 1 - Entrar no perfil; 2 - Criar perfil; 3 - Sair do sistema (FIGURA 1). Entrar no perfil permite acessar o menu com opções de usuário (FIGURA 2). Este menu tem: Historico de Filmes; Adicionar Filmes; Excluir Filmes; Filmes Recomendados; Mudar Notas; Excluir Perfil; Voltar ao menu principal. A opção Recomendar Filmes é a principal do sistema, é esta que utiliza Ciênca de Dados para recomendar um filme, através de predições, ao usuário (FIGURA 3). Criar perfil permite cadastrar um nome de usuário (”Nick”) e, opcionalmente, adicionar filmes que você assistiu, a partir do banco de dados do sistema com a nota classificatoria de 0.0 a 5.0 (FIGURA 4)

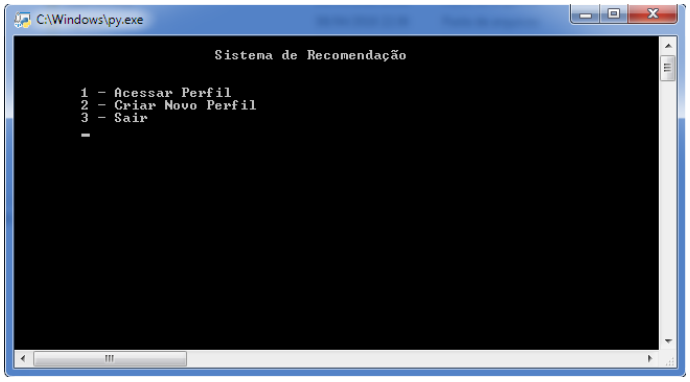

Figura 1. Menu Principal do Sistema

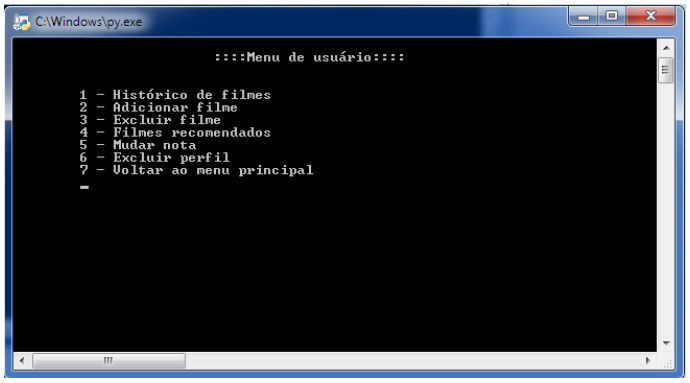

Figura 2. Menu de Usuário

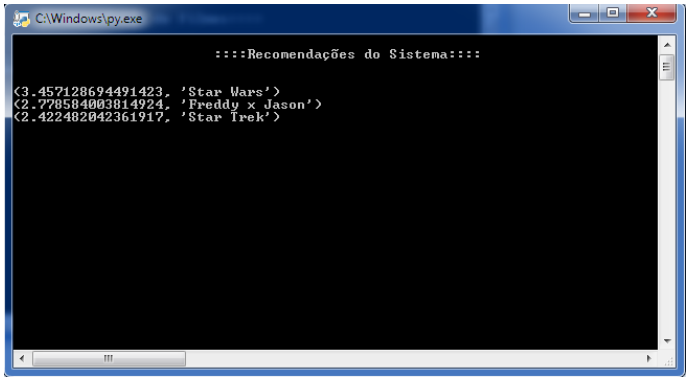

Figura 3. Recomendação de Filmes

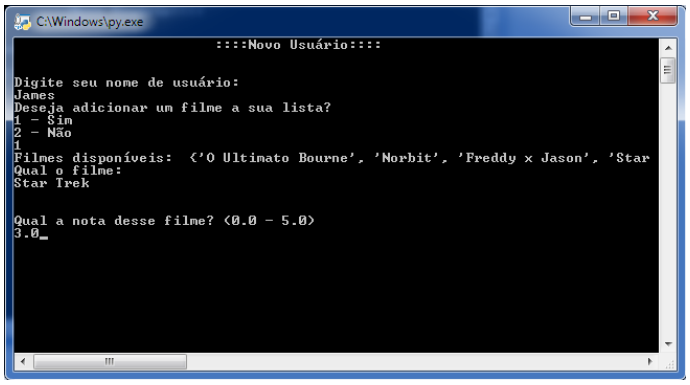

Figura 4. Menu Principal do Sistema

Referências

[1] Brian Wansink Aviva Musicus, Aner Tal. Eyes in the aisles - why is cap'n crunch looking down at my child? *Environment and Behavior*, 47, April 2014.

[2] Danielle Sandler dos Passos. Big data, data science e seus contributos para o avanço no uso da open source intelligence. *Sistemas Gestão*, 11, 2016.

[3] Lucas Coelho. Ciência de dados: O que é, conceito e definição, 2017.

[4] Everton Gago. Filtragem colaborativa, identificação de usuário e técnicas de ux para recomendação de conteúdo, 2017.

[5] Data Camp. Aprenda ciência de dados online, SI.

[6] Universidade de Stanford. Aprendizagem automática, SI.

[7] Organization Scikit Learn. scikit-learn - aprendizado de máquina em python, SI.

[8] Group MovieLens. Movielens, SI.

[9] Company SAS Company Software and Analytics Business Services. Machine learning - o que é e qual sua importância?, SI.

[10] Analytics Vidhya. Caminho de aprendizagem: seu mentor para se tornar um especialista em aprendizado de máquina, SI.