



Universidade de Brasília - UnB

Faculdade de Economia, Administração, Contabilidade e Gestão de Pol. Públicas - FACE

Departamento de Administração - ADM

# O USO DE ALGORITMOS DE RECOMENDAÇÃO NA SELEÇÃO DE DISCIPLINAS: UM ESTUDO DE CASO

LEANDRO MAGALHÃES LAMEGO

Brasília

2020



O USO DE ALGORITMOS DE RECOMENDAÇÃO NA SELEÇÃO DE  
DISCIPLINAS: UM ESTUDO DE CASO

Monografia apresentada ao Departamento de  
Administração - ADM como requisito parcial à  
obtenção do título de Bacharel em Administra-  
ção.

Orientador: Prof. Dr. Vinicius Amorim Sobreiro.

Ficha catalográfica elaborada automaticamente,  
com os dados fornecidos pelo(a) autor(a)

Lamego, Leandro Magalhães  
L228 O uso de algoritmos de recomendação na seleção de disciplinas: Um estudo de caso / Leandro Magalhães Lamego; orientador Vinicius Amorim Sobreiro. -- Brasília, 2020.  
62 p.

Monografia(Graduação - Administração) -- Universidade de Brasília, 2020.

1. Sistema de Recomendação. 2. Ensino Superior. 3. Escolha do Usuário. I. Sobreiro, Vinicius Amorim, orient. II. Título.

## DEDICATÓRIA

Dedico esse trabalho ao meu pai, Luis Henrique Moreira Lamego e a minha mãe Heloísa Helena Magalhães Lamego, que desde pequeno sempre acreditaram em mim apesar de todas dificuldades, sempre me proporcionando suporte, e estiveram sempre presentes. Sou profundamente grato à vocês.

Amo muito vocês, Leandro.



## AGRADECIMENTOS

Agradeço à minha família que sempre esteve presente na minha vida me ajudando e dando o suporte necessário para que eu chegasse até aqui. Um agradecimento em especial para o meu irmão Vitor Lamego, por me auxiliar em uma parte crítica do projeto isto é, me ajudou com o desenvolvimento de algumas ferramentas para facilitar meu trabalho.

Agradeço também a todos meus amigos por todos os momentos e suporte, vocês foram muito importantes para tornar esse período mais que especial e inesquecível.

Por último, agradeço ao meu orientador, Prof. Dr. Vinicius Amorim Sobreiro, pela sua disposição e paciência em ajudar e estar sempre preocupado com a qualidade deste trabalho.

LEANDRO MAGALHÃES LAMEGO





## EPÍGRAFE

*Qualquer tecnologia avançada é indistinguível da mágica*

*Arthur C. Clarke*

*O Dilema das Redes (Documentário)*



## RESUMO

A tomada de decisões é uma atividade que ocupa grande parte do dia a dia das pessoas e, ultimamente, vem se tornando uma tarefa árdua visto a grande quantidade de opções existentes. Isso se aplica também aos discentes do curso de Administração - ADM, da Universidade de Brasília - UnB, que no início de todo semestre precisam escolher que disciplina optativa cursar, fazendo com que muitos passem por dúvidas e indecisões. Atualmente, diversas ferramentas foram desenvolvidas para auxiliar na tomada de decisão das pessoas. Entre as existentes, se destaca o Sistema de Recomendação - S.R., pela sua simplicidade e eficácia. Conseqüentemente, o objetivo deste trabalho é propor a utilização de um S.R. no processo de escolha de disciplinas por parte do discente. Para tanto um S.R. foi testado considerando os discentes dos cursos de Administração, da Faculdade de Economia, Administração, Contabilidade, e Gestão de Políticas Públicas - FACE, da UnB. Os resultados obtidos indicam que a utilização de um S.R. na escolha de matérias auxilia os discentes tornando esse processo mais ágil e menos confuso.

**Palavras-chaves:** Sistema de Recomendação; Ensino Superior; Escolha do Usuário.



## ABSTRACT

Decision making is an activity that occupies a large part of people's daily lives and, lately, it has become an arduous task given the large number of existing options. This also applies to students of the Administration course - ADM, at the University of Brasilia - UnB, who at the beginning of every semester need to choose which elective course to take, causing many to go through doubts and indecision. Currently, several tools have been developed to assist in people's decision making. Among the existing ones, the Recommendation System - S.R. stands out, for its simplicity and effectiveness. Consequently, the objective of this work is to propose the use of an S.R. in the process of choosing subjects by the student. Therefore, an S.R. it was tested considering the students of the Administration courses, at the Faculty of Economics, Administration, Accounting, and Public Policy Management - FACE, from UnB. The results obtained indicate that the use of an S.R. when choosing subjects, it helps students making this process more agile and less confusing.

**Keywords:** Recommendation System; Higher Education, User Choise.



## RESUMEN

La toma de decisiones es una actividad que ocupa gran parte del día a día de las personas y, últimamente, se ha convertido en una ardua tarea dada la gran cantidad de opciones existentes. Esto también se aplica a los estudiantes del curso de Administración - ADM, de la Universidad de Brasilia - UnB, que al inicio de cada semestre deben elegir qué curso optativo tomar, lo que provoca que muchos pasen por dudas e indecisiones. Actualmente, se han desarrollado varias herramientas para ayudar en la toma de decisiones de las personas. Entre los existentes, destaca el Sistema de Recomendación - S.R., por su sencillez y eficacia. En consecuencia, el objetivo de este trabajo es proponer el uso de un S.R. en el proceso de elección de asignaturas por parte del alumno. Por lo tanto, un S.R. se evaluó considerando los alumnos de los cursos de Administración, de la Facultad de Economía, Administración, Contabilidad y Gestión de Políticas Públicas - FACE, de la UnB. Los resultados obtenidos indican que el uso de un S.R. a la hora de elegir asignaturas, ayuda a los estudiantes a hacer este proceso más ágil y menos confuso.

**Keywords:** Sistema de Recomendaciones; Enseñanza superior, Elección del usuario.





## LISTA DE FIGURAS

2.1	Adoção de tecnologia nos lares dos EUA. . . . .	7
2.2	Horas diárias gastas com mídia digital . . . . .	8
2.3	Diagrama de Venn. . . . .	12
3.1	Estrutura básica de um S.R. . . . .	22
3.2	Estrutura do S.R. para sugestão de matérias. . . . .	27
4.1	Ambiente <i>Spyder</i> . . . . .	36
4.2	Explorador de variáveis. . . . .	37
4.3	Executando o código. . . . .	37
4.4	Exemplo de resultados. . . . .	38
4.5	Gráfico da distribuição de acertos do S.R. (Pré-teste). . . . .	40
4.6	Resultados usuário 0. . . . .	41
4.7	Resultados para o usuário 1. . . . .	42
4.8	Resultados para o usuário 2. . . . .	43
4.9	Gráfico da distribuição de acertos do S.R. . . . .	46



## LISTA DE TABELAS

2.1	Crescimento do tempo em horas de uso diário em redes sociais por geração. . . . .	9
2.2	Média diária em horas de tempo em redes sociais por país . . . . .	9
2.3	Modelos de S.R's Híbridos . . . . .	18
2.4	Exemplos da utilização de S.R. . . . .	20
3.1	Equações de similaridade. . . . .	28
4.1	Lista de disciplinas. . . . .	39
4.2	Resumo dos resultados do pré-teste. . . . .	40
4.3	Resumo dos resultados (Parte I). . . . .	44
4.4	Resumo dos resultados parte (Parte II). . . . .	45
4.5	Relação entre número de aparecimentos e recomendações parte (Parte I). . . . .	47
4.6	Relação entre número de aparecimentos e recomendações parte (Parte II). . . . .	48



# SUMÁRIO

DEDICATÓRIA	v
AGRADECIMENTOS	vii
EPÍGRAFE	ix
RESUMO	xi
ABSTRACT	xiii
RESUMEN	xv
LISTA DE FIGURAS	xvii
LISTA DE TABELAS	xix
SUMÁRIO	xxi
I INTRODUÇÃO	I
2 REVISÃO DA LITERATURA	5
2.1 Surgimento da <i>Internet</i> . . . . .	5
2.2 <i>Boom</i> da <i>Internet</i> . . . . .	6
2.3 Surgimento do <i>Big Data</i> . . . . .	9
2.4 <i>MapReduce</i> . . . . .	11
2.5 <i>Data Science</i> . . . . .	12
2.6 Sistemas de Recomendação . . . . .	13
2.6.1 Surgimento do S.R. . . . .	13
2.6.2 Modelos de S.R. . . . .	14
2.6.3 Aplicações dos S.R's . . . . .	19
3 MÉTODO	21
3.1 Estrutura dos S.R's . . . . .	21
3.1.1 Identificação do Usuário . . . . .	22
3.1.2 Coleta de Dados . . . . .	22
3.1.3 Estratégias de Recomendação . . . . .	23
3.1.4 Visualização das Recomendações . . . . .	23
3.2 Similaridade entre Usuários . . . . .	24
3.3 Código . . . . .	25
4 RESULTADOS	35
4.1 Ambiente . . . . .	35
4.2 Interpretação de Resultados . . . . .	37

4.3	Lista . . . . .	38
4.4	Pré-teste . . . . .	38
4.5	Resultados . . . . .	39
4.5.1	Dados: Usuário 0 . . . . .	41
4.5.2	Dados: Usuário 1 . . . . .	42
4.5.3	Dados: Usuário 2 . . . . .	42
4.5.4	Resultados Individuais . . . . .	43
4.5.5	Avaliação de Resultados . . . . .	45
5	CONCLUSÃO	49
	APÊNDICE	52
	REFERÊNCIAS BIBLIOGRÁFICAS	56

# CAPÍTULO I

## INTRODUÇÃO

“Nada é mais difícil e, portanto, tão precioso, do que ser capaz de decidir”  
—Napoleão Bonaparte

A tomada de decisão é uma atividade que acompanha as pessoas no dia a dia, desde a hora que ela acorda, até a hora que vai dormir. Inúmeros tipos de decisões, com complexidades diferentes, tem que ser tomadas durante um dia, tanto qual caminho uma pessoa deve utilizar para chegar ao seu trabalho quanto que funcionário um *Chief Executive Officer* - CEO<sup>1</sup> deve demitir para reduzir custos da sua empresa. Para se ter uma ideia de quantas decisões podem ser tomadas, pode-se observar o número de escolhas que uma pessoa tem que pensar só para escolher o que comer durante um dia, que de acordo com Lang (2006, p. 1) se aproxima de 221 decisões em média.

De acordo com Slovic, Lichtenstein, Sarah, e Fischhoff (1998, p. 37), apesar da tomada de decisão poder envolver diversos tipos de assunto, basicamente todas decisões tem um objetivo em comum, que é definir a alternativa que maximiza o valor esperado dentre as decisões. Ao longo do tempo, muitos modelos matemáticos foram desenvolvidos para criar teorias de tomada de decisão, os mais modernos são: teoria da utilidade subjetiva esperada, teoria de aversão ao risco, teoria do prospecto, teoria do arrependimento, entre outros.

Percebe-se que a maioria das decisões tomadas não são opções binárias<sup>2</sup>, do tipo sim ou não, uma opção ou outra, na maioria das vezes são apresentadas uma grande variedade de escolhas, que devem ser ponderadas, descartadas e, no final, escolhidas. Conforme Schwartz (2009, pp. 120–126), em seu livro “O paradoxo da escolha”, essa quantidade de escolhas traz dois efeitos negativos nas pessoas. A primeira delas é a paralisia, na qual existem tantas opções para serem escolhidas que as pessoas acham difícil tomar essa decisão, que às vezes acabam postergando e deixando de escolher ou até mesmo, tomando decisões precipitadas. O segundo efeito é que o custo de oportunidade diminui a satisfação das escolhas, ou seja, mesmo conseguindo tomar

---

<sup>1</sup>Tradução livre: Diretor executivo.

<sup>2</sup>Binário é o adjetivo masculino que indica algo que tem duas unidades ou algo que é composto por dois elementos de informação.

uma decisão entre várias opções, a pessoa acaba menos satisfeita com o resultado dessa escolha, pois ela tende a compará-la com as opções descartadas.

Atualmente, a quantidade de informações, principalmente na *internet*, é muito grande, tornando ainda mais difícil o processo de tomada de decisão, tanto para o usuário, quanto para a pessoa que possui todos os dados dos usuários. O nome atribuído a essa quantidade de dados é *Big Data*, que conforme TAF (2012, p. 10) é o grande volume de informações rápidas, complexas e variadas que necessita de técnicas avançadas para coletá-las, armazená-las, distribuí-las, gerenciá-las e analisá-las.

Ainda, considerando as ideias da TAF (2012), na visão de Gandomi e Haider (2015, p. 138), o *Big Data* se resume em 3 principais dimensões, a saber:

- Volume: Se refere à magnitude das informações;
- Variedade: Se refere à heterogeneidade estrutural em um conjunto de dados;
- Velocidade: Se refere à velocidade em que os dados são gerados e que devem ser analisados e acionados.

Visando facilitar o processo de tomada de decisão, uma nova área disciplinar chamada Ciência de Dados ou *Data Science* surgiu para coletar um grande número de informações, que podem ser de diversas áreas de estudo, estruturá-las e, assim, gerar *insights*. Segundo Dhar (2013, p. 7), do ponto de vista da tomada de decisão, essa é a era na qual os computadores são inerentemente melhores tomadores de decisão que os humanos, pela capacidade de processamento de uma grande quantidade de informações. O autor até descreve que em algumas áreas, como finanças, os computadores já tomam a maioria das decisões de investimentos.

Existem inúmeras técnicas criadas pelos pesquisadores nessa área ou cientistas de dados, que são aplicadas nas mais diversas áreas de conhecimentos. Em seu livro, Grus (2015, pp. 5–11) apresenta algumas aplicações dessas técnicas, como, por exemplo, Inteligência Artificial - I.A., Propagandas Digitais - P.D., Reconhecimento de Imagem - R.I., Redes Neurais - R.N., Sistemas Recomendadores - S.R., entre outros. Visto sua simplicidade e o problema abordado neste trabalho, os S.R's foram escolhidos para serem aprofundados de forma mais específica.

Os S.R's foram desenvolvidos para melhorar a experiência do usuário que, hoje em dia, possui uma grande sobrecarga de informações oriundas de diversos tipos de serviços. De acordo com Ricci, Rokach, e Shapira (2010, p. 1), os S.R's são ferramentas e técnicas computacionais que fornecem sugestões de itens para usuários. Esses itens estão relacionados a uma série de tomada de decisões do usuário e retornam sugestões como qual livro ler, qual música escutar, qual produto escolher, entre outros.

Pode-se observar alguns tipos de S.R. em diferentes plataformas, como, por exemplo, na *Amazon*<sup>®</sup>, *Netflix*<sup>®</sup>, *Spotify*<sup>®</sup> etc. De acordo com MacKenzie, Meyer, e Noble (2013, pp. 4–5), os sistemas de recomendação são responsáveis por 35% das compras na *Amazon*<sup>®</sup> e 75% do que é assistido na *Netflix*<sup>®</sup>, mostrando ser um artifício bastante efetivo e expressivo nessas plataformas.

No âmbito universitário, ao longo de um curso de graduação, uma das principais tomadas de decisão é a escolha de que matérias optativas cursar ao longo do período acadêmico. Para fins de experimentações computacionais observando esse contexto, o caso dos discentes do Departamento de Administração - ADM, da Faculdade de Economia, Administração, Contabilidade, e Gestão de Políticas Públicas - FACE, da Universidade de Brasília - UnB, foi escolhido para ser estudado. Dado o contexto, a situação problema enfrentada pelos discentes pode ser abordada com a utilização de S.R's.

De acordo com o Projeto Pedagógico do curso de Administração, diurno e noturno, para completar a sua formação, um discente precisa completar um total de 200 créditos em seu currículo escolar, sendo que 110 são créditos de matérias obrigatórias e até 90 de matérias optativas



(ADM, 2018, p. 5). Esses 90 créditos são obtidos por meio de uma média de 22 matérias, que podem ser escolhidas dentro do próprio departamento do discente ou em departamentos próximos como, por exemplo, Economia, Contabilidade, Gestão Pública, entre outros.

No esforço de encontrar estas 22 matérias ao longo do curso, os alunos passam por muitas dificuldades. Uma delas é encontrar matérias adequadas a sua formação acadêmica, seja ela finanças, marketing, área pública etc. Isso faz com que os alunos passem muito tempo procurando essas matérias e tentando decidir quais são as melhores opções, o que se encaixa na situação de paralisia, indicada por Schwartz (2009, p. 145). Essa sobrecarga de informação também pode fazer com que os estudantes deixem matérias importantes para a sua formação passarem despercebidas.

Outro problema relacionado a essa escolha é a sobrecarga no sistema da UnB, pois quando é aberto o período de matrícula, um grande volume de discentes entram ao mesmo tempo no sistema e, como dito anteriormente, passam muito tempo pesquisando a respeito de diversas matérias causando instabilidade, site fora ar, lentidão etc.

Para evidenciar melhor o ponto de vista dos discentes, o relato de três alunos do Departamento de Administração é apresentado a seguir:

A primeira pergunta realizada foi a respeito da experiência de início de semestre no sistema *Matrícula Web*<sup>3</sup>, os entrevistados responderam o seguinte:

*“No geral acho o Matrícula Web um site instável e pouco intuitivo. Minha experiência muitas vezes não foi satisfatória.” - Túlio Torres do Val, 12/07/2020.*

*“Então, o sistema é muito ruim, mas ele cumpre o papel dele, minimamente.” - Lucas Lacerda 12/07/2020.*

*“Como aluno por 8 semestres utilizando o Matrícula Web, em cada início de semestre, tive uma boa experiência com o Matrícula Web. Alguns pontos poderiam ser esclarecidos que facilitariam a vida dos alunos, como que as matrículas nas matérias não são vinculadas ao pedido de matrícula nas matérias, causando sobrecarga no sistema principalmente nos primeiros dias.” - Matheus Pena Barbosa, 12/07/2020.*

*“Sobre o início de semestre no Matrícula Web, é saber as datas do período de matrícula, enfim, não tem muita comunicação sobre isso, sempre foi no boca a boca dos alunos e eu sempre fico sabendo pelos meus amigos, e isso dificulta bastante até pra gente se organizar. E outra questão é que, uma coisa que eu sempre tive comigo, é por que as datas não são fixas?” - Vinicius Zawadzki, 12/07/2020.*

A segunda pergunta questionava se o discente já passou por algum tipo de dificuldade ao acessar o *Matrícula Web*, de acordo com eles:

*“Os problemas mais constantes que tive no Matrícula Web foram sistema fora do ar e problemas no momento do processamento das matérias.” - Matheus Pena Barbosa, 12/07/2020.*

*“Algumas vezes sofri um pouco com a lentidão do site e com um tipo de erro que aparecia no momento de enviar a proposta das matérias escolhidas.” - Túlio Torres do Val, 12/07/2020.*

*“Tipo assim, o sistema sempre é muito lerdo, as vezes dá problema na fila de espera. Uma coisa que eu acho muito ruim e me irrita muito é que as vezes o sistema te diz que tem vaga na turma, aí você pega a matéria, tenta se matricular, e quando vai finalizar, o sistema te diz que tem fila de espera pra matéria. Pô, eu perdi um tempo aí, o sistema poderia dizer só que não tem vaga, tem fila de espera.” - Lucas Lacerda 12/07/2020.*

---

<sup>3</sup>Em termos simples, o *Matrícula Web* é o sistema utilizado na UnB para controlar o fluxo de disciplinas dos discentes.

*“Eu não me lembro de ter problema no sistema, sempre foi muito tranquilo, sempre que troquei de matéria ele cancelou na hora e me matriculou na que tinha vaga e na que não tinha vaga eu entrei na lista de espera tranquilo.” - Vinícius Zawadzki, 12/07/2020.*

A última pergunta questionava se o discente tinha alguma dificuldade para escolher suas matérias ou achar matérias que o agradavam. Os entrevistados responderam o seguinte:

*“Devido a grande quantidade de departamentos dentro da universidade, o aluno quando precisa procurar novas matérias tem dificuldade, de modo geral. Uma apresentação que destaque o departamento do aluno, departamentos com currículo similares, matérias obrigatórias pendentes ou matérias mais buscadas por alunos do curso podem auxiliar na escolha.” - Matheus Pena Barbosa, 12/07/2020.*

*“Acredito que essa maior dificuldade em escolher as matérias seja algo prejudicial para o planejamento dos estudantes. A enorme quantidade de matérias disponíveis confunde os alunos no momento da escolha.” - Túlio Torres do Val, 12/07/2020.*

*“As análises que o Matrícula Web faz pra entender qual matéria te ofertar, essa inteligência dele, é horrível, tipo assim, quase nunca me ofertam uma matéria que eu realmente precise pegar, em geral eu sempre tenho que cancelar tudo e ir atrás de outras matérias, e isso me demanda muito tempo porque são muitas para olhar, ver plano de matéria e tentar encaixar na sua grade.” - Lucas Lacerda 12/07/2020.*

*“Mas sobre a escolha de matérias, eu acho que a estrutura que é feita no site do Matrícula Web, na plataforma, é muito ruim, por que ela é dividida por departamentos, e eu não conheço todas as matérias dos departamentos, então pra eu escolher uma matéria eu tenho que ir departamento por departamento vendo as matérias ofertadas, vendo se alguma me interessa, acho que se tivesse um painel que mostrasse todas as matérias facilitaria muito também, pra eu escolher uma matéria mais diferente, uma matéria que eu nem fazia ideia que eu gostaria de fazer, as vezes ali aparecendo pra mim.” - Vinícius Zawadzki, 12/07/2020.*

Diante do contexto, esse trabalho **tem por objetivo verificar a possibilidade da aplicação de um S.R., baseado no usuário, na recomendação de disciplinas**, dessa forma, espera-se que os discentes farão escolhas mais assertivas e mais condizentes com sua formação acadêmica evitando alguns dos problemas citados anteriormente, como a paralisia e a incerteza devido ao custo de oportunidade. Também, como os alunos passarão menos tempo navegando no sistema da UnB para decidir quais matérias vão escolher, o mesmo ficará mais aliviado, e menos pessoas terão problemas de instabilidade.

O S.R. utilizado neste trabalho é baseado no livro *“Data Science do Zero, primeiras regras com python”*, de Grus (2015, pp. 267–274). No livro são apresentados alguns métodos de Sistemas Recomendadores, entre eles está recomendar o que é mais popular no geral, a filtragem colaborativa baseada no usuário, que indica itens de acordo com usuários similares a você e a filtragem colaborativa baseada em itens, que indica itens de acordo com similares aos seus interesses atuais.

O método escolhido foi a filtragem colaborativa baseada no usuário, que irá utilizar a técnica similaridade do cosseno para criar um vetor entre o discente e os usuários já cadastrados na base de dados e medir sua similaridade. Esse método foi escolhido, pois como os alunos de administração podem projetar o seu curso de acordo com algumas frentes, por exemplo, finanças, marketing, gestão de pessoas, focado em concurso etc. Consequentemente, seria lógico indicar matérias de alunos que tiveram uma grade curricular semelhante.

A seguir neste trabalho será apresentado no Capítulo 2 uma breve revisão da literatura a respeito de S.R., no contexto da *Data Science*, em seguida, no Capítulo 3 será abordado o método utilizado para o desenvolvimento do S.R. usado para recomendar matérias aos discentes

chegando ao Capítulo 4, no qual é apresentado o resultado da pesquisa e, por último, no Capítulo 5 são listadas as conclusões finais deste trabalho.



## CAPÍTULO 2

# REVISÃO DA LITERATURA

“Sem ideias, a evolução humana seria  
inexplicável.”  
—José Ingenieros.

Para entender melhor os conceitos que compõem a área de *Data Science*, uma revisão da literatura por partes é apresentada. Em um primeiro momento, o surgimento da *internet* e seu *boom* são abordados. Na sequência, o aparecimento dos *smartphones* e como isso também influenciou no *boom* da *internet* é contemplado. Considerando os conceitos previamente apresentados, uma conceituação de *Big Data* é expressa. Posteriormente, o desenvolvimento do *MapReduce*, para tratamento de dados, no âmbito de *Data Science*, é listado.

### 2.1 SURGIMENTO DA *INTERNET*

---

É um fato que o surgimento da *internet* foi um evento que mudou a forma como a sociedade vive. No entanto, a maioria dos seus usuários não entendem como a mesma funciona ou qual foi a ideia com que fez essa ferramenta surgir.

A ideia da *internet* em si pode parecer um pouco abstrata. Entretanto, autores como Vicentini, Lanzoni, Franzotti, e Yonenaga (2005, p. 1), definem ela como uma rede pública de comunicação de dados, com controle descentralizado e que utiliza o conjunto de protocolos, *Transmission Control Protocol - TCP*<sup>1</sup> e *Internet Protocol - IP*<sup>2</sup>, como base para a estrutura de comunicação e seus serviços de rede.

De acordo com Vicentini et al. (2005, pp. 2–3), a *internet* surgiu a partir de uma necessidade do Ministério de Defesa Americano em proteger os importantes e sigilosos dados militares durante a Guerra Fria.

---

<sup>1</sup>Tradução livre: Protocolo de Controle de Transmissão.

<sup>2</sup>Tradução livre: Protocolo de *internet*.

Mais precisamente, essa ideia veio a tona quando os militares começaram a se preocupar com um ataque no qual seus dados fossem parcialmente ou totalmente destruídos. Dessa forma, surgiu a ideia da criação de uma rede eletrônica aonde esses dados fossem espalhados em vários locais, evitando assim, um ataque concentrado e que no caso de modificação desses dados, os outros pontos fossem atualizados também.

Conforme visto no livro de Coffman e Odlyzko (2002, p. 28), em 1969, a *Advanced Reserch Projects Agency (ARPA)* inaugurou a chamada *ARPANet*, o que futuramente viria a se tornar *internet*. No ano de 1971, essa rede foi liberada para o uso de universidades nos EUA e, conseqüentemente, fez com que um modelo experimental do *e-mail* fosse criado e em 1973 as primeiras conexões internacionais foram criadas.

Ainda em seu livro, os autores Coffman e Odlyzko (2002, p. 31) informam que no ano de 1983, toda a parte militar se desvinculou da *ARPANet* e criou uma rede própria, a *MILNet*. Nesse momento, o conjunto de computadores interligados começaram a ser chamados de *internet*.

O ano de 1987 foi bem marcante para o impulsionamento e crescimento da *internet*, pois de acordo com L. W. Silva (2001, p. 1) foi o ano em que pela primeira vez essa rede foi liberada para seu uso comercial. Com tal liberação, várias empresas de provedoras de acesso começaram a surgir, e foi em 1992 que o Laboratório de Física de Partículas do *Conseil Européen pour la Recherche Nucléaire - CERN* inventou a *World Wild Web*, o que é conhecido popularmente como “WWW”, que, por seu turno, começou a ser utilizada para colocar informações ao alcance de qualquer usuário da *internet*.

A *internet* e a WWW são constantemente confundidas, mas não são sinônimos. Monteiro (2001, p. 29) define a WWW como “um espaço que permite a troca de informações multimídia (texto, som, gráficos e vídeo) através da estrutura da *internet*”. A WWW foi desenvolvida com o intuito de se tornar uma ferramenta de troca de informações mais amigável que as interfaces existentes, somente texto.

A partir desse ponto, a *internet* começou a crescer de forma exponencial. Para se ter uma noção de tal crescimento, de acordo com Coffman e Odlyzko (2002, p. 20), nos anos de 1995 e 1996, o número de usuários dobrava a cada 3 ou 4 meses, e a partir de 1997, esse número começou a dobrar a cada ano (até 2002, data de publicação do artigo).

Para acompanhar o crescimento de usuários, a tecnologia usada como base de sustentação da *internet* também teve que evoluir. No documentário *O Dilema das Redes*<sup>3</sup>, dirigido por Orłowski (2020), o entrevistado Randima Fernando<sup>3</sup> mostra que se for analisado de 1960 até hoje, será identificado que o poder de processamento aumentou um trilhão de vezes e que nenhuma outra tecnologia foi otimizada a esse nível.

De acordo com Monteiro (2001, p. 28), no Brasil, a *internet* começou a aparecer por volta do ano de 1995, quando o governo federal começou a atuar na implantação da infraestrutura necessária para a operação de empresas privadas provedoras de acesso aos usuários.

## 2.2 BOOM DA INTERNET

---

O acesso à *internet* é algo que vem se tornando cada vez mais fácil e, por isso, a cada dia é possível notar um crescimento nesta rede. Com objetivo de exemplificar esse fato, Monteiro (2001, p. 28) mostra que apenas em um ano, de 1996 até 1997, o número de usuários no Brasil cresceu aproximadamente 1000%, saindo de 170 mil usuários para 1,3 milhão.

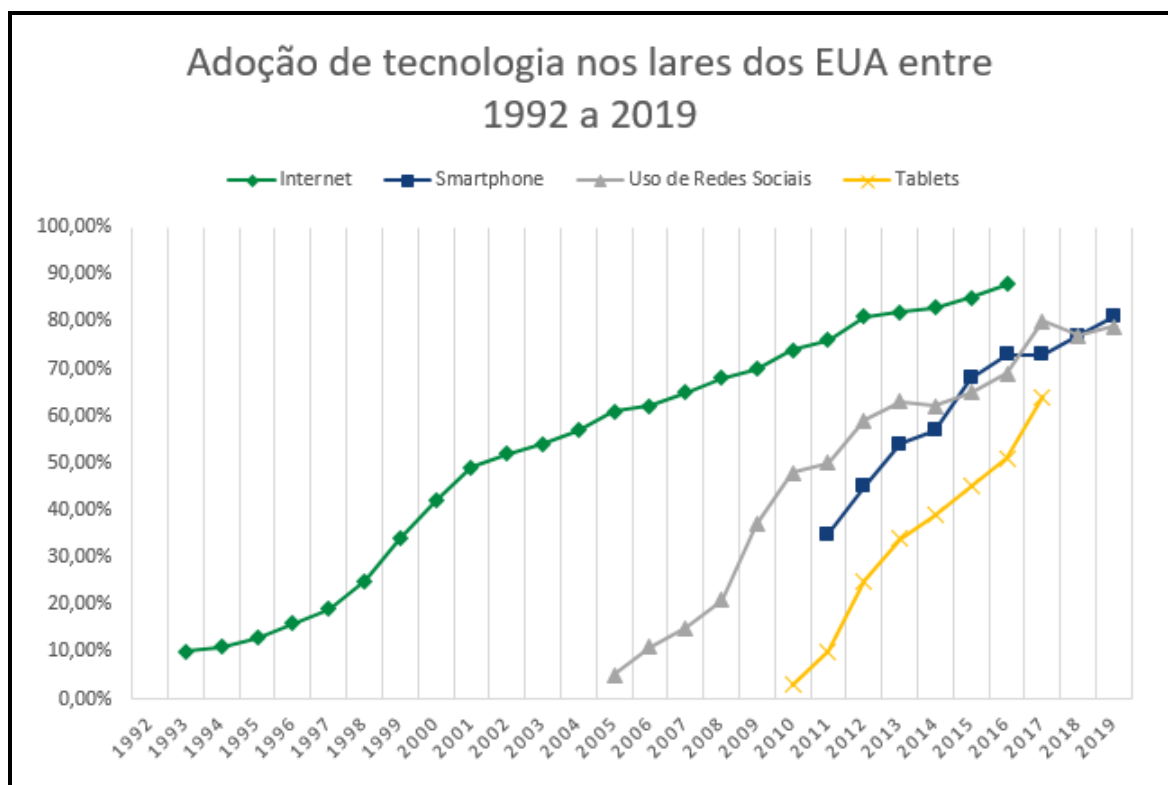
Um dos responsáveis pelo atual crescimento da *internet* são os famosos *SmartDevices*, que começaram a ser projetados há 26 anos, quando foi lançado o *Simon*, *smartphone* desenvol-

---

<sup>3</sup>Ex-gerente de produto da *Nvidia*<sup>®</sup>, ex-diretor executivo da *Mindful School*<sup>®</sup> e co-fundador do *Center for Humane Technology*<sup>®</sup>.

vido pela *International Business Machines Corporation - IBM*®. Ele possuía funções limitadas, mas foi um grande passo para a época. O próximo passo dado por essa indústria ocorreu com o lançamento do *StarTAC*®, desenvolvido pela Motorola®, em 1996, considerado um dos mais importantes aparelhos para a telefonia móvel. Em 2002, foi lançado um dos celulares mais famosos da história, o *BlackBerry*®, o primeiro do mercado a possuir um teclado *QWERTY*<sup>4</sup>. Por último, em 2007, houve o lançamento do primeiro *iPhone*®, da companhia *Apple*®, que revolucionou o mercado com seu *smartphone touch*.

Para entender melhor esse crescimento, Desjardins (2018, p. 1) mostra como essas tecnologias vêm crescendo nos últimos tempos, conforme apresentado na Figura 2.1:



**Figura 2.1:** Adoção de tecnologia nos lares dos EUA.

Fonte: Adaptado de Ali (2020, p. 1).

Esse crescimento se deu porque os *smartphones* são considerados uma parte essencial do dia a dia das pessoas, para mandar mensagens importantes, ver notícias, tirar fotos, acessar redes sociais, entre outras diversas possibilidades. Os autores Marty-Dugas e Smilek (2020, p. 1) citam que, desde 2015, a porcentagem de usuários de *smartphones* continuou a aumentar em vários países e diferentes faixas de idades, tanto nos jovens, que são considerados os *heavy users*<sup>5</sup> quanto no público com mais idade<sup>6</sup>.

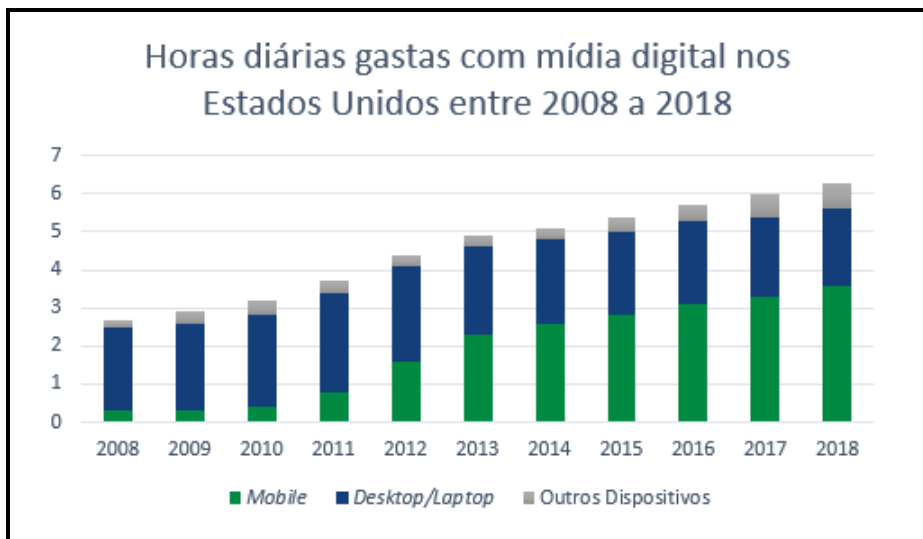
Para mostrar que esses aparelhos possuem uma expressividade significativa no dia a dia das pessoas, um estudo de 2016, realizado por Winnick (2016, p. 1), mostrou que um usuário comum costuma realizar uma média de 2.617 (Dois mil seiscentos e dezessete) toques diários em seus *smartphones* em 76 interações por dia, enquanto os *heavy users* possuem uma média de 5.427 (Cinco mil quatrocentos e vinte sete) em 132 interações. A instituição *Our World in Data*<sup>7</sup> realizou uma pesquisa para mostrar a evolução do tempo gasto diário em horas com mídias digitais, conforme na Figura 2.2:

<sup>4</sup>Layout de teclado para o alfabeto latino atualmente mais utilizado em computadores e máquinas de escrever.

<sup>5</sup>Termo destinado a pessoas que utilizam com muita frequência um certo produto ou serviço.

<sup>6</sup>Pessoas com idade igual ou superior a cinquenta anos.

<sup>7</sup>É uma publicação digital especializada em expor pesquisas empíricas e dados analíticos sobre mudanças nas condições de qualidade de vida ao redor do mundo.



**Figura 2.2:** Horas diárias gastas com mídia digital.  
**Fonte:** Adaptado de OWD (2019, p. 1).

No documentário *O Dilema das Redes*<sup>8</sup>, Orłowski (2020)<sup>8</sup> explica que a fixação das pessoas em seus *smartphones* e redes sociais está diretamente ligada a uma necessidade básica das pessoas de se conectarem. Ainda nesse documentário, a Dr. Anna Lembke<sup>9</sup> explica que durante a evolução, os humanos criaram mecanismos, como a liberação de dopamina como forma de recompensa por esse tipo de interação, e com as novas tecnologias, isso se tornou mais rápido e fácil. Entretanto, a interação com os *smartphones* pode trazer resultados negativos para a saúde das pessoas, como mostrado na pesquisa de Marty-Dugas e Smilek (2020, p. 5), mostrando que há evidências crescentes de que o uso do *smartphone* parece ser inversamente relacionado à saúde e bem-estar, sendo responsável muitas vezes por problemas como ansiedade e depressão.

Ainda no documentário *O Dilema das Redes*<sup>10</sup>, de Orłowski (2020)<sup>10</sup>, o entrevistado Tristan Harris<sup>11</sup> explica a vulnerabilidade da mente humana. Ele explica que nas grandes empresas tecnológicas existentes hoje, existe o seguinte pensamento: “como podemos usar tudo que sabemos sobre psicologia para persuadir as pessoas e transformar isso em tecnologia?”. O nome dado a essa prática é tecnologia persuasiva e ela tem como intuito ser aplicada ao extremo buscando a mudança de comportamento das pessoas. O *Instagram*<sup>12</sup> é utilizado como exemplo, aonde é mostrado que aquele movimento de puxar a tela para baixo e a tela atualizar com informações completamente novas em psicologia é chamado de reforço intermitente positivo, no qual basicamente, o usuário não sabe quando nem se vai conseguir algo novo, e isso é exatamente a psicologia adotada em caça-níqueis<sup>12</sup>.

A utilização dessas técnicas de persuasão e mudança de comportamento vêm fazendo com que os hábitos de consumo e navegação mudem dentro dessas redes. Em um estudo realizado pela *Global Web Index*<sup>13</sup>, apresentado, por Viens (2019), na Tabela 2.1, os dados mostram essa alteração no comportamento do consumidor.

Conforme apresentado nas Tabelas 2.1 e 2.2, é notável que o comportamento das gerações vem mudando ao longo do tempo, e junto com essa mudança, existe um aumento da utilização da *internet*, principalmente quando são observadas as redes sociais. Junto à esse crescimento é inevitável a produção de cada vez mais dados, o que originou o tema abordado na próxima

<sup>8</sup>A partir de 32 minutos de documentário.

<sup>9</sup>Diretora médica de medicina anti-dependência da Universidade de *Stanford*.

<sup>10</sup>A partir de 22 minutos de documentário.

<sup>11</sup>Ex-especialista em ética de design da *Google*<sup>®</sup> e Co-fundador do *Center for Humane Technology*<sup>®</sup>.

<sup>12</sup>Máquina de jogo que funciona por meio da introdução de moedas e que paga um prêmio.

<sup>13</sup>Empresa de pesquisa de mercado fundada por Tom Smith, em 2009, que fornece informações sobre o público para editores, agências de mídia e profissionais de marketing em todo o mundo.



seção.

Geração	2014	2015	2016	2017	2018
Global.	1:41	1:49	2:06	2:15	2:18
Geração Z.	1:43	2:13	2:27	2:53	2:55
<i>Millennials</i> .	2:07	2:12	2:32	2:39	2:38
Geração X.	1:25	1:32	1:47	1:48	1:49
<i>Baby Boomers</i> .	0:53	0:57	1:06	1:09	1:12

**Tabela 2.1:** Crescimento do tempo em horas de uso diário em redes sociais por geração.

**Fonte:** Adaptado de Viens (2019, p. 1).

País	Tempo
Asia.	2:13
Europa.	1:50
América Latina.	3:27
Oriente Médio e África.	3:03
América do Norte.	2:04

**Tabela 2.2:** Média diária em horas de tempo em redes sociais por país.

**Fonte:** Adaptado de Viens (2019, p. 1).

### 2.3 SURGIMENTO DO *BIG DATA*

Como apresentado nas seções anteriores, a *internet* e suas tecnologias vem crescendo de maneira rápida e, com isso, cada vez mais seus usuários interagem e geram informações. Atualmente, a *internet* possui 4,5 bilhões de usuários que geram inúmeros tipos de dados por dia. Petri (2013, p. 1) mostra que “em quinze minutos, a humanidade gera o triplo de informações disponíveis no acervo da Biblioteca do Congresso Americano, a maior do mundo”.

Outra pesquisa que mostra o volume de informações geradas foi realizada por Ali (2020, p. 1). De maneira resumida, ele indica que a cada minuto por dia no início de 2020:

- São criadas 319 contas novas no *Twitter*®;
- São compartilhadas 41.666.667 mensagens no *WhastApp*®;
- São enviados 500 horas de vídeo no *YouTube*®;
- São enviados 69.444 currículos pelo *Linkedin*®;
- São gastos 1.000.000 de dólares;
- São assistidos 404.444 horas de vídeos no *Netflix*®;
- São postados 347.222 *stories*<sup>14</sup> no *Instagram*®;
- São enviados 6.659 pacotes pela *Amazon*®.

<sup>14</sup>É um recurso que tem como objetivo melhorar a interação entre os usuários. Consiste na possibilidade de publicar fotos ou vídeos que ficam acessíveis por até 24 horas.

Como visto acima, milhões de dados são gerados por minuto e, com o passar do tempo, essas informações começaram a ser produzidas de forma desestruturada e caótica, seja por meio de cliques em anúncios, *tweets*, fotos em plataformas, vídeos assistidos, hábitos de consumo etc. Essa quantidade gigantesca de informações deu origem a um conceito amplamente estudado nos dias de hoje, o *Big Data*. Esse termo surgiu em meados de 1997 e uma das suas primeiras citações foi feita por Cox e Ellsworth (1997, p. 1) definem *Big Data* como:

*A visualização oferece um desafio interessante para os sistemas de computador: os conjuntos de dados geralmente são muito grandes, sobrecarregando as capacidades da memória principal, do disco local e até do disco remoto. Chamamos isso de problema de Big Data. Quando os conjuntos de dados não cabem na memória principal (no núcleo), ou quando não cabem nem mesmo no disco local, a solução mais comum é adquirir mais recursos.*

O que diferencia a *Big Data* de dados comuns é relacionado com o seu volume e a maneira desestruturada como eles são apresentados, de acordo com Chen, Mao, e Liu (2014, p. 1). Ainda de acordo com esse autor, o conceito de *Big Data* é abstrato e, por isso, muitas pessoas ainda possuem diferentes opiniões a respeito do tema, mas, em geral, *Big Data* significa conjuntos de dados que não puderam ser percebidos, adquiridos, gerenciados e processados por ferramentas tradicionais de TI e *software/hardware* em um tempo tolerável.

Apesar do conceito de *Big Data* variar bastante ao longo do tempo, os autores Ylijoki e Porras (2016, p. 73) comentam que o ano de 2001 foi um marco para tal definição, pois a abordagem de *Big Data* passou a ser tratada a partir de 3 dimensões, que são: volume, velocidade e variedade. Ainda em seu artigo, Ylijoki e Porras (2016, p. 74) realizaram um estudo para consolidar essa definição. No artigo, eles chegaram a conclusão que de uma pesquisa com 62 artigos, 95% deles apresentavam a dimensão volume como uma das característica chave para a definição de *Big Data*. Quanto as outras dimensões, variedade apareceu em 88% dos artigos e velocidade em 74%.

Muitos autores baseiam suas pesquisas de *Big Data* com base em seus pilares. Um deles é Lee (2017, p. 294) que, por sua vez, traz um aprofundamento em cada uma de suas dimensões e tenta defini-las, como mostrado abaixo:

- **Volume:** Refere-se à quantidade de dados que uma organização ou um indivíduo coleta e/ou gera. Embora atualmente o mínimo de 1 *terabyte* seja o limite de *Big Data*, o tamanho mínimo para se qualificar como *Big Data* é uma função do desenvolvimento de tecnologia. Atualmente, 1 *terabyte* armazena tantos dados quanto caberia em 1.500 CDs ou 220 DVDs, o suficiente para armazenar cerca de 16 milhões de fotografias do *Facebook*<sup>©</sup>. O comércio eletrônico, as mídias sociais e os sensores geram grandes volumes de dados não estruturados, como áudio, imagens e vídeo. Novos dados foram adicionados em uma taxa crescente à medida que mais dispositivos de computação são conectados à *internet*;
- **Velocidade:** Refere-se à velocidade com que os dados são gerados e processados. A velocidade dos dados aumenta com o tempo. Inicialmente, as empresas analisavam dados usando sistemas de processamento em lote devido à natureza lenta e cara do processamento de dados. Conforme a velocidade de geração e processamento de dados aumentou, o processamento em tempo real tornou-se uma norma para aplicativos de computação. A capacidade aprimorada de *streaming*<sup>15</sup> de dados de dispositivos conectados continuará a acelerar a velocidade;
- **Variedade:** Refere-se ao número de tipos de dados. Os avanços tecnológicos permitem

---

<sup>15</sup>Forma de distribuição digital, em oposição à descarga de dados. A difusão de dados, geralmente em uma rede, por meio de pacotes, é frequentemente utilizada para distribuir conteúdo multimídia por meio da *internet*.

que as organizações gerem vários tipos de dados estruturados, semi-estruturados e não estruturados. Texto, foto, áudio, vídeo, dados de cliques e dados de sensores são exemplos de dados não estruturados, que não possuem a estrutura padronizada necessária para uma computação eficiente. Os dados semi-estruturados não estão em conformidade com as especificações do banco de dados relacional, mas podem ser especificados para atender a certas necessidades estruturais dos aplicativos. À medida que novas técnicas de análise são desenvolvidas, os dados não estruturados são gerados em uma taxa muito mais rápida do que os dados estruturados.

Para se entender melhor, o *Big Data* é um conceito amplo que se divide em algumas vertentes como, por exemplo, o *Big Social Data*, que de acordo com Seo, Kim, Lee, e Baik (2017, pp. 135–136) surgiu por meio dos Serviços de Redes Sociais - S.R.S, aonde os usuários podem expressar suas opiniões de maneira irrestrita e compartilhar seus interesses com outras pessoas. Essa participação espontânea de usuários em S.R.S resulta na geração de enormes quantidades de dados com várias características.

## 2.4 MAPREDUCE

---

O surgimento do *Big Data* foi algo que forçou a indústria da tecnologia a criar novas técnicas para processar tais dados, pois as técnicas existente não eram suficientes para processar e gerar informações com os dados apresentados. O desafio era criar uma técnica capaz de armazenar, manipular e gerar *insights* inteligentes sobre essa quantia enorme de dados desestruturados, principalmente de forma clusterizada, conforme apontam Ghazi e Gangodkar (2015, p. 46).

Dessa forma, o desenvolvimento da técnica chamada *MapReduce*, apresentada por Dean e Ghemawat (2004, p. 1) foi a precursora nesse tema. De acordo com os autores, o *MapReduce* é um modelo de programação e uma implementação associada para processar e gerar grandes conjuntos de dados. Os usuários especificam uma função de mapa (*Map*) que processa um par de chave/valor para gerar um conjunto de pares de chaves/valores intermediários e uma função de redução (*Reduce*) que mescla todos os valores intermediários associados à mesma chave intermediária.

De acordo com Goldman, Kon, Junior, Polato, e Pereira (2012, p. 4), o desenvolvimento do *MapReduce* possibilitou otimizar a indexação e catalogação dos dados sobre as páginas na *internet* e suas ligações, pois essa técnica permite dividir um grande problema em vários pedaços e distribuí-los em diversos computadores.

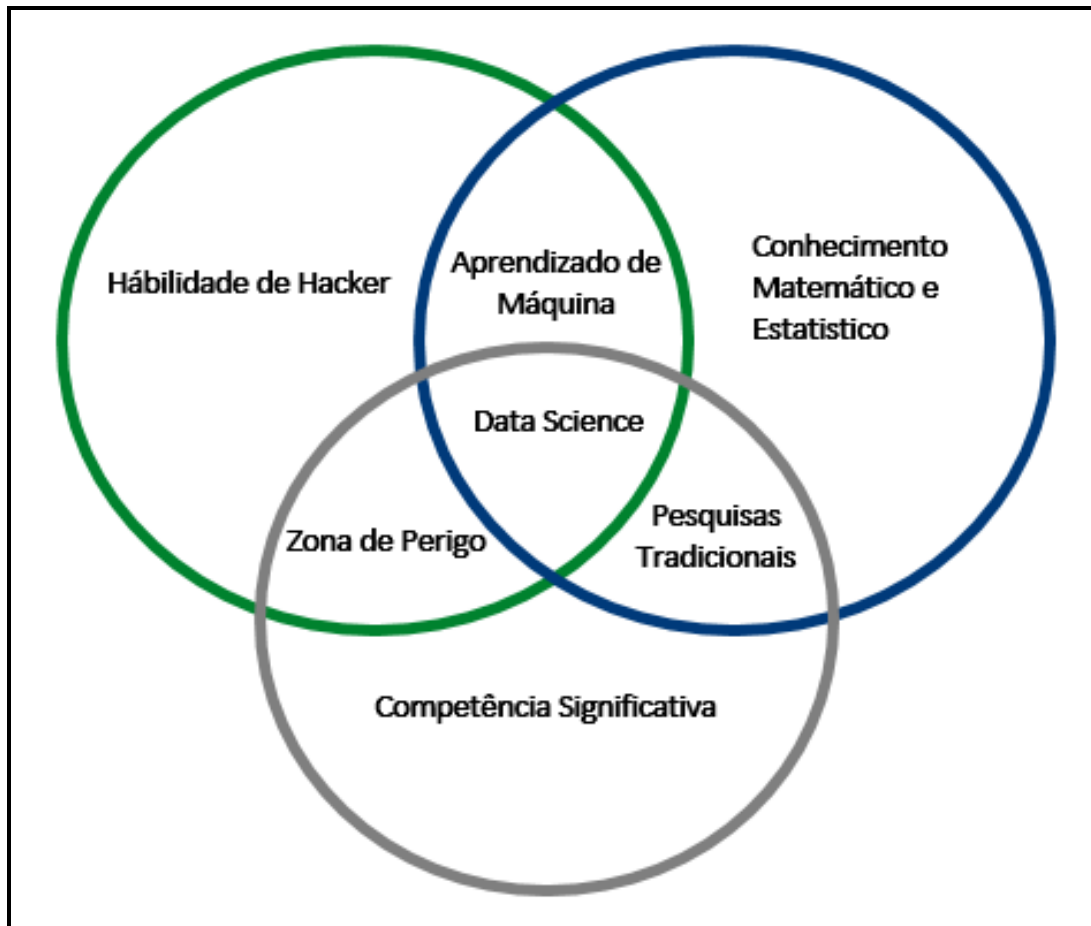
De acordo com Lins (2016, p. 5), esse modelo se tornou famoso, pois com ele dois problemas foram corrigidos, o da paralelização, que é executar partes diferentes de um algoritmo em unidades de processamento diferentes e obter um menor tempo de execução e o problema da computação distribuída. Em suma, no *MapReduce*, o programa vai até os dados e não o contrário.

Com o desenvolvimento do *MapReduce* várias implementações da mesma começaram a ser utilizadas para o tratamento de *Big Data*. O mais famoso deles é o *Hadoop* que, de acordo com Ghazi e Gangodkar (2015, p. 46), é a implementação de código aberto mais popular do modelo de programação *MapReduce*. *Apache Hadoop* é uma estrutura de *software* para computação confiável, escalonável, paralela e distribuída. *MapReduce*, e seu projeto de código aberto existente chamado *Hadoop*, permite o processamento paralelo de grande quantidade de dados e partição automática de dados, distribuição de dados, tolerância a falhas e gerenciamento de balanceamento de carga que finalmente resulta em computação confiável e escalável.

A implementação dessa tecnologia abriu porta para o que é conhecido hoje como *Data Science* ou Ciência de Dados, apresentado na Seção 2.5.

Após o desenvolvimento do *MapReduce*, a análise de dados se tornou uma tarefa menos onerosa e para aproveitar esses dados e conseguir analisá-los foi necessário a criação de uma nova área de conhecimento. De acordo com Grus (2015, p. 2), o *Data Science* surgiu para extrair conhecimento de dados desorganizados e, para isso, foram desenvolvidas inúmeras técnicas e ferramentas como, por exemplo, Inteligência Artificial - I.A, Aprendizado de Máquina - A.M, Sistemas de Recomendação - S.R, Redes Neurais - R.N, *Naive Bayes* - N.B entre outros.

Para algumas pessoas, esse conceito de *Data Science* pode parecer um pouco abstrato, então um Diagrama de Venn<sup>16</sup> foi desenvolvido por Conway (2010, p. 1) para apresentar melhor esse conceito, conforme demonstrado na Figura 2.3:



**Figura 2.3:** Diagrama de Venn.

Fonte: Adaptado de Conway (2010, p. 1).

De acordo com Conway (2010, p. 1), conforme apresentado na Figura 2.3, os três pilares que resultam na Ciência de Dados são:

- **Habilidade de *hacker***<sup>17</sup>: Dados são uma mercadoria negociada eletronicamente, por meio da *internet*. É crucial que pessoas nesse mercado tenham noções de programação e manipulação de dados, entendendo operações vetorizadas e pensando algoritmicamente;

<sup>16</sup>De acordo com Chow e Ruskey (2004, p. 466), em 1880, John Venn introduziu uma notação para representar proposições lógicas usando curvas. Esses diagramas de Venn eram instâncias especializadas de uma notação mais geral para representar relações de conjuntos.

<sup>17</sup>Qualquer pessoa que se dedique intensamente em alguma área específica da computação e descobre utilidades além das previstas nas especificações originais, de acordo com Kleinknecht (2003, p. 5), considerado um dos mais brilhantes em seu campo, o termo *hacker* tornou-se sinônimo de ser um mago da tecnologia ou virtuoso do computador.

- **Conhecimento matemático e estatístico:** Depois de adquirir os dados e estruturá-los, é necessário gerar *insights* desses. Para isso, é necessário aplicar métodos matemáticos e estatísticos apropriados, o que requer uma familiaridade com essas ferramentas;
- **Competência significativa:** Ciência lida com questões de mundo e do dia a dia. Sem um conhecimento de mundo, ou um conhecimento crítico, a pessoa não possui questões e problemas motivadores que ele possa testar solucionar por meio de métodos e estatística.

Além desses pilares, existem alguns subconjuntos de conceitos apresentados por Conway (2010, p. 1), formados pela junção de dois pilares, como:

- **Aprendizado de máquina:** Formado pela junção de habilidade *hacker* e conhecimento de estatística e matemática, esse subconjunto mostra que apenas esses dois pilares não são suficientes para formar *Data Science*, pois ciência trata de descoberta e construção de conhecimento, o que requer algumas questões motivadoras sobre o mundo e hipóteses que podem ser trazidas aos dados e testadas com métodos estatísticos;
- **Pesquisas tradicionais:** De acordo com autor, a mistura de conhecimento de estatística e matemática junto com competência significativa é aonde se encontram a maioria dos pesquisadores tradicionais. Pesquisadores em nível de doutorado passam a maior parte de seu tempo adquirindo experiência nessas áreas, mas muito pouco tempo aprendendo sobre tecnologia
- **Zona de perigo:** Para definir essa área, o autor cita que é nela aonde se encaixam aquelas pessoas que “Sabem suficiente para serem perigosas”, no sentido de que essas pessoas são capazes de extrair e estruturar dados de algum campo que eles possuem conhecimento, mas não entendem o que, no final, os coeficientes querem dizer. Essa sobreposição de habilidades dá às pessoas a capacidade de criar o que parece ser uma análise legítima, sem qualquer compreensão de como eles chegaram lá ou o que eles criaram.

Com o surgimento dessa nova área, muitas ferramentas foram desenvolvidas para poder coletar esses dados desestruturados e gerar informações valiosas. Entre essas ferramentas, esta o S.R. que, de acordo com Katarya e Verma (2017, p. 105), são ferramentas de filtragem de informações cuja função é avaliar as classificações de usuários e itens, predominantemente de *Big Data*, para recomendar seus gostos e isso faz com que essa técnica seja uma parte essencial de *websites* e *e-commerces*.

## 2.6 SISTEMAS DE RECOMENDAÇÃO

---

De acordo com S. Yang, Korayem, AlJadda, Grainger, e Natarajan (2017, p. 38) com a sua ascensão em destaque, os S.R.’s têm aliviado bastante a sobrecarga de informações para seus usuários, fornecendo sugestões personalizadas para inúmeros tipos de produtos e serviços. De acordo com eles, desde a década de 1990, não apenas novas teorias de S.R. vêm sendo apresentadas, mas também, seu *software* vem sendo desenvolvido, envolvendo vários domínios para sua aplicação como, por exemplo, “*e-govern*”, “*e-business*”, “*e-commerce*”.

### 2.6.1 SURGIMENTO DO S.R.

---

No ano de 1992, os autores Goldberg, Nichols, Oki, e Terry (1992) apresentaram em seu trabalho o que seria considerado um dos primeiros S.R. desenvolvidos, o *Tapestry*, no seu artigo

“*Using Collaborative Filtering to Weave an Information Tapestry*”. De maneira mais específica, nesse artigo, Goldberg et al. (1992, pp. 61–63) apresentam um sistema de correio eletrônico experimental, *Tapestry*, desenvolvido pela *Xerox Palo Alto Research Center*<sup>18</sup>, que tinha como objetivo oferecer *e-mails* de forma mais personalizada para os usuários, pois os mesmos vinham recebendo um grande volume de documentos. Para isso, os autores apresentam também um modelo de filtragem que hoje em dia é amplamente conhecida e usada, a filtragem colaborativa que, de acordo com os autores significa simplesmente que as pessoas colaboram para ajudar umas às outras a realizar a filtragem, registrando suas reações aos documentos que possuíam.

Outros autores que vieram mencionar os S.R's nos anos 90 foram Resnick e Varian (1997, p. 56), que realizaram um estudo comparando alguns sistemas já existentes na época. No texto, os autores definem S.R. como um sistema típico aonde as pessoas fornecem recomendações como entradas, que são agregadas e direcionadas aos destinatários apropriados. É válido ressaltar que foi nessa época que algumas preocupações que são estudadas até hoje começaram a surgir a respeito do tema, a saber:

- **Incentivos:** De acordo com Resnick e Varian (1997, p. 57), os S.R's são alimentados tanto por informações implícitas (que são coletadas facilmente pelos sistemas), quanto informações explícitas. Essas informações explícitas são as avaliações dos usuários a respeito do item, e o problema de incentivos é como motivar os usuários a preencherem e contribuir com esse tipo de informação e não somente se aproveitarem das recomendações dos sistemas;
- **Privacidade Pessoal:** Resnick e Varian (1997, p. 57) dizem que em geral, quanto mais informações os indivíduos tiverem sobre as recomendações, melhor serão capazes de avaliá-las. No entanto, as pessoas podem não querer que seus hábitos ou pontos de vista sejam amplamente conhecidos. Alguns sistemas de recomendação permitem a participação anônima ou sob pseudônimo, mas esta não é uma solução completa, pois algumas pessoas podem desejar uma combinação intermediária de privacidade e crédito atribuído por seus esforços.

Como antigamente a manutenção dos S.R's era cara, uma das preocupações de Resnick e Varian (1997, p. 58) era a questão da monetização. Ele possuía dúvidas de como gerar receita com esses sistemas, se a melhor forma era cobrar os destinatários, cobrando taxas dos proprietários dos itens avaliados, entre outros. Atualmente, o custo de manutenção desses sistemas não é elevado como antigamente, e como mostrado no primeiro capítulo deste trabalho, os S.R's são responsáveis por grande parte da receita de grandes empresas, cobrindo assim, seus eventuais custos.

## 2.6.2 MODELOS DE S.R.

---

Para se entender melhor os S.R's, é importante entender primeiramente as formas que esses mecanismos podem ser desenvolvidos. De acordo com Li, Zhang, Wang, Chen, e Pan (2017, p. 46), os S.R's são separados em três categorias baseadas em como os mesmos são construídos, são elas:

- Baseado em conteúdo (“*Content-based*”);
- Filtragem colaborativa (“*Collaborative Filtering*”);
- S.R's híbridos (“*Hybrid approaches*”).

---

<sup>18</sup>Foi uma importante divisão de pesquisa da *Xerox Corporation*<sup>©</sup> baseada em Palo Alto.

Um dos primeiros autores a comentarem o assunto foram Herlocker, Konstan, Borchers, e Riedl (1999, p. 1) que abordaram as duas principais maneiras de se construir um S.R., uma delas é a ferramenta de S.R. com base em conteúdo, elas basicamente selecionam as informações certas para as pessoas certas, comparando as representações do conteúdo contido nos documentos com as representações do conteúdo no qual o usuário está interessado. Esse tipo de ferramenta provou ser bem efetiva em localizar documentos textuais relevantes.

No entanto, existe a maneira de se desenvolver um S.R. que os autores julgam a mais popular que são os sistemas automatizados de filtragem colaborativa, que funcionam coletando julgamentos humanos (conhecidos como classificações) para itens em um determinado domínio e combinando pessoas que compartilham as mesmas necessidades de informação ou os mesmos gostos. Os usuários de um sistema de filtragem colaborativa compartilham seus julgamentos analíticos e opiniões sobre cada item que consomem, para que outros usuários do sistema possam decidir melhor qual item consumir. Ainda em 2017, J. Yang et al. (2017) citam que os S.R.'s com filtragem colaborativa baseada em conteúdo melhoram a precisão das previsões baseadas somente em conteúdo.

### **Filtragem Colaborativa**

Dentre essas maneiras de se construir um S.R., a filtragem colaborativa mostrou ser a mais usada e bem-sucedida segundo os autores. Seo et al. (2017, p. 55) abordam que a filtragem colaborativa tem um papel importante nos S.R.'s, e isso se dá pelo seu sucesso em plataformas como *Youtube*®, *Amazon*®, *Netflix*®, etc e inspirados nesses sistemas, muitos S.R.'s baseados em filtragem colaborativa para previsão de Qualidade de Serviço - QoS foram criados.

De acordo com Bobadilla, Hernando, Ortega, e Bernal (2011, p. 14609) simulando cada etapa do nosso próprio comportamento na medida do possível, o processo de filtragem colaborativa do S.R. seleciona primeiro o grupo de usuários de uma amostra que é mais semelhante ao usuário e, em seguida, fornece um grupo de recomendações de elementos que o usuário não tem classificados (assumindo, desta forma, que são novos) e que foram classificados como os melhores pelo grupo de usuários com gostos semelhantes ao do usuário.

Herlocker et al. (1999, p. 1) ainda ressaltam que a filtragem colaborativa oferece três vantagens adicionais importantes para a filtragem de informações que não são fornecidas pela filtragem baseada em conteúdo:

1. Suporte para filtragem de item cujo conteúdo não é facilmente analisado por processo automatizado;
2. A capacidade de filtrar itens com base na qualidade e gostos;
3. A capacidade de fornecer recomendações casuais.

Seo et al. (2017, pp. 135-136) também comentam sobre a técnica de filtragem colaborativa, dizendo que a filtragem colaborativa entrega recomendações para usuários analisando suas características individuais a fim de utilizar as informações de outros usuários que são muito semelhantes a eles. Os autores ainda comentam a existência de duas maneiras de se coletar dados na *Web*, sendo elas explícitas (avaliações dos usuários) e implícita (quantas vezes o usuário escutou certa música, quais *downloads* foram feitos), e que para a filtragem colaborativa, a melhor maneira é a implícita, pois os resultados são mais personalizados.

No entanto, algumas falhas são apresentadas por Balabanovi e Shoham (1997, p. 3), que mostram por exemplo: Se um novo item aparecer no banco de dados, não há como recomendá-lo a um usuário até que mais algumas informações sobre ele sejam obtidas, por meio de outro usuário classificando-o ou especificando a quais outros itens ele é semelhante. Assim, se o número de usuários for pequeno em relação ao volume de informações no sistema (porque há um banco

de dados muito grande ou que muda rapidamente), há o perigo de a cobertura das avaliações se tornar muito esparsa, diminuindo a coleção de itens recomendáveis. Um segundo problema é simplesmente que, para um usuário cujos gostos são incomuns em comparação com o resto da população, não haverá nenhum outro usuário que seja particularmente semelhante, levando a recomendações ruins.

## S.R. Baseado em Conteúdo

Apesar da filtragem colaborativa ser a mais popular entre os S.R's ainda há modelos que utilizam a abordagem baseada em conteúdo. Como Herlocker et al. (1999, p. 1) mencionou, essa abordagem é interessante para S.R's que tem o foco em localizar elementos textuais.

De acordo com Pazzani e Billsus (2007, pp. 325–326), os sistemas de recomendação com base em conteúdo são sistemas que recomendam um item a um usuário com base em uma descrição do item e um perfil dos interesses do usuário. Esses sistemas podem ser usados em uma variedade de domínios, desde páginas da *web* de recomendação, artigos de notícias, restaurantes, programas de televisão e itens à venda. Embora os detalhes de vários sistemas sejam diferentes, os sistemas de recomendação baseados em conteúdo compartilham um meio de descrever os itens que podem ser recomendados, um meio para criar um perfil do usuário que descreve os tipos de itens que o usuário gosta e um meio de comparar itens com o perfil do usuário para determinar o que recomendar. Os sistemas de recomendação baseados em conteúdo analisam as descrições dos itens para identificar os itens que são de interesse particular para o usuário.

Para os autores Balabanovi e Shoham (1997, p. 2), a abordagem baseada em conteúdo para recomendação tem suas raízes na comunidade de recuperação de informações e emprega muitas das mesmas técnicas. Os documentos de texto são recomendados com base na comparação entre seu conteúdo e o perfil do usuário. As estruturas de dados para ambos são criadas usando recursos extraídos do texto dos documentos.

Além dessa definição, Balabanovi e Shoham (1997, p. 2) também trouxeram um conceito importante para os S.R's baseados em conteúdo, que foi o *feedback* de relevância, que conforme o autor:

*“Se o usuário gostou de uma página, pesos para as palavras extraídas para escolher aquela página podem ser somados aos pesos para as palavras correspondentes no perfil do usuário. Além de ser simples e rápido, é conhecido empiricamente por fornecer resultados melhores em um sistema normal recuperação de informações.”<sup>19</sup>*

No entanto nessa modalidade também são apresentadas algumas falhas pelos autores Balabanovi e Shoham (1997, p. 2). De maneira geral, os autores mostram que, esse tipo de S.R. entregam análises superficiais dos objetos de pesquisa. Em alguns domínios, os itens não são passíveis de quaisquer métodos de extração de recursos úteis com a tecnologia da época. Mesmo para documentos de texto, as representações capturam apenas alguns aspectos do conteúdo, e há muitos outros que influenciam a experiência do usuário. Um segundo problema, que tem sido estudado extensivamente neste domínio e em outros, é o de sobre-especialização. Quando o sistema só pode recomendar itens com alta pontuação em relação ao perfil de um usuário, o usuário fica restrito a ver itens semelhantes aos já avaliados.

De acordo com os autores Pereira e Varma (2016, p. 281), as vantagens da utilização desse método são a aprendizagem do perfil é fácil, a qualidade melhora com o tempo e ele considera *feedback* implícito, enquanto suas desvantagens são que ele não supera completamente o problema de sobre-especialização e o problema da repetição de itens já avaliados.

---

<sup>19</sup>Tradução livre.



## S.R's híbridos

Como apresentado acima, uma variedade de técnicas para S.R's foi desenvolvida, e com a evolução desses métodos e da tecnologia, foi observado que eventualmente, uma junção desses métodos em alguns casos poderia gerar um S.R. mais eficiente, e dessa forma surgiram os S.R's híbridos.

De acordo com Burke (2002, p. 339), os S.R's híbridos combinam duas ou mais técnicas de recomendação para obter melhor desempenho com menos desvantagens de qualquer um. Mais comumente, a filtragem colaborativa é combinada com alguma outra técnica. Além das técnicas apresentadas anteriormente, é de suma importância abordar também duas técnicas não tão populares mas que são muito importantes para o desenvolvimento dos S.R's híbridos, de acordo com Burke (2002, pp. 333–334), são elas:

- **Demográfico:** têm como objetivo categorizar o usuário com base em atributos pessoais e fazer recomendações com base em classes demográficas;
- **Baseada na Utilidade:** Não tentam construir generalizações de longo prazo sobre seus usuários, mas sim baseiam seus conselhos em uma avaliação da correspondência entre a necessidade de um usuário e o conjunto de opções disponíveis. Recomendadores baseados em utilitários fazem sugestões com base em um cálculo da utilidade de cada objeto para o usuário.

Essa junção criou algumas modalidades de S.R's Híbridos, que é apresentada por Burke (2002, pp. 339–343), como pode-se observar a seguir:

- **Ponderado:** O recomendador híbrido ponderado é aquele em que a pontuação de um item recomendado é calculada a partir dos resultados de todas as técnicas de recomendação disponíveis presentes no sistema;
- **Comutação:** Um híbrido de comutação aumenta a sensibilidade de nível de item para a estratégia de hibridização. O sistema usa algum critério para alternar entre as técnicas de recomendação;
- **Misto:** Quando for prático fazer um grande número de recomendações simultaneamente, pode ser possível usar um híbrido misto, no qual as recomendações de mais de uma técnica são apresentadas juntas;
- **Cascata:** Ao contrário dos métodos de hibridização anteriores, o híbrido em cascata envolve um processo em etapas. Nesse sentido, uma técnica de recomendação é empregada primeiro para produzir uma classificação grosseira de candidatos e uma segunda técnica aprimora a recomendação entre o conjunto de candidatos.

A seguir, é apresentado na Tabela 2.3 alguns exemplos de autores que desenvolveram alguns modelos de S.R's híbridos.

Modelos	Ponderado	Misto	Comutação	Cascata
FC/BC	Miranda et al. (1999).	Smyth e Cotter (2000).	Billsus e Pazzani (2000).	Balabanovic (1997).
FC/DM	Pazzani (1999).			
FC/BU	Towle e Quinn (2000).		Tran e Cohen (2000).	
BC/DM	Pazzani (1999).			

**Tabela 2.3:** Modelos de S.R's Híbridos.

**Legenda:** FC: Filtragem Colaborativa, BC: Baseado em Conteúdo, DM: Demográfico, BU: Baseado na Utilidade.

**Fonte:** Adaptado de Burke (2002, p. 345).

Como pode-se observar na Tabela 2.3, o modelo híbrido ponderado é o mais popular dentre as possibilidades. Miranda et al. (1999, p. 2) desenvolveu o *P-Tango*, um S.R. ponderado que misturava a filtragem colaborativa com a filtragem baseada em conteúdo. Nesse caso, um S.R. para um jornal *online* foi desenvolvido, no qual inicialmente os pesos tanto para a filtragem colaborativa quanto para a baseada em usuário eram iguais, mas iam ajustando gradualmente a ponderação à medida que as previsões sobre as classificações do usuário são confirmadas ou não.

Já os autores Pazzani (1999, p. 395) desenvolveram de uma forma diferente. Combinando filtragem colaborativa e filtragem demográfica, os autores desenvolveram um S.R. híbrido que não usa pontuações numéricas, mas trata a saída de cada recomendador (colaborativo, baseado em conteúdo e demográfico) como um conjunto de votos, que são então combinados em um esquema de consenso.

O último caso dos modelos ponderados é apresentado por Towle e Quinn (2000, p. 74), que apresentam um S.R. ponderado baseado na filtragem colaborativa e baseado na utilidade. Nesse caso, os autores desenvolveram um S.R. que é alimentado por informações explícitas do usuário, pois acreditavam que as informações adicionais fornecidas pelos modelos de usuário e produto podem dar ao sistema vantagem em tarefas de recomendação difíceis e também aliviar o problema do avaliador inicial e das classificações esparsas experimentado pelos sistemas de recomendação atuais.

A fim de trabalhar em cima do problema da sobrecarga de informações, os autores Smyth e Cotter (2000, p. 53) desenvolveram um S.R. misto aonde eles utilizaram as técnicas da filtragem colaborativa baseada em conteúdo (com base em descrições textuais de programas de TV) e a filtragem colaborativa baseada no usuário (suas preferências). O resultado do trabalho dos autores foi a criação do *Personalized Television Listings - PTL* que é um S.R. inteligente, que aprende automaticamente sobre as preferências de visualização de TV de usuários individuais, de modo a fornecer-lhes guias diários de TV altamente customizados e personalizados.

Já na abordagem de comutação, pode-se observar dois trabalhos diferentes. Primeiramente, Billsus e Pazzani (2000, p. 148) apresentam o *Dailylearner*, um S.R. que utiliza as técnicas de filtragem colaborativa e a baseada no usuário. Esse modelo funciona da seguinte forma: inicialmente é utilizada a técnica baseada no usuário e os resultados são analisados. Caso o programa julgue esses resultados como não satisfatórios, ele utiliza a filtragem colaborativa como segunda alternativa.

Ainda utilizando a abordagem de comutação, os autores Tran e Cohen (2000, p. 78) propuseram um S.R. para auxiliar aplicativos de comércio eletrônico (*E-commerce*). Após diversas pesquisas os autores chegaram a conclusão que as técnicas de filtragem colaborativa e baseada na utilidade são as mais comuns entre os S.R's existentes nesse setor, por isso, decidiram desenvolver um S.R que trouxesse essas duas abordagens. Esse sistema analisa os itens individualmente e decide qual dessas duas abordagens é melhor para ele.

Por último, os autores Balabanovic (1997, p. 379) desenvolveram o *Fab*, um S.R. que traz o modelo de cascata. Nele, a filtragem colaborativa é utilizada para criar um primeiro ranqueamento das opções e então a filtragem baseada em conteúdo toma a decisão final dentro das opções.

Além dos métodos apresentados acima, existe uma etapa crucial para o desenvolvimento de qualquer S.R. De acordo com S. Yang et al. (2017, p. 38), essa etapa é a aplicação da similaridade do cosseno, aonde é criado dois vetores para poder calcular a semelhança de dois itens. Apesar da similaridade do cosseno ser uma das mais famosas, não é a única maneira de se calcular a similaridade entre dois usuários ou itens. Em seu estudo, Seo et al. (2017, p. 136) apresentam algumas formas alternativas, como:

- **PCC:** *Pearson Correlation Coefficient* apresentado em Herlocker et al. (1999);
- **JMSD:** *Jaccard Mean Square Difference* apresentado em Bobadilla, Ortega, Hernando, e Gutiérrez (2013);
- **NHSM:** *New Heuristic Similarity Model* apresentado em Lai, Liu, e Liu (2013).

Para esse trabalho, como sendo a mais popular, a similaridade do cosseno foi escolhida para compor o S.R. desenvolvido. Esse assunto será abordado com mais profundidade no Capítulo 3.

### 2.6.3 APLICAÇÕES DOS S.R's

---

Muitos imaginam os S.R's funcionando apenas em *e-commerce* e redes sociais, pois realmente são as plataformas mais comuns aonde são usados esse tipo de ferramenta. No entanto, tais métodos vem sendo desenvolvidas para uma grande diversidade de serviços, dos mais variados, indo de ramos como a medicina, até seleção de trabalho. Na Tabela 2.4 são apresentados alguns modelos de S.R. como pode-se ver abaixo:

Autor	Ano	Uso dos Sistemas Recomendadores
Afzal et al. (2017).	2017	Propõem um método automatizada de aquisição de conhecimento com um modelo de conhecimento compreensível, denominada CKM-CT. Esse método recomenda e prevê o plano de tratamento adequado para pacientes com câncer de cabeça e pescoço com base nas informações recuperadas dos documentos clínicos.
Benabderrahmane, Mellouli, Lamolle, e Paroubek (2017).	2017	Sugerem um S.R. baseado em um modelo híbrido, combinando tanto classificação semântica modular quanto previsão de séries temporais combinado com uma plataforma de <i>Big Data</i> para desenvolver um S.R. para recrutamento e seleção, que ultimamente vêm crescendo bastante. Esse tipo de processo é denominado “ <i>E-recruitment</i> ”.
J. Yang et al. (2017).	2017	O objetivo do artigo é estudar e desenvolver um S.R. para multimídias com o sistema de armazenamento em nuvem. Os autores focam também nos problemas de segurança e performances mencionados anteriormente.
Ma, Ma, Li, Jiang, e Gao (2018).	2018	Projetar uma estrutura de preservação de privacidade baseada em confiança para recomendação descentralizada de amigos em redes sociais <i>online</i> (ARMOR).
Su, Xiao, Liu, Zhang, e Zhang (2017).	2017	O objetivo desse trabalho é apresentar uma abordagem confiável de previsão de <i>Quality of Service - QoS</i> utilizando totalmente os dados de <i>QoS</i> observados de usuários semelhantes confiáveis e serviços semelhantes.
Tian, Zheng, Wang, Zhang, e Wu (2019).	2019	O autor projeta um sistema de recomendação personalizado para bibliotecas universitárias baseado em algoritmo de recomendação híbrido para resolver o problema da dificuldade na escolha de livros e melhorar a taxa de utilização dos recursos da biblioteca.
Amato, Moscato, Picariello, e Piccialli (2019).	2019	Neste trabalho de pesquisa, o autor apresenta um novo sistema de recomendação de aplicações de <i>Big Data</i> capaz de fornecer recomendações com base nas interações entre usuários e os conteúdos multimídia gerados em uma ou mais redes de mídia social. O sistema proposto depende de uma abordagem “centrada no usuário”.
Mao, Zhao, Lin, e Herrera-Viedma (2019).	2019	Neste artigo, os autores propõem um novo método de recomendação de tipo de loja que sugere um tipo de loja adequado com base em informações de várias fontes coletadas de um site de avaliação de negócios, um sistema de navegação baseado em localização e uma operadora móvel.
Massai, Nesi, e Pantaleo (2019).	2019	Neste artigo, os autores propõem o <i>Praval</i> (Assistente pessoal virtual ciente de localização), um mecanismo de assistência semântica para sugerir Pontos de Interesse (POIs) locais e serviços por meio da análise de consultas em linguagem natural dos usuários, a fim de estimar a necessidade de informação e o potencial geográfico referências expressas pelos usuários.
Dong, Zeng, Koehl, e Zhang (2020).	2020	Neste artigo, os autores propõem um S.R. de design interativo e baseado no conhecimento ( <i>JKDRS</i> ) para esquemas de design de produtos de moda personalizados relevantes com suas demonstrações virtuais para um consumidor específico. Este sistema permite a interação iterativa entre a demonstração virtual do produto e o conhecimento e percepção profissional do designer, a fim de encontrar a melhor solução de design existente, ou seja, combinação de elementos básicos de vestuário.

**Tabela 2.4:** Exemplos da utilização de S.R.

# CAPÍTULO 3

## MÉTODO

“A tecnologia move o mundo.”  
—Steve Jobs.

Como indicado no Capítulo 1, o objetivo deste capítulo é explicar o método utilizado para o desenvolvimento do S.R. utilizado neste trabalho, assim como algumas particularidades envolvendo o mesmo.

Para a realização do SR desenvolvido nesse trabalho, a linguagem de programação escolhida foi *Python*<sup>1</sup> pela sua simplicidade, velocidade e sua versatilidade. O código utilizado foi uma adaptação de Grus (2015, pp. 267–274) apresentado em seu livro *Data Science do Zero - Primeiras regras com o Python*. Ainda é oportuno destacar que o *Python* apresenta diversas vantagens, sendo pela sua sintaxe, que se aproxima bastante da linguagem natural, quanto pela sua velocidade de processamento.

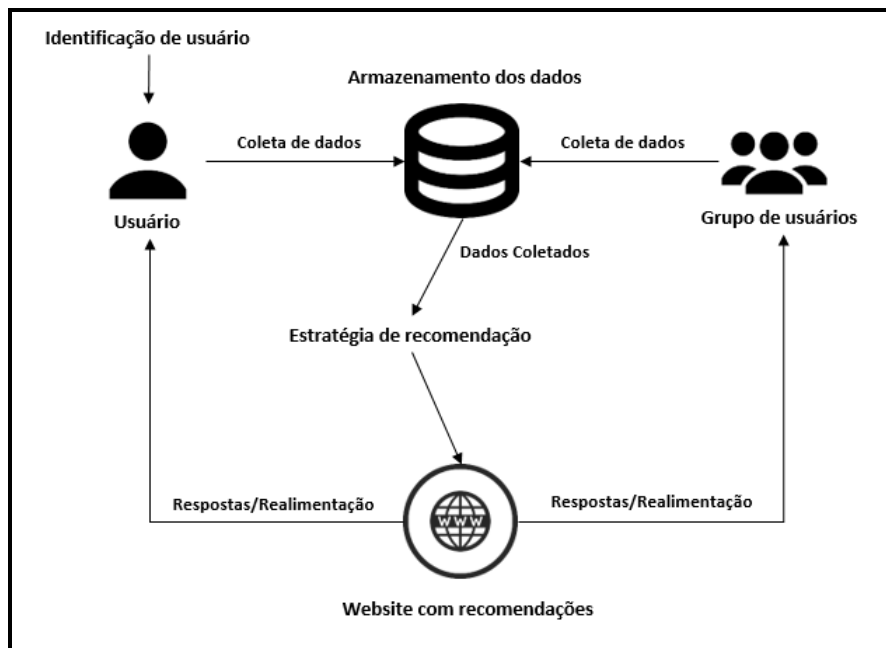
### 3.1 ESTRUTURA DOS S.R’S

---

Os S.R’s são ferramentas presentes no nosso dia a dia e, como muito bem discutido no Capítulo 2, existem diversas possibilidades para o seu desenvolvimento, seja mudando na sua parte de coleta de dados, no seu algoritmo, etc. Entretanto, de acordo com Schafer, Konstan, e Riedl (1999, p. 160), existe um escopo padrão para esse desenvolvimento. Para os autores, a estrutura básica de um S.R. é composta por 4 etapas, que são elas: a) identificação do usuário; b) coleta de informações; c) estratégias de recomendação e d) visualização das recomendações. Esse esquema é melhor representado na Figura 3.1:

---

<sup>1</sup>De acordo com Rossum (2003, p. 6), o *Python* é uma linguagem de programação, criada em 1991 por Guido van Rossum, projetada com a filosofia de enfatizar a importância do esforço do programador sobre o esforço computacional. Prioriza a legibilidade do código sobre a velocidade ou expressividade.



**Figura 3.1:** Estrutura básica de um S.R.  
**Fonte:** Adaptado de Schafer et al. (1999, p. 160).

Após observar a Figura 3.1, é possível entender melhor como o fluxo de informações é realizado dentro de um S.R. Para ficar mais claro, as etapas apresentadas acima são explicadas nas subseções apresentadas a seguir.

### 3.1.1 IDENTIFICAÇÃO DO USUÁRIO

Ao contrário do que possa ser esperado, não necessariamente um S.R. precisa da identificação do usuário para poder funcionar. O S.R. pode simplesmente recomendar o item mais vendido do site para todos os usuários por exemplo. No entanto essa é uma prática não muito utilizada, pois o verdadeiro foco dos S.R's é realizar cada vez mais recomendações personalizadas, então é de suma importância que ao realizar uma recomendação, o usuário seja identificado. Tomando isso em conta, no caso do S.R. proposto neste trabalho, a identificação do usuário acontece por meio *login* do estudante no sistema de matrículas da UnB.

### 3.1.2 COLETA DE DADOS

Como previamente indicado, os S.R's atualmente focam em personalização e por isso coletam todos os tipos de dados como cliques, avaliações, buscas, entre outros para recomendar itens que mais se encaixem em seu perfil. O objetivo principal de um S.R. é a coleta de informações e, com isso, facilitar a escolha ou tomada de decisão do usuário, ao recomendar um item que pode ser um conteúdo, um produto ou até mesmo uma pessoa. As informações que serão coletadas dizem respeito a três elementos, sendo eles:

- Os itens que serão recomendados;
- O usuário que será entregue a recomendação;
- Outros usuários que já interagiram com a plataforma.

De acordo com Seo et al. (2017, pp. 175–176), essas informações podem ser obtidas de duas maneiras, a saber:

- **Implícitas:** são produzidos de forma espontânea pelos usuários e têm a ver com o comportamento dele durante a navegação, como cliques, buscas realizadas e/ou tempo de permanência em alguma página, etc;
- **Explícitas:** consistem em informação concedida pelos usuários, geralmente diante de alguma pergunta ou solicitação, como comentários e avaliações.

A personalização das recomendações sempre vão depender da qualidade e quantidade dos dados. Quanto mais dados o sistema tiver em relação a um usuário e pessoas semelhantes a ele, melhores e mais personalizadas podem ser as recomendações.

Observando essa premissa, para o S.R. proposto neste trabalho, os dados utilizados como banco de dados foram os históricos escolares de ex-alunos e alunos no período final da sua graduação. E quanto ao dado do usuário, ao acessar o S.R. o mesmo teria o seu histórico escolar coletado.

### 3.1.3 ESTRATÉGIAS DE RECOMENDAÇÃO

---

Existem inúmeras estratégias de recomendação e elas variam de acordo com o objetivo do S.R. Uma estratégia comumente utilizada atualmente é a clusterização. De acordo com Garofalakis, Rastogi, Seshadri, e Shim (1999, p. 5), a clusterização é compreendida como um agrupamento de dados em categorias. Isso é particularmente desejável no contexto de dados, no qual é útil agrupar clientes semelhantes para fins de recomendações direcionadas.

Ainda nesse contexto, outra estratégia é apresentada por Schafer et al. (1999, p. 162), a chamada *business-rule* no qual o S.R. pode utilizar informações demográficas para direcionar a recomendação de um item para um segmento do mercado. Nesse caso, a recomendação não necessariamente será focada em seus interesses ou necessidades pessoais, mas sim em acontecimentos demográficos.

Garofalakis et al. (1999, p. 6), por último, ressalva que uma estratégia comumente utilizada é a da correlação. Essa técnica é baseada na associação de itens que tendem a serem escolhidos juntos em uma compra, podendo ser utilizados na identificação de padrões. A correlação também pode ser definida por meio de regras de associação, que demonstram como produtos e serviços se relacionam uns com os outros.

A correlação foi escolhida como estratégia de recomendação nesse trabalho, já que após criar um vetor das matérias, elas são facilmente correlacionadas podendo assim mostrar ao S.R. se as matérias costumam ser feitas juntas ou não.

Além disso, ainda no campo da estratégia de recomendação, a filtragem colaborativa baseada no usuário foi escolhida dentre as outras opções pois, o intuito do S.R. desenvolvido no trabalho é recomendar matérias similares a outros alunos que tenham o perfil similar do mesmo.

### 3.1.4 VISUALIZAÇÃO DAS RECOMENDAÇÕES

---

A visualização também é uma etapa fundamental para a uma recomendação de sucesso. As recomendações aos usuários devem ser apresentadas de maneira que as mesmas possam ser facilmente visualizadas e compreendidas.

De acordo com Barcellos, Musa, Brandão, e Warpechowski (2007, p. 3) nessa etapa também é de suma importância identificar os níveis de recomendação. Ainda para os mesmos autores, a recomendação pode vir em três níveis, que são eles:

- **Não-Recomendação:** As recomendações são iguais para todos os usuários. Como, por exemplo, uma lista com os produtos mais vendidos para todos os seus clientes;

- **Recomendação Efêmera:** As recomendações são baseadas inteiramente na navegação de um único usuário e não utiliza informações das navegações anteriores do mesmo;
- **Recomendação Persistente:** As recomendações são baseadas no reconhecimento do usuário, e sugere produtos que são do seu interesse, com base nas suas navegações anteriores.

### 3.2 SIMILARIDADE ENTRE USUÁRIOS

Além das etapas apresentadas acima, para começar o desenvolvimento de qualquer S.R., é de suma importância que seja definido uma maneira de se calcular a similaridade entre usuários, itens ou ambos. Como apresentado no Capítulo 2, existem diversas maneiras de se realizar esse cálculo.

Para isso, é necessário que seja definido um grau de correlação entre os itens ou usuários. De acordo com MM (2012, p. 69), a correlação é um método de avaliar uma possível associação linear bidirecional entre duas variáveis contínuas. A correlação é medida por um método estatístico chamado coeficiente de correlação, que representa a força da associação linear entre as variáveis em questão. É um indicador adimensional que assume um valor no intervalo de -1 a +1. Um coeficiente de correlação de zero indica que não existe nenhuma relação linear entre duas variáveis contínuas e um coeficiente de correlação de -1 ou +1 indica uma relação linear perfeita. A força do relacionamento pode estar em qualquer lugar entre -1 e +1. Quanto mais forte a correlação, mais próximo fica o coeficiente de correlação de  $\pm 1$ . Se o coeficiente for um número positivo, as variáveis estão diretamente relacionadas, ou seja, conforme o valor de uma variável aumenta, o valor da outra também tende a aumentar. Se, por outro lado, o coeficiente é um número negativo, as variáveis estão inversamente relacionadas, ou seja, à medida que o valor de uma variável aumenta, o valor da outra diminui.

Uma técnica tradicional para a determinação da correlação é o cálculo do coeficiente de correlação de *Pearson* que, ainda de acordo com MM (2012, p. 70), é denotado como  $\rho$  para um parâmetro populacional e como  $r$  para uma estatística de amostra. É usado quando ambas as variáveis em estudo são normalmente distribuídas. Esse coeficiente é afetado por valores extremos, que podem exagerar ou diminuir a força do relacionamento e, portanto, é inadequado quando uma ou ambas as variáveis não são normalmente distribuídas. Pode-se notar a utilização desse coeficiente no trabalho de Herlocker et al. (1999, p. 2), aonde o autor apresenta uma estrutura algorítmica para realizar filtragem colaborativa e novos elementos algorítmicos que aumentam a precisão dos algoritmos de previsão colaborativa. Para uma correlação entre as variáveis  $x$  e  $y$ , o coeficiente de correlação de *Pearson*, da amostra, é apresentada na Equação 3.1.

$$p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} \quad (3.1)$$

Outra forma muito comum de se calcular um coeficiente de correlação é por meio do indicador de *Spearman* ou  $\rho$  de *Spearman* que de acordo com Spearman (1904, p. 77) é uma medida não paramétrica de correlação de postos (dependência estatística entre a classificação de duas variáveis). O coeficiente avalia com que intensidade a relação entre duas variáveis pode ser descrita pelo uso de uma função monótona<sup>2</sup>. Em termos simples, seu cálculo é apresentado na Equação 3.2, aonde  $p$  é o  $\rho$ ,  $n$  é o número de observações e  $d$  é a diferença entre os dois

<sup>2</sup>De acordo com Oliveira (2020, p. 6), no escopo da matemática, uma função entre dois conjuntos ordenados é monótona quando ela preserva (ou inverte) a relação de ordem. Quando a função preserva a relação, ela é chamada de função crescente. Quando ela inverte a relação, ela é chamada de função decrescente.



postos de cada observação, a saber:

$$p = 1 - \frac{6 \sum_{i=1}^n d^2}{n(n^2 - 1)} \quad (3.2)$$

De acordo com Suryakant e Mahara (2016, p. 452), em suma, a diferença entre os coeficientes de *Pearson* e *Spearman* é que a correlação de *Pearson* avalia a relação linear entre duas variáveis contínuas. Uma relação é linear quando a mudança em uma variável é associada a uma mudança proporcional na outra variável e a correlação de *Spearman* avalia a relação monotônica entre duas variáveis contínuas ou ordinais. Em uma relação monotônica, as variáveis tendem a mudar juntas mas não necessariamente a uma taxa constante. O coeficiente de correlação de *Spearman* baseia-se nos valores classificados de cada variável, em vez de os dados brutos.

Por último, também comumente utilizada para o desenvolvimento de S.R's é a similaridade do cosseno que, de acordo com Grus (2015, p. 269), é uma medida da similaridade entre dois vetores num espaço vetorial que avalia o valor do cosseno do ângulo compreendido entre eles. Dados dois vetores  $v$  e  $w$  ela mede o “ângulo” entre  $v$  e  $w$ . Se  $v$  e  $w$  apontarem para a mesma direção, então o numerador e o denominador são iguais e sua similaridade é igual a 1. Se  $v$  e  $w$  apontam para direções opostas, sua similaridade é igual a -1. E se  $v$  é o sempre que  $w$  não é (vice-versa) então sua similaridade é 0. Essa técnica é comumente utilizada em mineração de textos para de estabelecer uma métrica de semelhança entre textos. Tal conceito é apresentado na Equação 3.3.

$$p = \frac{\sum_{i=1}^n v_i w_i}{\sqrt{\sum_{i=1}^n v_i^2} \sqrt{\sum_{i=1}^n w_i^2}} \quad (3.3)$$

Na Tabela 3.1 é apresentado uma adaptação de Suryakant e Mahara (2016, p. 452) aonde são apresentados os principais coeficientes adaptados para S.R's e seus devidos pontos fortes e fracos<sup>3</sup>.

### 3.3 CÓDIGO

---

Primeiramente, essa seção apresentará um Pseudocódigo do algoritmo utilizado neste trabalho. De acordo com Fernandes e Fernandes (2014, p. 17), o pseudocódigo serve para evitar alguns problemas de interpretação na linguagem natural devido ao seu grau de complexidade relativamente elevado. Em seguida, o código será desmembrado em partes para que possam ser feitas as devidas observações.

---

<sup>3</sup>De acordo com Penrose (2008, p. 59), a cardinalidade de um conjunto é uma medida do “número de elementos do conjunto”.

---

**Algoritmo 1:** Pseudocódigo

---

**Input:** Histórico escolar de todos os usuários.

**Output:** Recomendação de matérias para o próximo semestre.

```
1 Importar o pacote de matemática;
2 def similaridade do cosseno;
3 Criar uma lista com valores únicos de todas as matérias apresentadas pelos usuários;
4 def vetor do usuário;
5 if Usuário já possui a matéria then
6 |   Retornar 1;
7 else
8 |   Retornar 0;
9 end
10 def Similaridade entre os usuários;
11 def Comando para trazer a lista de recomendações de matérias para usuário x;
12 Criar uma lista ordenada;
13 Excluir interesses já existentes;
```

---

Como apresentado no Capítulo 1, muitos alunos de Administração consideram o início de semestre um período bastante conturbado, devido ao grande volume de alunos acessando o sistema da universidade. Pensando nisso, o código apresentado no Algoritmo 1 foi desenvolvido para ajudar ambos os lados, melhorando a experiência dos alunos ao receberem recomendações mais personalizadas e direcionadas e ajudando a universidade a melhorar cada vez mais o seu sistema, tendo em vista os pontos positivos de *Python*, apresentados no início deste capítulo.

Para o desenvolvimento desse código, o autor Grus (2015, pp. 269-272) foi também utilizado como referência. A estratégia de filtragem colaborativa, apresentada pelo mesmo autor, foi tomada em conta e escolhida para o desenvolvimento do S.R. Portanto, o objetivo deste S.R. é indicar matérias baseadas nos registros e nas opções de usuários semelhantes ao usuário atual.

Visando uma visualização melhor do S.R. proposto, um esquema similar ao apresentado por Schafer et al. (1999, p. 160) foi desenvolvido e apresentado na Figura 3.2:

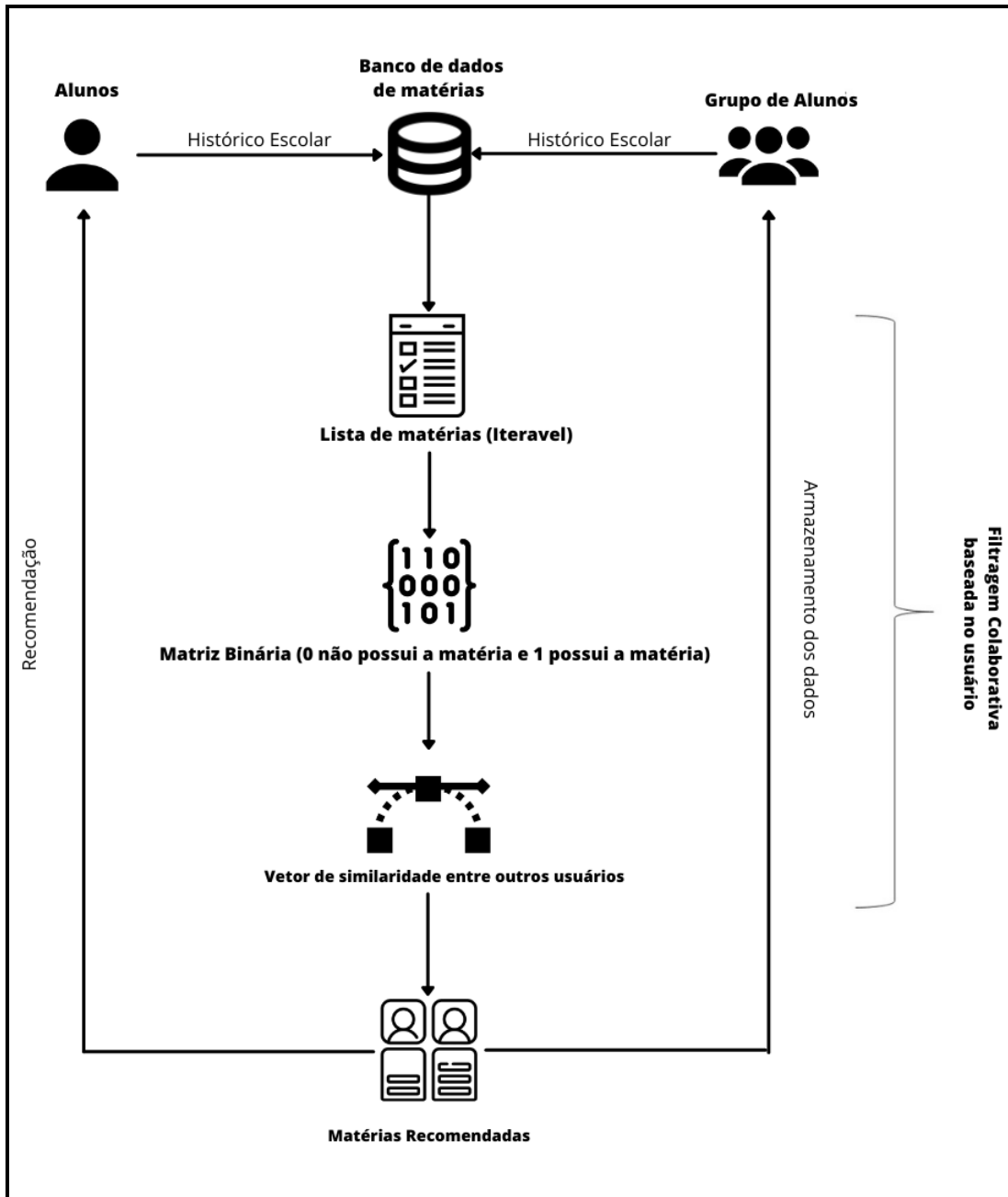


Figura 3.2: Estrutura do S.R. para sugestão de matérias.

Coeficiente	Fórmula	Desvantagens
<i>Pearson.</i>	$\text{sim}(u, u') = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{u',i} - \bar{r}_{u'})}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \cdot \sqrt{\sum_{i \in I} (r_{u',i} - \bar{r}_{u'})^2}}$	Sofre de problemas de poucos itens co-avaliados. Produz alta ou baixa similaridade, mesmo se houver uma diferença significativa nas classificações.
<i>Spearman.</i>	$\text{sim}(u, u') = 1 - \frac{6 \sum_{i \in I} \text{rank}(r_{u,i})^2 - \text{rank}(r_{u',i})^2}{ I ( I ^2 - 1)}$	Resulta, as vezes, em alta similaridade, mesmo que as avaliações não sejam semelhantes.
<i>Cosseno.</i>	$\text{sim}(u, u') = \frac{\sum_{i \in I} (r_{u,i}) \cdot (r_{u',i})}{\sqrt{\sum_{i \in I} (r_{u,i})^2} \cdot \sqrt{\sum_{i \in I} (r_{u',i})^2}}$	Sofre de problemas de poucos itens co-avaliados. Dá alta similaridade. Produz alta similaridade, mesmo se houver uma diferença significativa nas classificações.

**Tabela 3.1:** Equações de similaridade.

**Fonte:** Adaptado de Suryakant e Mahara (2016, p. 452).

**Nota:** I representa o conjunto de itens,  $r_{u,i}$  avaliação dada ao item i pelo usuário u,  $\bar{r}_u$  é a média de avaliação do usuário u e |I| é a cardinalidade de itens co-avaliados.

Primeiramente, foi necessário a coleta de dados para serem utilizados como base para o S.R. Logo, foram coletados 51 históricos escolares de alunos da Universidade de Brasília - UnB, formados ou no estágio final de formação, do curso de graduação de Administração. Além disso, para se iniciar o código foi necessário importar o pacote de matemática para *Python* e já deixar definido a similaridade do cosseno para uso futuro. Tal raciocínio é apresentado no Algoritmo 2, aonde  $H_x$  é o histórico escolar do usuário  $x$ :

---

**Algoritmo 2:** Coleta de dados, importação de pacotes e definição da similaridade do cosseno.

---

```
1 users_interests = [(H.1),(H.2),...,(H.x)];
2 from numpya import dotb;
3 import math;
4 def cosine_similarity(v, w):
5     | return dot(v, w) / math.sqrt(dot(v, v) × dot(w, w));
```

---

<sup>a</sup>Pacote para a linguagem *Python* que suporta *arrays* e matrizes multidimensionais, possuindo uma larga coleção de funções matemáticas para trabalhar com estruturas.

<sup>b</sup>Fornecer uma função para realizar o produto escalar de duas matrizes. Se ambas as matrizes “a” e “b” são matrizes unidimensionais, a função `dot()` realiza o produto interno dos vetores.

A próxima etapa para o desenvolvimento do algoritmo é criar uma lista com todas as matérias que estão presentes no banco de dados, para que dessa forma, seja possível futuramente realizar a identificação das matérias que um usuário possui ou não. Para isso, todos os itens repetidos no banco de dados são excluídos para que só sobre itens únicos, conforme observa-se abaixo:

---

**Algoritmo 3:** Criando lista com todas as matérias presentes.

---

```
1 unique_interests=
2     | sorteda(list({interest
3     | for users_interests in users_interests
4     | for interest in users_interests}));
```

---

<sup>a</sup>Define uma nova lista ordenada à partir de um iterável.

Ao gerar uma lista com valores únicos de todas as matérias presentes nos históricos escolares foi obtido um valor de 47 matérias. Em seguida, é necessário produzir um vetor “interesse” de 0 e 1 no qual o 0 significa que o usuário nunca realizou aquela matéria e 1 representa que o usuário já a realizou, como apresentado no Algoritmo 4.

---

**Algoritmo 4:** Definindo vetor para cada usuário listando 1 para interesse possuído e 0 para não.

---

```
1 from typinga import List;
2 def make_user_interest_vector(user_interests: List[str]) -> List[int]:
3     return [1 if interest in user_interests else 0
4             for interest in unique_interests];
5 user_interest_vectors =
6     [make_user_interest_vector(user_interests)
7     for user_interests in users_interests];
```

---

<sup>a</sup>É *built-in* a partir do *Python* 3.5 e traz um conjunto de tipos para serem utilizados em definições mais complexas.

Após definido os vetores do usuário, é possível criar uma matriz e dentro da mesma, realizar uma das partes mais importantes em todo esse processo, que é calcular a similaridade entre os usuários por meio da similaridade do cosseno, como pode-se observar no Algoritmo 5.

---

**Algoritmo 5:** Definindo a similaridade dos usuários.

---

```
1 user_similarities = [[cosine_similarity(interest_vector_i, interest_vector_j)
2                       for interest_vector_j in user_interest_vectors]
3                       for interest_vector_i in user_interest_vectors];
```

---

Nesse ponto, `user_similarities[i]` é um vetor de similaridades do usuário `i` para cada outro usuário. Com isso, uma função foi escrita para encontrar os usuários mais parecidos ao usuário pesquisado e ao mesmo tempo ordenar a lista com os usuários mais similares, retirar o usuário procurado da lista e retirar usuários com similaridade 0. Conforme apresentado no Algoritmo 6.

---

**Algoritmo 6:** Encontrando outros usuários com similaridade não zero e ordenando a lista de resultados.

---

```
1 def most_similar_user_to(user_id):
2     pairs = [(other_user_id, similarity)
3             for other_user_id, similarity in
4             enumerate(user_similarities[user_id])
5             if user_id != other_user_id and similarity > 0];
6     return sorted(pairs,
7                  key=lambda pair: pair[-1],
8                  reverse=True);
```

---

Por último, para se calcular e entregar ao usuário as recomendações, a estratégia é que para cada matéria, pode-se somar similaridades de usuário dos outros usuários interessados, converter em uma lista ordenada e então excluir matérias já realizadas, como devidamente apresentado no Algoritmo 7.

---

**Algoritmo 7:** Definindo matérias recomendadas em uma lista ordenada e única.

---

```
1 from collections import defaultdict;
2 def user_based_suggestions(user_id, include_current_interests=False):
3     suggestion = defaultdict(float)
4     for other_user_id, similarity in most_similar_user_to(user_id):
5         for interest in users_interests[other_user_id]:
6             suggestion[interest] += similarity
7     suggestion = sorted(suggestion.items()),
8         key=lambda pair: pair[-1],
9         reverse=True)
10    if include_current_interests:
11        return suggestion
12    else:
13        return [(suggestion, weight)
14                for suggestion, weight in suggestion
15                if suggestion not in users_interests[user_id]].
```

---

Logo, quando colocado no console o comando `user_based_suggestions(Número_do_usuario)` o S.R. irá te retornar as recomendações de matérias de acordo com a filtragem colaborativa baseada no usuário. Para facilitar a visualização do código como um todo, é apresentado a seguir a sua versão completa.

---

**Algoritmo 8:** Código completo (parte I).

---

```
1 users_interests = [(H1),(H2),...,(Hx)];
2 from numpy import dot;
3 import math;
4 def cosine_similarity(v, w):
5     return dot(v, w) / math.sqrt(dot(v, v) × dot(w, w));
6 unique_interests=
7     sorted(list({interest
8         for users_interests in users_interests
9         for interest in users_interests}));
10 def make_user_interest_vector(user_interests: List[str]) -> List[int]:
11     return [1 if interest in user_interests else 0
12         for interest in unique_interests];
13 user_interest_vectors=
14     [make_user_interest_vector(user_interests)
15     for user_interests in users_interests];
16 user_similarities = [[cosine_similarity(interest_vector_i, interest_vector_j)
17     for interest_vector_j in user_interest_vectors]
18     for interest_vector_i in user_interest_vectors];
19 def most_similar_user_to(user_id):
20     pairs = [(other_user_id, similarity)
21         for other_user_id, similarity in
22         enumerate(user_similarities[user_id])
23         if user_id != other_user_id and similarity > 0];
24     return sorted(pairs ,
25         key=lambda pair: pair[-1],
26         reverse=True)
```

---



---

**Algoritmo 9:** Código completo (parte 2).

---

```
1 from collections import defaultdict;
2 def user_based_suggestions(user_id, include_current_interests=False):
3     suggestion = defaultdict(float)
4     for other_user_id, similarity in most_similar_user_to(user_id):
5         for interest in users_interests[other_user_id]:
6             suggestion[interest] += similarity
7     suggestion = sorted(suggestion.items()),
8         key=lambda pair: pair[-1],
9         reverse=True)
10    if include_current_interests:
11        return suggestion
12    else:
13        return [(suggestion, weight)
14                for suggestion, weight in suggestion
15                if suggestion not in users_interests[user_id]]
```

---



# CAPÍTULO 4

## RESULTADOS

“Se tornou aparentemente óbvio que  
nossa tecnologia excedeu nossa  
humanidade”  
—Albert Einstein.

Como dito no Capítulo 3, 51 históricos<sup>1</sup> escolares foram usados para começar o preenchimento do banco de dados do S.R. proposto neste trabalho. Observando esse contexto, o objetivo desse capítulo é apresentar os resultados obtidos pelo S.R., assim como realizar testes para que seja possível a análise da eficiência da técnica proposta. Além disso, é importante que o ambiente de desenvolvimento utilizado seja apresentado para o total entendimento de como todos os processos para identificação dos resultados foram realizados.

### 4.1 AMBIENTE

---

Como dito anteriormente, o código utilizado no S.R. proposto por esse trabalho foi escrito na linguagem de programação *Python* versão 3.7.6. O ambiente de desenvolvimento utilizado foi baixado por meio do pacote *Anaconda*<sup>®</sup>, que de acordo com Inc. (2014) é uma distribuição das linguagens de programação *Python* e *R* para computação científica (ciência de dados, aplicativos de aprendizado de máquina, processamento de dados em grande escala, análise preditiva, etc.), que visa simplificar o gerenciamento e implantação de pacotes.

Dentre esses pacotes, é encontrada a IDE<sup>2</sup> *Spyder*, que de acordo com seu criador Ray-

---

<sup>1</sup>Nota do autor: A título de explicação foi consultado a coordenação na data 18/03/2021 a fim de identificar quantos alunos são formados em média pelo curso de ADM, que são 75 por semestre. Temos 51 históricos, fazendo com que tivéssemos uma média de 68% do total. É importante destacar que alguns alunos não disponibilizaram o histórico acadêmico por questão de sigilo.

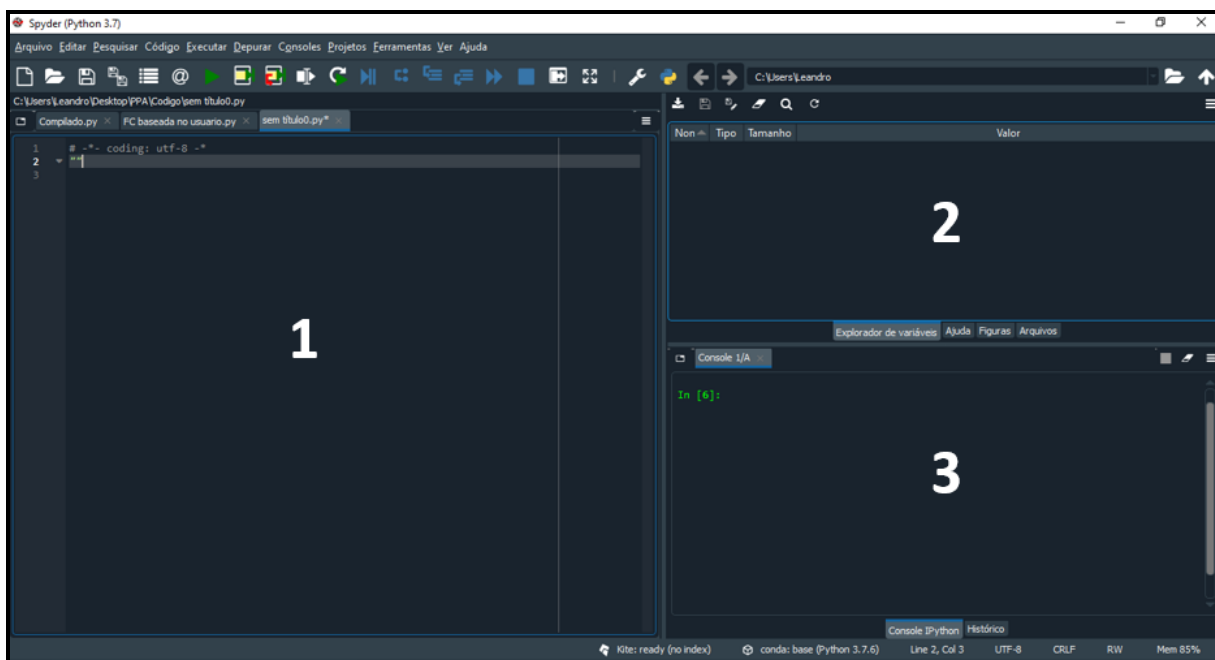
<sup>2</sup>De acordo com Vieira, Raabe, e Zeferino (2009, p. 2), IDE é a sigla em inglês para *Integrated Development Environment*, ou seja, ambiente de desenvolvimento integrado, um *software* que combina ferramentas comuns de desenvolvimento em uma única interface gráfica do usuário, facilitando o desenvolvimento de aplicações.

baut (2009) é num ambiente científico gratuito e de código aberto<sup>3</sup> escrito em *Python*, para *Python* e projetado por e para cientistas, engenheiros e analistas de dados. Ele apresenta uma combinação única de edição avançada, análise, depuração e funcionalidade de criação de perfil de uma ferramenta de desenvolvimento abrangente com a exploração de dados, execução interativa, inspeção profunda e belas capacidades de visualização de um pacote científico.

Segundo Raybaut (2009), esses são alguns dos atributos oferecidos pelo *Spyder* que a tornam uma poderosa ferramenta de desenvolvimento:

- **Editor:** O Editor *multilíngue* do *Spyder* integra uma série de ferramentas poderosas prontas para uma experiência de edição eficiente e fácil. Os principais recursos do editor incluem destaque de sintaxe (pigmentos), código em tempo real e análise de estilo;
- **Console:** Permite que você execute comandos e insira, interaja e visualize dados dentro de qualquer número de intérpretes *IPython* completos. Cada console é executado em um processo separado, permitindo que você execute *scripts*<sup>4</sup>, interrompa a execução e reinicie ou encerre um *shell*<sup>5</sup> sem afetar os outros ou o próprio *Spyder*, e testar facilmente seu código em um ambiente limpo sem interromper sua sessão primária;
- **Explorador de Variáveis:** O explorador de variáveis permite que você navegue e gerencie interativamente os objetos gerados executando seu código. Ele fornece informações sobre o nome, tamanho, tipo e valor de cada objeto.

Na Figura 4.1, é possível notar o editor ocupando a parte esquerda da tela (1), o console a parte direita superior (2) e o explorador de variáveis na direita inferior (3).



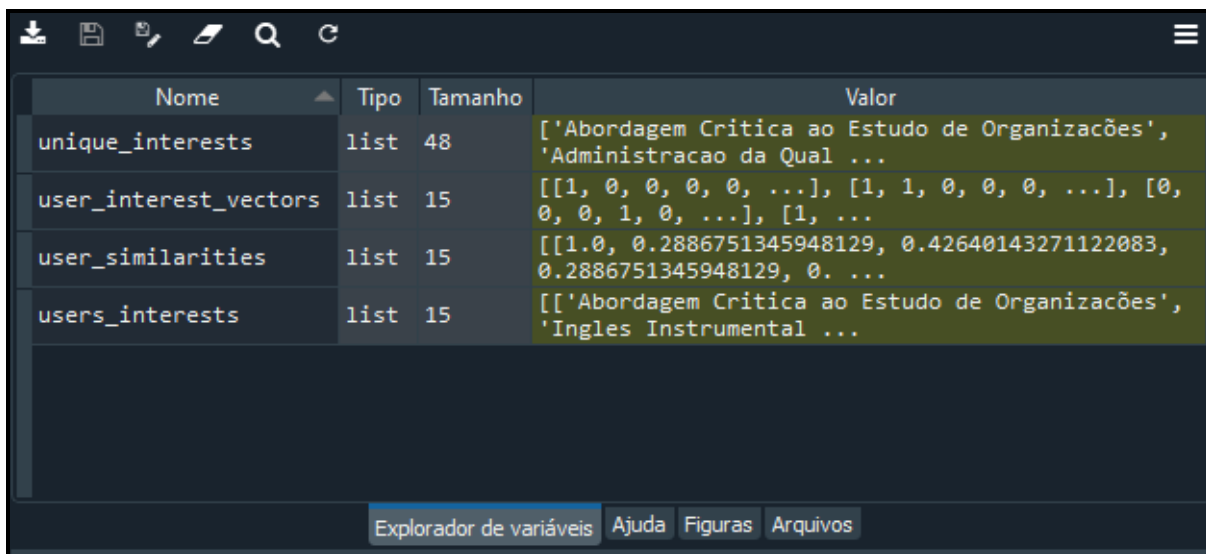
**Figura 4.1:** Ambiente *Spyder*.

<sup>3</sup>De acordo com L. A. Silva (2003), o conceito de código aberto compreende um modelo de desenvolvimento criado em 1998, que promove o licenciamento livre para o design ou esquematização de um produto, e a redistribuição universal desses, com a possibilidade de livre consulta, examinação ou modificação do produto, sem a necessidade de pagar uma licença comercial, promovendo um modelo colaborativo de produção intelectual.

<sup>4</sup>Nota do autor: Sequências de códigos de uma linguagem de programação.

<sup>5</sup>De acordo com Newham (2005, p. 2), em termos gerais, em computação, um *shell* é uma interface de usuário para acessar os serviços de um sistema operacional.

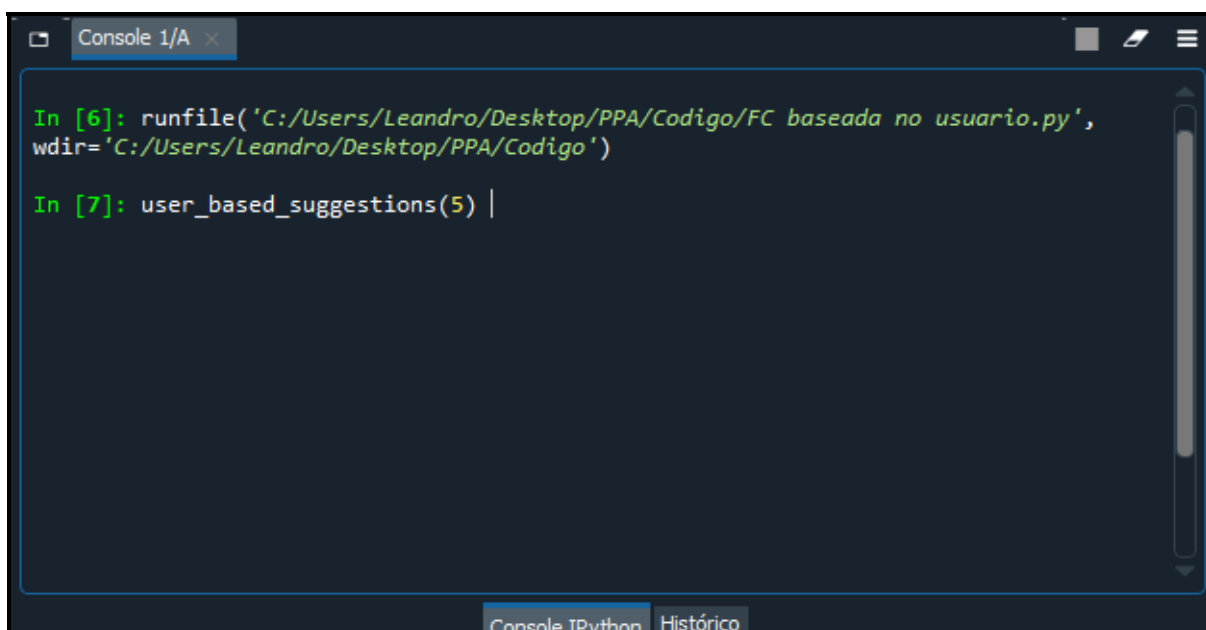
Apresentado a interface, é necessário uma breve explicação do fluxo de processamento para o entendimento de como é feita a utilização e a interpretação dos resultados do S.R. Sendo assim, o código precisa ser executado no sinal de *start* na parte superior da tela. Após executado, será possível verificar que o ambiente explorador de variáveis será preenchido com todos os objetos do código, conforme apresentado na Figura 4.2.



Nome	Tipo	Tamanho	Valor
unique_interests	list	48	['Abordagem Critica ao Estudo de Organizações', 'Administracao da Qual ...
user_interest_vectors	list	15	[[1, 0, 0, 0, 0, ...], [1, 1, 0, 0, 0, ...], [0, 0, 0, 1, 0, ...], [1, ...
user_similarities	list	15	[[1.0, 0.2886751345948129, 0.42640143271122083, 0.2886751345948129, 0. ...
users_interests	list	15	['Abordagem Critica ao Estudo de Organizações', 'Ingles Instrumental ...

Figura 4.2: Explorador de variáveis.

Como observado na Figura 4.2, é possível notar todos os objetos criados pelo código, como também seu tipo, e seu tamanho respectivamente. Em seguida, para gerar as recomendações é necessário digitar o comando `user_based_suggestions` no console e entre parênteses o número do usuário desejado, como mostrado na Figura 4.3:



```

In [6]: runfile('C:/Users/Leandro/Desktop/PPA/Codigo/FC baseada no usuario.py',
wdir='C:/Users/Leandro/Desktop/PPA/Codigo')

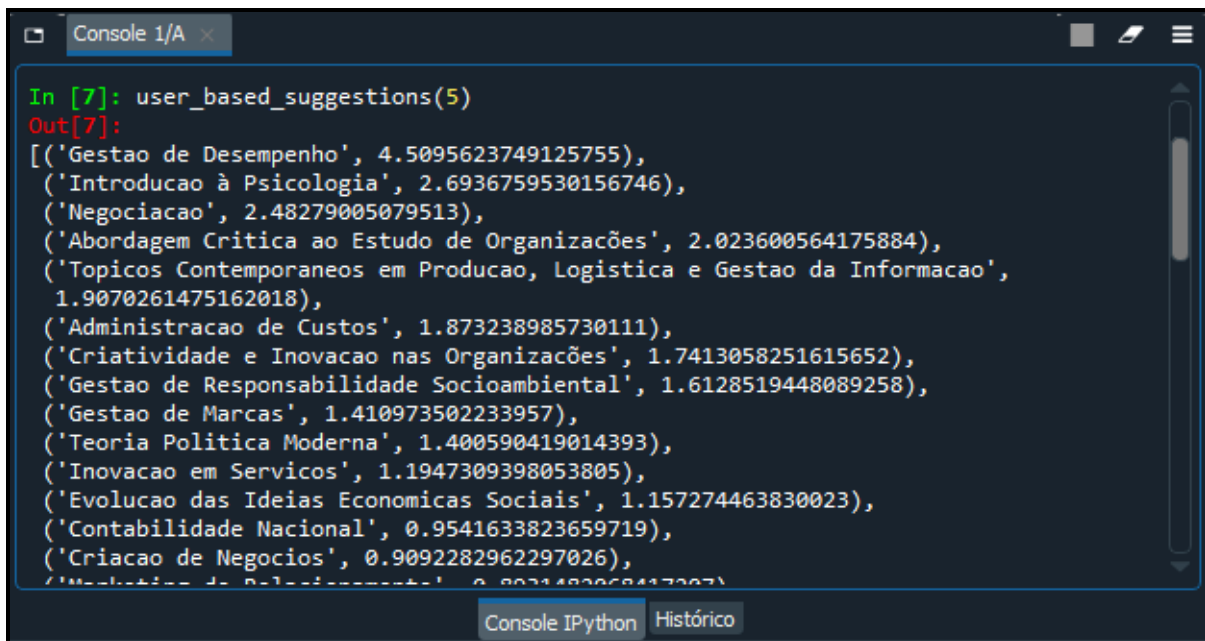
In [7]: user_based_suggestions(5) |

```

Figura 4.3: Executando o código.

Finalmente, o código será executado e o S.R. irá retornar todas as matérias recomendadas para o usuário solicitado. Os resultados apresentados são listados no seguinte formato,

('Administração da Qualidade', 3.071530698791393) aonde primeiro é a matéria recomendada, seguida pelo seu peso gerado a partir da similaridade do cosseno. Na Figura 4.4 é apresentado um exemplo dos resultados obtidos.



```
In [7]: user_based_suggestions(5)
Out[7]:
[('Gestao de Desempenho', 4.5095623749125755),
 ('Introducao à Psicologia', 2.6936759530156746),
 ('Negociacao', 2.48279005079513),
 ('Abordagem Critica ao Estudo de Organizações', 2.023600564175884),
 ('Topicos Contemporaneos em Producao, Logistica e Gestao da Informacao',
 1.9070261475162018),
 ('Administracao de Custos', 1.873238985730111),
 ('Criatividade e Inovacao nas Organizações', 1.7413058251615652),
 ('Gestao de Responsabilidade Socioambiental', 1.6128519448089258),
 ('Gestao de Marcas', 1.410973502233957),
 ('Teoria Politica Moderna', 1.400590419014393),
 ('Inovacao em Servicos', 1.1947309398053805),
 ('Evolucao das Ideias Economicas Sociais', 1.157274463830023),
 ('Contabilidade Nacional', 0.9541633823659719),
 ('Criacao de Negocios', 0.9092282962297026),
 ('Marketing de Relacionamento', 0.8021482008417307)]
```

Figura 4.4: Exemplo de resultados.

### 4.3 LISTA

Após realizar a filtragem das matérias, apresentada no Algoritmo 3, é definida a seguinte lista de matérias para `unique_interests`, que é um conjunto de todas as matérias apresentadas nos históricos escolares, conforme apontado Tabela 4.1.

Consequentemente, observado na Tabela 4.1, dentro dos 51 históricos escolares obtidos como base, é possível observar 69 matérias distintas, é importante ressaltar que as matérias acima são todas optativas, as quais o aluno tem a opção de realizá-las durante a graduação ou não. Dessa forma, o S.R. irá trabalhar inicialmente recomendando uma destas matérias que constam na lista.

### 4.4 PRÉ-TESTE

No início do trabalho, um teste foi realizado com os primeiros 15 históricos escolares obtidos a fim de testar o S.R. podendo consertar suas falhas, verificar sua eficiência e testar sua viabilidade de uso. Os resultados desse pré-teste são apresentados na Tabela 4.2.

Após a realização do pré-teste foi verificada a viabilidade da utilização do S.R. desenvolvido para este trabalho e também nota-se um bom desempenho do mesmo, aonde ao final o sistema recomendou 73% das vezes pelo menos uma matéria retirada conforme visto na Figura 4.5.

<b>Disciplinas</b>	
Análise da Liquidez.	Inglês Instrumental 1.
Análise Econômico-Financeira 1.	Inovação em Serviços.
Análise Institucional.	Inovação no Setor Público.
Auditoria 1.	Internacionalização de Empresas e Gestão de Negócios.
Avaliação da Eficiência e Produtividade.	Introdução à Ciência da Computação.
Avaliação de Projetos de Investimento.	Introdução à Filosofia.
Avaliação de Treinamento e de Desenvolvimento.	Introdução à Psicologia.
Avaliação e Monitoramento da Estratégia Organizacional.	Introdução à Sociologia.
Comportamento do Consumidor.	Legislação Administrativa.
Comportamento Humano e Trabalho.	Legislação Social.
Comportamento Organizacional.	Legislação Tributária.
Contabilidade Comercial.	Língua Alemã 1.
Contabilidade Geral 2.	Língua de Sinais Brasileira.
Contabilidade Nacional.	Língua Espanhola 1.
Criação de Negócios.	Língua Italiana 1.
Criatividade e Inovação nas Organizações.	Marketing de Relacionamento.
Custos.	Marketing de Serviços.
Economia Brasileira.	Marketing Digital.
Ergonomia.	Marketing Socialmente Responsável.
Estado, Governo e Sociedade.	Mercado Financeiro e de Capitais.
Evolução das Ideias Econômicas Sociais.	Negociação.
Finanças 2.	Pesquisa de Marketing.
Formação Econômica do Brasil.	Programação Fiscal e Financeira.
Fundamentos da Administração Pública.	Sistema de Informações Contábeis.
Fundamentos de Políticas Públicas.	Teoria Política Moderna.
Gestão da Inovação.	Tópicos Avançados em Computadores.
Gestão de Compras Públicas.	Tópicos Contemporâneos em Administração 2.
Gestão de Desempenho.	Tópicos Contemporâneos em Estratégia.
Gestão de Marcas.	Tópicos Contemporâneos em Estratégia e Inovação em Organização.
Gestão de Processos.	Tópicos Contemporâneos em Finanças.
Tópicos Contemporâneos em Produção, Logística e Gestão da Informação - TCPLGI.	

**Tabela 4.1:** Lista de disciplinas.

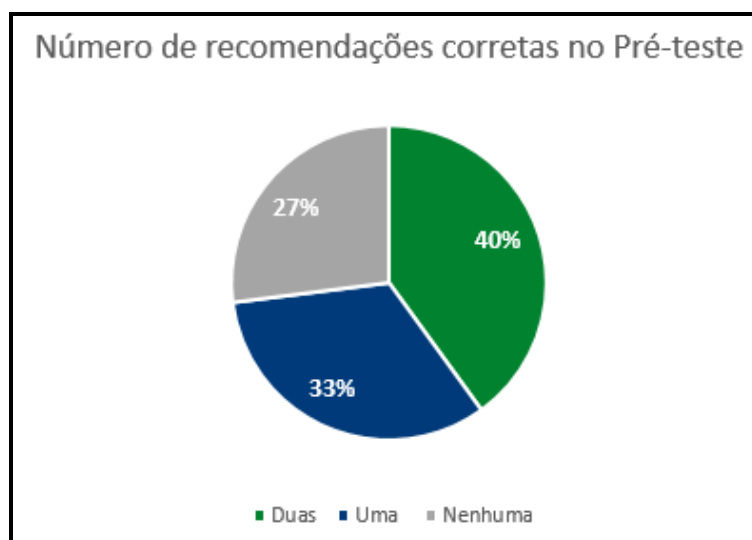
#### 4.5 RESULTADOS

A título de testes, um procedimento foi adotado para que o S.R. desenvolvido pudesse ser testado. Todos os 51 usuários foram submetidos a um teste no S.R., mas para isso, duas matérias foram retiradas de seus históricos escolares para que ao final fosse possível observar se o S.R. recomendou pelo menos uma das matérias entre as suas cinco primeiras recomendações.

Para facilitar na montagem das listas de cada usuário e evitar a parcialidade, um sistema foi desenvolvido para retirar aleatoriamente as matérias de cada usuário e em seguida, criar a lista sem as mesmas, conforme observado no Pseudocódigo apresentado no Algoritmo 4.1 ou pelo código completo no Anexo.

Usuário	Primeira disciplina removida	Segunda disciplina removida	Acertos
0	Negociação I.	Gestão da Inovação.	1
1	Marketing Digital.	Criatividade e Inov. nas Organizações.	0
2	Gestão de Desempenho.	Mercado Financeiro e de Capitais.	2
3	Negociação.	Criação de Negócios.	0
4	Administração da Qualidade.	Tópicos Cont. em Administração 2.	2
5	Marketing Digital.	Análise Institucional.	1
6	Economia Brasileira.	Mercado Financeiro e de Capitais.	1
7	Comportamento do Consumidor.	Administração de Custos.	1
8	Introdução à Psicologia.	Gestão da Inovação.	2
9	Inglês Instrumental I.	TCPLGI.	2
10	Administração de Custos.	Gestão da Inovação.	1
11	Introdução à Psicologia.	Gestão de Processos.	2
12	Administração da Qualidade.	Formação Econômica do Brasil.	2
13	Gestão de Compras Públicas.	Inglês Instrumental I.	0
14	Criação de Negócios.	Gestão de Marcas.	0

**Tabela 4.2:** Resumo dos resultados do pré-teste.



**Figura 4.5:** Gráfico da distribuição de acertos do S.R. (Pré-teste).

---

**Algoritmo 10:** Pseudocódigo 2.

---

**Input:** Histórico escolar de todos os usuários.

**Output:** Lista de matérias retiradas e listas testes.

- 1 Importar o pacote de *random*;
  - 2 Importar o pacote de *copy*;
  - 3 **def** Remover 2 matérias;
  - 4 Remover 2 matérias da lista de cada usuário aleatoriamente;
  - 5 **for** *x* **in** *array*<sup>a</sup> de matérias;
  - 6 Remover dois(*x*,*i*);
  - 7 **Print** número do usuário e matérias retiradas;
  - 8 **Print** lista de teste para cada usuário;
- 

<sup>a</sup>Segundo Grus (2015), um *array* é a estrutura de dados que armazena uma coleção de elementos de tal forma que cada um dos elementos possa ser identificado por, pelo menos, um índice ou uma chave.



A partir do processo previamente apresentado, 51 listas novas foram criadas, cada uma com sua respectiva matéria retirada, logo já é possível realizar os testes. No entanto, é importante ressaltar que no *Python*, os caracteres com acentos e “ç” não são registrados, e na linguagem a sequência numérica começa a ser contada a partir de 0. Nas subseções 4.5.1, 4.5.2 e 4.5.3 são mostrados 3 exemplos de como os resultados são apresentados pelo S.R.

#### 4.5.1 DADOS: USUÁRIO O

Aluno o:

- **Matérias:**

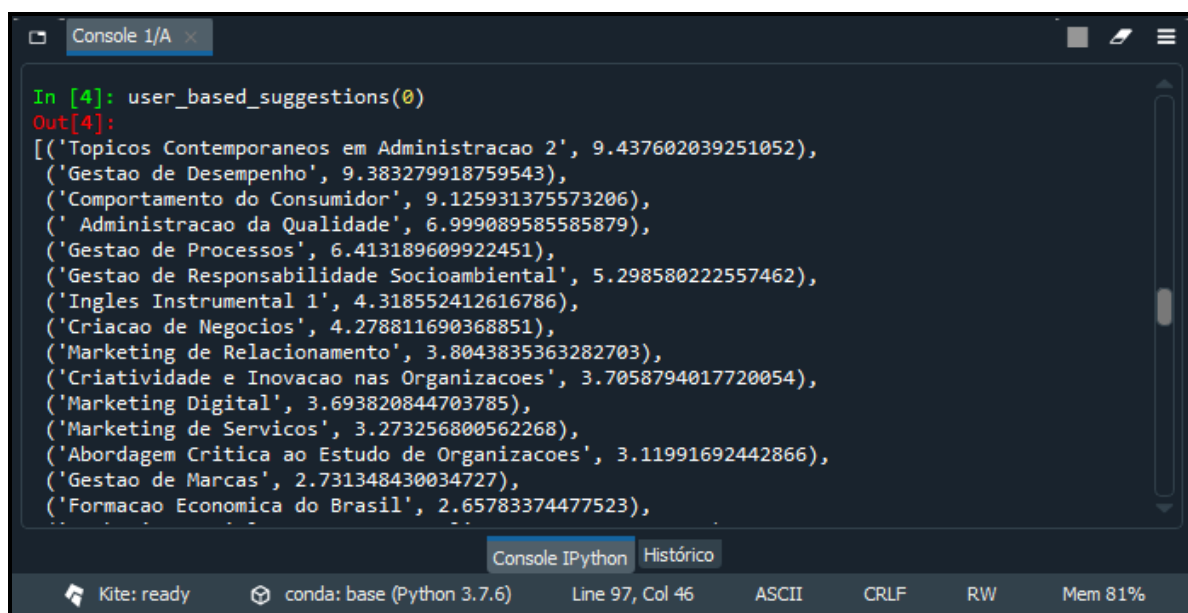
1. Inglês Instrumental;
2. Introdução à Psicologia;
3. Gestão da Inovação;
4. Negociação;
5. Tópicos Contemporâneos em Produção, Logística e Gestão da Informação.

- **Retiradas:** Inglês Instrumental I e Gestão de Desempenho;

- **Recomendadas:**

1. Tópicos Contemporâneos em Administração 2 - 9,44;
2. Gestão de Desempenho - 9,38;
3. Comportamento do Consumidor - 9,12;
4. Administração da Qualidade - 6,99;
5. Gestão de Processos 6,41.

- **Número de Acertos:** 1



```
Console 1/A x
In [4]: user_based_suggestions(0)
Out[4]:
[('Tópicos Contemporaneos em Administracao 2', 9.437602039251052),
 ('Gestao de Desempenho', 9.383279918759543),
 ('Comportamento do Consumidor', 9.125931375573206),
 (' Administracao da Qualidade', 6.999089585585879),
 ('Gestao de Processos', 6.413189609922451),
 ('Gestao de Responsabilidade Socioambiental', 5.298580222557462),
 ('Ingles Instrumental 1', 4.318552412616786),
 ('Criacao de Negocios', 4.278811690368851),
 ('Marketing de Relacionamento', 3.8043835363282703),
 ('Criatividade e Inovacao nas Organizacoes', 3.7058794017720054),
 ('Marketing Digital', 3.693820844703785),
 ('Marketing de Servicos', 3.273256800562268),
 ('Abordagem Critica ao Estudo de Organizacoes', 3.11991692442866),
 ('Gestao de Marcas', 2.731348430034727),
 ('Formacao Economica do Brasil', 2.65783374477523),
 ...]
```

Figura 4.6: Resultados para o usuário o.

Aluno 1:

• **Matérias:**

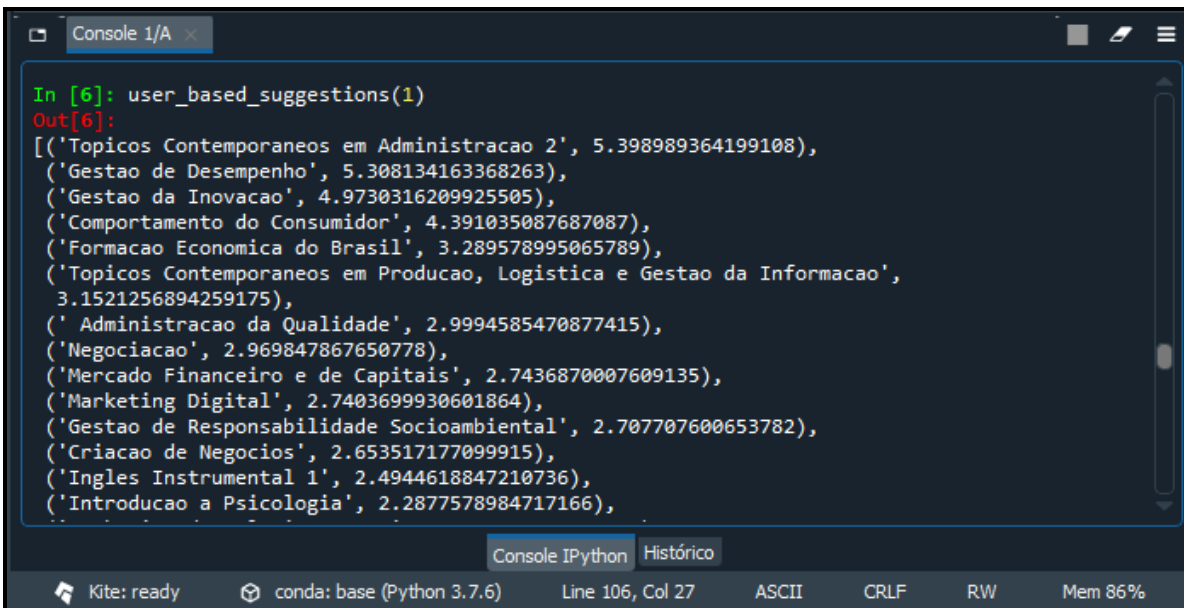
1. Abordagem Crítica ao Estudo de Organizações;
2. Criatividade e Inovação nas Organizações;
3. Marketing Digital;
4. Administração da Qualidade;
5. Gestão de Processos.

• **Retiradas:** Marketing Digital e Abordagem Critica ao Estudo de Organizações;

• **Recomendadas:**

1. Tópicos Contemporâneos em Administração 2 - 5,40;
2. Gestão de Desempenho - 5,30;
3. Gestão da Inovação - 4,97;
4. Comportamento do Consumidor - 4,39;
5. Formação Econômica do Brasil - 3,29.

• **Número de Acertos:** 0



```
In [6]: user_based_suggestions(1)
Out[6]:
[('Tópicos Contemporâneos em Administração 2', 5.398989364199108),
 ('Gestão de Desempenho', 5.308134163368263),
 ('Gestão da Inovação', 4.9730316209925505),
 ('Comportamento do Consumidor', 4.391035087687087),
 ('Formação Econômica do Brasil', 3.289578995065789),
 ('Tópicos Contemporâneos em Produção, Logística e Gestão da Informação',
 3.1521256894259175),
 ('Administração da Qualidade', 2.9994585470877415),
 ('Negociação', 2.969847867650778),
 ('Mercado Financeiro e de Capitais', 2.7436870007609135),
 ('Marketing Digital', 2.7403699930601864),
 ('Gestão de Responsabilidade Socioambiental', 2.707707600653782),
 ('Criação de Negócios', 2.653517177099915),
 ('Inglês Instrumental 1', 2.4944618847210736),
 ('Introdução a Psicologia', 2.2877578984717166),
```

Figura 4.7: Resultados para o usuário 1.

Aluno 2:

• **Matérias:**

- I. Inglês Instrumental I;

2. Tópicos Contemporâneos em Administração 2;
3. Criatividade e Inovação nas Organizações;
4. Mercado Financeiro e de Capitais;
5. Análise Econômico Financeira;
6. Gestão de Marcas;
7. Análise da Liquidez;
8. Gestão da Inovação;
9. Auditoria I;
10. Gestão de Desempenho;
11. Gestão de Processos.

- **Retiradas:** Tópicos Contemporâneos em Administração 2 e Gestão de Desempenho;

- **Recomendadas:**

1. Comportamento do Consumidor - 7,28;
2. Gestão de Desempenho - 6,70;
3. Tópicos Contemporâneos em Administração 2 - 6,81;
4. Administração da Qualidade - 5,80;
5. Negociação - 4,96.

- **Número de Acertos:** 2

```

In [8]: user_based_suggestions(2)
Out[8]:
[('Comportamento do Consumidor', 7.282480098774478),
 ('Gestao de Desempenho', 6.8962787200179365),
 ('Topicos Contemporaneos em Administracao 2', 6.815302194582161),
 (' Administracao da Qualidade', 5.580219093839333),
 ('Negociacao', 4.962485441559634),
 ('Marketing Digital', 4.6746654404153745),
 ('Topicos Contemporaneos em Producao, Logistica e Gestao da Informacao',
 4.5243997012310375),
 ('Gestao de Responsabilidade Socioambiental', 3.9706626367016447),
 ('Introducao a Psicologia', 3.4129306429754),
 ('Criacao de Negocios', 3.36870364275769),
 ('Formacao Economica do Brasil', 3.2383873033851414),
 ('Marketing de Relacionamento', 2.56432025650564),
 ('Marketing de Servicos', 2.1739860531453585),
 ('Administracao da Qualidade', 2.067814482813594),
 ('Abandono Critico no Estudo de Organizacao', 1.8407642216303607)]

```

Figura 4.8: Resultados para o usuário 2.

#### 4.5.4 RESULTADOS INDIVIDUAIS

Para facilitar a visualização dos resultados, considerando os 51 alunos, as Tabelas 4.3 e 4.4 foram desenvolvidas contendo o número do usuário, matérias retiradas e a quantidade de acertos.

Usuário	Primeira disciplina removida	Segunda disciplina removida	Acertos
0	Inglês Instrumental 1.	Gestão de Desempenho.	1
1	Marketing Digital.	Abordagem Crítica ao Estudo de Organizações.	0
2	Tópicos Contemporâneos em Administração 2.	Gestão de Desempenho.	2
3	Abordagem Crítica ao Estudo de Organizações.	Criação de Negócios.	0
4	Gestão da Inovação.	Tópicos Contemporâneos em Produção, Logística e Gestão da Informação.	2
5	Inglês Instrumental 1.	Gestão de Processos.	1
6	Evolução das Ideias Econômicas Sociais.	Finanças 2.	0
7	Comportamento do Consumidor.	Mercado Financeiro e de Capitais.	1
8	Introdução à Psicologia.	Gestão de Responsabilidade Socioambiental.	0
9	Administração da Qualidade.	Fundamentos de Políticas Públicas.	1
10	Gestão de Responsabilidade Socioambiental.	Negociação.	2
11	Gestão da Inovação.	Mercado Financeiro e de Capitais.	1
12	Gestão da Inovação.	Introdução à Psicologia.	2
13	Gestão de Processos.	Contabilidade Nacional.	1
14	Comportamento Organizacional.	Gestão de Responsabilidade Socioambiental.	1
15	Gestão de Desempenho.	Finanças 2.	1
16	Introdução à Psicologia.	Tópicos Contemporâneos em Produção, Logística e Gestão da Informação.	1
17	Marketing Digital.	Gestão da Inovação.	1
18	Negociação.	Comportamento do Consumidor.	2
19	Tópicos Contemporâneos em Administração 2.	Comportamento do Consumidor.	2
20	Tópicos Avançados em Computadores.	Avaliação da Eficiência e Produtividade.	0
21	Criatividade e Inovação nas Organizações.	Marketing Digital.	0
22	Tópicos Contemporâneos em Produção, Logística e Gestão da Informação.	Inglês Instrumental 1.	0
23	Tópicos Contemporâneos em Produção, Logística e Gestão da Informação.	Marketing de Serviços.	1
24	Inglês Instrumental 1.	Gestão de Responsabilidade Socioambiental.	2
25	Abordagem Crítica ao Estudo de Organizações.	Tópicos Contemporâneos em Administração 2.	1
26	Inglês Instrumental 1.	Negociação.	1
27	Introdução à Psicologia.	Comportamento do Consumidor.	2
28	Inglês Instrumental 1.	Legislação Social.	1
29	Tópicos Contemporâneos em Administração 2.	Comportamento do Consumidor.	2
30	Gestão de Responsabilidade Socioambiental.	Tópicos Contemporâneos em Administração 2.	1

**Tabela 4.3:** Resumo dos resultados (Parte I).

Usuário	Primeira disciplina removida	Segunda disciplina removida	Acertos
31	Introdução à Psicologia.	Tópicos Contemporâneos em Administração 2.	1
32	Gestão de Processos.	Gestão de Desempenho.	2
33	Tópicos Contemporâneos em Finanças.	Criatividade e Inovação nas Organizações.	0
34	Gestão de Processos.	Marketing de Serviços.	1
35	Criatividade e Inovação nas Organizações.	Criação de Negócios.	1
36	Gestão de Responsabilidade Socioambiental.	Gestão da Inovação.	1
37	Comportamento do Consumidor.	Gestão da Inovação.	2
38	Administração Pública e Gestão Social.	Gestão da Inovação.	1
39	Administração da Qualidade.	Comportamento do Consumidor.	2
40	Formação Econômica do Brasil.	Gestão de Desempenho.	1
41	Tópicos Contemporâneos em Administração 2.	Marketing Socialmente Responsável.	1
42	Administração da Qualidade.	Administração de Custos.	1
43	Comportamento do Consumidor.	Tópicos Contemporâneos em Estratégia.	1
44	Legislação Social.	Inovação no setor Público.	0
45	Formação Econômica do Brasil.	Gestão da Inovação.	1
46	Administração da Qualidade.	Administração Pública e Gestão Social.	1
47	Marketing de Relacionamento.	Gestão da Inovação.	1
48	Negociação.	Gestão de Processos.	2
49	Tópicos Contemporâneos em Administração 2.	Inglês Instrumental I.	1
50	Criatividade e Inovação nas Organizações.	Gestão de Responsabilidade Socioambiental.	1

**Tabela 4.4:** Resumo dos resultados parte (Parte II).

#### 4.5.5 AVALIAÇÃO DE RESULTADOS

O primeiro ponto a se observar é a efetividade do S.R. desenvolvido. Como visto nas Tabelas 4.3 e 4.4, o S.R. recomendou ambas as matérias retiradas 25% das vezes, uma das matérias retiradas 57% das vezes e nenhuma das matérias retiradas 18% das vezes, ou seja, o sistema recomendou pelo menos uma das matérias retiradas do histórico escolar dos alunos 82% das vezes que foi executado, conforme observado na Figura 4.9.

Um ponto importante de ser avaliado é que o número de recomendações de uma determinada matéria tende a ser diretamente proporcional ao número de vezes que esta matéria aparecia em históricos escolares, ou seja, quanto mais uma matéria aparecia nos históricos, mais ela era recomendada. Nas Tabelas 4.4 e 4.5 pode-se observar essa relação.

Como pode-se observar nas Tabelas 4.5 e 4.6, as matérias mais recomendadas são exatamente aquelas que mais aparecem no currículo dos discentes. Isso é de se esperar, já que o sistema irá avaliar mais vezes aquela matéria dentro dos currículos, sendo assim, atribuindo uma nota maior a eles quando for realizar a similaridade do cosseno.

Como a abordagem escolhida para a realização deste trabalho foi a filtragem colaborativa baseada no usuário, é de se esperar o viés acima, pois, de uma maneira simplificada, o S.R. baseado no usuário tem a função de recomendar, no caso deste trabalho, matérias que usuários



**Figura 4.9:** Gráfico da distribuição de acertos do S.R.

similares ao discente tenham cursado. Se a maioria dos usuários cursam determinada matéria, é um caminho lógico indicá-la para os usuários que não cursaram ela ainda.

Quanto ao número de históricos, espera-se que com o aumento de informações no banco de dados, o S.R. se torne cada vez mais eficiente. Isso é demonstrado quando o pré-teste é comparado com o teste final, onde no primeiro, haviam 15 históricos escolares e uma taxa de acerto de 73% de pelo menos uma matéria retirada recomendada e já no segundo houve uma taxa de acerto de 82% de pelo menos uma matéria retirada recomendada.

<b>Matéria</b>	<b>Históricos</b>	<b>Recomendada</b>	<b>Retirada</b>
Gestão da Inovação.	36	23	9
Administração da Qualidade.	31	27	4
Comportamento do Consumidor.	31	28	8
Gestão de Desempenho.	31	25	5
Tópicos Contemporâneos em Administração 2.	31	27	8
Gestão de Processos.	25	29	5
Negociação.	22	26	4
Criatividade e Inovação nas Organizações.	21	8	4
Gestão de Responsabilidade Socioambiental.	18	14	7
Inglês Instrumental I.	18	12	7
Tópicos Contemporâneos em Produção, Logística e Gestão da Informação.	18	22	4
Introdução à Psicologia.	17	9	5
Marketing Digital.	17	5	3
Criação de Negócios.	15	0	2
Formação Econômica do Brasil.	13	0	2
Marketing de Relacionamento.	12	0	1
Análise Institucional.	11	0	0
Administração de Custos.	10	0	1
Gestão de Marcas.	10	0	0
Marketing de Serviços.	10	0	2
Mercado Financeiro e de Capitais.	10	0	2
Abordagem Crítica ao Estudo de Organizações.	9	0	3
Marketing Socialmente Responsável.	9	0	1
Economia Brasileira.	8	0	0
Evolução das Ideias Econômicas Sociais.	7	0	1
Gestão de Compras Públicas.	7	0	0
Internacionalização de Empresas e Gestão de Negócios.	7	0	0
Gestão por Competência.	6	0	0
Inovação em Serviços.	6	0	0
Administração Pública e Gestão Social.	5	0	2
Gestão de Projetos.	4	0	0
Fundamentos de Políticas Públicas.	3	0	1
Inovação no Setor Público.	3	0	1
Pesquisa de Marketing.	3	0	0
Tópicos Contemporâneos em Estratégia.	3	0	0

**Tabela 4.5:** Relação entre número de aparecimentos e recomendações parte (Parte I).

<b>Matéria</b>	<b>Históricos</b>	<b>Recomendada</b>	<b>Retirada</b>
Tópicos Contemporâneos em Finanças.	3	0	1
Análise Econômico-Financeira 1.	2	0	0
Avaliação e Monitoramento da Estratégia Organizacional.	2	0	0
Comportamento Organizacional.	2	0	1
Contabilidade Nacional.	2	0	1
Finanças 2.	2	0	2
Governo e Administração no Brasil.	2	0	0
Legislação Social.	2	0	2
Legislação Tributária.	2	0	0
Língua de Sinais Brasileira.	2	0	0
Língua Espanhola 1.	2	0	0
Teoria Política Moderna.	2	0	0
Análise da Liquidez.	1	0	0
Auditoria 1.	1	0	0
Avaliação da Eficiência e Produtividade.	1	0	1
Avaliação de Projetos de Investimento.	1	0	0
Avaliação de Treinamento e de Desenvolvimento.	1	0	0
Comportamento Humano e Trabalho.	1	0	0
Contabilidade Comercial.	1	0	0
Contabilidade Geral 2.	1	0	0
Custos.	1	0	0
Ergonomia.	1	0	0
Estado, Governo e Sociedade.	1	0	0
Fundamentos da Administração Pública.	1	0	0
Introdução à Ciência da Computação.	1	0	0
Introdução à Filosofia.	1	0	0
Introdução à Sociologia.	1	0	0
Legislação Administrativa.	1	0	0
Língua Alemã 1.	1	0	0
Língua Italiana 1.	1	0	0
Programação Fiscal e Financeira.	1	0	0
Sistema de Informações Contábeis.	1	0	0
Tópicos Avançados em Computadores.	1	0	1
Tópicos Contemporâneos em Estratégia e Inovação em Organização.	1	0	0

**Tabela 4.6:** Relação entre número de aparecimentos e recomendações parte (Parte II).



## CAPÍTULO 5

### CONCLUSÃO

“Se queres prever o futuro, estuda o  
passado”  
—Confúcio

A tomada de decisão é um fator presente na vida de todas as pessoas, todos os dias existem em média 221 tipos de tomada de decisão que devem ser feitos por cada um. Atualmente, esse trabalho vem ficando cada vez mais complicado devido a quantidade de informações que chegam a todos por conta da tecnologia. Foi constatado que quanto mais opções são entregues ou apresentadas aos indivíduos, mais complicado e menos prazeroso é o resultado dessa decisão. Tendo esse fator em mente, diversas técnicas na área de *Data Science* vêm sendo desenvolvidas para ajudar as pessoas com suas decisões no seu dia a dia. Uma dessas ferramentas desenvolvidas e amplamente usada pelas maiores redes sociais e sites de vendas é o Sistema de Recomendação - S.R., criado justamente para recomendar os itens que o usuário precisa, sem o mesmo ter que gastar muito do seu tempo procurando ou ponderando entre outros itens.

Pensando em uma realidade mais próxima, os discentes do Departamento de Administração - ADM da Universidade de Brasília - UnB foram escolhidos como objeto de estudo para este trabalho. Todo início de semestre uma das maiores dificuldades desses discentes é a escolha de matérias optativas que combinem com o seu perfil, para isso, esse trabalho teve como objetivo verificar a possibilidade da aplicação de um S.R. baseado no usuário na recomendação de disciplinas. Para entender os S.R's, é preciso voltar para o surgimento da *internet*, onde tudo começou. Em 1969, um conceito inicial do que viria a ser a *internet* foi desenvolvida pelo Ministério de Defesa Americano para proteger suas informações durante a Guerra Fria, a chamada *ARPANet*. Após a liberação dessa rede para o uso público, a *ARPANet* foi sendo desenvolvida até virar o que atualmente é conhecido como *internet*.

Após a criação da *internet*, a cada ano que se passava o número de usuários só crescia, tendo épocas onde a cada três meses esse número dobrava. Com esse crescimento de usuários era inevitável que cada vez mais o número de dados também crescesse. Esse número de dados fez com que um conceito muito usado atualmente surgisse, o *Big Data*, que nada mais é do que conjuntos de dados que não puderam ser percebidos, adquiridos, gerenciados e processados por ferramentas tradicionais de TI e *software/hardware* em um tempo tolerável. Até os dias de hoje,

o conceito de *Big Data* ainda é muito discutido pelos profissionais da área, no entanto, chegou-se a conclusão de que 3 dimensões estão diretamente ligadas a ela, que são: volume (refere-se à quantidade de dados que uma organização ou um indivíduo coleta e/ou gera.), velocidade (refere-se à velocidade com que os dados são gerados e processados) e variedade (refere-se ao número de tipos de dados).

Junto com o *Big Data*, a técnica que desencadeou o estudo em *Data Science* foi desenvolvida, o chamado *Map Reduce*, um modelo de programação e uma implementação associada para processar e gerar grandes conjuntos de dados. O desenvolvimento dessa técnica possibilitou otimizar a indexação e catalogação dos dados sobre as páginas na internet e suas ligações, pois essa técnica permite dividir um grande problema em vários pedaços e distribuí-los em diversos computadores.

Após a criação do *Map Reduce*, veio a criação do que é chamado hoje de *Data Science*, que de acordo com alguns autores, surgiu para extrair informações de dados desorganizados. Para alguns autores, são necessários 3 pilares necessários para uma pessoa ser reconhecida como um cientista de dados, que são: competência significativa, conhecimento matemático e estatístico e habilidade de hacker. Dentre as ferramentas desenvolvidas por essa nova área da ciência, o S.R. é uma delas, e que vem sendo usado amplamente por diversos tipos de serviço e inúmeras aplicações diferentes.

Em 1992 um dos precursores dos S.R's foi desenvolvido, o chamado *Tapestry*, que tinha como objetivo oferecer e-mails de forma mais personalizada para os usuários, pois os mesmos vinham recebendo um grande volume de documentos. Ainda nos anos 90, já começaram a surgir pesquisas em relação as preocupações que esses tipos de sistema poderiam trazer, muitos deles experienciados no desenvolvimento desse trabalho como, por exemplo, o problema da privacidade pessoal, onde dizem que em geral, quanto mais informações os indivíduos tiverem sobre as recomendações, melhor serão capazes de avaliá-las. No entanto, as pessoas podem não querer que seus hábitos ou pontos de vista sejam amplamente conhecidos, trazendo esse aspecto para o ponto de vista do trabalho, muitos discentes não gostariam de expor suas notas.

Mesmo sendo criado há bastante tempo, o *Tapestry* apresentou o que hoje é amplamente conhecido e usado no desenvolvimento de S.R's, que é a filtragem colaborativa, ainda bastante usada nos dias de hoje. Existem inúmeras formas de desenvolver um S.R., seja pelas abordagens mais comuns como a filtragem colaborativa baseada no usuário e a filtragem colaborativa baseada em itens que selecionam primeiro o grupo de usuários de uma amostra que é mais semelhante ao usuário e, em seguida, fornece um grupo de recomendações de elementos que o usuário não tem classificados e que foram classificados como os melhores pelo grupo de usuários com gostos semelhantes ao do usuário, os S.R's baseados em conteúdo, que recomendam um item a um usuário com base em uma descrição do item e um perfil dos interesses do usuário, como também por métodos mais específicos, como S.R's híbridos, que combinam duas ou mais técnicas de recomendação para obter melhor desempenho com menos desvantagens de qualquer um.

Para esse trabalho, a técnica escolhida foi a filtragem colaborativa com base no usuário, justamente por encontrar discentes com perfis parecidos e recomendar matérias semelhantes. Outro ponto que deve ser lembrado, pois é um dos pontos chaves do S.R. é a estratégia utilizada para calcular a similaridade do usuário e dos demais. Alguns exemplos são bastante citados na literatura, como o coeficiente de correlação de *Pearson*, o coeficiente de correlação de *Spearman* e o escolhido para esse trabalho, a similaridade do cosseno, que cria um vetor entre o usuário principal e o resto dos usuários para então verificar sua similaridade.

Quanto ao desenvolvimento, a linguagem escolhida para desenvolver o código foi o *Python*, pois apesar de ser complicado para um iniciante desenvolver do zero, a partir do momento que é compreendida a linguagem ela se torna de simples entendimento, proporciona uma maior velocidade de processamento e sintaxe amigável. Já a IDE utilizada foi o *Spyder* por ser

gratuito e ao mesmo tempo, oferecer todas as funcionalidades integradas em uma interface só, como o editor, o console e o explorador de variáveis. Nessa IDE é possível desenvolver o código e rodar os resultados nele mesmo.

Para esta pesquisa, foi necessário a coleta de 51 históricos escolares, o que demandou certo tempo, pois muitas pessoas não se sentiam confortáveis em compartilhar esse tipo de informação<sup>1</sup>. Além disso, é oportuno destacar que é complicado achar esse número de pessoas que estão próximas a se formar.

A técnica utilizada para a realização dos testes foi coletar esse banco de dados, retirar duas matérias de cada discente, e ao final verificar se o S.R. indicaria pelo menos uma das matérias que foram retiradas de sua lista.

Primeiramente, foi realizado um teste do código para verificar se tudo estava ocorrendo como planejado, para isso, foram utilizados 15 históricos escolares. Ao final do teste, foi constatado que o S.R. estava funcionando da maneira desejada e, como resultado, ele possuía uma taxa de acerto aonde recomendou pelo menos uma das matérias retiradas, 73%. Conseqüentemente quando os 51 históricos foram obtidos, foi realizado o teste final do S.R. desenvolvido, aonde a taxa de acerto foi de 82% das vezes o S.R. indicou pelo menos uma das matérias retiradas de seus históricos escolares.

Esse resultado se mostra bastante positivo, pois além de ser uma alta taxa de acertos, mostra que com o aumento da amostragem também houve um aumento da taxa de acerto, isso pode indicar que conforme o banco de dados for sendo alimentado, o S.R. ficará cada vez mais preciso.

Ao final, os testes foram bastante positivos, mas é importante reforçar que uma das limitações foram os números de históricos obtidos, pois como não havia o acesso ao banco de dados da universidade, o número de históricos se limitou a alunos ainda em processo de formação e no estágio final do curso, e como dito anteriormente, muitos ainda apresentavam algum tipo de resistência para compartilhar esse tipo de informação.

Espera-se que esse S.R. possa ser cada vez mais trabalhado para poder ajudar tanto os discentes na hora de escolher suas matérias optativas, para não precisarem passar tanto tempo pesquisando e em dúvida e, também, a UnB, fazendo com que as pessoas passem menos tempo dentro da plataforma e, dessa forma, diminua as chances de erros no site. Além disso, esse estudo foi feito especialmente no Departamento de Administração - ADM, da referida universidade, porém, com mais alguns estudos e refinamentos no código, o mesmo pode ser estendido para outros departamentos ou até mesmo a universidade toda.

Para trabalhos futuros, foi deixado junto ao código original, no Apêndice, uma etapa adicional na qual é possível realizar a filtragem colaborativa baseada no item (diferente deste trabalho, que utilizou a filtragem colaborativa baseada no usuário). Seria interessante realizar um estudo comparando os resultados de ambos modelos para verificar qual melhor se encaixa na situação demonstrada. Por fim, um ponto para ser comparado entre esses sistemas é a relação mostrada pelas Tabelas 4.4 e 4.5 entre número de aparecimentos em históricos e recomendações, para ver se a abordagem colaborativa baseada no item também é influenciada por esse viés.

Com o avanço da tecnologia a discussão de temas como “as máquinas vão substituir os seres humanos” vêm crescendo cada vez mais, causando até um certo grau de preocupação em algumas pessoas. No caso dos S.R's não acredito que os mesmos possam substituir o trabalho humano, mas sim servir como uma ferramenta no dia a dia visando acelerar certas demandas, no caso deste trabalho, o papel de um coordenador de curso. No caso, uma demanda que ocupa muito tempo dos coordenadores no início de semestre são alunos procurando matérias para suas grades. Com o uso de um S.R., espera-se que essa demanda diminua drasticamente, já que as matéria já vão ser recomendadas para os alunos antes mesmo do semestre começar. Dessa forma,

---

<sup>1</sup>O autor registra aqui seu agradecimento a todos os colegas que disponibilizaram os históricos escolares para a elaboração deste trabalho.

esse S.R. nada mais é do que uma ferramenta, desenvolvida para auxiliar e facilitar o dia a dia dos seus utilizadores.

## CÓDIGO PRINCIPAL

```
1 ##### lista com interesses dos usuarios #####
2
3 users_interests = [Base de Dados]
4
5
6 ### importando funcao "Counter" da biblioteca #####
7
8 from collections import Counter
9
10 ### Criando uma lista com os interesses mais populares ###
11
12 popular_interests = Counter(interest
13                             for users_interests in users_interests
14                             for interest in users_interests).most_common()
15
16 ##### Criando uma função para trazer os interesses mais populares tirando
17 ##### os interesses ja listados por tal usuario #####
18
19 def most_popular_new_interests(users_interests, max_results=5):
20     sugestao = [(interest, frequency)
21                 for interest, frequency in popular_interests
22                 if interest not in users_interests]
23     return sugestao[:max_results]
24
25 ##### definindo a funcao similaridade do cosseno para uso futuro
26
27
28 from numpy import dot
29 import math
30
31 def cosine_similarity(v, w):
```

```

32     return dot(v, w) / math.sqrt(dot(v, v) * dot(w, w))
33
34
35
36
37
38 ##### PARTE 2 - FILTRAGEM COLABORATIVA BASEADA NO USUARIO
39
40         ##### Separando itens da lista
41
42 unique_interests = sorted(list({interest
43                             for users_interests in users_interests
44                             for interest in users_interests}))
45
46         ##### Definindo vetor para cada usuario
47         ##### Listando 1 para interesse possuido e 0 para não
48
49 from typing import List
50
51 def make_user_interest_vector(user_interests: List[str]) -> List[int]:
52
53     return [1 if interest in user_interests else 0
54            for interest in unique_interests]
55
56 user_interest_vectors = [make_user_interest_vector(user_interests)
57                          for user_interests in users_interests]
58
59 user_similarities = [[cosine_similarity(interest_vector_i,
60                                       interest_vector_j)
61                      for interest_vector_j in user_interest_vectors]
62                      for interest_vector_i in user_interest_vectors]
63
64 def most_similar_user_to(user_id):
65     pairs = [(other_user_id, similarity)
66             for other_user_id, similarity in
67             enumerate(user_similarities[user_id])
68             if user_id != other_user_id and similarity > 0]
69     return sorted(pairs,
70                  key=lambda pair: pair[-1],
71                  reverse=True)
72
73 from collections import defaultdict
74
75 def user_based_suggestions(user_id, include_current_interests=False):
76
77     ##### Soma das similaridades
78
79     suggestion = defaultdict(float)
80     for other_user_id, similarity in most_similar_user_to(user_id):

```

```

80     for interest in users_interests[other_user_id]:
81         suggestion[interest] += similarity
82
83         ##### Converter em uma lista ordenada
84
85     suggestion = sorted(suggestion.items(),
86                         key=lambda pair: pair[-1],
87                         reverse=True)
88
89     #e (talvez) exclui interesses ja existentes
90     if include_current_interests:
91         return suggestion
92     else:
93         return [(suggestion, weight)
94                 for suggestion, weight in suggestion
95                 if suggestion not in users_interests[user_id]]

```

## CÓDIGO PARA REMOVER MATÉRIAS E MONTAR LISTAS

---

```

1  import random
2  import copy
3
4      ##### Remover 2 matérias do histórico de cada aluno #####
5
6  def removeTwo(x, i):
7      retiradas = []
8      tamanho = len(x)
9      pos1 = 0
10     pos2 = 0
11     while(pos1 == pos2):
12         pos1 = random.randint(0,tamanho-2)
13         pos2 = random.randint(0, tamanho-2)
14
15     retiradas.append(arrayCopiaMaterias[i][pos1])
16     retiradas.append(arrayCopiaMaterias[i][pos2])
17     arrayRetiradas.append(retiradas)
18
19     del(arrayCopiaMaterias[i][pos1])
20     if pos1 > pos2:
21         del(arrayCopiaMaterias[i][pos2])
22     else:
23         del(arrayCopiaMaterias[i][pos2-1])
24
25 arrayMaterias = #Base de dados completa sem matérias retiradas#
26 arrayRetiradas = []
27 i = 0
28
29     ##### Criar listas para cada discente com as matérias já retiradas #####
30

```

```

31 for x in arrayMaterias:
32     arrayCopiaMaterias = #Base de dados completa sem matérias retiradas#
33
34     removeTwo(x, i)
35     print("\n\nALUNO %d" %i)
36     i = i + 1
37     print('[')
38     k = 0
39     for x in arrayCopiaMaterias:
40         tam = len(x)
41         if k != 50:
42             print("%s," %x)
43             k = k + 1
44         else:
45             print(x)
46
47     print(']')
48
49 print("RETIRADAS")
50 i = 0
51 for x in arrayRetiradas:
52     print('ALUNO %d' %i)
53     i = i+1
54     print(x)
55     print('\n')

```



## REFERÊNCIAS BIBLIOGRÁFICAS

- ADM. (2018). *Proposta de mudança no projeto político pedagógico do curso de administração: Turno noturno* (Tech. Rep. No. 1). Brasília: Departamento de Administração - ADM, da Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas - FACE, da Universidade de Brasília - UnB. (Acessado em 29/06/2020, no site do Departamento de Administração - ADM)
- Afzal, M., Hussain, M., Khan, W. A., Ali, T., Lee, S., Huh, E.-N., ... Hydari, M. A. (2017, March). [Comprehensible knowledge model creation for cancer treatment decision making](#). *Computers in Biology and Medicine*, 82(1), 119–129.
- Ali, A. (2020, September). *Here's What Happens Every Minute on the Internet in 2020* (techreport No. 1). Visual Capitalism.
- Amato, F., Moscato, V., Picariello, A., & Piccialli, F. (2019, April). [SOS: A multimedia Recommender System for Online Social Networks](#). *Future Generation Computer Systems*, 93(1), 914–923.
- Balabanovi, M., & Shoham, Y. (1997, March). Combining content-based and collaborative recommendation. *Communications of the ACM*, 1(1), 9.
- Balabanovic, M. (1997, February). [An adaptive Web page recommendation service](#). *Association for Computing Machinery*, 1(1), 378–385.
- Barcellos, C. D., Musa, D. L., Brandão, A. L., & Warpechowski, M. (2007, December). Sistema de Recomendação Acadêmico para Apoio à Aprendizagem. *Centro Interdisciplinar de Novas Tecnologias na Educação*, 1(1), 10.
- Benabderrahmane, S., Mellouli, N., Lamolle, M., & Paroubek, P. (2017, March). [Smart4Job: A Big Data Framework for Intelligent Job Offers Broadcasting Using Time Series Forecasting and Semantic Classification](#). *Big Data Research*, 7(1), 16–30.
- Billsus, D., & Pazzani, M. (2000, June). [User Modeling for Adaptive News Access](#). *User Modeling and User-Adapted Interaction*, 10(2/3), 147–180.
- Bobadilla, J., Hernando, A., Ortega, F., & Bernal, J. (2011, November). [A framework for collaborative filtering recommender systems](#). *Expert Systems with Applications*, 38(12), 14609–14623.
- Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013, July). [Recommender systems survey](#). *Knowledge-Based Systems*, 46(1), 109–132.
- Burke, R. (2002, November). [Hybrid Recommender Systems: Survey and Experiments](#). *User Modeling and User-Adapted Interaction*, 12(4), 331–370.
- Chen, M., Mao, S., & Liu, Y. (2014, January). [Big Data: A Survey](#). *Mobile Networks and Applications*, 19(2), 171–209.

- Chow, S., & Ruskey, F. (2004). [Drawing Area-Proportional Venn and Euler Diagrams](#). *International Symposium on Graph Drawing*, 2912(1), 466–477.
- Coffman, K. G., & Odlyzko, A. M. (2002). *Growth of the Internet* (4th ed.; R. Gate, Ed.). Minneapolis, Minnesota: Elsevier.
- Conway, D. (2010). *The Data Science Venn Diagram* (techreport No. 1). Drew Conway Data Consulting.
- Cox, M., & Ellsworth, D. (1997, July). *Application-Controlled Demand Paging for Out-of-Core Visualization* (techreport No. 1). NASA Ames Research Center.
- Dean, J., & Ghemawat, S. (2004). *MapReduce: Simplified Data Processing on Large Clusters* (techreport Nos. 1–13). Google, Inc.
- Desjardins, J. (2018, February). *The Rising Speed of Technological Adoption - Acessado em 21/02/2021* (techreport No. 1). Our World In Data.
- Dhar, V. (2013, December). [Data science and prediction](#). *Communications of the ACM*, 56(12), 64–73.
- Dong, M., Zeng, X., Koehl, L., & Zhang, J. (2020, November). [An interactive knowledge-based recommender system for fashion product design in the big data environment](#). *Information Sciences*, 540(1), 469–488.
- Fernandes, N. M. M. C., & Fernandes, W. L. (2014). *Lógica de programação: Pseudocódigo* (1st ed., Vol. 1-69; C. de Autores, Ed.) (No. 1). Mococa: Fundação Biblioteca Nacional. (1)
- Gandomi, A., & Haider, M. (2015, April). [Beyond the hype: Big data concepts, methods, and analytics](#). *International Journal of Information Management*, 35(2), 137–144.
- Garofalakis, M. N., Rastogi, R., Seshadri, S., & Shim, K. (1999). [Data mining and the Web](#). *Proceedings of the Second International Workshop on Web Information and Data Management - WIDM '99*.
- Ghazi, M. R., & Gangodkar, D. (2015). [Hadoop, MapReduce and HDFS: A Developers Perspective](#). *Procedia Computer Science*, 48(1), 45–50.
- Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992, December). [Using collaborative filtering to weave an information tapestry](#). *Communications of the ACM*, 35(12), 61–70.
- Goldman, A., Kon, F., Junior, F. P., Polato, I., & Pereira, R. d. F. (2012, January). [Apache Hadoop: Conceitos teóricos e práticos, evolução e novas possibilidades](#). *Instituto de Matemática e Estatística - Universidade de São Paulo*(1).
- Grus, J. (2015). *Data Science from scratch: First principles with Python* (1st ed., Vol. 1; A. Books, Ed.). Sebastopol, CA: O'Reilly. (315)
- Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999, August). [An Algorithmic Framework for Performing Collaborative Filtering](#). *ACM SIGIR Forum*, 51(2), 227–234.
- Inc., A. (2014, January). *Who is Anaconda?* (techreport). Austin, Texas: Anaconda. (Acessado em 02/03/2021)
- Katarya, R., & Verma, O. P. (2017, July). [An effective collaborative movie recommender system with cuckoo search](#). *Egyptian Informatics Journal*, 18(2), 105–112.
- Kleinknecht, S. W. (2003, July). *Hacking Hackers: Ethnographic Insights into the Hacker Subculture-Definition, Ideology and Argot*. *McMaster University*, 1(1), 200.
- Lai, C.-H., Liu, D.-R., & Liu, M.-L. (2013). [Recommendations Based on Different Aspects of Influences in Social Media](#). *International Conference on Electronic Commerce and Web Technologies*, 1(1), 194–201.
- Lang, S. S. (2006, December). *'Mindless autopilot' drives people to dramatically underestimate how many daily food decisions they make, Cornell study finds* (techreport). Ithaca, Nova York: Cornell University.

- Lee, I. (2017, May). [Big Data: Dimensions, evolution, impacts, and challenges](#). *Business Horizons*, 60(3), 293–303.
- Li, G., Zhang, Z., Wang, L., Chen, Q., & Pan, J. (2017, May). [One-class collaborative filtering based on rating prediction and ranking prediction](#). *Knowledge-Based Systems*, 124(1), 46–54.
- Lins, J. P. S. (2016, February). *Hadoop - MapReduce* (Tech. Rep. No. 1). Centro de Informática da Universidade Federal de Pernambuco.
- Ma, X., Ma, J., Li, H., Jiang, Q., & Gao, S. (2018, February). [ARMOR: A trust-based privacy-preserving framework for decentralized friend recommendation in online social networks](#). *Future Generation Computer Systems*, 79(1), 82–94.
- MacKenzie, I., Meyer, C., & Noble, S. (2013, October). *How retailers can keep up with consumers* (techreport No. 1). McKinsey & Company.
- Mao, X., Zhao, X., Lin, J., & Herrera-Viedma, E. (2019, February). [Utilizing multi-source data in popularity prediction for shop-type recommendation](#). *Knowledge-Based Systems*, 165(1), 253–267.
- Marty-Dugas, J., & Smilek, D. (2020, October). [The relations between smartphone use, mood, and flow experience](#). *Personality and Individual Differences*, 164(1).
- Massai, L., Nesi, P., & Pantaleo, G. (2019, January). [PAVAL: A location-aware virtual personal assistant for retrieving geolocated points of interest and location-based services](#). *Engineering Applications of Artificial Intelligence*, 77(1), 70–85.
- Miranda, T., Claypool, M., Claypool, M., Gokhale, A., Gokhale, A., Mir, T., ... Sartin, M. (1999, June). Combining Content-Based and Collaborative Filters in an Online Newspaper. *Worcester Polytechnic Institute*, 1(1), 1–12.
- MM, M. (2012, September). [Statistics corner: A guide to appropriate use of correlation coefficient in medical research](#). *Malawi Medical Journal: The Journal of Medical Association in Malawi*, 24(3), 69–71.
- Monteiro, L. (2001, Setembro). A Internet como meio de comunicação: Possibilidades e Limitações. *INTERCOM – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação*, 1(1), 27–97.
- Newham, C. (2005). *Learning the bash Shell* (1st ed., Vol. 1-333) (No. 1). O'Reilly Media. (1)
- Oliveira, E. C. D. (2020). Sobre a clássica função de Mittag-Leffler. *Revista Matemática Universitária*, 1(1), 1–25.
- Orlowski, J. (2020, September). *The Social Dilemma* (audiocd No. 1). The social dilema, LLC.
- OWD. (2019, January). *Daily hours spent with digital media, United States, 2008 to 2018* (techreport No. 1). Our World in Data (OWD). (Acessado em 27/09/2020)
- Pazzani, M. J. (1999, December). [A Framework for Collaborative, Content-Based and Demographic Filtering](#). *Artificial Intelligence Review*, 13(5/6), 393–408.
- Pazzani, M. J., & Billsus, D. (2007). *The Adaptive Web* (Vol. 4321). Springer-Verlag GmbH.
- Penrose, R. (2008). *The Road to Reality* (Vol. 59-61; Vintage Books USA, Ed.) (No. 1). Random House UK Ltd.
- Pereira, N., & Varma, S. (2016, January). Survey on Content Based Recommendation System. In *International Journal of Computer Science and Information Technologies* (Ed.), (Vol. 7, pp. 281–284).
- Petri, A. (2013, May). *O berço do Big Data* (techreport No. 2348). Revista Veja.
- Raybaut, P. (2009, January). *Spyder* (techreport). Spyder - The Scientific Python Development Environment. (Acessado em 03/03/2021)
- Resnick, P., & Varian, H. R. (1997, March). [Recommender systems](#). *Communications of the ACM*, 40(3), 56–58.

- Ricci, F., Rokach, L., & Shapira, B. (2010). *Introduction to Recommender Systems Handbook* (1st ed., Vol. 1-35) (No. 1). Springer, Boston, MA: Springer US. (33)
- Rossum, G. V. (2003). *An introduction to Python* (1st ed., Vol. 1-115). Network Theory Ltd. Bristol.
- Schafer, J. B., Konstan, J., & Riedi, J. (1999). [Recommender systems in e-commerce](#).
- Schwartz, B. (2009). *The Paradox of Choice* (1st ed., Vol. 1; H. Perennial, Ed.) (No. 304). Nova York - NY: Harper Perennial.
- Seo, Y.-D., Kim, Y.-G., Lee, E., & Baik, D.-K. (2017, March). [Personalized recommender system based on friendship strength in social network services](#). *Expert Systems with Applications*, 69(1), 135–148.
- Silva, L. A. (2003, Outubro). *O movimento do código aberto* (techreport). Viva o Linux. (Acessado em 03/03/2021)
- Silva, L. W. (2001, Agosto). *Internet foi criada em 1969 com o nome de “Arpanet” nos EUA* (techreport). São Paulo, Brasil: Folha de São Paulo.
- Slovic, P., Lichtenstein, Sarah, & Fischhoff, B. (1998). Decision Making. *Stevens Handbook of Experimental Psychology, Wiley*.
- Smyth, B., & Cotter, P. (2000, April). [A personalised TV listings service for the digital TV age](#). *Knowledge-Based Systems*, 13(2–3), 53–59.
- Spearman, C. (1904, January). [The Proof and Measurement of Association between Two Things](#). *The American Journal of Psychology*, 15(1), 72.
- Su, K., Xiao, B., Liu, B., Zhang, H., & Zhang, Z. (2017, January). [TAP: A personalized trust-aware QoS prediction approach for web service recommendation](#). *Knowledge-Based Systems*, 115(1), 55–65.
- Suryakant, & Mahara, T. (2016). [A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment](#). *Procedia Computer Science*, 89(1), 450–456.
- TAF. (2012). *Demystifying Big Data - A practical guide to transforming the business of government* (techreport No. 39). 601 Pennsylvania Avenue, N.W. North Building, Suite 600 Washington, D.C.: TechAmerica Foundation's - TAF.
- Tian, Y., Zheng, B., Wang, Y., Zhang, Y., & Wu, Q. (2019). [College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm](#). *Procedia CIRP*, 83(1), 490–494.
- Towle, B., & Quinn, C. (2000, January). Knowledge Based Recommender Systems Using Explicit User Models. *Proceedings of the AAAI Workshop on Knowledge-Based Electronic Markets*, 1(1), 74–77.
- Tran, T., & Cohen, R. (2000, January). Hybrid recommender systems for electronic commerce. *Knowledge-Based Electronic Markets, Papers from the AAAI Workshop*, 40(1), 78–84.
- Vicentini, L., Lanzoni, E., Franzotti, V., & Yonenaga, W. (2005, Setembro). [Introdução da tecnologia de voz sobre IP em redes corporativas](#). *Congresso Brasileiro de Ensino de Engenharia*, 1(1), 8.
- Vieira, P. V., Raabe, A. L. A., & Zeferino, C. A. (2009, January). Bipide: Ambiente de Desenvolvimento Integrado para Utilização dos Processadores BIP no Ensino de Programação. *Simpósio Brasileiro de Informática na Educação*, 1(1), 9.
- Viens, A. (2019, September). *Visualizing Social Media Use by Generation* (techreport No. 1). Visual Capitalist. (Acessado em 04/10/2020 em <https://www.visualcapitalist.com/visualizing-social-media-use-by-generation/>)
- Winnick, M. (2016, June). *Putting a Finger on Our Phone Obsession, Mobile touches: A study on how humans use technology* (techreport No. 1). Dscout.

- Yang, J., Wang, H., Lv, Z., Wei, W., Song, H., Erol-Kantarci, M., ... He, S. (2017, May). [Multimedia recommendation and transmission system based on cloud platform](#). *Future Generation Computer Systems*, 70(1), 94–103.
- Yang, S., Korayem, M., AlJadda, K., Grainger, T., & Natarajan, S. (2017, November). [Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach](#). *Knowledge-Based Systems*, 136(1), 37–45.
- Ylijoki, O., & Porras, J. (2016, May). [Perspectives to Definition of Big Data: A Mapping Study and Discussion](#). *Journal of Innovation Management*, 4(1), 69–91.

