



Machine Learning para Sistemas de Recomendação

Gabriel Moreira • Lead Data Scientist • gabrielpm@ciandt.com

sli.do: #ciandttechsummit

Gabriel Moreira

@gspmoreira

- Lead Data Scientist na CI&T
- Aluno de Doutorado no ITA - pesquisa com ênfase em *Deep Learning* e Sistemas de Recomendação
- Mais de quinze anos de experiência em desenvolvimento de software

<https://about.me/gspmoreira>





A vida é curta!



Recomendações sociais



Recomendações por interação





“Muitas vezes, as pessoas não sabem o que elas querem até que você mostre a elas.”



*Steve Jobs
Apple*

“Estamos saindo da Era da Informação e entrando na Era da Recomendação.”



Chris Anderson
Autor do Livro “A cauda longa”

Sistemas de recomendação são responsáveis por...



70% da home page da Amazon é dedicada a recomendações



$\frac{2}{3}$ dos filmes assistidos

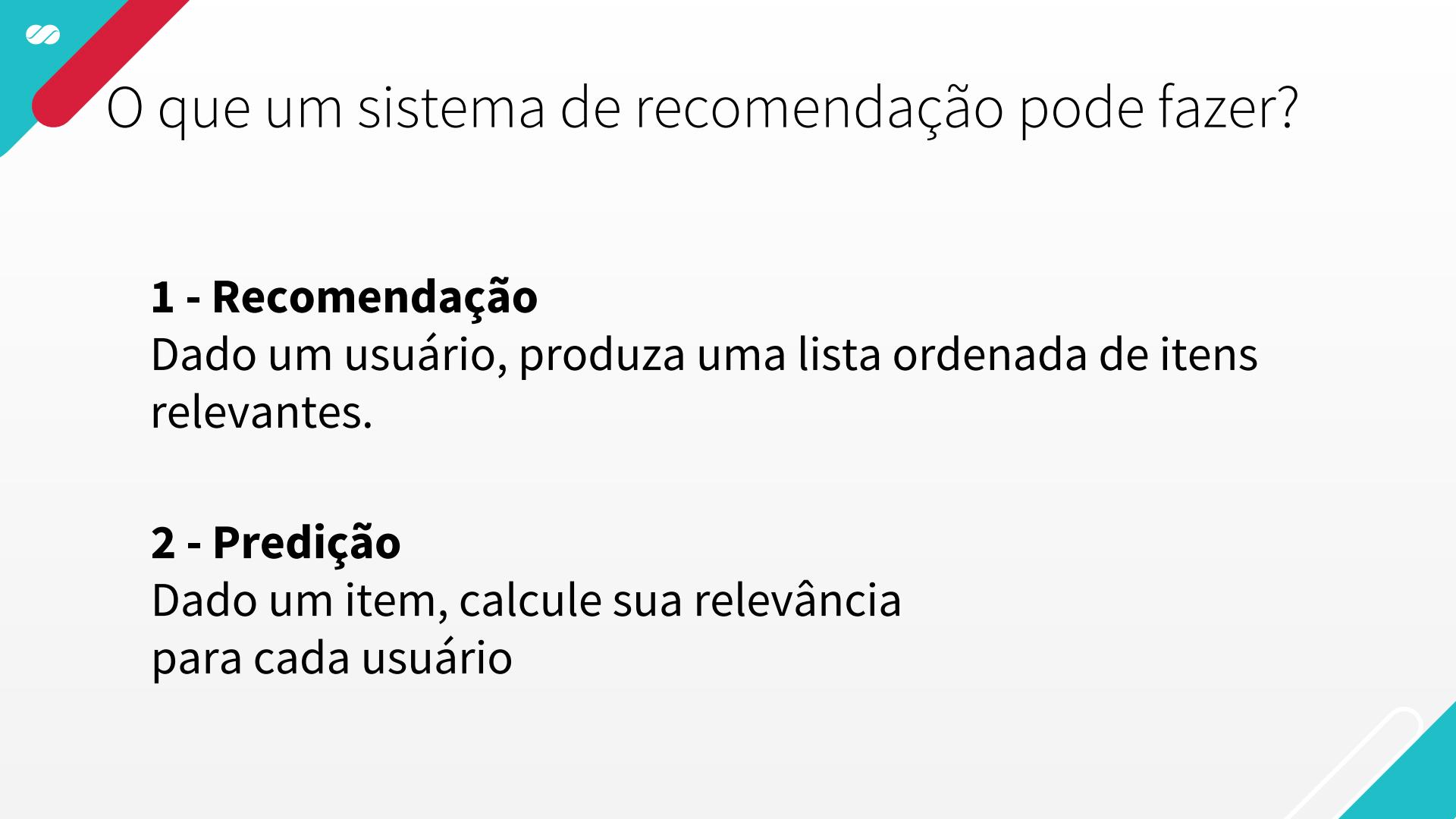


Recomendações geram 38% mais clicks



O que mais eu posso recomendar?

products
tags
professionals
courses
musics movies
jobs books
papers girlfriends
investiments restaurants
 videos
dressing



O que um sistema de recomendação pode fazer?

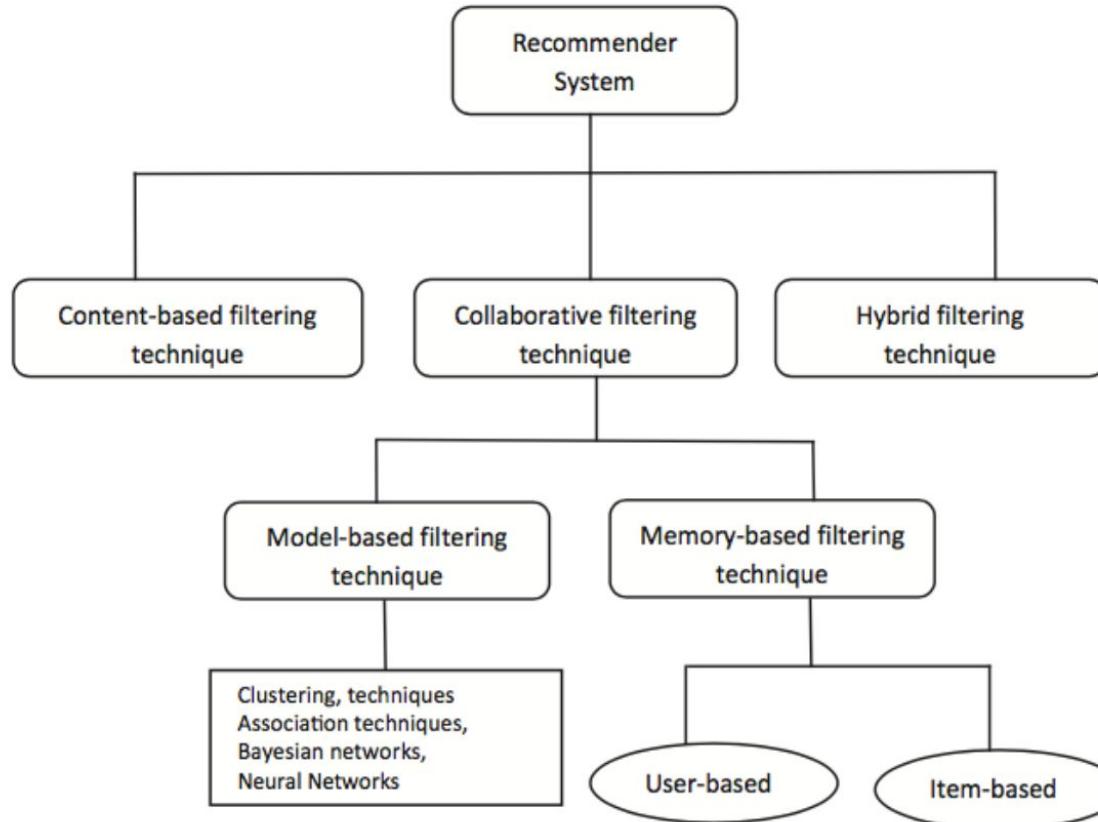
1 - Recomendação

Dado um usuário, produza uma lista ordenada de itens relevantes.

2 - Predição

Dado um item, calcule sua relevância para cada usuário

Abordagens de recomendação tradicionais



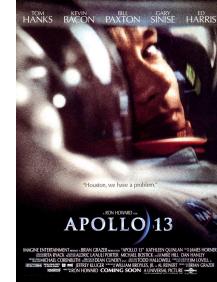
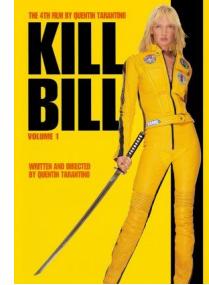
Fonte: "Recommendation systems: Principles, methods and evaluation"



Filtragem Colaborativa

Filtragem Colaborativa baseada em Memória

Similaridade entre Usuários (k -nearest neighbors)



Gosta

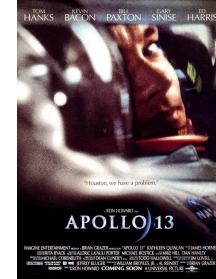
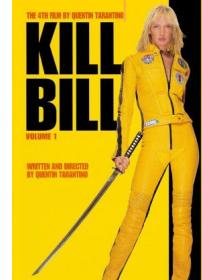
Recomenda



Interesses similares

Filtragem Colaborativa baseada em Memória

Similaridade entre Itens (*k-nearest neighbors*)



Quem gosta de A também gosta de B

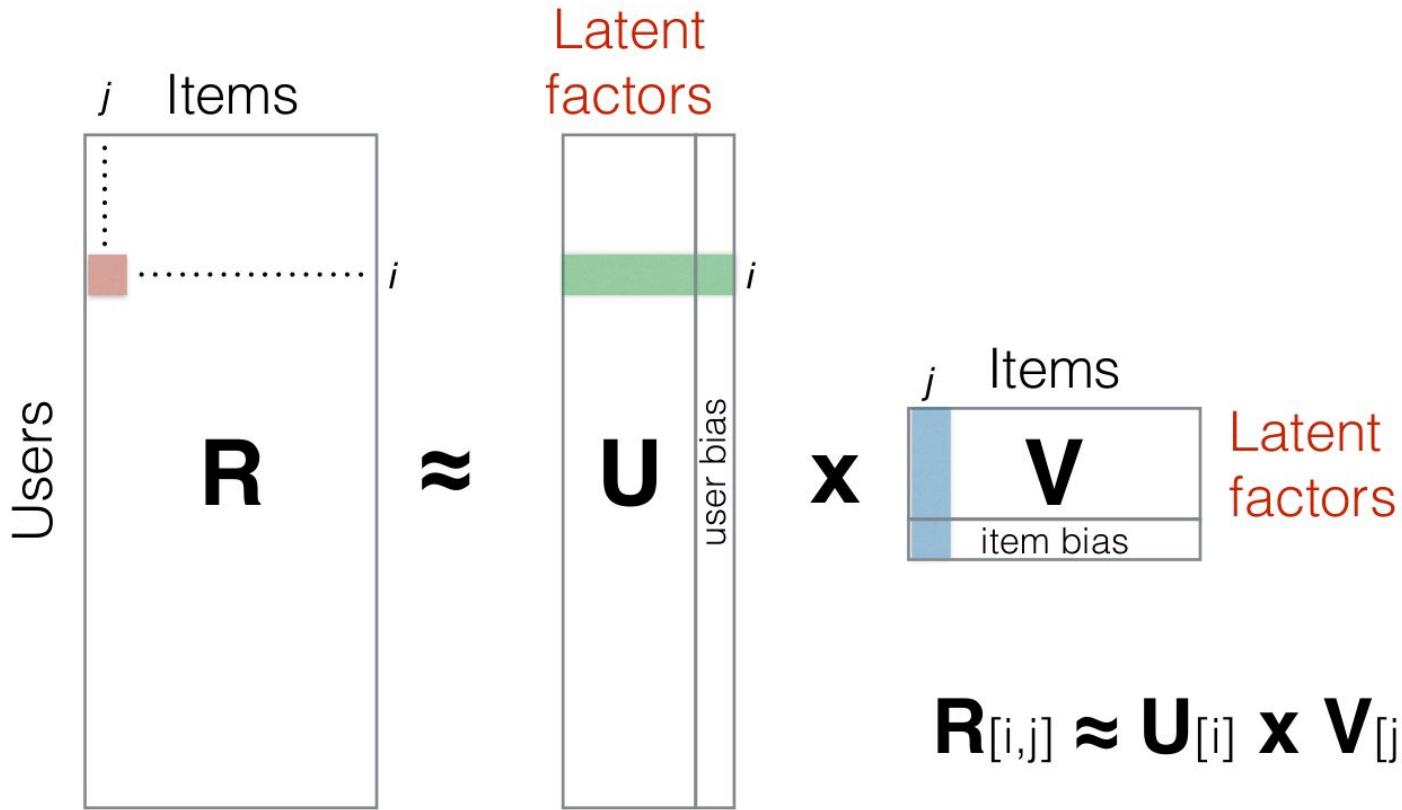
Gosta

Gosta

Recomenda



Filtragem Colaborativa baseada em Fatoração de Matrizes





Filtragem Colaborativa

Vantagens

- Funciona para qualquer tipo de item, pois ignora atributos

Desvantagens

- Precisa de um mínimo de usuários e itens para calcular similaridades
- Não pode recomendar itens que ainda não foram consumidos pelos usuários

Frameworks de Filtragem Colaborativa



Java



Python / Scala



Python



Java



.NET

Exemplo de Filtragem Colaborativa usando Mahout

User,Item,Rating1,

15,4.0
1,16,5.0
1,17,1.0
1,18,5.0
2,10,1.0
2,11,2.0
2,15,5.0
2,16,4.5
2,17,1.0
2,18,5.0
3,11,2.5

input.csv

```
1 // Loads user-item ratings
2 DataModel model = new FileDataModel(new File("input.csv"));
3 // Defines a similarity metric to compare users (Person's correlation coefficient)
4 UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
5 // Threshold the minimum similarity to consider two users similar
6 UserNeighborhood neighborhood = new ThresholdUserNeighborhood(0.1, similarity,
7 model);
8 // Create a User-Based Collaborative Filtering recommender
9 UserBasedRecommender recommender = new
10 GenericUserBasedRecommender(model, neighborhood, similarity);
11 // Return the top 3 recommendations for userId=2
12 List recommendations = recommender.recommend(2, 3);
```

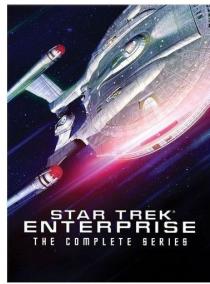


Filtragem baseada em Conteúdo



Filtragem baseada em Conteúdo

Conteúdo similar (ex: atores)



Gosta

Não gosta

Recomenda



Filtragem Baseada em Conteúdo

Vantagens

- Não depende de outros usuários
- Podem recomendar itens novos
- Recomendações são facilmente explicáveis

Desvantagens

- Super-especialização
- Extrair atributos de áudio, filmes e imagens não é trivial

Vetorização de Texto

Representa cada documento como um vetor de *features*. Cada posição no vetor representa a relevância de uma palavra (token) no documento.

- *BoW (Bag of words)*
- *TF-IDF (Term Frequency - Inverse Document Frequency)*

	D1	D2	D3	D4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

Matriz Documento-Token - *Bag of Words*

Vetorização de Texto

TF-IDF - Termos mais “relevantes” em um documento consistem em termos frequentes em um documento que são raros em outros documentos.

f_{ij} = frequency of term (feature) i in doc (item) j

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

n_i = number of docs that mention term i

N = total number of docs

$$IDF_i = \log \frac{N}{n_i}$$

TF-IDF score: $w_{ij} = TF_{ij} \times IDF_i$

Exemplo de TF-IDF com scikit-learn

```
from sklearn.feature_extraction.text import TfidfVectorizer  
vectorizer = TfidfVectorizer(max_df=0.5, max_features=1000,  
                            min_df=2, stop_words='portuguese')  
tfidf_corpus = vectorizer.fit_transform(text_corpus)
```

		<i>tokens</i>										
		face	pessoa	guia	faca	gato	peixe	durma	micro	festa	cento	...
		0	1	2	3	4	5	6	7	8	9	
documentos		D1		0.05					0.25			
		D2	0.02			0.32				0.45		
		...										

Exemplo de uma matriz esparsa de TF-IDF

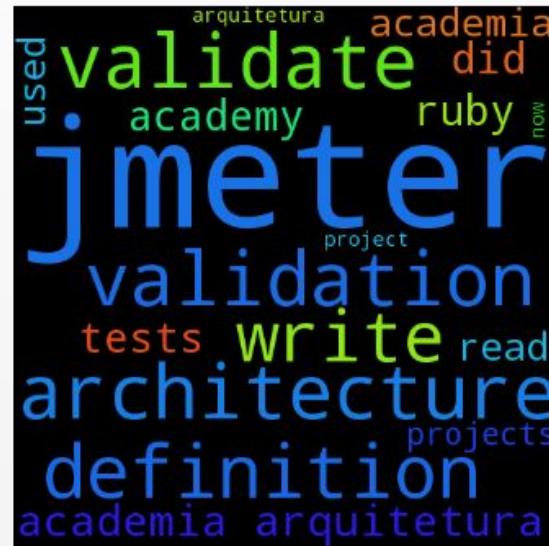
Vetorização de Texto

Exemplo:

“Did you ever wonder how great it would be if you could **write** your **jmeter tests** in **ruby**? This project aims to do so. If you use it on your project just let me know. On the Architecture Academy you can read how jmeter can be used to **validate** your Architecture. **definition** | architecture **validation** | **academia de arquitetura**”

Termos relevantes (TF-IDF)

Termos (<i>unigrams e bigrams</i>)	peso
jmeter	0.466
architecture	0.380
validate	0.243
validation	0.242
definition	0.239
write	0.225
academia arquitetura	0.218
academy	0.216
ruby	0.213
tests	0.209



Vetorização de Texto

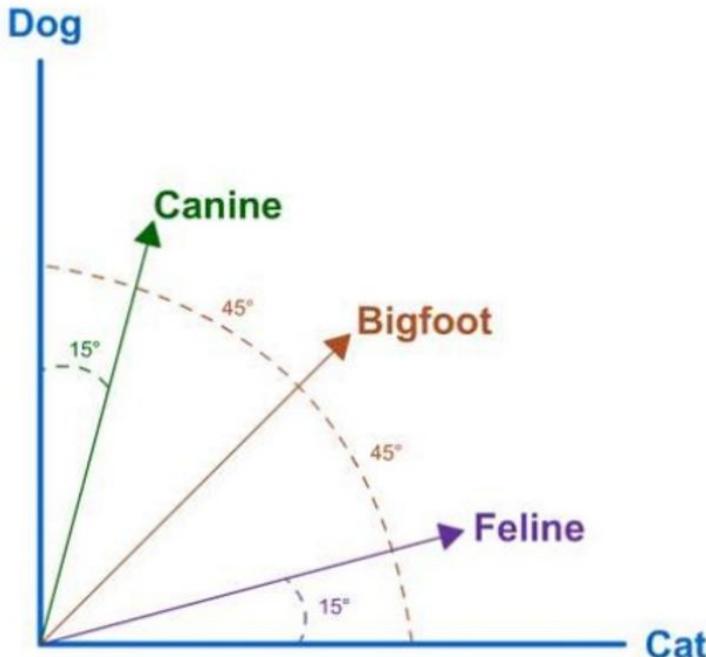
Visualização da média dos vetores TF-IDF dos posts no Google+ for Work (CI&T)



Mais detalhes sobre social e textual analytics em http://bit.ly/python4ds_nb

Cosine Similarity

Métrica de similaridade calculada como o cosseno do ângulo entre os vetores



```
from sklearn.metrics.pairwise import cosine_similarity  
cosine_similarity(tfidf_matrix[0:1], tfidf_matrix)
```

Exemplo com scikit-learn

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos \theta$$

$$\cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Similaridade entre pessoas

GABRIEL MOREIRA INTERESTS



Gabriel Moreira

Product Engineer Master
gabrielpm@ciandt.com

Activities on Google+

Analysed Posts: **42**

Average interactions on Gabriel Moreira posts: **3.5**

Main terms* in G+ (posts, comments and +1s)



* Click in the terms to search for people interested in them.

CI&T people with similar interests



Gilmar Souza (40%) - Posts



Henrique Souza (28%) - Posts



Mars Cyrillo (26%) - Posts



Johann Vivot (24%) - Posts



Fabio Fogliarini Brolesi (24%) - Posts



Rubens Barreto (23%) - Posts



Fulvio P. Parmejani (21%) - Posts



Fabio da Silva Santos (20%) - Posts



Lucas Arruda (20%) - Posts



Mauricio Pedroso (19%) - Posts



Felipe Antonio Souza Dewulf (19%) - Posts

Exemplo de tokens relevantes de uma pessoa e similaridade com outras pessoas



Filtragem Híbrida

Sistemas de Recomendação Híbridos

Algumas abordagens...

Composto

Agrega recomendações de uma cadeia de algoritmos.

Ponderado

Cada algoritmo tem um peso e a ordenação final das recomendações é definida por uma média ponderada

Integrados

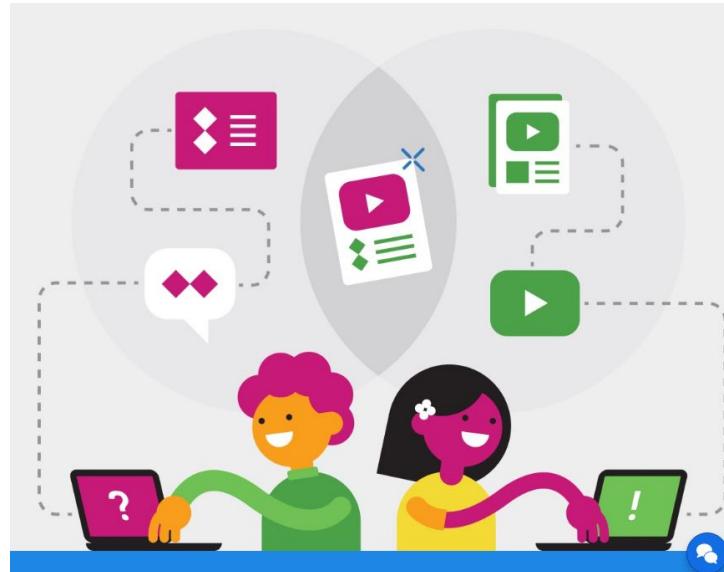
Filtragem Colaborativa e Filtragem Baseada em Conteúdo integradas em uma função objetivo (loss function). Ex:

- Matrix Factorization + Topic Modeling
- Multi-view Deep neural networks to model user interactions and textual items content (word embeddings)

Case

Smart Canvas[©]

Colaboração Corporativa



<http://www.smartcanvas.com/>

Recomendação de conteúdo

Smart Canvas Discover

Suggestions based on what people like you posted or read

Google Sheets ✓ ...
[OKR] D1 Q2/2016 edited 8 days ago

Google ✓ ...
Research at Google

VentureBeat ✓ ...
How Google's AI paved the way for the next generation of bots

YouTube ✓ ...
SpaceX Falcon 9 - Successful Drone Ship Landing - 8th April 2016

Suggestions based on your reads about machine learning, google, ci&t

Google Cloud Platform... ✓ ...
Google takes Cloud Machine Learning service mainstream

Google ✓ ...
Google: Machine Learning For Spam Detection & Search Quality Is Coming

TechCrunch ✓ ...
Google launches new machine learning platform

Tech Insider ✓ ...
'Machine learning' is a revolution as big as the internet or personal computers

Suggestions based on your reads about smart canvas

TN 2016 Goals: SAAS users
TN 2016 Goals: SAAS users

Smart Canvas Backlog and Ideas
Smart Canvas Backlog and Ideas edited on Feb 28

Improving the Google Drive for Work experience - Accept and

Smart Canvas Release Notes
Mars edited on Mar 9

+

Tópicos de interesse



Gabriel Moreira (gabrielpm)

#algorithm #computer #python #machine_learning

✉ gabrielpm@ciandt.com

LESS ^

Busca por pessoas interessadas ou experts

← machine learning |  

 Machine Learning
 Language Learning
 Bayesian machine learning
 Add insights with machine learning
 The Machine Learning GCP Spectrum
 Gabriel Moreira (gabrielpm) - gabrielpm@ciandt.com
 Gilmar Souza (gilmarj) - Head of Machine Learning - +55 19 33333333 - gilmarj@ciandt.com
 Machine Learning Practitioners

Pessoas similares

← People related



People that are also interested in **artificial intelligence, google**



Mars Cyrillo

VP Products and Cognitive Computing



Eduardo Sangion

Product Planner @D1



Lucas Persona

Chief Digital Evangelist

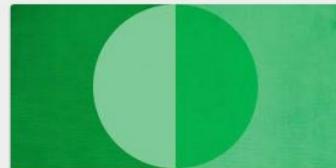


Daniel Viveiros

Head of Product Engineering



People that are also interested in **machine learning**



Gabriel Moreira



Mars Cyrillo

VP Products and Cognitive Computing



Lucas Arruda

Software Architect



Cesar Gon

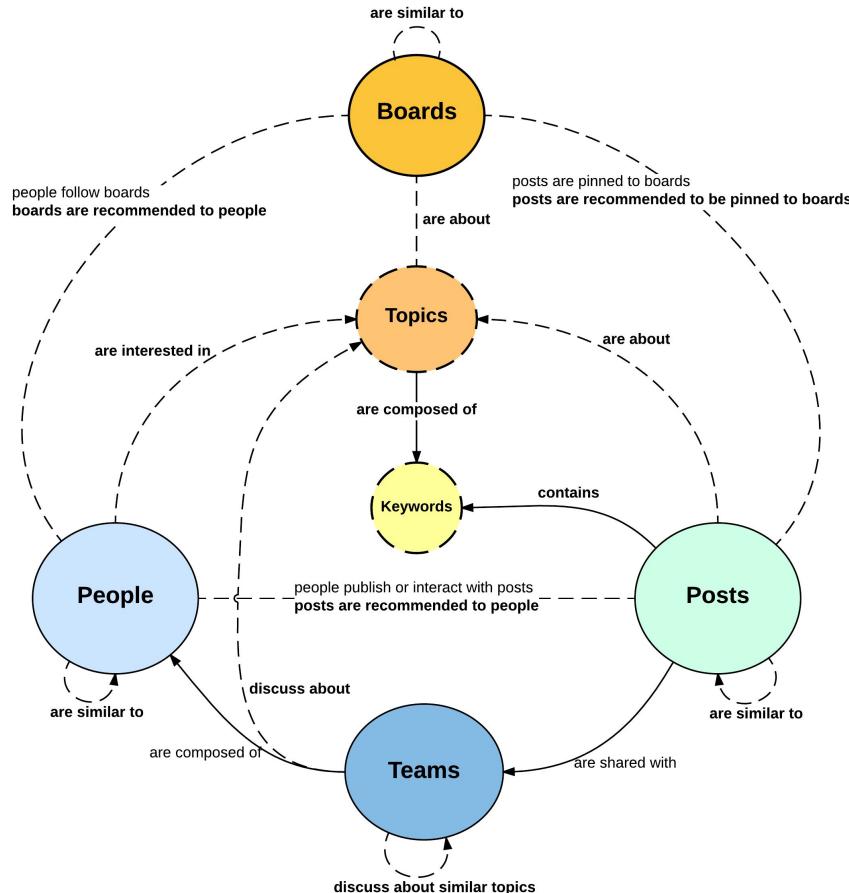
CEO





Como funciona?

Modelo de Grafo



Smart Canvas © Graph Model:
Linhas tracejadas e rótulos em
negrito são os relacionamentos
inferidos por Machine Learning,
usando técnicas de Sistemas de
Recomendação e Modelagem de
Tópicos

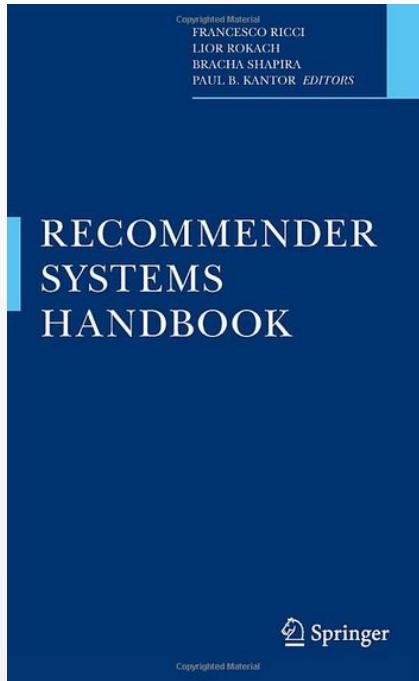
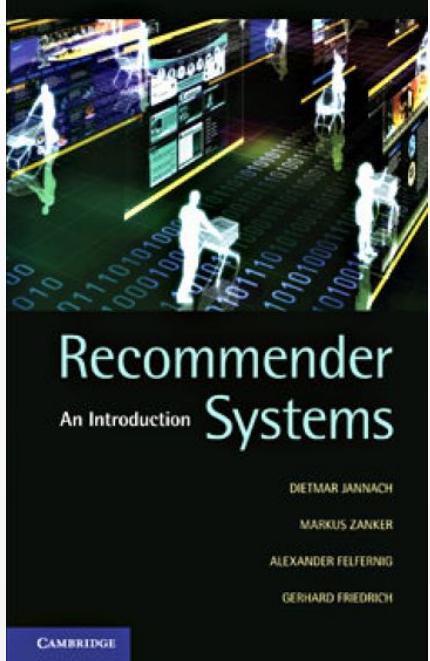
Dataset do CI&T Deskdrop no Kaggle!

The screenshot shows the Kaggle dataset page for 'Articles sharing and reading from CI&T DeskDrop'. The top navigation bar includes 'Search kaggle', 'Competitions', 'Datasets', 'Kernels', 'Discussion', 'Jobs', and a user profile icon. The main title 'Articles sharing and reading from CI&T DeskDrop' is displayed with a subtitle 'Logs of users interactions on shared articles for content Recommender Systems'. Below the title, it shows 'Gabriel Moreira · last updated 2 months ago'. The page features tabs for 'Overview' (highlighted in blue), 'Data', 'Kernels', 'Discussion', 'Activity', and 'Settings'. A 'Download (9 MB)' button and a 'New Kernel' button are also present. The 'Tags' section includes 'internet', 'human-computer interaction', 'web sites', 'medium', and 'featured', with an 'Add Tag' button. The 'Top Contributors' section lists 'Gabriel Moreira' as 1st contributor. The 'Kernels' section, highlighted with a red box, contains three entries: 'DeskDrop Articles Topic Mod...' (6 votes, run 2 months ago), 'DeskDrop datasets EDA' (3 votes, run 2 months ago), and 'Recommender Systems in Py...' (2 votes, run a month ago). The 'Discussion' section indicates 'There are no conversations yet.' with a 'Start one' button. The 'Description' and 'Context' sections provide additional details about the dataset.

- 12 meses de logs
(Mar. 2016 - Feb. 2017)
- ~ 73k interações de usuários
- ~ 3k artigos públicos
compartilhados na plataforma

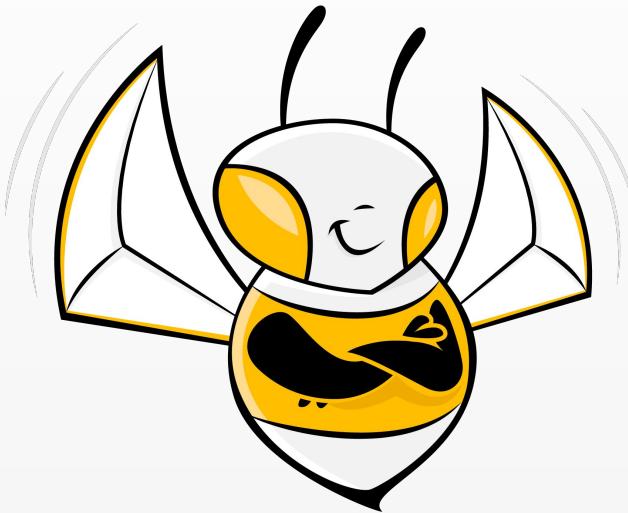
bit.ly/dataset4recsys

Referências



The ACM Conference Series on
Recommender Systems

recsys.acm.org/



MACHINE
LEARNING

C I & T

Perguntas?

Gabriel Moreira
Lead Data Scientist
gabrielpm@ciandt.com
[@gspmoreira](https://twitter.com/gspmoreira)



Obrigado!

