

INF 493/**792**

Tópicos especiais III

Introdução à mineração de dados

Aula I: Regras do jogo e introdução

Professor

Giovanni Comarela

CCE 303b

e-mail: gcom@ufv.br (gcom.ufv@gmail.com)

Áreas de Interesse:

- Análise e medição em redes de computadores
- Mineração de dados
- Estatística
- Estudo de redes sociais (redes complexas)
- Algoritmos
- Ciência da Computação

Disciplina

Aulas

Segunda-feira: 10:00 às 12:00

Quinta-feira: 8:00 às 10:00

Local: PVA 254

PVAnet

Slides

Calendário

Recursos

Tutoriais

Dúvidas (dúvidas não serão elucidadas por e-mail – **Piazza**)

Pré-requisitos

Probabilidade e estatística

Álgebra linear

Algoritmos

Presença

Obrigatória

Sobre as aulas

Em maioria teóricas

Slides + quadro

- Slides estarão no PVAnet
- Notas sobre material no quadro é de responsabilidade do aluno
- Sugestão: cada aula tem um escriba. Notas podem publicadas no PVAnet

Avaliações

Duas provas teóricas (50%)

Trabalho prático (40%)

Listas de exercícios (10%)

Nota final será acrescida de uma constante C

Como serão os projetos?

Vai depender do número e interesse dos alunos matriculados... Mais notícias em breve.

Posso gravar a aula?

SIM! Para uso pessoal e desde que não perturbe o andamento da aula

Vocês **NÃO** têm minha autorização para distribuir ou publicar material gravado em aula

Ambiente interativo

Dúvidas durante as aulas são bem-vindas

Ambiente inclusivo e participativo

Intolerâncias não serão toleradas!

- Sexismo
- Racismo
- Homofobia

Em resumo: piadas e comentários que tenham poder exclusivo devem ser evitados

Preciso saber Inglês?

Inglês é a língua franca da Ciência

Habilidade de **ler** textos científicos na língua inglesa é
requerida

É encorajado que listas e relatórios sejam escritos em
inglês (não obrigatório)

Plágio

Apenas um comentário

Não paguem pra ver!

A punição será máxima, dentro das regras da UFV

Alunos da pós?

Livros

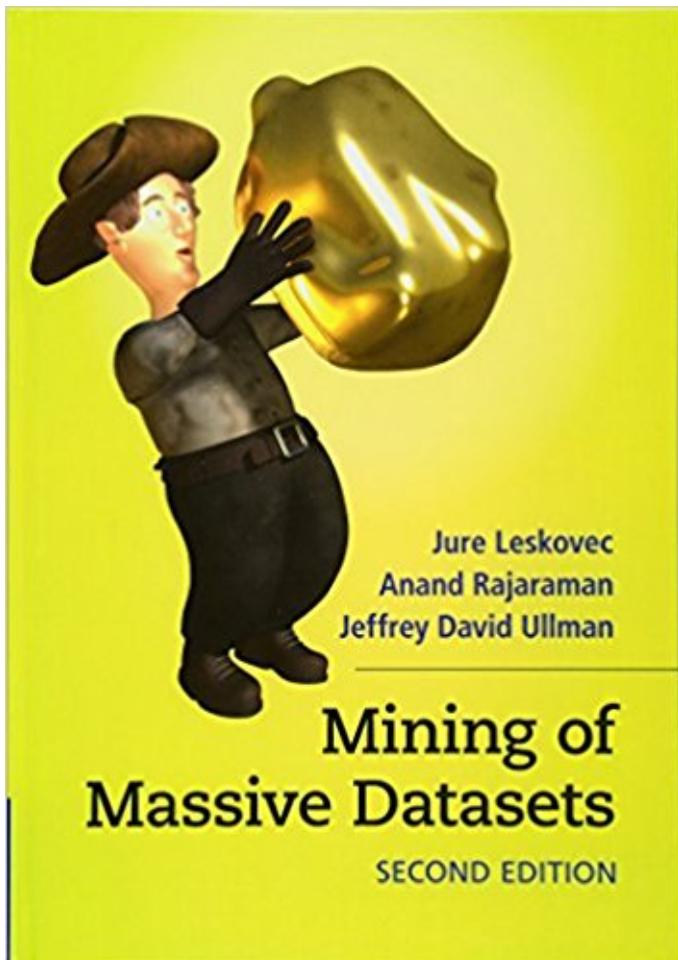
Aulas serão baseadas em vários livros

A compra dos livros não é obrigatória

Vários estão disponíveis na Web gratuitamente

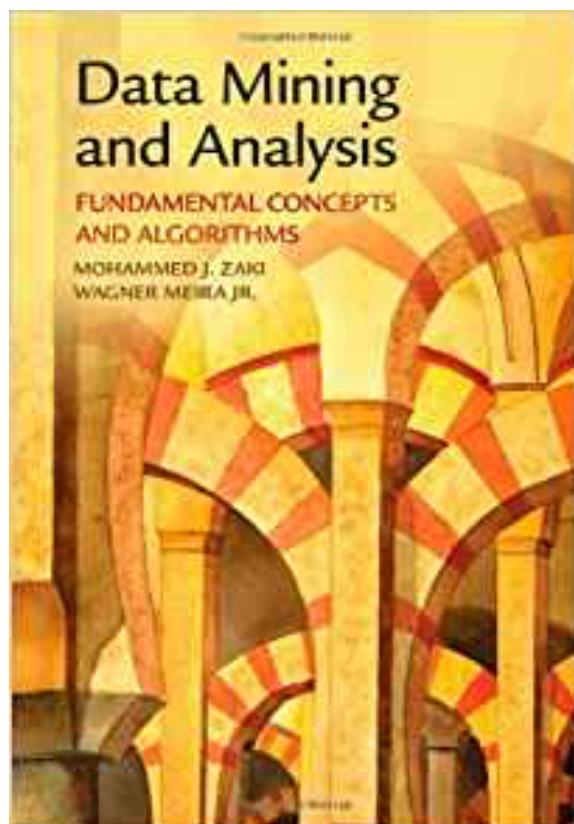
- Incluindo os slides (que serão reusados nessa disciplina)

Mining of Massive Datasets

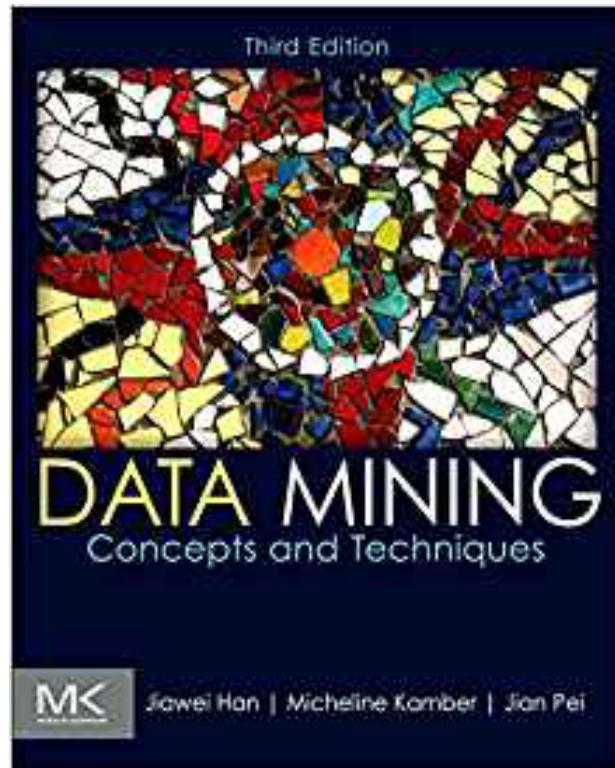


<http://www.mmds.org>

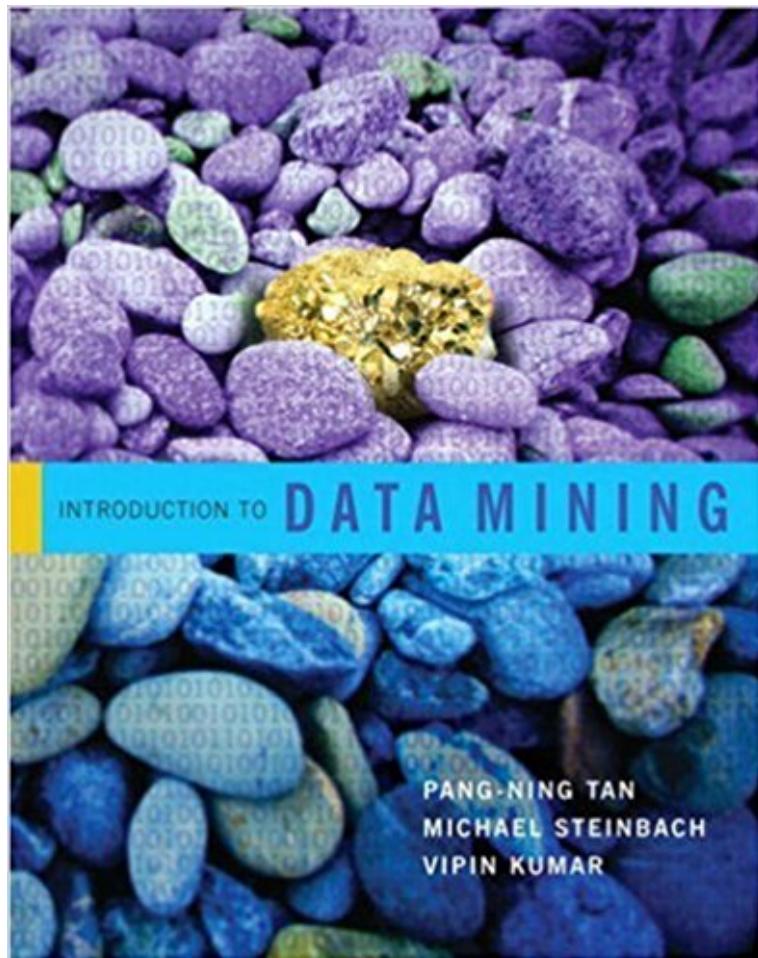
Data Mining and Analysis: Fundamental concepts and Algorithms



Data Mining: Concepts and techniques



Introduction to data mining



O que é mineração de dados?

Interseção com várias outras disciplinas

- Estatística
- Aprendizado de Máquina
- Inteligência Artificial
- Banco de dados
- Processamento paralelo e distribuído
- Teoria da computação
- Algoritmos

O que é mineração de dados?

Dada uma grande quantidade de dados

O objetivo é descobrir padrões e modelos que:

Válidos: se aplicam a novas observações

Úteis: tenha algum significado ou utilidade

Inesperados: não sejam óbvios

Entendíveis: podem ser interpretados por humanos

O que é mineração de dados?

Métodos descritivos

- Encontrar padrões que descrevam os dados
 - Exemplo: agrupamento

Métodos preditivos

- Usar certas variáveis para prever valores não conhecidos ou futuros de outras variáveis
 - Exemplo: Recomendação e Classificação

(exemplos no notebook)

Temas principais

- 1 – Representação de dados,
- 2 – Funções de distância e métricas
- 3 – Busca de objetos similares
- 4 – Redução de dimensionalidade
- 5 – Busca de elementos relevantes em grafos
- 6 – Algoritmos de agrupamento
- 7 – Algoritmos de classificação
- 8 – Reconhecimento de padrões frequentes
- 9 – Sistemas de recomendação

Havendo tempo, outros temas serão incluídos
Sugestões serão bem-vindas

Cronograma

(disponível no PVAnet)

Recomendação

Secure | https://www.netflix.com/browse/genre/83

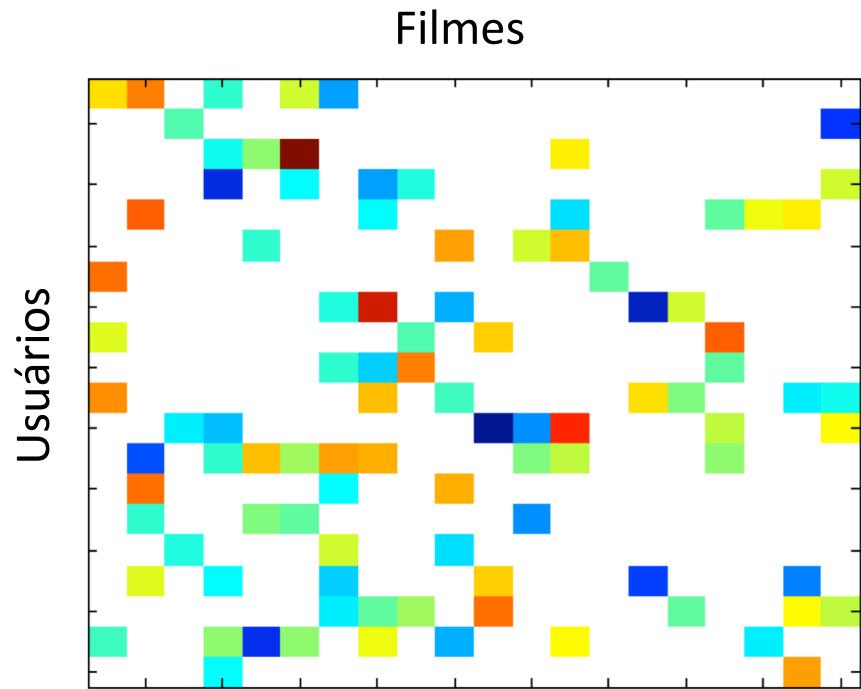
TV Shows GENRES

Because you watched Arrested Development

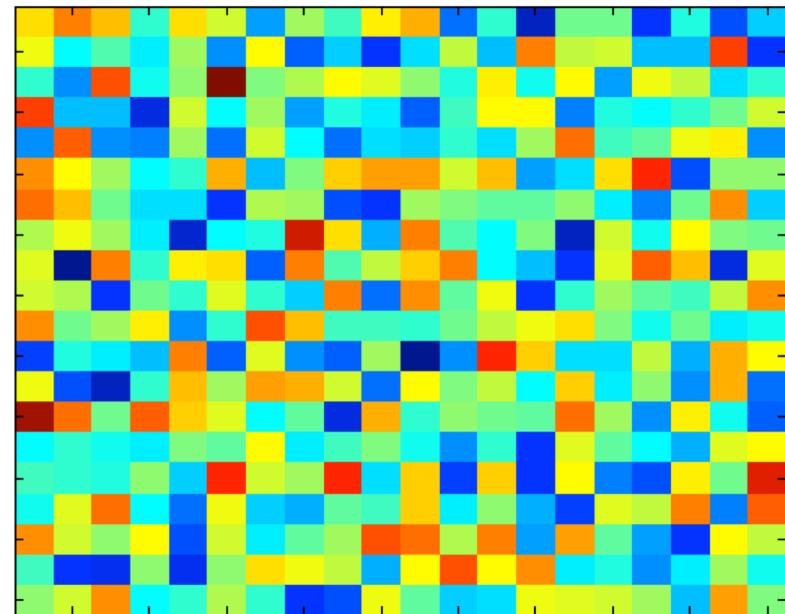
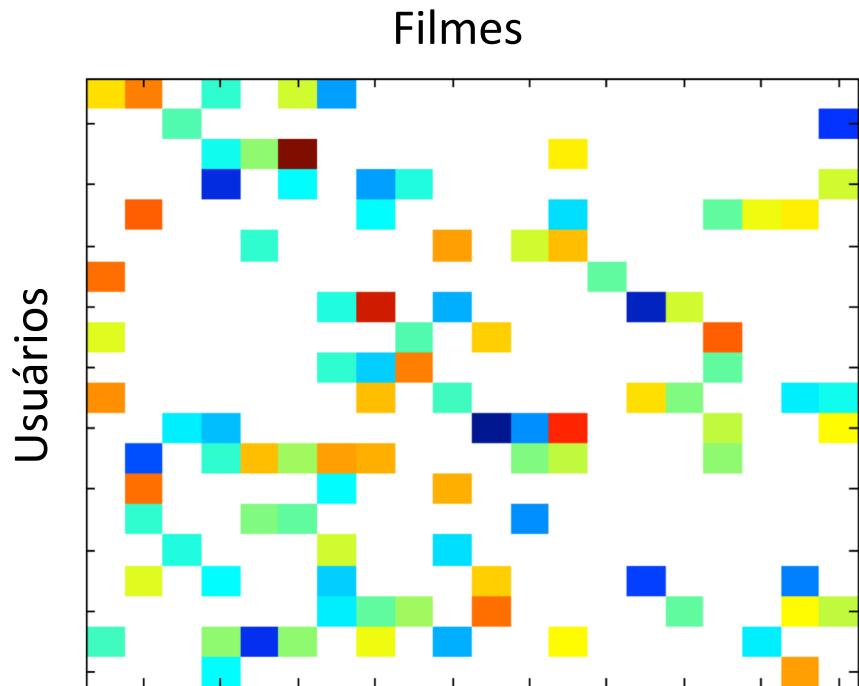
Because you watched Spartacus

Top Picks for Giovanni

Recomendação



Recomendação



Propaganda na Web

Google

sofá dois lugares

All Images Shopping Videos News More

About 595,000 results (0.35 seconds)

Shop for sofá dois lugares on Google

Sofá-cama Casal Premium Sued...
R\$911.99
Mobly

Sofá 2 Lugares Mariah Pé Palit...
R\$771.39
Mobly

Sofá Retrô Elegancy 2...
R\$899.99
Elegancy Design

Sofá 2 Lugares New Passion...
R\$691.59
Mobly

Sofá 2 Lugares é Na Tok&Stok | Vem Pra Loja, Vem Agora
Ad www.tokstok.com.br/Sofá/Tok&Stok ▾ 0800 701 0161
Soluções de Decoração Para Toda a Sua Casa. Vem Correndo Aproveitar!
Produtos Ecosociais · Itens de Designers · Coleções Exclusivas · Acesse o Portal ·

Ótimas Ofertas de Sofás
Diversos Modelos de Sofás
Em Até 10X S/ Juros. Aproveite!

Ótimas Ofertas de Pufes
Diversos Modelos de Pufes
Em Até 10X S/ Juros. Aprc

Sofá 2 Lugares - Lojas Americanas
<https://www.americanas.com.br> › Móveis › Sofá
Sofá 2 Lugares com preço baixo é na Americanas.com! Os melhores modelos e ma
Lugares em oferta. Aproveite agora!

Sofá 2 Lugares - Lar, confortável, lar | Mobly
<https://www.mobly.com.br/moveis/sofas-2-lugares/> ▾ Translate this page
Encontre na Mobly o sofá 2 lugares ideal para sua sala. Diversos modelos com pre

Itens relevantes

Google search results for "Giovanni Comarela". The search bar shows the query. Below it, a navigation bar includes "All", "Images", "Maps", "News", "Videos", "More", "Settings", and "Tools". The search results section starts with a link to Google Scholar Citations, followed by links to his Boston University page, his dpi@ufv page, his academic citations page, his dblp profile, his Escavador page, his Twitter account, and his LinkedIn professional profile.

About 81,500 results (0.40 seconds)

Giovanni Comarela - Google Scholar Citations

scholar.google.com/citations?user=onfmt40AAAAJ&hl=en ▾ [Translate this page](#)

New kid on the block: Exploring the google+ social graph. G Magno, G Comarela, D Saez-Trumper, M Cha, V Almeida. Proceedings of the 2012 Internet Measurement Conference, 159-170, 2012. 70, 2012. Finding trendsetters in information networks. D Saez-Trumper, G Comarela, V Almeida, R Baeza-Yates, F Benevenuto.

Giovanni Comarela - Boston University

cs.people.bu.edu/gcom/ ▾

This page is no longer being maintained! Please go to my new page <http://www.dpi.ufv.br/~gcom/>. Giovanni Comarela. PhD Student. Boston University · Computer Science Department. Advisor: Professor Mark Crovella Email: gcom at bu dot edu. Academic History. 2009 - BSc - Computer Science - UFES - Brazil. Advisor: ...

Giovanni Comarela - dpi@ufv

www.dpi.ufv.br/~gcom/ ▾ [Translate this page](#)

Contact. gcom@ufv.br +55 31 3899 1776. Address. Prof. Giovanni Comarela Universidade Federal de Viçosa Departamento de Informática – Campus da UFV Viçosa – MG – Brasil 36570-900. Page generated 2017-12-21 16:43:39 -02, by jemdoc.

Giovanni Comarela - Citações do Google Acadêmico

<https://scholar.google.com.br/citations?user=onfmt40AAAAJ...> ▾ [Translate this page](#)

Ginga-NCL em Dispositivos Portáteis: Uma Implementação para a Plataforma Android. GD Ferreira, G Nogueira, G Comarela, F Fabris, M Martinello, JG P Filho . 9, 2010. Detecting Unusually-Routed ASes: Methods and Applications. G Comarela, E Terzi, M Crovella. Proceedings of the 2016 ACM on Internet Measurement ...

dblp: Giovanni Comarela

dblp.uni-trier.de ▾ Persons

Dec 10, 2017 - List of computer science publications by Giovanni Comarela.

Giovanni Ventorim Comarela | Escavador

<https://www.escavador.com/sobre/.../giovanni-ventorim-comarela> ▾ [Translate this page](#)

Giovanni Ventorim Comarela - Atualmente é estudante de doutorado em ciência da computação pela Universidade de Boston (Boston University). Possui mestrado em ciência da computação pela Universidade Federal de Minas Gerais e graduação em ciência da computação pela Universidade Federal do Espírito Santo.

Giovanni Comarela (@gcomarela) | Twitter

<https://twitter.com/gcomarela> ▾

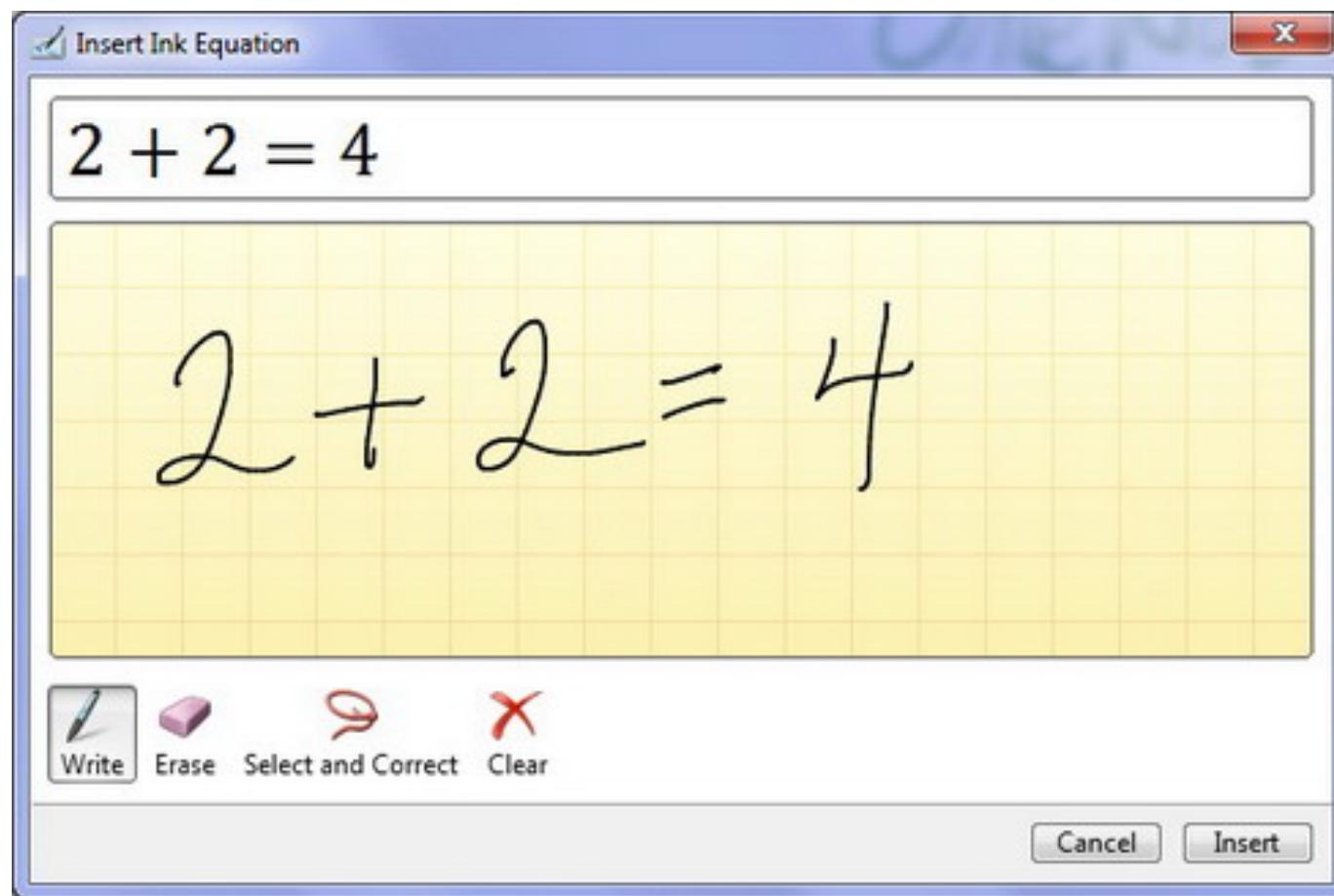
The latest Tweets from Giovanni Comarela (@gcomarela)

Giovanni Comarela | Professional Profile - LinkedIn

<https://www.linkedin.com/in/giovanni-comarela-40b51110>

Boston, Massachusetts - Research Assistant - Boston University

Classificação



Análise de redes sociais



Produtos comprados em conjunto

amazon Try Prime Electronics

Departments Your Pickup Location Browsing History Giovanni's Amazon.com Today's Deals Gift C

Computers Laptops Tablets Desktops Monitors

prime now This product is eligible

Electronics > Computers & Accessories > Data Storage > Internal Solid State Drives



Samsung
Samsung 850 EVO 250GB
★★★★★ 17,731 customer reviews
Amazon's Choice for "250gb ssd"

List Price: \$149.99
Price: **\$84.99 & FREE Shipping**. Det:
You Save: \$65.00 (43%)
Get \$40 off instantly: Pay \$44.99 upon app
✓prime | Try Fast, Free Shipping
In Stock.
Want it Tuesday, March 6? Order within 7
Ships from and sold by Amazon.com. Gift-
Size: **250 GB**
1 TB 1TB 2 TB 2TB 4
Style: **SSD Only**
SSD w/ Mount Bundle **SSD Only**

Frequently bought together



Total price: **\$95.81**

Add all three to Cart

Add all three to List

- This item:** Samsung 850 EVO 250GB 2.5-Inch SATA III Internal SSD (MZ-75E250B/AM) **\$84.99**
- Corsair Dual SSD Mounting Bracket 3.5" CSSD-BRKT2 **\$6.99**
- Monoprice 18-Inch SATA III 6.0 Gbps Cable with Locking Latch and 90-Degree Plug - Blue **\$3.83**

Ferramentas vs. Teoria

Há varias ferramentas que implementam soluções para os problemas anteriores

- Weka <https://www.cs.waikato.ac.nz/ml/weka/>
- Várias bibliotecas em Python

É possível utilizar essas ferramentas sem conhecimento teórico dos métodos?

- SIM

Objetivo do curso é mais que isso...

Problemas de Aquecimento!

Problema 1

Seja S um fluxo de números inteiros

Há em S um elemento s que ocorre com frequência maior que $|S| / 2$

Como encontrar s ?

Problema 1 – Mais frequente

Seja S um fluxo de números inteiros

Há em S um elemento s que ocorre com frequência maior que $|S| / 2$

Como encontrar s ?

Exemplo:

$$S = \{A, A, A, A, A, A, B, B, C, D\}$$

$$s = A$$

Solução trivial

Armazene os elementos de **S** em um vetor **V**

- 1) **x** = NULL
- 2) Para cada elemento **v** de **V**:
- 3) Conte quantas vezes **v** aparece em **V**
- 4) Se **v** for o elemento mais frequente:
- 5) **x** = **v**
- 6) retorno **x**

Sugestões para melhorar essa solução trivial?

Qual a complexidade?

- Espaço
- Tempo

Solução não trivial

- 1) **x** = primeiro item de **s**
- 2) **cont** = 1
- 3) Para **s** em **s**: /* segundo em diante */
- 4) Se **s** == **x**:
- 5) **cont** += 1
- 6) Senão:
- 7) Se **cont** == 0:
- 8) **cont** = 1
- 9) **x** = **s**
- 10) Senão:
- 11) **cont** -= 1
- 12) Retorne **x**

Exemplo 1

```
1) x = primeiro item de s
2) cont = 1
3) Para s em s:
4)   Se s == x:
5)     cont += 1
6)   Senão:
7)     Se cont == 0:
8)       cont = 1
9)     x = s
10)    Senão:
11)      cont -= 1
12) Retorne x
```

| s | s | x | cont |
|---|---|---|------|
| A | | A | 1 |
| A | A | | 2 |
| A | A | | 3 |
| A | A | | 4 |
| A | A | | 5 |
| A | A | | 6 |
| B | B | | 5 |
| B | B | | 4 |
| C | C | | 3 |
| D | D | | 2 |

Exemplo 2

```
1) x = primeiro item de s
2) cont = 1
3) Para s em s:
4)   Se s == x:
5)     cont += 1
6)   Senão:
7)     Se cont == 0:
8)       cont = 1
9)     x = s
10)    Senão:
11)      cont -= 1
12) Retorne x
```

| s | s | x | cont |
|---|---|---|------|
| A | | A | 1 |
| B | B | | 0 |
| A | A | | 1 |
| B | B | | 0 |
| A | A | | 1 |
| C | C | | 0 |
| A | A | | 1 |
| D | D | | 0 |
| A | A | | 1 |
| A | A | | 2 |

Exemplo 3

```
1) x = primeiro item de s
2) cont = 1
3) Para s em s:
4)   Se s == x:
5)     cont += 1
6)   Senão:
7)     Se cont == 0:
8)       cont = 1
9)     x = s
10)    Senão:
11)      cont -= 1
12) Retorne x
```

| s | s | x | cont |
|---|---|---|------|
| A | | A | 1 |
| B | B | | 0 |
| B | B | B | 1 |
| A | A | | 0 |
| C | C | C | 1 |
| D | D | | 0 |
| A | A | A | 1 |
| A | A | | 2 |
| A | A | | 3 |
| A | A | | 4 |

Complexidade

Tempo: linear

Espaço: logarítmica

Por que funciona?

A ideia da geral é que se você pudesse parear itens de forma que itens distintos estivessem no mesmo par e então remover tais pares, o único item que sobreviveria é s

$$S = \{A, A, A, A, A, A, B, B, C, D\}$$

Por que funciona?

```
1) x = primeiro item de s
2) cont = 1
3) Para s em s:
4)   Se s == x:
5)     cont += 1
6)   Senão:
7)     Se cont == 0:
8)       cont = 1
9)       x = s
10)  Senão:
11)    cont -= 1
12) Retorne x
```

O algoritmos apresentado faz exatamente isso!

Repare que:

- x é o valor mais recente que ainda não foi pareado a um elemento diferente
- cont representa quantos vezes x deve ser pareado

Problema 2 – Maior que mediana

Seja V um vetor de tamanho N (muito grande)

Encontre em v , em V , que *provavelmente* seja maior que a mediana de V

Mediana?

Solução trivial?

Solução trivial

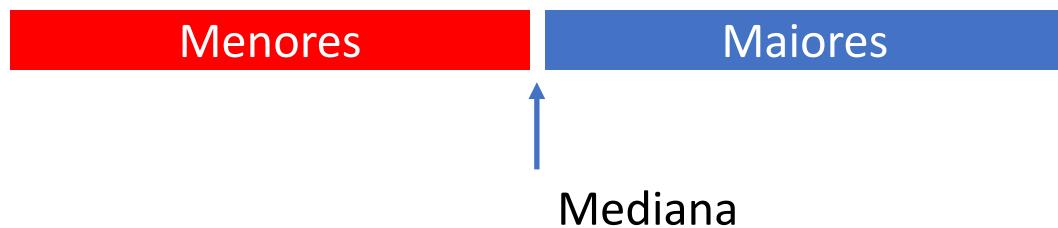
Encontre o maior elemento de V e o retorne.

Custo de tempo linear

Solução menos trivial

Tome vantagem que apenas queremos um elemento **provavelmente** maior que a mediana

Suponha que V esteja ordenado



Solução menos trivial

Selecione aleatória e independentemente k elementos de V e retorne o maior deles

Exemplo de algoritmos aleatorizado

- Frequentes em mineração de dados
- *Las Vegas* (sempre certo, nem sempre retorna)
- *Monte Carlo* (sempre retorna, às vezes errado)

Funciona corretamente? Com que probabilidade?

Problema 3 - Amostragem

Seja V um vetor de tamanho N

Encontre uma amostra aleatória S (tamanho n) de V (sem repetição)

Amostra Aleatória: elementos de V aparecem em S com probabilidade n / N .

Solução “trivial”

- 1) Permute aleatoriamente os elementos de V , obtendo V'
- 2) Retorne os n primeiros elementos de V'

Por que funciona?

Solução “trivial”

Passos 1 e 2 podem ser feitos em tempo linear

Desvantagens:

Assume que **V** se **N** são conhecidos

Duas passagens sobre os dados

Solução 2 (nada trivial)

- 1) Para i de 1 até n :
 $\quad \quad s[i] = v[i]$
- 2) Para i de $n + 1$ até N :
 $\quad \quad j = \text{item aleatório de } \{1, \dots, i\}$
 $\quad \quad \text{Se } j \leq n:$
 $\quad \quad \quad \quad s[j] = v[i]$

Solução 2

Vantagens

- Não precisa saber o tamanho de V
- Única passagem sobre os dados
- Tempo linear

Como provar que é correto?

Solução 2 (nada trivial)

- 1) Para i de 1 até n :
 $\quad \quad \quad s[i] = v[i]$
- 2) Para i de $n + 1$ até N :
 $\quad \quad \quad j = \text{item aleatório de } \{1, \dots, i\}$
 $\quad \quad \quad \text{Se } j \leq n:$
 $\quad \quad \quad \quad \quad \quad s[j] = v[i]$

Mineração de dados

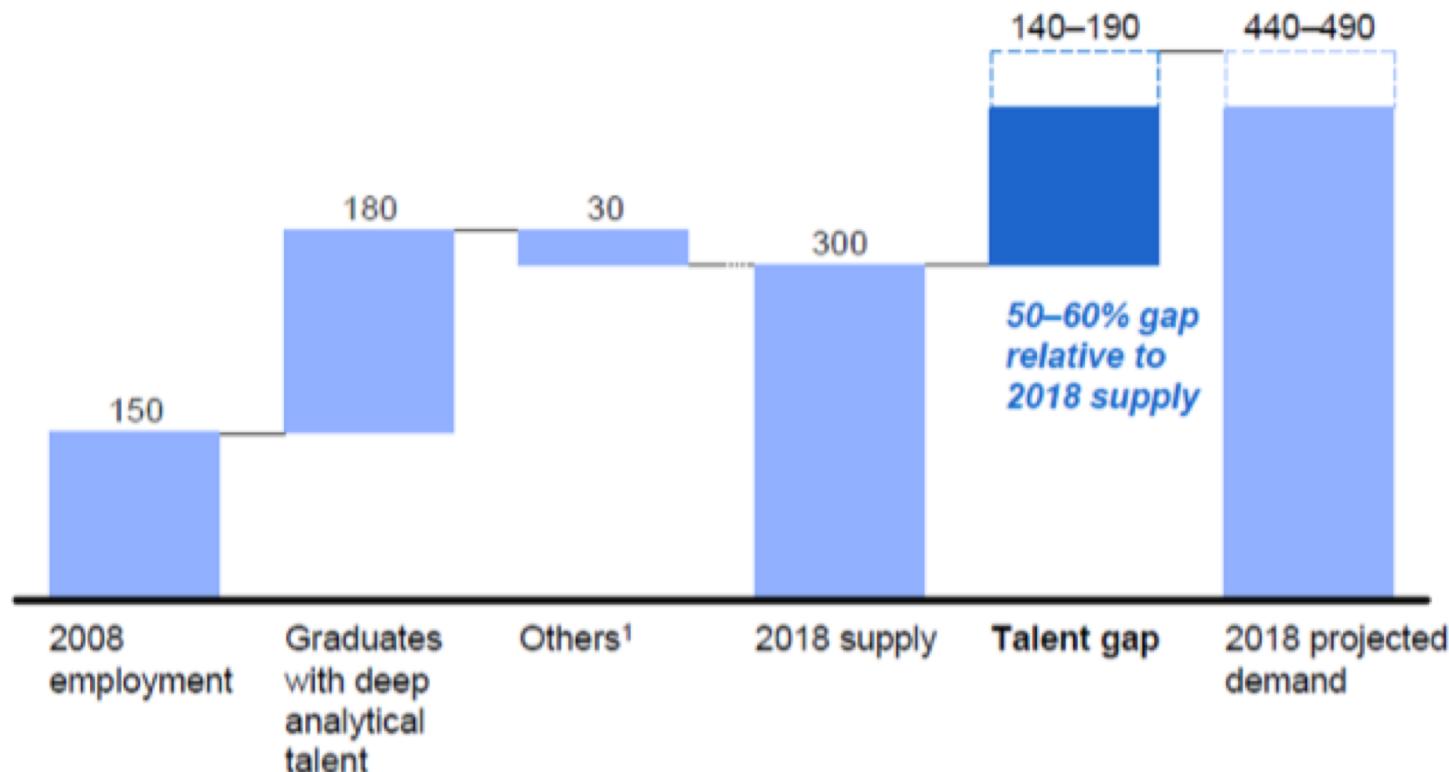
É só rodar código e programar?

Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be
50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Referências

<http://www.mmds.org>

<https://www.cs.bu.edu/~evimaria/cs565-13.html>

https://en.wikipedia.org/wiki/Reservoir_sampling

<http://minimallysufficient.github.io/2015/08/01/reservoir-sampling.html>