

INF 493/792  
Tópicos especiais III  
Introdução à mineração de dados

Aula II: Representação de dados,  
funções de dissimilaridade e métricas

# Avisos Gerais

Alunos da pós:

- Solicitei a criação de INF 790
- INF 792 + INF 790 = 4 créditos!

PVANet

Piazza

# Objetivo de hoje

Representações e tipos de dados

Funções de dissimilaridade

Métricas

# Conjuntos de dados

## Coleção de **objetos de dados**

Representam uma *entidade*

Exemplos:

Dados de vendas: clientes, lojas, produtos

Dados médicos: pacientes, médicos, tratamentos

Dados universitários: estudantes, professores, matérias

Objetos de dados também são chamados de:

*Amostras, exemplos, instâncias, pontos, objetos, tuplas*

Objetos são descritos por **atributos**

# Atributos

Outros nomes para **atributo** chamado de

Dimensão

Característica (*feature*)

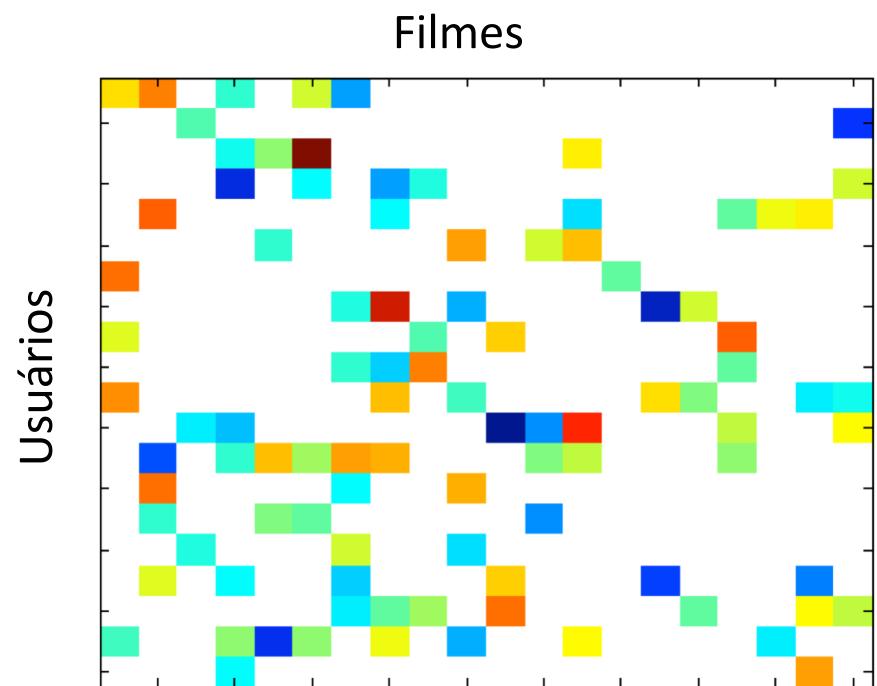
Variável

Em uma organização matricial

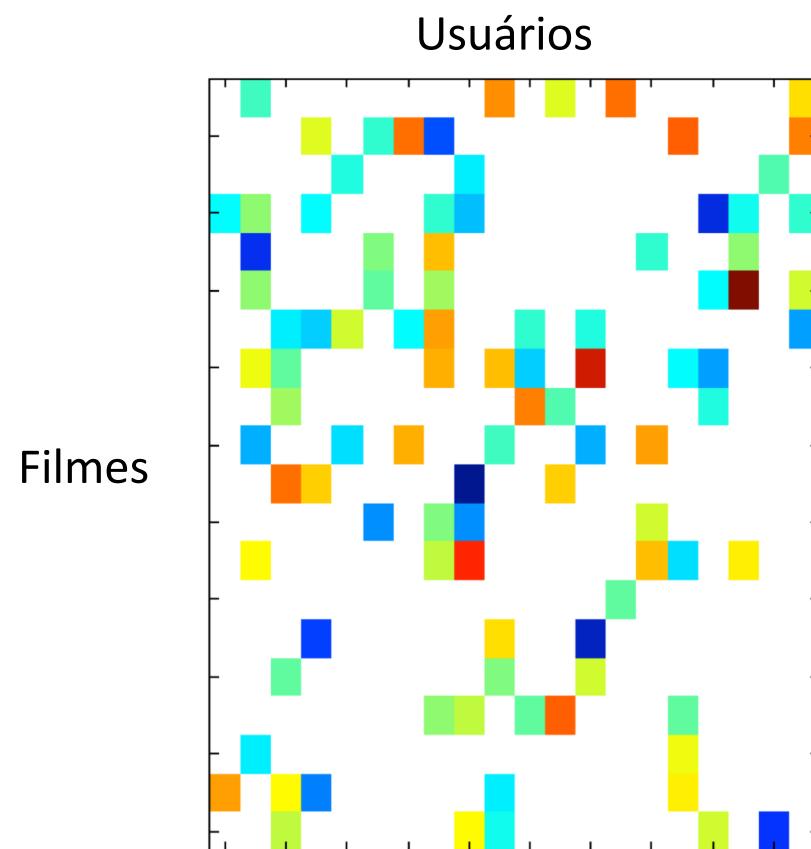
Objetos são representados em linhas

Atributos são representados em colunas

# Exemplo de representação matricial



# Mas nada impede que...



Atributos na maioria dos casos  
são de dois tipos

Categóricos ou nominais

Numéricos

# Atributos categóricos (nominais)

São categorias, estados ou “nomes de coisas”

## Exemplos

- Estado civil: {casado, solteiro}
- Cor dos olhos: {azul, verde, castanho}

Podem ser ordinais. Exemplo

- Tamanho de roupa: {pequeno, médio, grande}

# Atributos binários

Tipo especial de atributo categórico com apenas dois estados, **0** e **1**

Em muitas aplicações/algoritmos, atributos binários pode ser considerados números

# De categórico para binário

Nome	Cor dos olhos
Maria	Castanho
João	Verde
José	Castanho

Nome	Castanho	Verde	Azul
Maria	1	0	0
João	0	1	0
José	1	0	0

# Atributos numéricos

Quando é necessário quantificar uma grandeza

Podem ser:

Discretos

O possível conjunto de valores é **finito** ou **infinito e enumerável**

Representados por variáveis inteiras (ou binárias)

Ex: número de dependentes

Contínuos

Assumem valores pertencente ao conjunto dos números reais

Representados pro variáveis do tipo ponto flutuante

Ex: altura e peso

Nem sempre são uma matriz

# Grafos

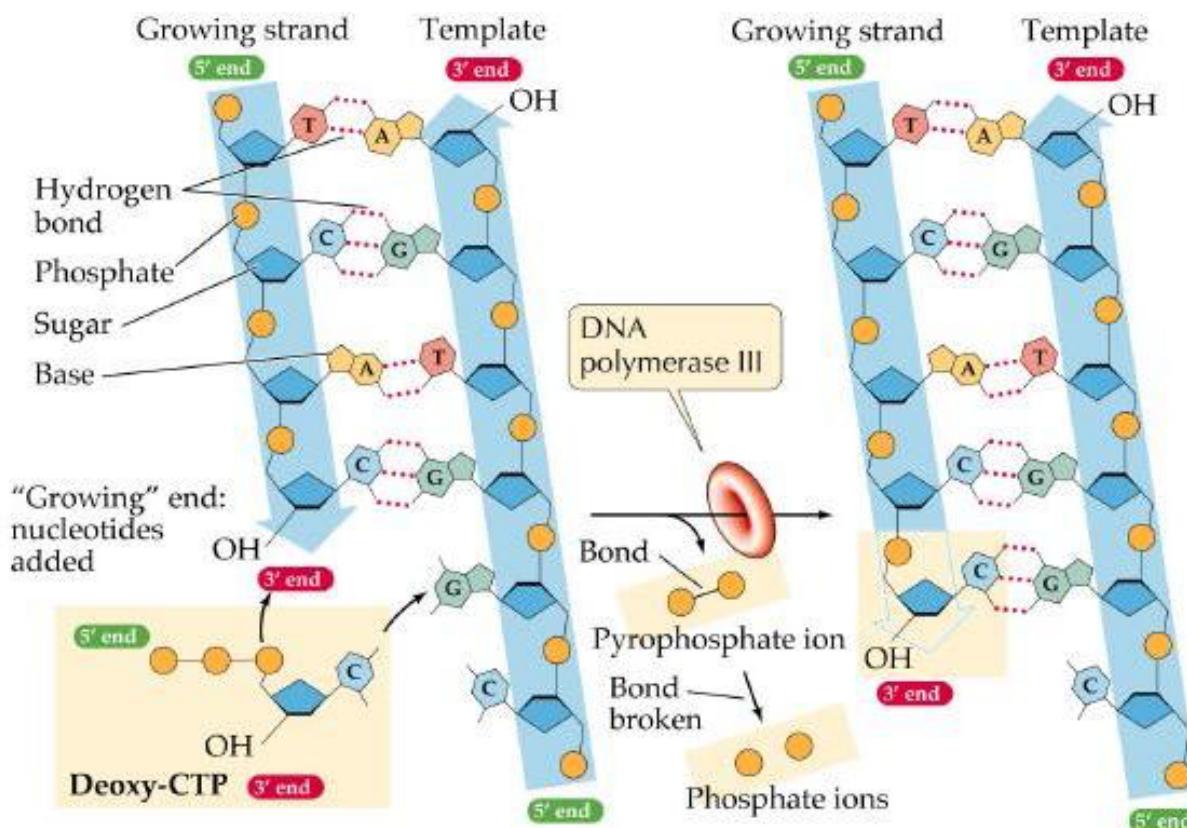
Par de conjuntos **V** e **E** (vértices e arestas)

Formas mais comuns de representação

- Matriz de adjacências
- Lista de adjacências

Exemplos?

# Sequências



# Texto

Suponha que você tenha uma coleção de documentos (e.g. várias páginas Web)

Qual é uma forma razoável de representar essa base de dados?

# Texto

Suponha que você tenha uma coleção de documentos (e.g. várias páginas Web)

Qual é uma forma razoável de representar essa base de dados?

Matriz Documento-Termo

# Matrix documento termo

A = “Eu gosto de mineração de dados”

B = “Eu odeio mineração de dados”

Doc\termo	eu	gosto	odeio	mineracao	dados
A	1	1	0	1	1
B	1	0	1	1	1

Repare as transformações!

Problemas? O que ocorre com preposições?

# TF-IDF

***Term Frequency – Inverse Document Frequency***

Dada uma coleção de documentos, o quanto importante é uma palavra pra um documento específico

# TF – Term Frequency

$tf(t, d)$  = frequência do termo  $t$  no documento  $d$

Geralmente é normalizado pelo número de termos no documento

# IDF – Inverse Document Frequency

$$\text{idf}(t, D) = \log(N / D_t)$$

Onde:

**D** é a coleção de documentos

**N** é o número de documentos em **D**

**D<sub>t</sub>** é o número de documentos em que **t** aparece

# TF-IDF

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

# TF-IDF

Várias outras variações

<https://en.wikipedia.org/wiki/Tf–idf>

# TF-IDF

**Document 1**

Term	Term Count
this	1
is	1
a	2
sample	1

**Document 2**

Term	Term Count
this	1
is	1
another	2
example	3

# TF-IDF (termo “this”)

Document 1	
Term	Term Count
this	1
is	1
a	2
sample	1

Document 2	
Term	Term Count
this	1
is	1
another	2
example	3

$$\text{tf}("this", d_1) = \frac{1}{5} = 0.2$$

$$\text{tf}("this", d_2) = \frac{1}{7} \approx 0.14$$

$$\text{idf}("this", D) = \log\left(\frac{2}{2}\right) = 0$$

$$\text{tfidf}("this", d_1) = 0.2 \times 0 = 0$$

$$\text{tfidf}("this", d_2) = 0.14 \times 0 = 0$$

# TF-IDF (termo “example”)

Document 1	
Term	Term Count
this	1
is	1
a	2
sample	1

Document 2	
Term	Term Count
this	1
is	1
another	2
example	3

$$\text{tf}("example", d_1) = \frac{0}{5} = 0$$

$$\text{tf}("example", d_2) = \frac{3}{7} \approx 0.429$$

$$\text{idf}("example", D) = \log\left(\frac{2}{1}\right) = 0.301$$

$$\text{tfidf}("example", d_1) = \text{tf}("example", d_1) \times \text{idf}("example", D) = 0 \times 0.301 = 0$$

$$\text{tfidf}("example", d_2) = \text{tf}("example", d_2) \times \text{idf}("example", D) = 0.429 \times 0.301 \approx 0.13$$

# TF-IDF

Pode-se criar uma matriz (Documento-termo) onde os valores são dados por TF-IDF ao invés de simplesmente a frequência (TF) dos termos.

Como resolver o problema do grande número de termos?

# Medidas de dissimilaridade

Têm o objetivo de quantificar a grau de **diferença** entre dois **objetos** (ou **atributos**)

São necessárias em vários algoritmos que veremos ao longo do semestre

# Distância euclidiana

Dados  $\mathbf{x}$  e  $\mathbf{y}$ , vetores do espaço real  $n$ -dimensional

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

# Distância euclidiana

Exemplo:

$$\mathbf{x} = (1, 2)$$

$$\mathbf{y} = (-2, 1)$$

# Distância euclidiana

Desvantagem

Difícil de relativizar

E se uma das dimensões é muito maior que as outras?

# Distância de Manhattan

Dados  $\mathbf{x}$  e  $\mathbf{y}$ , vetores do espaço real  $n$ -dimensional

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i|$$

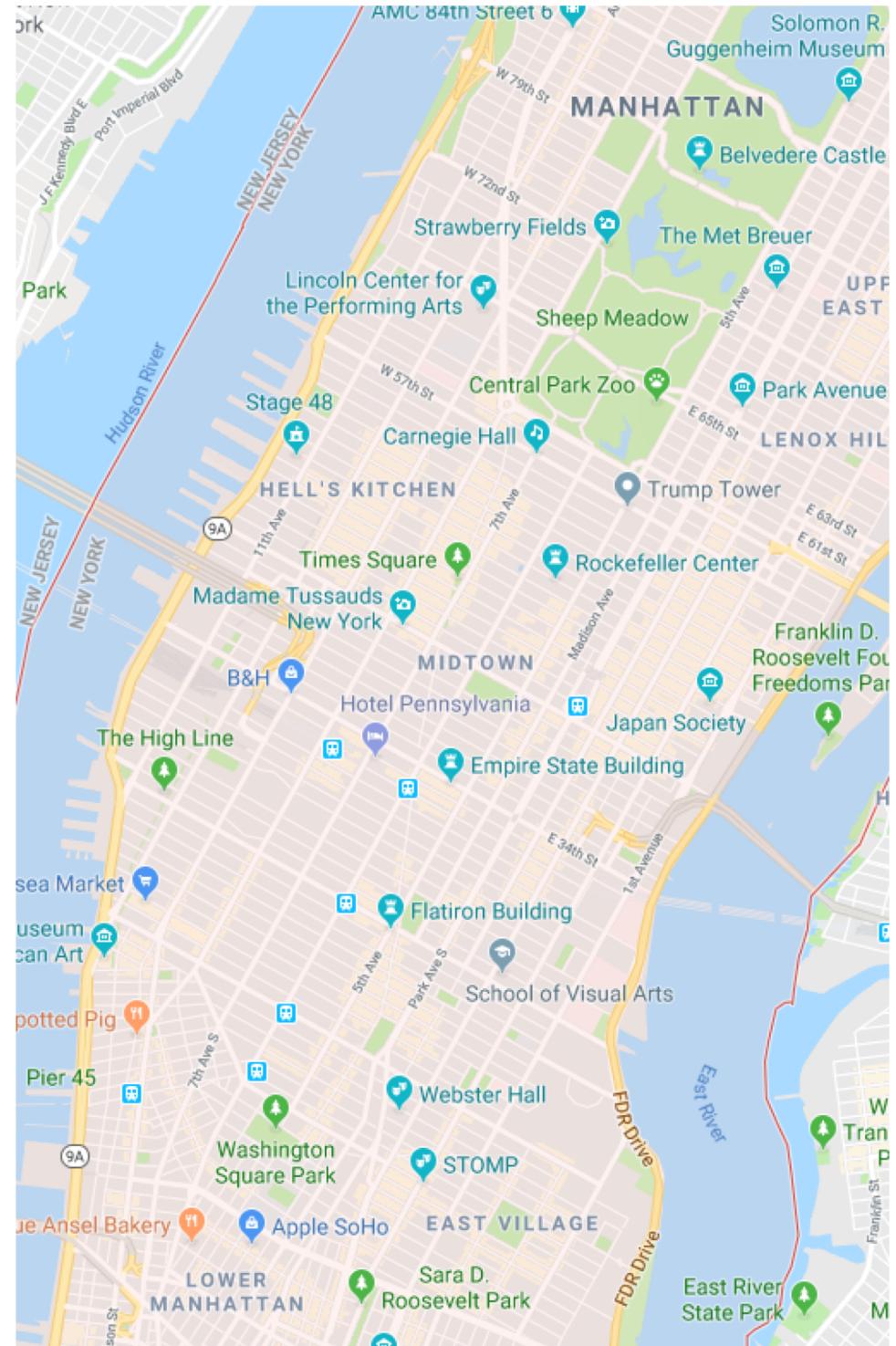
# Distância de Manhattan

Exemplo:

$$\mathbf{x} = (1, 2)$$

$$\mathbf{y} = (-2, 1)$$

# Distância de Manhattan



# Distância de Minkowsky

Dados  $\mathbf{x}$  e  $\mathbf{y}$ , vetores do espaço real  $n$ -dimensional

$$d(x, y) = \left( \sum_{i=1}^n (x_i - y_i)^r \right)^{\frac{1}{r}}$$

Onde  $r > 0$  é um parâmetro

# Distância do cosseno

Dados  $\mathbf{x}$  e  $\mathbf{y}$ , vetores do espaço real  $\mathbf{n}$ -dimensional

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}$$

onde

$$x \cdot y = \sum_{i=1}^n x_i y_i \quad \|x\| = \sqrt{\sum_{i=1}^n x_i^2}$$

# Distância do cosseno

Também chamada de distância angular.

Tem valor dado por  $1 - \text{cosseno}$  do ângulo entre  $x$  e  $y$

# Distância do cosseno

Exemplos:

$$\mathbf{x} = (1, 2), (1, 2), (0, 1), (0, 1)$$

$$\mathbf{y} = (-2, 1), (2, 4), (1, 1), (2, 2)$$

# Distância de hamming

Dados  $x$  e  $y$ , vetores **binários** de mesmo tamanho

$d(x, y)$  é o número de posições que  $x$  e  $y$  differem

# Distância de hamming

Hamming distance = 3 —

<i>A</i>	1	0	1	1	0	0	1	0	0	1
			‡			‡		‡		
<i>B</i>	1	0	0	1	0	0	0	0	1	1

# Distância de hamming

## Desvantagem

Vetores similares são completamente diferentes após uma operação simples de deslocamento (shift)

# Distância de Jaccard

Dados dois **conjuntos** x e y

$$d(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

# Distância de Jaccard

Exemplo

$$\mathbf{A} = \{1,2,3,4,5\}$$

$$\mathbf{B} = \{3,4\}$$

$$J(\mathbf{A}, \mathbf{B}) =$$

# Distância de Jaccard

Como visualizar?

Conjuntos também podem ser representados como vetores binários!

Há uma generalização conhecida como **Distância generalizada de Jaccard**

# Métrica

Uma medida de dissimilaridade  $\mathbf{d}(., .)$  é uma métrica se:

$$d(x, y) \geq 0$$

$$d(x, y) = d(y, x)$$

$$d(x, y) = 0 \iff x = y$$

$$d(x, y) \leq d(x, z) + d(z, y)$$

Propriedades: não-negatividade, simetria, isolamento, desigualdade triangular

# Distância de hamming

Prove que esta é (ou não) uma métrica

**5 minutos**

Dados  $x$  e  $y$ , vetores **binários** de mesmo tamanho  
 $d(x, y)$  é o número de posições que  $x$  e  $y$  differem

# Distância de Manhattan

Prove que esta é (ou não) uma métrica

**5 minutos**

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

# Distância de Manhattan

E se eu falar que  $|a + b| \leq |a| + |b|$ , sendo **a** e **b** números reais

# Distância Euclidiana

Prove que esta é (ou não) uma métrica

**5 minutos**

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

# Distância Euclidiana

E se eu der a desigualdade de Cauchy-Schawrz?

# Distância Euclidiana ao quadrado

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

# Distância euclidiana modificada

$$d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$$

# Distância do cosseno

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}$$

$$x \cdot y = \sum_{i=1}^n x_i y_i \quad \|x\| = \sqrt{\sum_{i=1}^n x_i^2}$$

# Distância de Jaccard

$$d(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

# Notação

Em alguns textos:

dissimilaridade e distância são a mesma coisa

Em outros textos:

Distância e métrica são a mesma coisa

# Matriz de dados vs. Matriz de distâncias

# Leitura recomendada

[http://www.dataminingbook.info/pmwiki.php/Main/  
BookResources](http://www.dataminingbook.info/pmwiki.php/Main/BookResources)

Capítulos 1, 2, 3 e 4

Sobre Mineração, dados e summarização de dados  
(algumas partes são complexas)

# Leitura recomendada

[http://hanj.cs.illinois.edu/bk3/bk3\\_slidesindex.htm](http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm)

Seção 2.4

Sobre funções de distância e métricas

Aula baseada em slides de:

<http://www.mmds.org>

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

[http://hanj.cs.illinois.edu/bk3/bk3\\_slidesindex.htm](http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm)

<http://www.dataminingbook.info/pmwiki.php/Main/BookResources>