Fatoração de Matrizes no problema de Coagrupamento com sobrepreposição de colunas

Lucas Fernandes Brunialti
Orientadora: Profa. Dra. Sarajane Marques Peres

Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

lucas.brunialti@usp.br
sarajane@usp.br

31 de agosto de 2016

Agenda

Introdução

Conceitos Fundamentais

Algoritmos de FM não-negativas para agrupamento e coagrupamento

Fatoração de matrizes não-negativas para coagrupamento com sobreposição de colunas

Experimentos

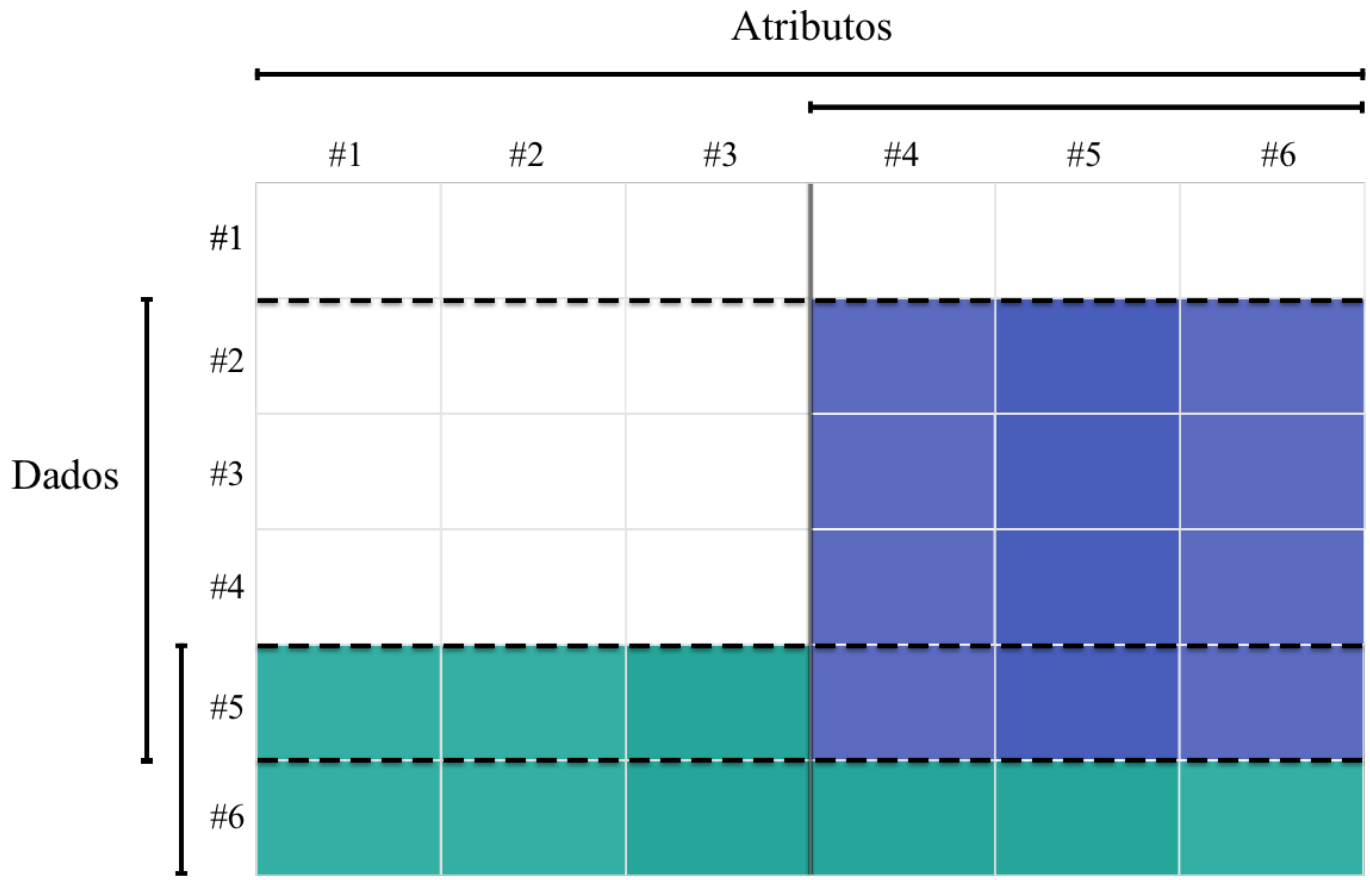
Conclusão

Seção 1

Introdução

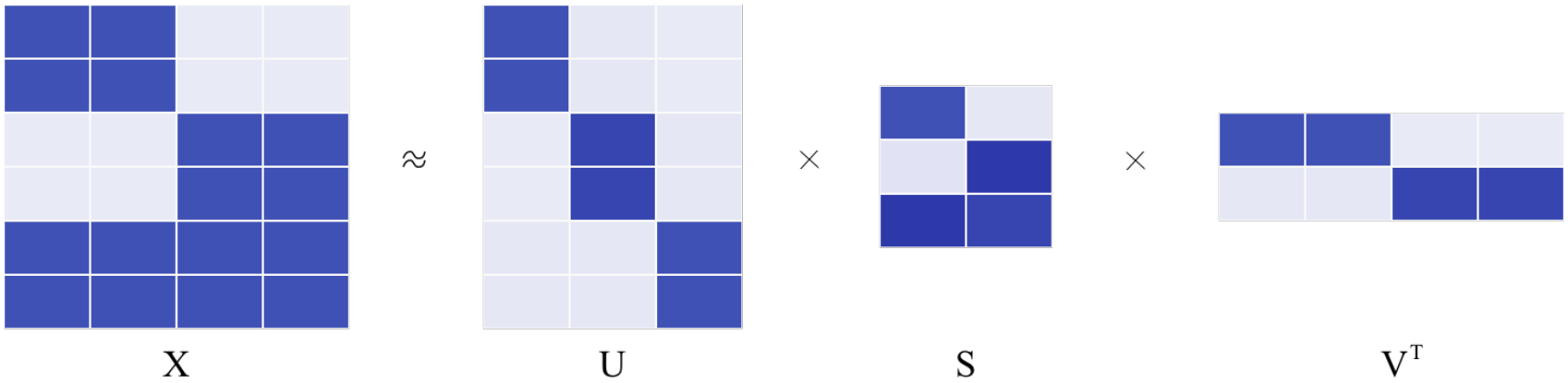
- Agrupamento
  - Coleção de dados: matriz com n objetos e m características
  - Organiza uma coleção de dados em grupos
  - Dados em um mesmo grupo são similares
  - Estratégias: particional, hierárquica, baseada em densidade, etc
- Coagrupamento
  - Uma estratégia: formar grupos considerando a **similaridade parcial**, **similaridade por partes** ou **resconstrução por partes**
  - Análises mais refinadas

Coagrupamento



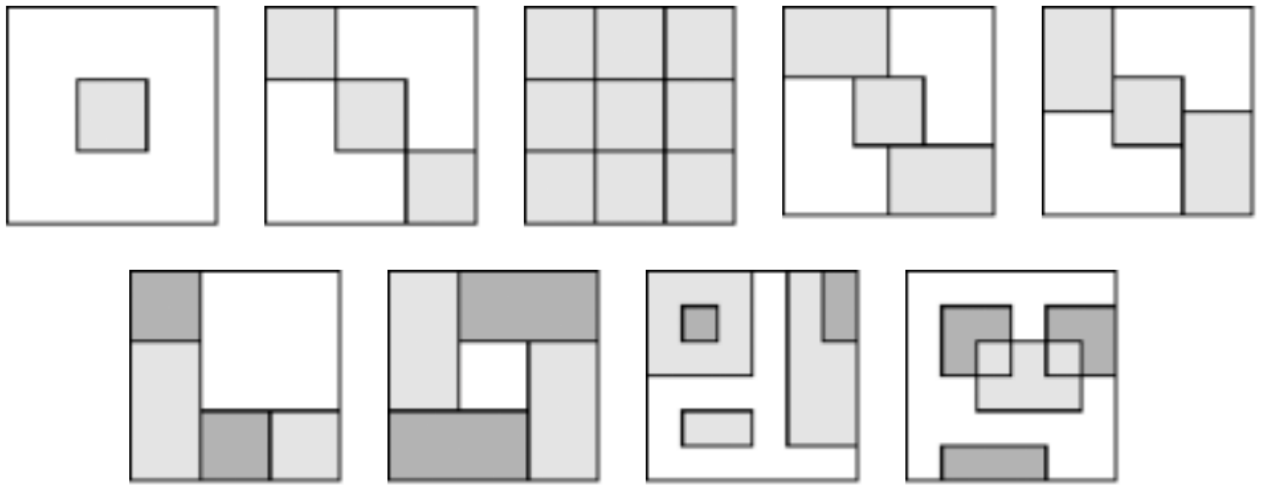
- Visão tradicional:
  - 2 cogrupos (submatrizes azul e verde)

Fatoração de matrizes não-negativas



- U particionamento de linhas
- V particionamento de colunas
- S relação entre as partes
- Pós-processamento para obtenção dos cogrupos

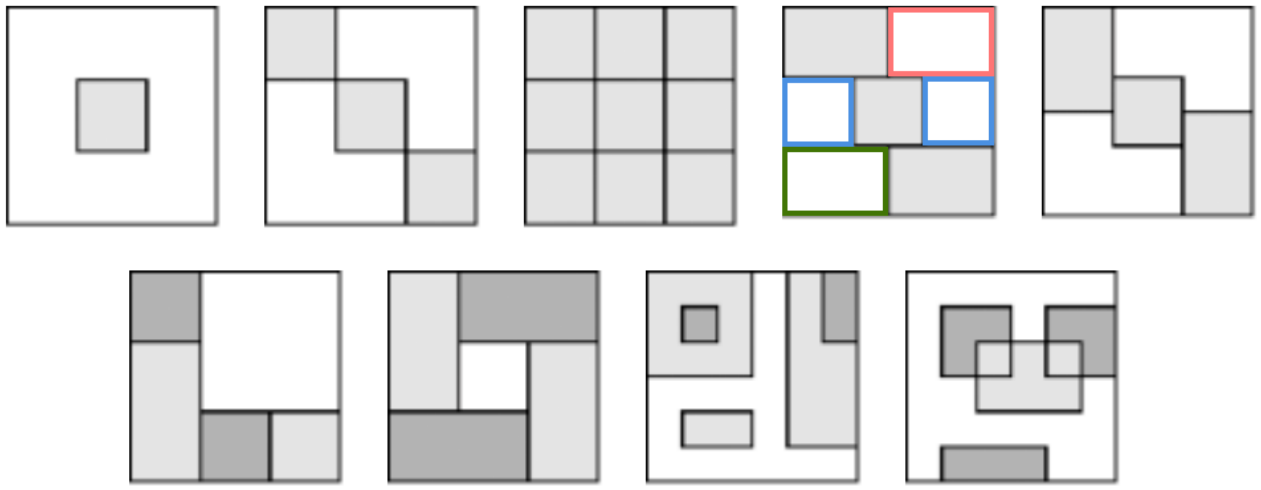
Estruturas de cogrupos I



Estruturas de cogrupos:

- caso 3: Coagrupamento em FM
- caso 4: Coagrupamento com sobreposição de colunas em FM (proposto)

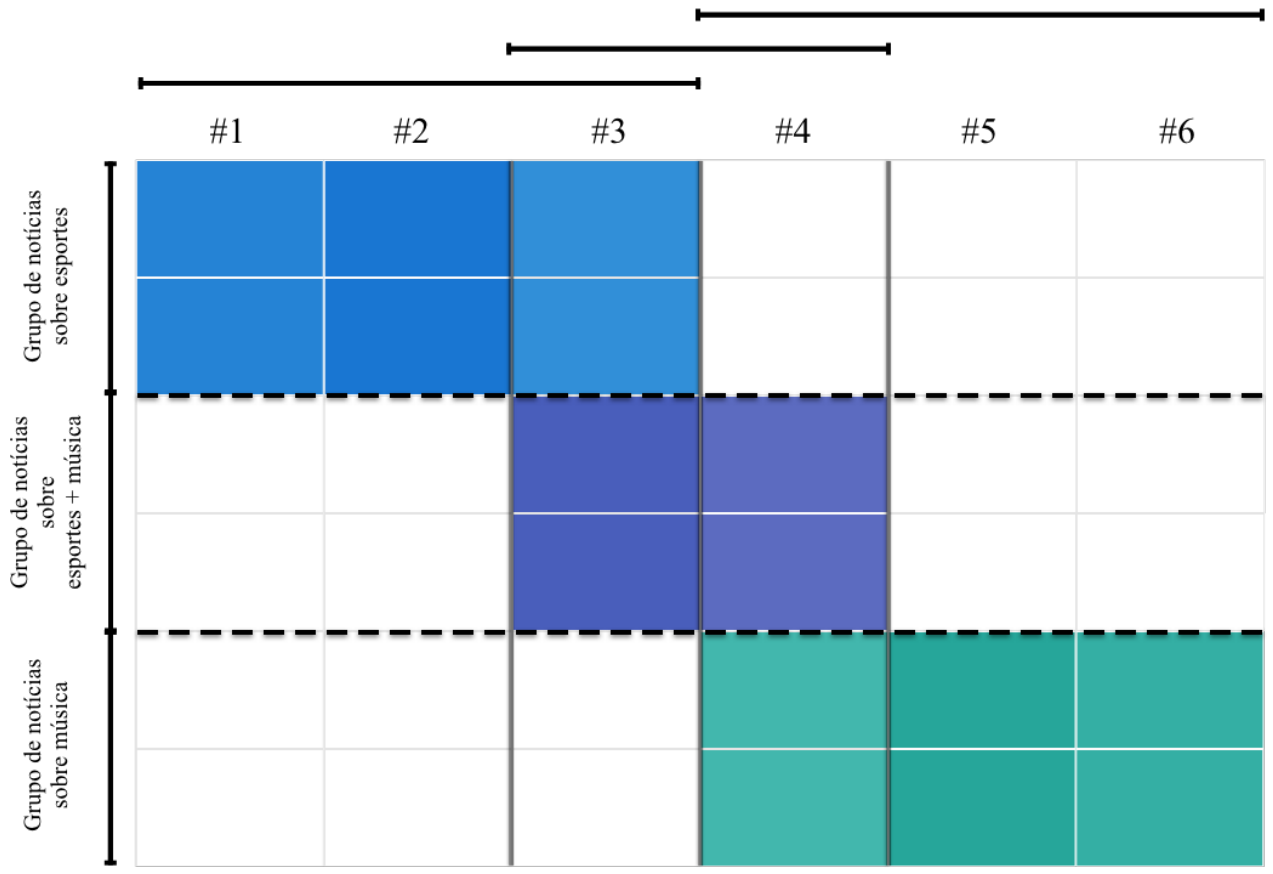
Estruturas de cogrupos II



- Particionamento

Aplicação de coagrupamento de notícias

- Aplicação de mineração de textos
- Coleção de documentos $\rightarrow$ matriz de dados
- Contagem de palavras para cada documento



Problema e Hipótese

Definição do problema

- Definição do problema de coagrupamento com **sobreposição** de colunas em FM
- Soluções algorítmicas:
  - Resolver coagrupamento
  - **sobreposição** de cogrupos de colunas de forma adequada

Hipótese

- **Sobreposição** de cogrupos de colunas pode ser resolvida por:

$$X \approx g(U, S, V_{(1)}, \ldots, V_{(k)})$$

- Permite que grupos de colunas sejam independentes

Objetivos I

Objetivo geral

- Proposição do problema de coagrupamento com sobreposição de colunas
- Estratégias algorítmicas para resolver $X \approx g(U, S, V_{(1)}, \ldots, V_{(k)})$:
  - *OvNMTF*
  - *BinOvNMTF*
- Avaliando em termos de:
  - capacidade de quantização e reconstrução
  - capacidade de agrupamento
  - capacidade de extração de informação (interpretabilidade)

Objetivos II

Objetivos específicos

- Derivação formal para o *OvNMTF* e *BinOvNMTF*
- Aplicação do *OvNMTF* e *BinOvNMTF* em:
  - ambientes controlados (bases de dados sintéticas)
  - contexto de aplicação real (análise de dado textuais)
- Novo conjunto de dados de notícias em língua portuguesa

Seção 2

Definição de agrupamento

Definição

- Matriz de dados: $X \in \mathbb{R}^{n \times m}$
- Conjunto de vetores de linhas:

$$\mathcal{N} = \{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{n\cdot}\}$$

- Objetivo:
  - Particionar $\mathcal{N}$:

$$\mathcal{K}_p \subseteq \mathcal{N}, \forall p \in \{1, \ldots, k\}$$

  - Resultado:

$$\mathscr{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_k\}$$

Definição de coagrupamento

Definição

- Matriz de dados: $X \in \mathbb{R}^{n \times m}$
- Conjunto de vetores de linhas:

$$\mathcal{N} = \{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{n\cdot}\}$$

- Conjunto de vetores de colunas

$$\mathcal{M} = \{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot m}\}$$

- Objetivo:
  - Encontrar subconjuntos de $\mathcal{N}$:

$$\mathcal{K}_p \subseteq \mathcal{N}, \forall p \in \{1, \ldots, k\}$$

  - Encontrar subconjuntos de $\mathcal{M}$:

$$\mathcal{L}_q \subseteq \mathcal{M}, \forall q \in \{1, \ldots, l\}$$

  - Encontrar submatrizes de X:

$$X_{\mathcal{K}_p \mathcal{L}_q}$$

Definição de coagrupamento em FM não-negativas

Definição

- Matriz de dados: $X \in \mathbb{R}_{+}^{n \times m}$
- Conjunto de vetores de linhas:

$$\mathcal{N} = \{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{n\cdot}\}$$

- Conjunto de vetores de colunas

$$\mathcal{M} = \{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot m}\}$$

- Objetivo:
  - Particionar $\mathcal{N}$:

$$\mathcal{K}_p \subseteq \mathcal{N}, \forall p \in \{1, \ldots, k\}$$

  - Particionar $\mathcal{M}$:

$$\mathcal{L}_q \subseteq \mathcal{M}, \forall q \in \{1, \ldots, l\}$$

  - Resultado: $\mathscr{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_k\}$ $\mathscr{L} = \{\mathcal{L}_1, \ldots, \mathcal{L}_l\}$

Seção 3

Problema *k-means* (forma fatoração)

$$\begin{aligned} \mathcal{F}_1(U, C) \quad = \quad & \min_{U,C} \quad && \|X - UC\|_F^2 \\ & \text{suj. a} && U \in \Psi^{n \times k}, \\ & && C \in \mathbb{R}^{k \times m}, \\ & && \sum_{p=1}^{k} u_{ip} = 1, \forall i \end{aligned}$$

Fuzzy K-means

Problema *fuzzy k-means* ($w = 2$)

$$\begin{aligned} \mathcal{F}_2^{w=2}(U, C) \quad = \quad & \min_{U,C} && \|X - UC\|_F^2 \\ & \text{suj. a} && U \in \mathbb{R}^{n \times k}, \\ & && C \in \mathbb{R}^{k \times m}, \\ & && \sum_{p=1}^{k} u_{ip} = 1, \forall i \end{aligned}$$

FM não-negativas para coagrupamento (BVD)

Problema *BVD*

$$\begin{aligned} \mathcal{F}_3(U,S,V) \quad = \quad & \min_{U,S,V} \quad \|X - USV^T\|_F^2 \\ & \text{suj. a} \quad U \geq 0, S \geq 0, V \geq 0 \end{aligned}$$

FM não-negativas ortogonal tripla (ONMTF)

Problema *ONMTF*

$$\begin{aligned} \mathcal{F}_6(U,S,V) \quad = \quad & \min_{U,S,V} \quad \|X - USV^T\|_F^2 \\ & \text{suj. a} \quad U \geq 0, S \geq 0, V \geq 0, \\ & \qquad\quad U^T U = I, \\ & \qquad\quad V^T V = I \end{aligned}$$

Fatoração tripla rápida de matrizes não-negativas (FNMTF)

Problema *FNMTF*

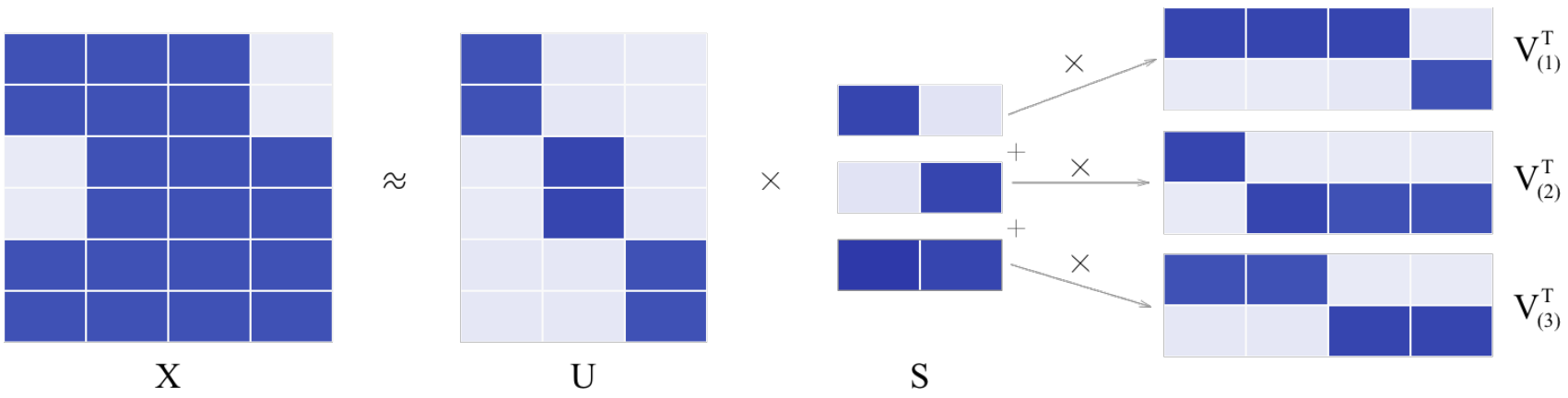
$$\begin{aligned} \mathcal{F}_7(U,S,V) \quad = \quad & \min_{U,S,V} \quad \|X - USV^T\|_F^2 \\ & \text{suj. a} \quad U \in \Psi^{n\times k}, \\ & \qquad\quad V \in \Psi^{m\times l}, \\ & \qquad\quad \sum_{p=1}^{k} u_{ip} = 1, \forall i, \\ & \qquad\quad \sum_{q=1}^{l} v_{jq} = 1, \forall j \end{aligned}$$

Resumo das fatorações da literatura

	Fatoração	**Compactação**	**Restrições**
k-means	$X \approx UC$	$n + km$	$U \in \Psi^{n \times k}, C \in \mathbb{R}^{k \times m}, \sum_{p=1}^{k} u_{ip} = 1$
fuzzy k-means	$X \approx UC$	$nk + km$	$U \in \mathbb{R}^{n \times k}, C \in \mathbb{R}^{k \times m}, \sum_{p=1}^{k} u_{ip} = 1$
BVD	$X \approx USV^T$	$nk + kl + ml$	$U \geq 0, S \geq 0, V \geq 0$
ONMTF	$X \approx USV^T$	$nk + kl + ml$	$U \geq 0, S \geq 0, V \geq 0, U^T U = I, V^T V = I$
FNMTF	$X \approx USV^T$	$n + kl + m$	$U \in \Psi^{n \times k}, V \in \Psi^{m \times l}, \sum_{p=1}^{k} u_{ip} = 1, \sum_{q=1}^{l} v_{jq} = 1$

Seção 4

Fatoração de matrizes não-negativas para coagrupamento com sobreposição de colunas



- k particionamentos para as colunas (V_p)
- independência

Definição de fatoração de matrizes não-negativas para coagrupamento com sobreposição de colunas

Definição

- Matriz de dados: $X \in \mathbb{R}_{+}^{n \times m}$
- Conjunto de vetores de linhas: $\mathcal{N} = \{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{n\cdot}\}$
- Conjunto de vetores de colunas: $\mathcal{M} = \{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot m}\}$

- Objetivo:
  - Particionar $\mathcal{N}$:

$$\mathcal{K}_p \subseteq \mathcal{N}, \forall p \in \{1, \ldots, k\}$$

  - Particionar $\mathcal{M}$ k vezes:

$$\mathcal{L}_{pq} \subseteq \mathcal{M}, \forall p \in \{1, \ldots, k\}, \forall q \in \{1, \ldots, l\}$$

  - Resultado:

$$\mathscr{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_k\}$$

$$\mathscr{L} = \{\mathcal{L}_{11}, \ldots, \mathcal{L}_{1l}, \ldots, \mathcal{L}_{k1}, \ldots \mathcal{L}_{kl}\}$$

Fatoração Tripla de Matrizes Não-negativas com Sobreposição (OvNMTF)

Problema *OvNMTF*

$$\mathcal{F}_6(U, S, V_{(1)}, \ldots, V_{(k)}) = \min_{U,S,V_{(1)},\ldots,V_{(k)}} \|X - U\sum_{p=1}^{k} I_{(p)} S V_{(p)}^T\|_F^2$$
$$\text{suj. a} \quad U \geq 0, S \geq 0, \; V_{(p)} \geq 0, \quad \forall p$$

- $I_{(p)} \in \{0,1\}^{k \times k}$ matriz seletora

Fatoração Binária Tripla de Matrizes Não-negativas com Sobreposição (BinOvNMTF)

Problema *BinOvNMTF*

$$\mathcal{F}_7(U, S, V_{(1)}, \ldots, V_{(k)}) = \min_{U,S,V_{(1)},\ldots,V_{(k)}} \|X - U\sum_{p=1}^{k} I_{(p)} S V_{(p)}^T\|_F^2$$
$$\text{suj. a} \quad U \in \Psi^{n\times k},$$
$$V_{(p)} \in \Psi^{m\times l}, \quad \forall p,$$
$$\sum_{p=1}^{k} u_{ip} = 1, \forall i,$$
$$\sum_{q=1}^{l} v_{(p)_{jq}} = 1, \forall p, j$$

Resumo das fatorações

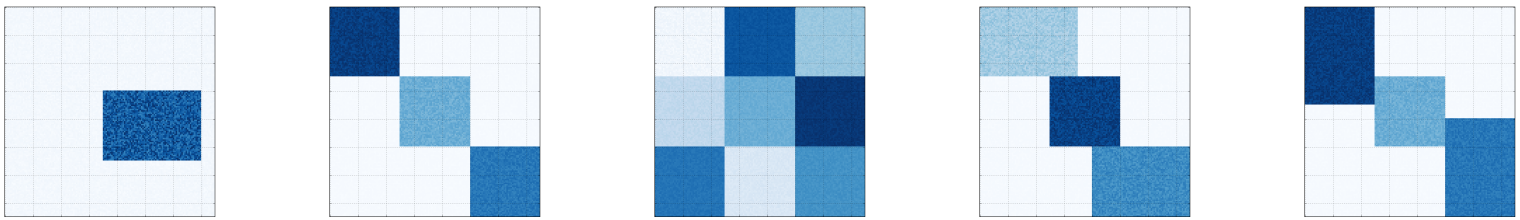
	Fatoração	Compactação	Restrições
k-means	$X \approx UC$	$n + km$	$U \in \Psi^{n \times k}, C \in \mathbb{R}^{k \times m}, \sum_{p=1}^{k} u_{ip} = 1$
fuzzy k-means	$X \approx UC$	$nk + km$	$U \in \mathbb{R}^{n \times k}, C \in \mathbb{R}^{k \times m}, \sum_{p=1}^{k} u_{ip} = 1$
BVD	$X \approx USV^T$	$nk + kl + ml$	$U \geq 0, S \geq 0, V \geq 0$
ONMTF	$X \approx USV^T$	$nk + kl + ml$	$U \geq 0, S \geq 0, V \geq 0, U^T U = I, V^T V = I$
FNMTF	$X \approx USV^T$	$n + kl + m$	$U \in \Psi^{n \times k}, V \in \Psi^{m \times l}, \sum_{p=1}^{k} u_{ip} = 1, \sum_{q=1}^{l} v_{jq} = 1$
OvNMTF	$X \approx g(U, S, V_{(1)}, \ldots, V_{(k)})$	$nk + kl + klm$	$U \geq 0, S \geq 0, V_{(p)} \geq 0$
BinOvNMTF	$X \approx g(U, S, V_{(1)}, \ldots, V_{(k)})$	$n + kl + km$	$U \in \Psi^{n \times k}, V_{(p)} \in \Psi^{m \times l}, \sum_{p=1}^{k} u_{ip} = 1, \sum_{q=1}^{l} v_{(p)_{jq}} = 1$

Seção 5

Experimentos com bases de dados sintéticas

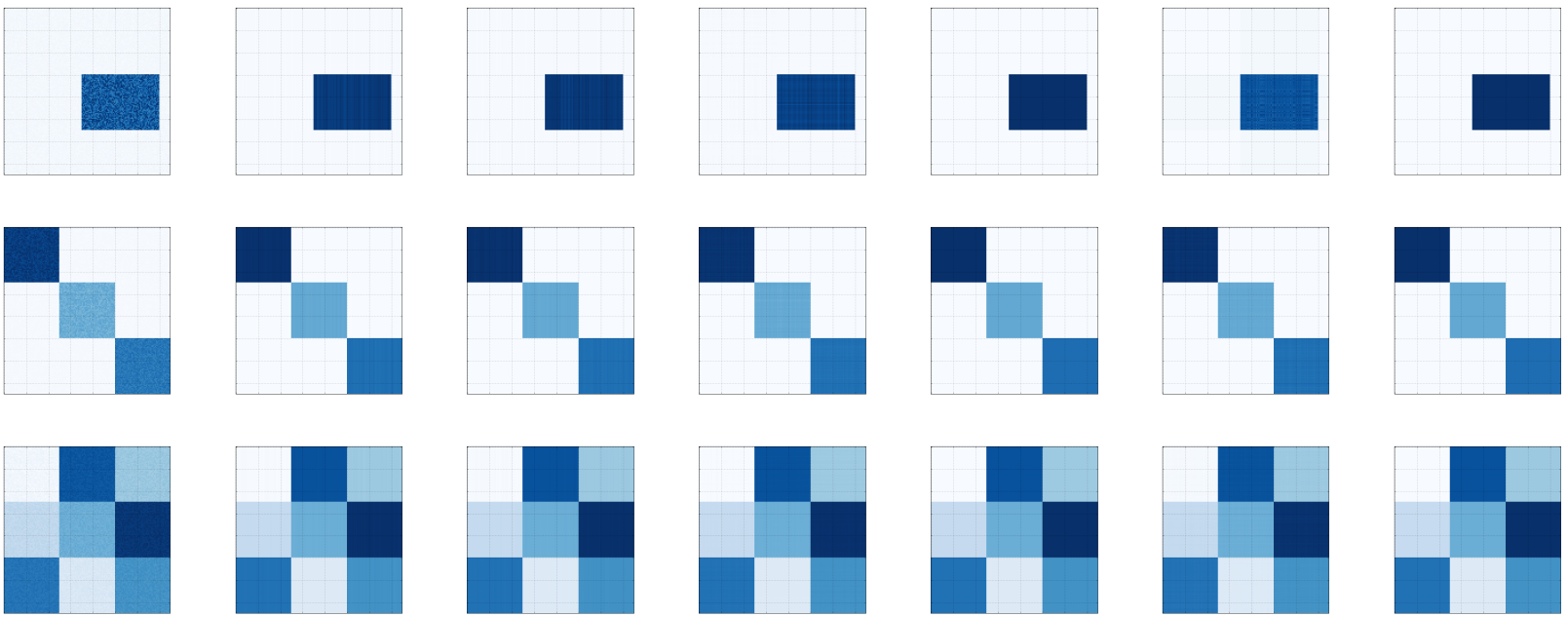
- Experimentos quantitativos com bases de dados sintéticas
  - análise da capacidade de reconstrução
  - análise da capacidade de quantização
- Experimentos quantitativos com bases de dados reais
  - Bases: *IG*, *IG toy*, *NIPS*
  - Análise da capacidade de agrupamento (Índice de Rand e NMI)
- Experimentos qualitativos com *IG toy*
  - Análise da capacidade de extração de informação (interpretabilidade)

Experimentos com bases de dados sintéticas



(a) (b) (c) (d) (e)

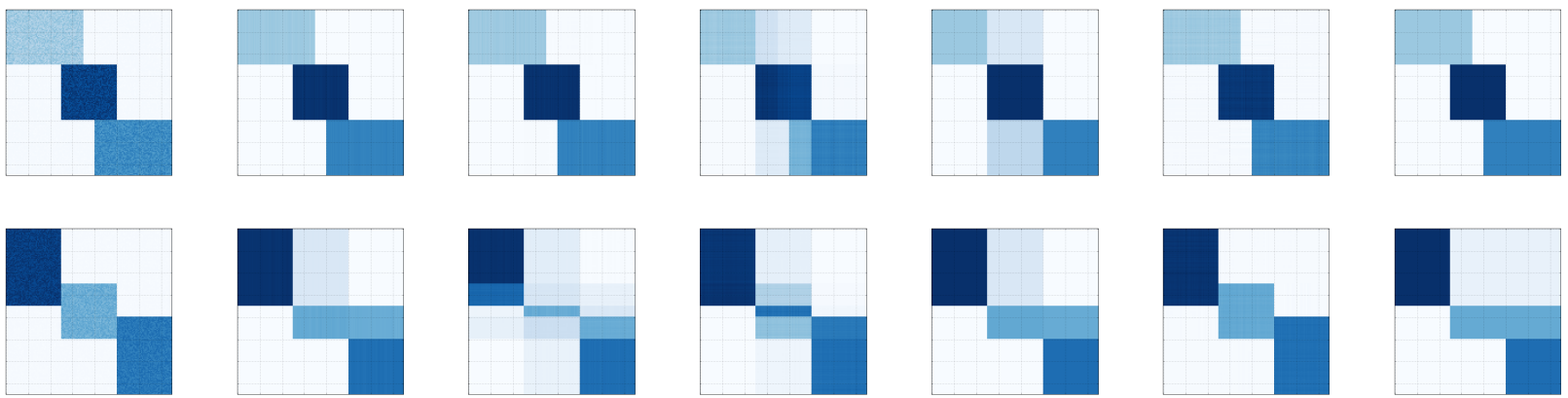
Reconstrução para as bases (a), (b) e (c)



original kmeans fkmeans ONTMF FNMTF OvNMTF BOvNMTF

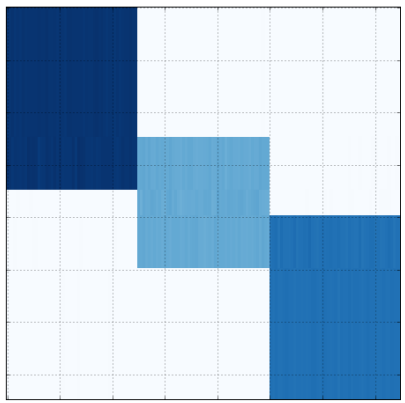
Reconstrução para as bases (d) e (e)

- $k = l = 3$ (*kmeans*, *fkmeans*, *ONMTF* e *FNMTF*)
- $k = 3$ e $l = 2$ (*OvNMTF* e *BinOvNMTF*)

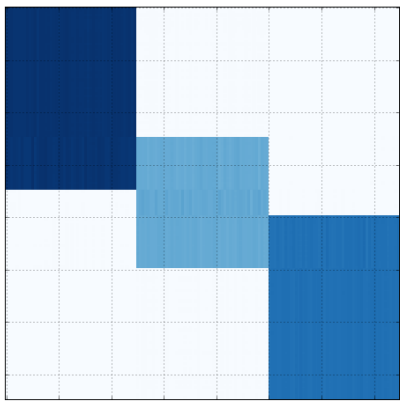


Reconstrução a partir dos resultados do algoritmo *k-means* e *fuzzy k-means*

- $k = 5$ para (e)



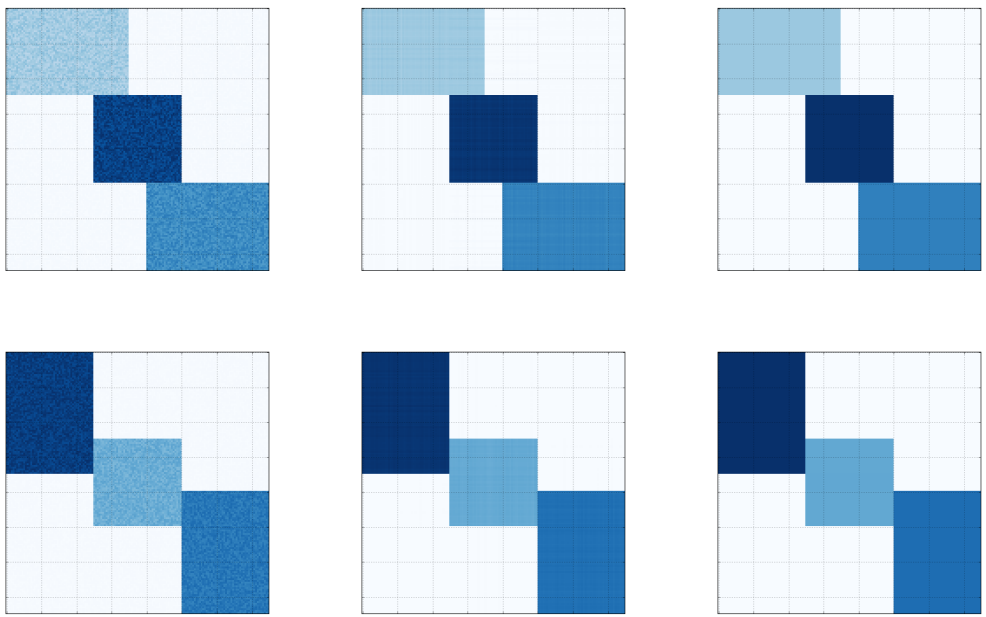
k-means



fuzzy k-means

Reconstrução a partir dos resultados do algoritmo *ONMTF* e *FNMTF*

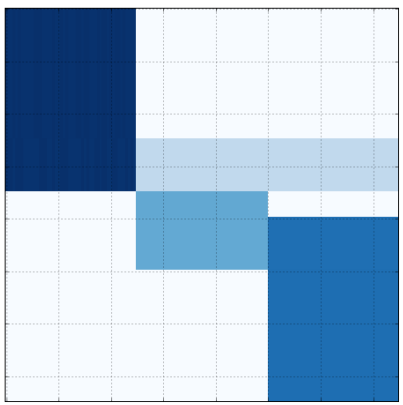
- $k = 5$ para (d)
- $l = 5$ para (e)



original ONTMF FNMTF

Reconstrução a partir dos resultados do algoritmo *BinOvNMTF*

- $k = 5$ para (e)



Análise da capacidade de quantização

	base (a)	**base (b)**	**base (c)**	**base (d)**	**base (e)**
k-means	30.336, 0	62.776, 5	184.992, 8	**79.238, 5**	2.886.245, 5
fuzzy k-means	29.402, 8	63.768, 5	183.991, 4	**78.340, 0**	2.307.168, 6
ONMTF	30.555, 4	60.794, 8	184.255, 9	579.136, 7	781.131, 6
FNMTF	30.924, 7	64.636, 1	186.224, 4	1.634.328, 5	2.881.172, 0
OvNMTF	30.439, 8	61.863, 2	178.886, 8	**75.533, 6**	**75.931, 2**
BinOvNMTF	31.239, 0	63.660, 4	187.579, 8	**79.968, 0**	3.160.391, 0

- *OvNMTF* preserva informação de sobreposição
- *BinOvNMTF* preserva informação de sobreposição de colunas (d)
- *ONMTF* preserva parte da informação de sobreposição

Experimentos com bases de dados reais

Bases de dados

- *NIPS*
  - Trabalhos acadêmicos do período de 2001 a 2003
  - Rotulados por áreas técnicas de forma desbalanceada
  - Foram usadas 9 das 13 áreas técnicas
- *IG*
  - Notícias publicadas no período de 2 de janeiro de 2012 à 11 de outubro de 2014
  - Rotulados por canal que compreende o assunto da notícia
  - Notícias distribuídas em 13 canais de forma desbalanceada
  - Notícias com mais de 200 caracteres no corpo

Base de dados	**# Palavras únicas**	**# Total de palavras**	**# Documentos**	**# Grupos**	**Esparsidade**
NIPS14-17	6.881	746.826	555	9	0,804
IG	19.563	1.187.334	4.593	13	0,987
IG toy	6.764	70.169	300	3	0,965

- Pré-processamento *IG*: tokenização, *stopwords*
- Representação textual: TF, TF-IDF, TF-normalizado, TF-IDF-normalizado

Análises quantitativas

Configuração dos experimentos

- Número de grupos de documentos:
  - *NIPS*: $k = 9$
  - *IG*: $k = 13$
  - *IG toy*: $k = 3$
- Número de grupos de termos (algoritmos *ONMTF*, *FNMTF*, *OvNMTF* e *BinOvNMTF*):
  - *NIPS*: $l \in \{6, 9, 12, 15, 18\}$
  - *IG*: $l \in \{7, 10, 13, 16, 19\}$
  - *IG toy*: $l \in \{2, 3, 4, 5, 6\}$
- Representações: *tf*, *tfidf*, tf_{norm} e $tfidf_{norm}$
- 10 rodadas para cada combinação de parâmetros

Análises quantitativas - Resultados IG toy

Índice de Rand médio ($k = 3$)

Algoritmo	tf	tf_{norm}	$tfidf_{norm}$	$tfidf$
k-means	$0,7017$	$0,7086$	$0,4701$	$0,3869$
fuzzy k-means	$0,4966$	$0,4970$	$0,4673$	$0,4390$
ONMTF	$0,3372 : l = 5$	$0,6479 : l = 5$	$0,5717 : l = 3$	$0,1758 : l = 3$
FNMTF	$0,2615 : l = 3$	$0,2590 : l = 3$	$0,1535 : l = 6$	$0,1543 : l = 3$
OvNMTF	$\mathbf{0,7466} : l = 4$	$\mathbf{0,7487} : l = 3$	$\mathbf{0,6755} : l = 6$	$\mathbf{0,6674} : l = 6$
BinOvNMTF	$0,4360 : l = 5$	$0,4818 : l = 5$	$0,2943 : l = 5$	$0,4079 : l = 3$

Informação mútua normalizada média ($k = 3$)

Algoritmo	tf	tf_{norm}	$tfidf_{norm}$	$tfidf$
k-means	$0,7169$	$0,7134$	$0,5467$	$0,4668$
fuzzy k-means	$0,0694$	$0,5421$	$0,4701$	$0,0836$
ONMTF	$0,3704 : l = 5$	$0,6720 : l = 5$	$0,6411 : l = 3$	$0,2143 : l = 3$
FNMTF	$0,2770 : l = 3$	$0,2734 : l = 3$	$0,2039 : l = 6$	$0,1690 : l = 3$
OvNMTF	$\mathbf{0,7257} : l = 4$	$\mathbf{0,7288} : l = 3$	$\mathbf{0,7033} : l = 6$	$\mathbf{0,6964} : l = 6$
BinOvNMTF	$0,4975 : l = 5$	$0,5500 : l = 5$	$0,3559 : l = 5$	$0,4343 : l = 3$

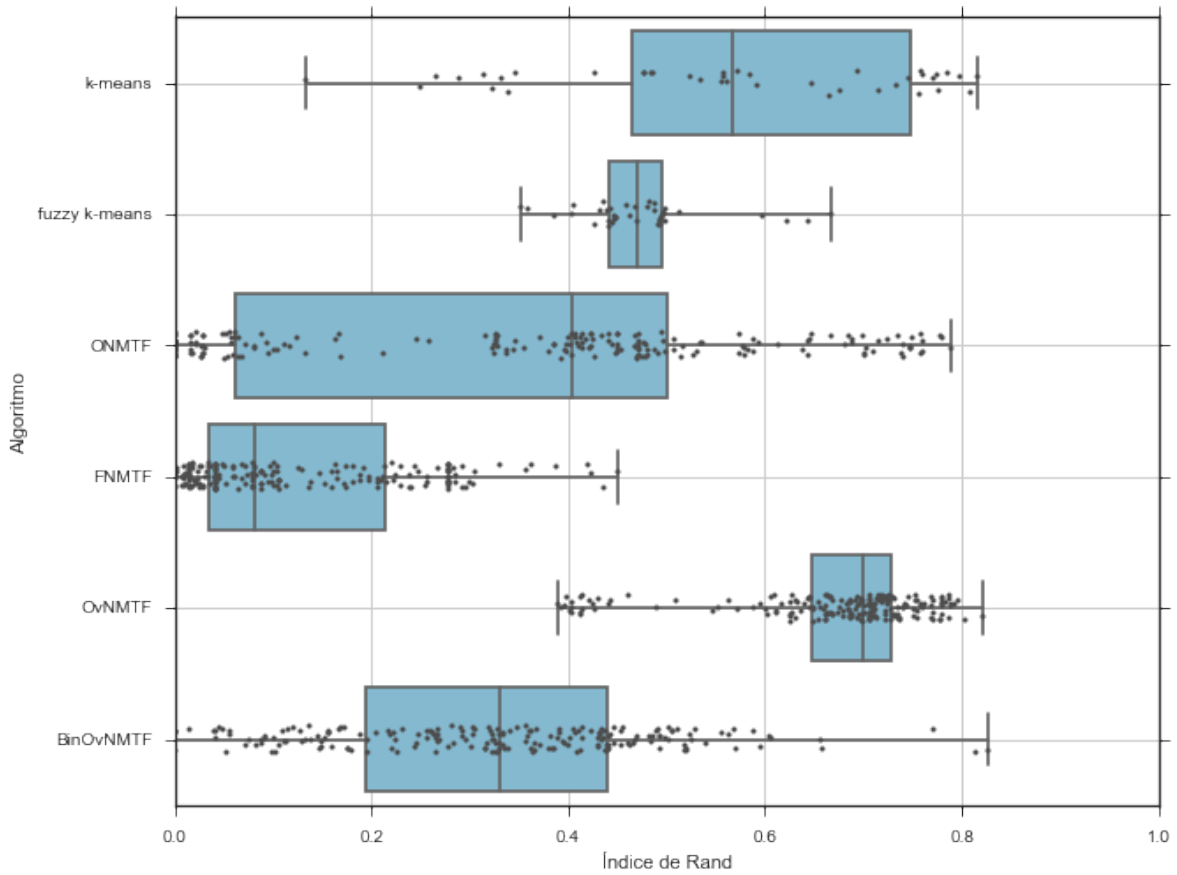
Análises quantitativas - Resultados IG toy

Melhores (máximos) resultados ($k = 3$)

Algoritmo	**Índice de Rand**	**Informação Mútua Normalizada**
k-means	$0,8152$: tf_{norm}	$\mathbf{0,8110}$: tf_{norm}
fuzzy k-means	$0,6669$: tf_{norm}	$0,6785$: tf
ONMTF	$0,7885$: $l = 5$, $tfidf_{norm}$	$0,7719$: $l = 5$, $tfidf_{norm}$
FNMTF	$0,4502$: $l = 4$, $tfidf$	$0,5041$: $l = 4$, $tfidf$
OvNMTF	$0,8208$: $l = 2$, tf_{norm}	$0,7855$: $l = 2$, tf_{norm}
BinOvNMTF	$\mathbf{0,8261}$: $l = 3$, $tfidf$	$0,8024$: $l = 3$, $tfidf$

Análises quantitativas - Resultados IG toy

Distribuições dos valores do índice de Rand



Análises quantitativas - Resultados IG

Índice de Rand médio ($k = 13$)

Algoritmo	tf	tf_{norm}	$tfidf_{norm}$	$tfidf$
k-means	$0,3137$	$0,3049$	$0,2750$	$0,2784$
fuzzy k-means	$0,1694$	$0,1619$	$0,2429$	$0,2662$
ONMTF	$0,1437 : l = 16$	$0,1802 : l = 19$	$0,1184 : l = 7$	$0,1279 : l = 7$
FNMTF	$0,2327 : l = 19$	$0,2399 : l = 19$	$0,2165 : l = 19$	$0,2124 : l = 13$
OvNMTF	$0,3384 : l = 10$	$0,3455 : l = 16$	$\mathbf{0,3554} : l = 16$	$\mathbf{0,3534} : l = 7$
BinOvNMTF	$\mathbf{0,3784} : l = 16$	$\mathbf{0,3591} : l = 7$	$0,2807 : l = 19$	$0,2868 : l = 19$

Informação Mútua Normalizada média ($k = 13$)

Algoritmo	tf	tf_{norm}	$tfidf_{norm}$	$tfidf$
k-means	$0,5235$	$0,5240$	$0,5361$	$0,5350$
fuzzy k-means	$0,2548$	$0,2518$	$0,3769$	$0,3929$
ONMTF	$0,4186 : l = 19$	$0,4312 : l = 19$	$0,4338 : l = 19$	$0,4416 : l = 19$
FNMTF	$0,4412 : l = 19$	$0,4492 : l = 19$	$0,4518 : l = 19$	$0,4593 : l = 19$
OvNMTF	$0,4930 : l = 7$	$0,5001 : l = 16$	$\mathbf{0,5493} : l = 16$	$\mathbf{0,5451} : l = 7$
BinOvNMTF	$\mathbf{0,5563} : l = 16$	$\mathbf{0,5424} : l = 7$	$0,5423 : l = 19$	$0,5327 : l = 19$

Análises quantitativas - Resultados IG

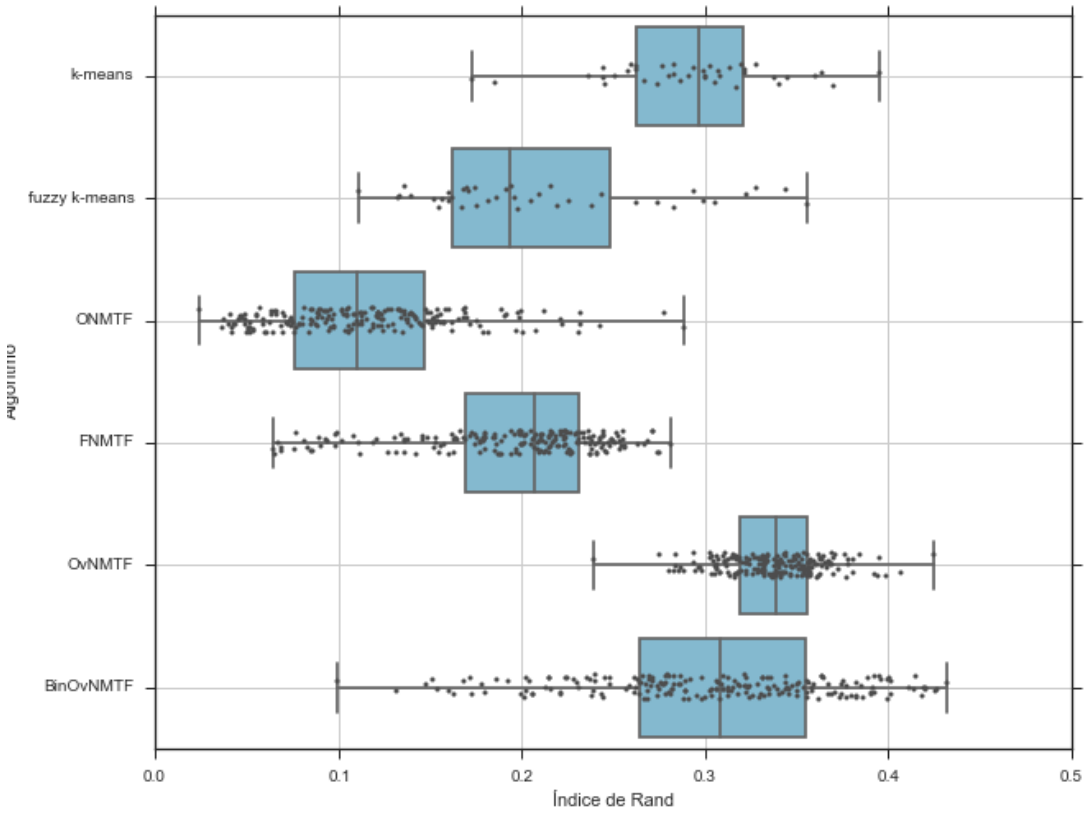
Melhores resultados ($k = 13$)

Algoritmo	**Índice de Rand**	**Informação Mútua Normalizada**
k-means	$0,3955$: tf	$0,5826$: $tfidf_{norm}$
fuzzy k-means	$0,3557$: $tfidf$	$0,4365$: $tfidf_{norm}$
ONMTF	$0,2884$: $l = 10$, tf_{norm}	$0,4938$: $l = 16$, $tfidf$
FNMTF	$0,2813$: $l = 16$, $tfidf_{norm}$	$0,5047$: $l = 19$, $tfidf$
OvNMTF	$0,4251$: $l = 10$, $tfidf_{norm}$	$0,5778$: $l = 16$, $tfidf_{norm}$
BinOvNMTF	$\mathbf{0,5743}$: $l = 10$, tf	$\mathbf{0,6064}$: $l = 7$, $tfidf_{norm}$

- *BinOvNMTF* apresenta melhores resultados quando k é grande

Análises quantitativas - Resultados IG

Distribuições dos valores do índice de Rand



Análises quantitativas - Resultados NIPS

Índice de Rand médio ($k = 9$)

Algoritmo	*tf*	tf_{norm}	$tfidf_{norm}$	*tfidf*
k-means	$0,1573$	$0,1527$	$0,1519$	$0,1368$
fuzzy k-means	$0,1223$	$0,1240$	$0,1736$	$0,1882$
ONMTF	$0,1579 : l = 6$	$0,1352 : l = 15$	$0,1318 : l = 9$	$0,1442 : l = 18$
FNMTF	$0,1293 : l = 18$	$0,1325 : l = 18$	$0,2128 : l = 18$	$0,2199 : l = 18$
OvNMTF	$0,1672 : l = 6$	$0,1641 : l = 9$	$0,1742 : l = 12$	$0,1711 : l = 9$
BinOvNMTF	$\mathbf{0,2247} : l = 9$	$\mathbf{0,2118} : l = 15$	$\mathbf{0,2811} : l = 6$	$\mathbf{0,2813} : l = 15$

Informação Mútua Normalizada média ($k = 9$)

Algoritmo	*tf*	tf_{norm}	$tfidf_{norm}$	*tfidf*
k-means	$0,3226$	$0,3106$	$0,3506$	$0,3476$
fuzzy k-means	$0,1876$	$0,1929$	$0,2575$	$0,2496$
ONMTF	$0,2930 : l = 15$	$0,2832 : l = 18$	$0,3361 : l = 18$	$0,3441 : l = 18$
FNMTF	$0,2272 : l = 18$	$0,2312 : l = 18$	$0,3109 : l = 18$	$0,3017 : l = 18$
OvNMTF	$0,3090 : l = 6$	$0,3092 : l = 9$	$0,3541 : l = 12$	$0,3493 : l = 15$
BinOvNMTF	$\mathbf{0,3255} : l = 15$	$\mathbf{0,3139} : l = 9$	$\mathbf{0,4013} : l = 12$	$\mathbf{0,4009} : l = 15$

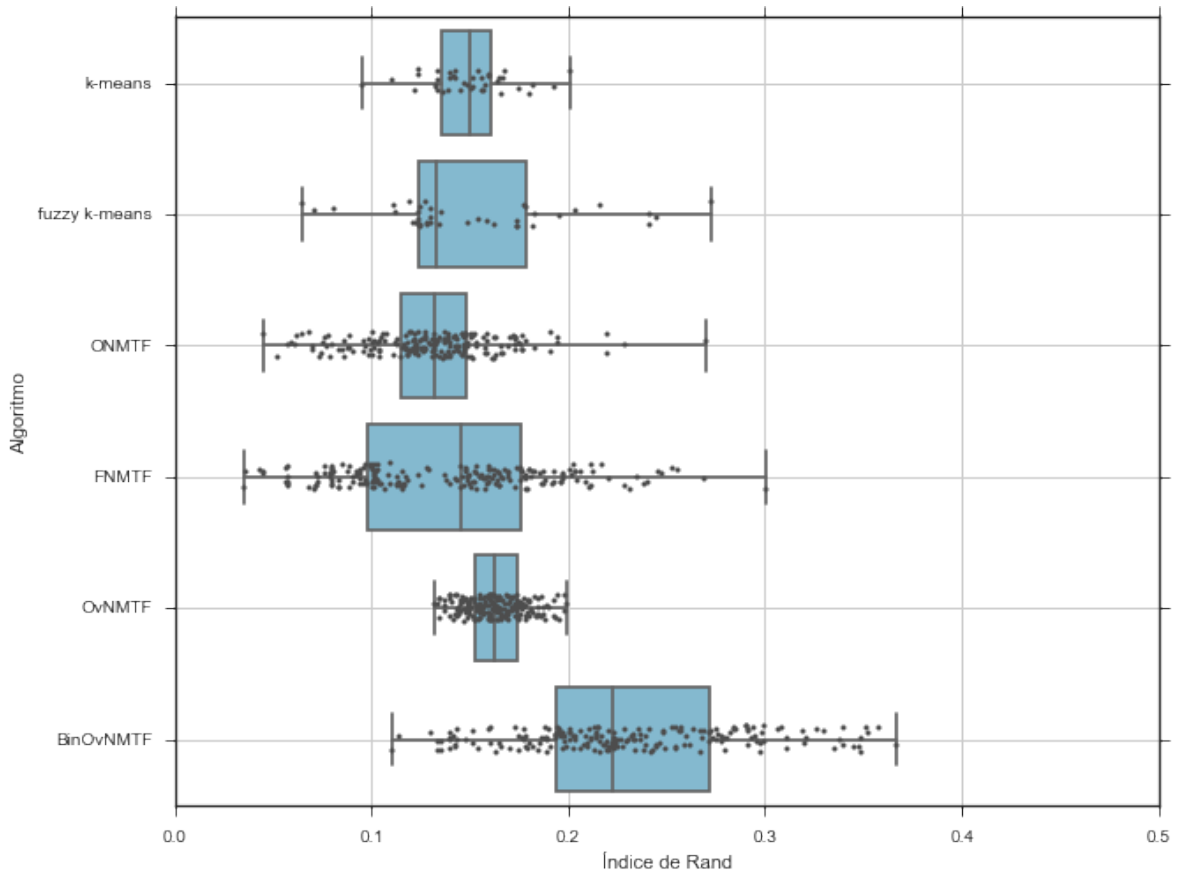
Análises quantitativas - Resultados NIPS

Melhores resultados ($k = 9$)

Algoritmo	**Índice de Rand**	**Informação Mútua Normalizada**
k-means	$0,2006$: $tfidf_{norm}$	$0,3952$: $tfidf_{norm}$
fuzzy k-means	$0,2728$: $tfidf$	$0,3115$: $tfidf$
ONMTF	$0,2704$: $l = 18$, $tfidf$	$0,3997$: $l = 18$, $tfidf$
FNMTF	$0,3009$: $l = 15$, $tfidf$	$0,3744$: $l = 18$, $tfidf_{norm}$
OvNMTF	$0,1992$: $l = 9$, $tfidf$	$0,3870$: $l = 9$, $tfidf$
BinOvNMTF	$\mathbf{0,3670}$: $l = 15$, $tfidf$	$\mathbf{0,4589}$: $l = 9$, $tfidf_{norm}$

Análises quantitativas - Resultados NIPS

Distribuições dos valores do Índice de Rand



Análises Qualitativas

- Algoritmos *ONMTF* e *OvNMTF*
- Base de dados *IG toy*
  - composto por 100 notícias de cada um dos três canais:
    - esportes
    - jovem
    - arena

Análises Qualitativas - ONMTF

- melhor modelo gerado na análise quantitativa:
  - $k = 3$
  - $l = 5$
  - representação $tfidf_{norm}$

Matriz S normalizada para o algoritmo *ONMTF* com $k = 3$ e $l = 5$

	CP #1	**CP #2**	**CP #3**	**CP #4**	**CP #5**
CN "esportes"	$0,0$	$\mathbf{0,5}$	$0,1$	$0,0$	$\mathbf{0,4}$
CN "arena"	$0,0$	$0,05$	$0,05$	$\mathbf{0,9}$	$0,0$
CN "jovem"	$\mathbf{0,4}$	$0,1$	$\mathbf{0,5}$	$0,0$	$0,0$

Análises Qualitativas - ONMTF

Top-15 palavras para cada cogrupo de palavras

CP #1 *"esportes radicais"*	**CP #2** *"futebol"*	**CP #3** *"esportes em geral"*	**CP #4** *"games"*	**CP #5** *"futebol"*
skate	real	anos	jogos	gol
surfe	breno	mundial	xbox	madrid
bob	time	brasil	playstation	gols
burnquist	barcelona	etapa	wii	messi
circuito	partida	brasileiro	jogo	euro
games	equipe	jovem	of	guardiola
mineirinho	minutos	rio	console	ronaldo
slater	jogador	dias	sony	itália
rampa	campeonato	música	legends	cristiano
medina	liga	pedro	nintendo	bola
manobras	futebol	atleta	game	bayern
mega	casa	americano	league	pontos
megarampa	temporada	gente	one	espanhol
kelly	grupo	esporte	novo	zagueiro
prova	feira	campeão	ps	atacante

Análises Qualitativas - ONMTF

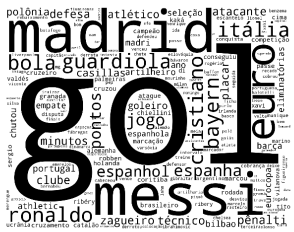
Visualização em nuvem de palavras das top-100 palavras

(a) CP #1 *"esportes radicais"*

(b) CP #2 *"futebol"*

(c) CP #3 *"esportes em geral"*

(d) CP #4 *"games"*



(e) CP #5 *"futebol"*

- palavra "jogo" com semântica de games vs semântica de futebol (interdependência)

Análises Qualitativas - OvNMTF

- melhor modelo gerado na análise quantitativa:
  - $k = 3$
  - $l = 2$
  - representação tf_{norm}

Matriz S normalizada para o algoritmo *OvNMTF* com $k = 3$ e $l = 2$

CN	**"arena"**	**0.38**	*CP #1*	**0.62** *CP #2*
CN	**"jovem"**	**0.46**	*CP #3*	**0.54** *CP #4*
CN	**"esportes"**	**0.94**	*CP #5*	0.06 *CP #6*

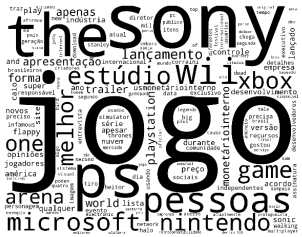
Análises Qualitativas - OvNMTF

Top-20 palavras para cada cogrupo de palavras

CP #1 *"games"*	**CP #2** *"games"*	**CP #3** *"esportes em geral"*	**CP #4** *"esportes radicais + música"*	**CP #5** *"futebol"*	**CP #6** *"futebol"*
jogos	jogo	games	anos	time	breno
sony	of	jovem	skate	real	casa
ps	playstation	brasileiro	mundial	feira	gol
the	game	paulo	brasil	final	bayern
pessoas	novo	dia	surfe	gols	minutos
wii	console	mundo	música	madrid	clube
microsoft	xbox	ano	rio	jogador	partida
nintendo	games	esporte	conta	tempo	técnico
estúdio	league	vai	dias	pontos	título
one	legends	janeiro	primeira	grupo	livre
arena	brasil	bem	final	liga	jogo
melhor	além	burnquist	atleta	fez	meia
apenas	nova	além	pessoas	brasileiro	volta
lançamento	jogadores	gente	casa	jogadores	segunda
versão	dia	bob	paulista	campo	técnica
opiniões	lançado	série	ficou	rodada	casillas
caio	usmonetáriointerno	etapa	fim	três	espanha
site	personagens	história	usmonetáriointerno	cristiano	equipe
and	feira	melhor	melhores	copa	argentino
forma	dois	circuito	amigos	deixe	semana

Análises Qualitativas - OvNMTF

Visualização em nuvem das top-100 palavras



(a) #1 "games"

(b) #2 "games"

(c) #3 "esportes em geral"

(d) #4 "esportes radicais + música"

(e) #5 "futebol"

(f) #6 "futebol"

- palavra "jogo" pode assumir semântica diferente (independência)

Seção 6

Conclusão

- Resultados mostraram a superioridade ou equivalência dos algoritmos propostos
  - Reconstrução e capacidade de quantização do espaço
  - Capacidade de agrupamento
  - Capacidade de agrupamento (nas bases de dados *IG toy*, *IG* e *NIPS*)
  - Geração de informação
- As fatorações propostas tem potencial para lidar com cogrupos com subreposição de colunas
  - independência de cogrupos de colunas

Contribuições

- proposição do problema e um algoritmo para *OvNMTF*, com derivação formal das regras de atualização
- proposição do problema e um algoritmo para *BinOvNMTF*, com derivação formal das regras de atualização
- interpretação semântica para o novo problema de fatoração
- construção das bases de dados de notícias *IG* e *IG toy* no idioma português (brasileiro)

Desvantagens, Limitações e Trabalhos Futuros

Desvantagens e Limitações:

- tempo de execução dos algoritmos
- validade dos resultados considerando testes estatísticos
- convergência dos algoritmos desenvolvidos

Trabalhos Futuros:

- estudar a determinação do parâmetro k
- estudar com mais detalhes a vantagem da independência dos cogrupos de colunas diante de contextos de aplicação
- estudar o efeito da estratégia de aplicação de múltiplas matrizes ao problema *NMF*
- estudar com maior profundidade as restrições de ortogonalidade dos algoritmos

Fatoração de Matrizes no problema de Coagrupamento com sobrepreposição de colunas

Lucas Fernandes Brunialti
Orientadora: Profa. Dra. Sarajane Marques Peres

Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

lucas.brunialti@usp.br
sarajane@usp.br

31 de agosto de 2016