Mineração de Textos

Cristiane Neri Nobre

Mineração de textos

Os estudos em Aprendizado de Máquina normalmente trabalham com dados **estruturados**

- ✓ Por exemplo, uma tabela em um banco de dados relacional

Entretanto, uma grande quantidade de informação é armazenada em textos, que são dados s**emiestruturados**

Podemos dizer que a **mineração de texto** é uma especialização da **mineração de dados**

Mineração de textos

Uma grande quantidade de toda informação disponível atualmente encontra-se sob a forma de textos (ou documentos) semiestruturados, tais como livros, artigos, manuais, e-mails e a Web

- ✓ Estima-se que mais de 80% de todas as informações eletrônicas produzidas no mundo sejam de dados não estruturados

O termo **semiestruturado** indica que os dados não são completamente estruturados nem completamente sem estrutura

Mineração de textos

Um documento pode conter alguns atributos estruturados:

- Título, autor(es), data da publicação, assunto

mas também contém alguns elementos textuais sem estrutura

- Resumo e conteúdo

Mineração de textos

Mineração de Textos (Text Mining - TM)
tem como objetivo tratar essa **informação semistruturada**

Apesar desta fonte de recursos ser atrativa e de fácil acesso, a extração automática de informação útil a partir dela é um desafio uma vez que os resumos estão em linguagem natural

Mineração de textos

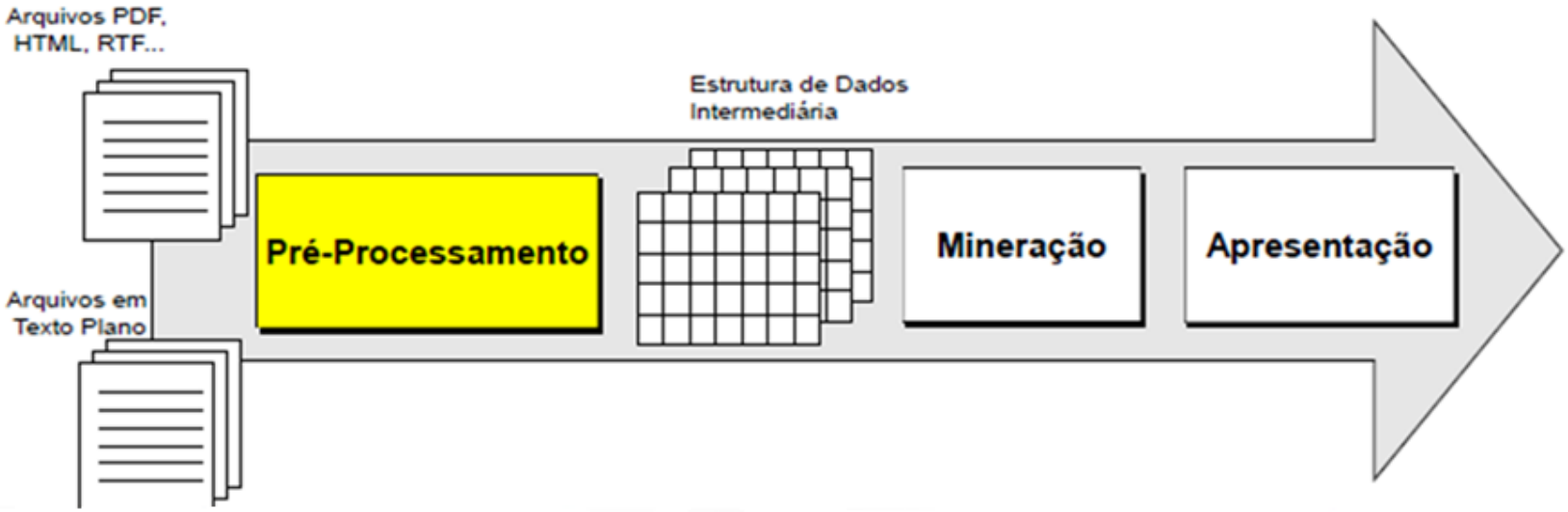
O objetivo da **Mineração de Textos** é o processamento de informação textual, extraindo **índices numéricos** significativos a partir do texto e então tornar esta informação acessível para os programas disponíveis nos sistemas de mineração de dados.

Mineração de textos

Podem ser analisadas palavras, agrupamentos de palavras, ou mesmo documentos entre si através das suas similaridades ou de suas relações com outras variáveis de interesse num projeto de mineração de textos.

Mineração de textos

O objetivo na fase inicial do projeto é “transformar textos em números (índices significativos)”, que podem então ser incorporados em outras análises tais como problemas **supervisionados** ou **não supervisionados**.



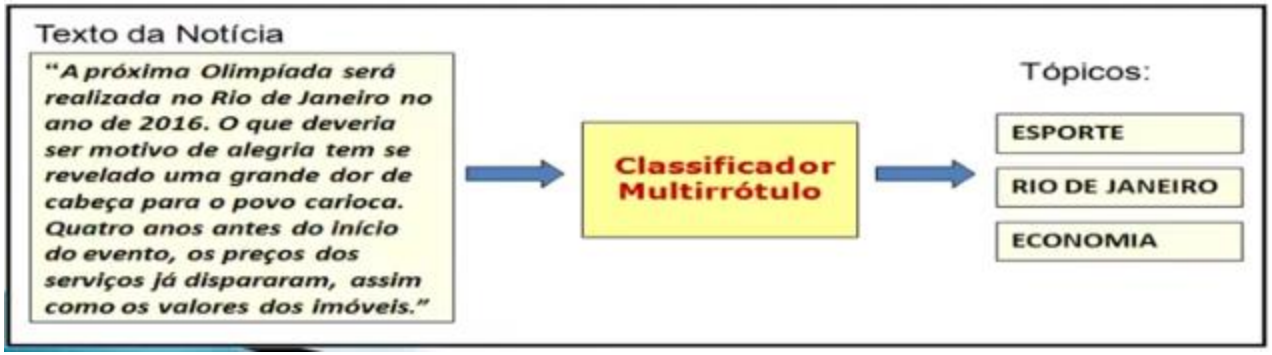
Mineração de textos

Algumas **aplicações típicas** para mineração de textos:

- Processamento automático de mensagens, “e-mails”, conteúdo de redes sociais, etc
- Análise de sentimentos
- Classificação de documentos
- Detecção de fraudes
- Filtros de Spam
- Análise de questões abertas em questionários

Mineração de textos

Problemas de textos, vistos como problemas de classificação:



Mineração de textos

Mais apropriado para um grande número de textos de tamanho médio ou pequeno.

Não deve ser tratado como uma caixa preta.

- A **intervenção do analista** é necessária.

Soluções não podem ser importadas de outra língua.

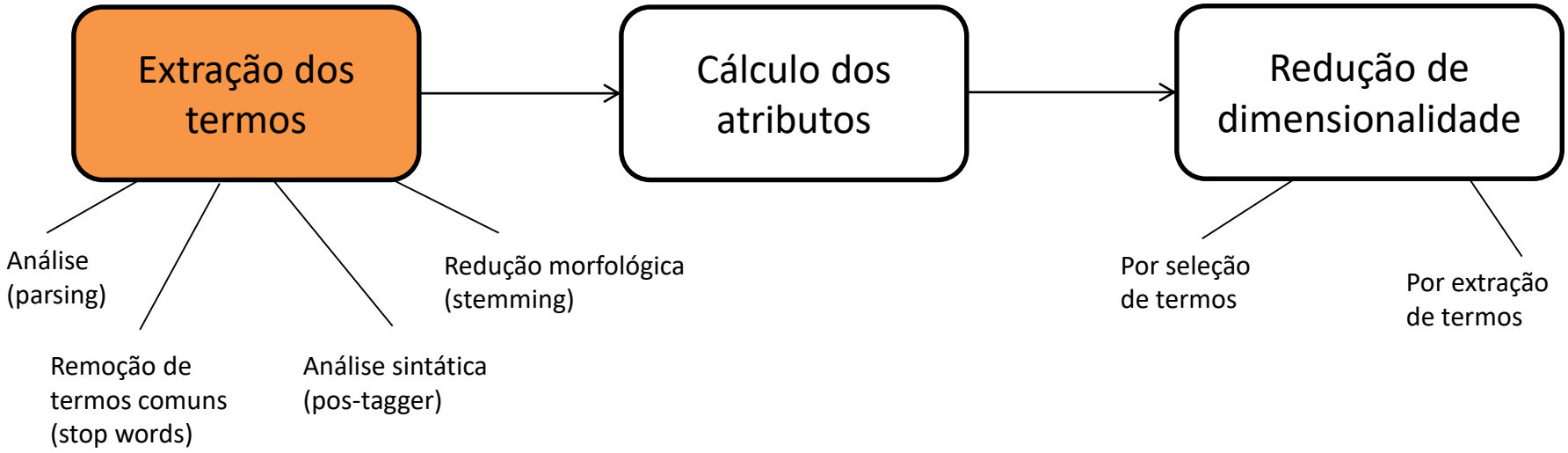
Abordagens da Mineração de textos

Estatística
Frequência dos termos, ignorando informações semânticas

Processamento de linguagem natural.
Interpretação sintática e semântica das frases

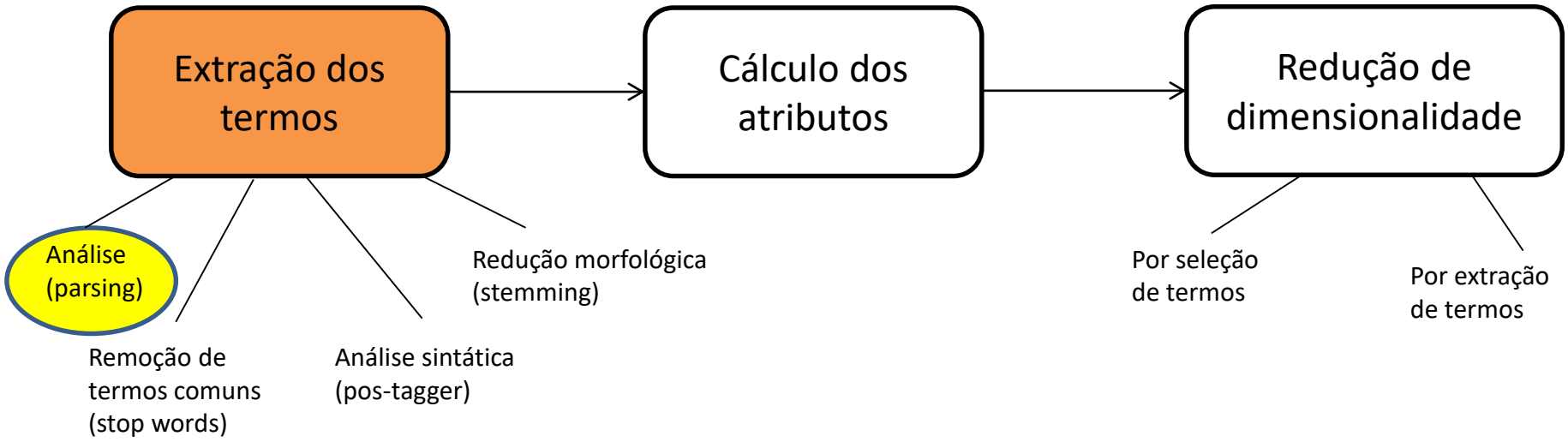
Mineração de textos

O que deverá ser feito na etapa de pré-processamento do texto?



Mineração de textos

O que deverá ser feito na etapa de pré-processamento do texto?



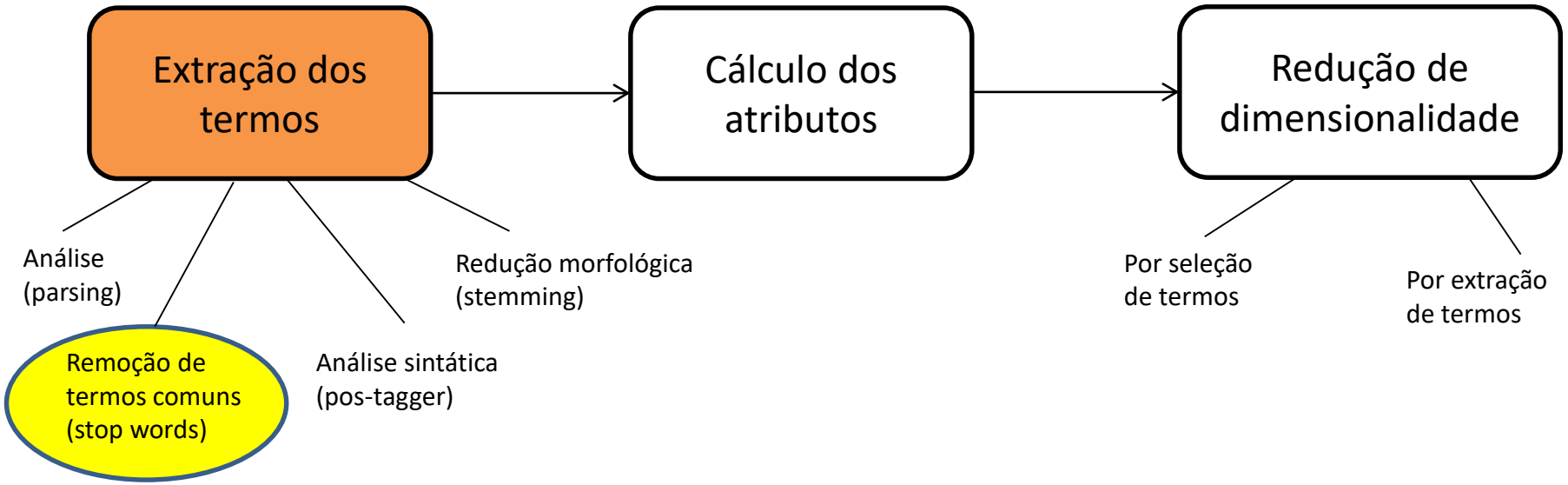
Mineração de textos – Análise de Parsing

Fragmentar o texto original com base no conceito de “termo” adotado

Remoção das marcações

Normalização da estrutura dos documentos fracamente estruturados

Mineração de textos



Mineração de textos – Remoção de termos comuns

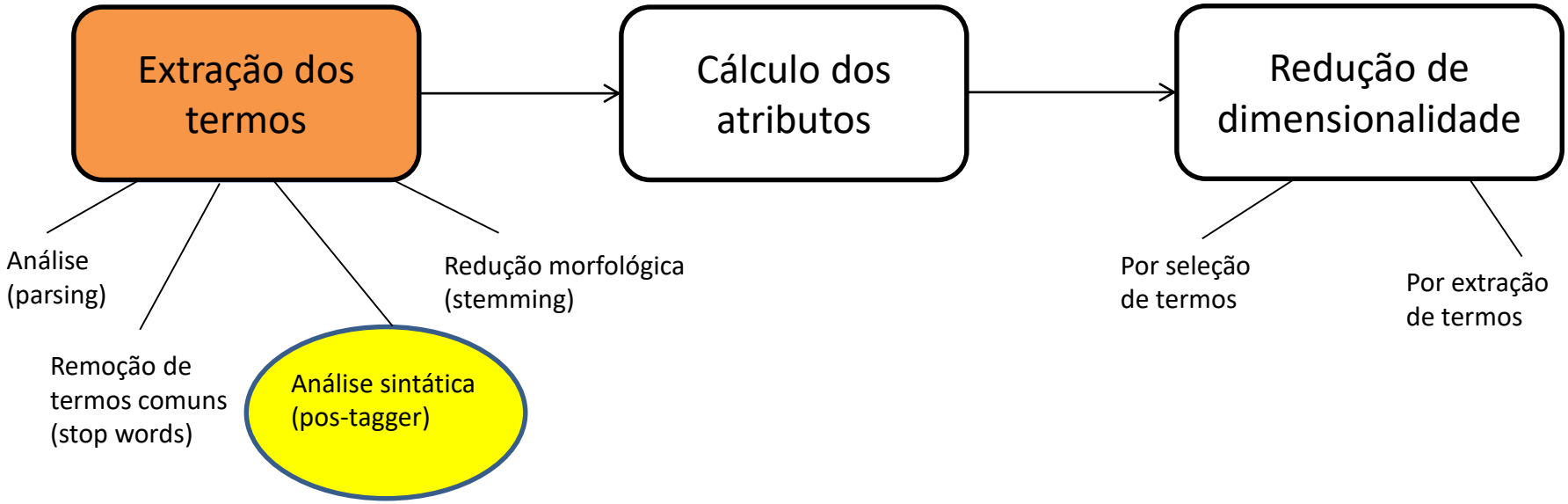
Um sistema de MT geralmente associa uma **stop list** com um conjunto de documentos

Uma **stop list** é um conjunto de palavras que são consideradas "irrelevantes"

- Normalmente inclui artigos, preposições, conjunções

A **stop list** pode variar entre conjuntos de documentos (mesma área, mesma língua)

Mineração de textos



Mineração de textos – Análise sintática

Definir o tipo gramatical dos termos presentes no vetor de termos.

Lidar com a ambiquidade – posição da palavra no texto

- ✓ A sua atitude prova o seu caráter (verbo)
- ✓ A prova estava difícil (substantivo)

Mineração de textos – Análise sintática

Um sistema de MT deve considerar a ocorrência de sinonímia e polissemia

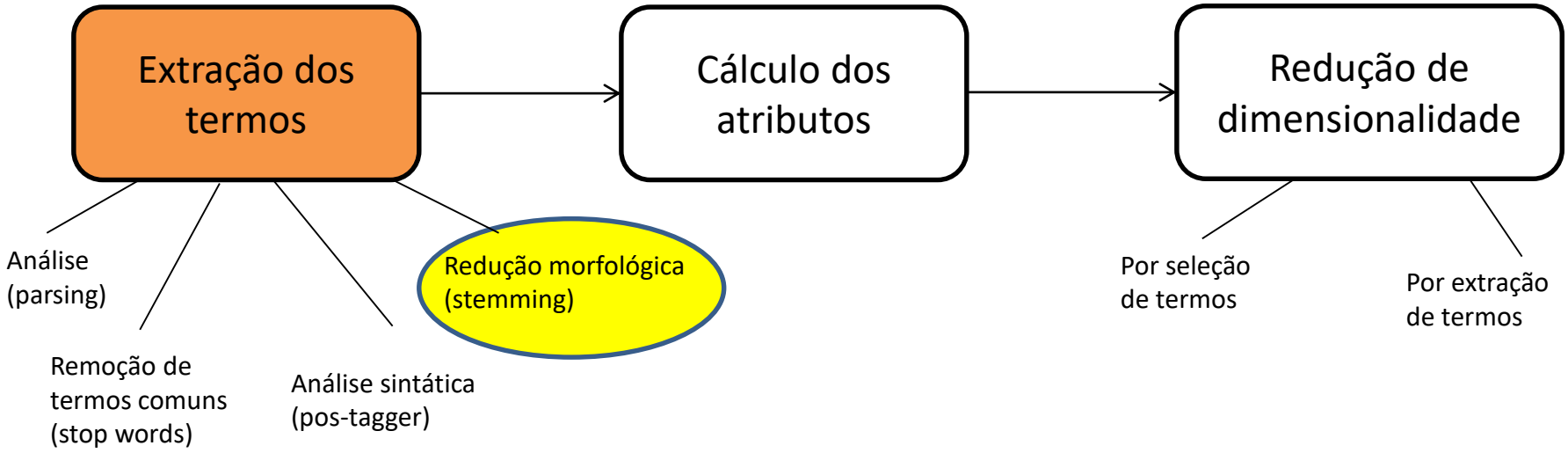
Sinonímia: uma palavra possui vários sinônimos

- Carro, automóvel, veículo

Polissemia: uma mesma palavra tem diferentes significados, dependendo do contexto

- Mineração (textos?), mineração (carvão?)
- Exame (teste?), exame (médico?)

Mineração de textos



Mineração de textos – Análise morfológica (Stem)

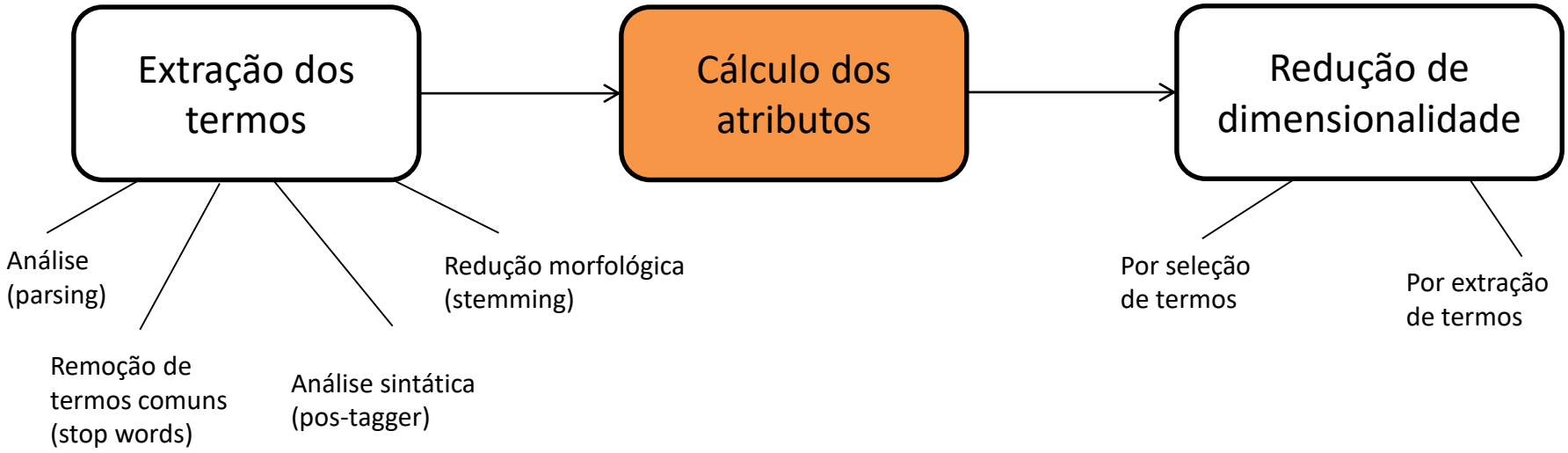
Um grupo de diferentes palavras podem compartilhar um mesmo radical (*stem*)

Um sistema de TM precisa identificar grupos de palavras nas quais as palavras em um mesmo grupo são pequenas variações sintáticas umas das outras

- Droga, drogas, drogado, drogaria

Com essa identificação, é possível armazenar apenas o *stem*

Mineração de textos



Mineração de textos – cálculo de atributos

É a determinação de quais atributos devem representar ou estar presentes na representação de um texto

Mineração de textos – cálculo de atributos

Iniciando com um conjunto de ***n*** documentos e ***t*** termos, é possível modelar cada documento como um vetor ***v*** no espaço ***t***-dimensional $\Re^t$

Os vetores podem ser binários, onde **0** indica que um determinado termo não ocorre no documento e **1** caso contrário

Os vetores podem conter a frequência (absoluta ou relativa) de cada termo no documento

Frequência de termos tf

Número de vezes que o termo *t* ocorre na coleção de documentos *d*

Frequência absoluta não é uma boa opção:

- Um documento com 10 ocorrências de um termo é mais relevante que somente uma ocorrência do termo.
- Mas não 10 vezes mais relevante!

Relevância não deve crescer proporcionalmente com frequência

Exemplo

• Exemplo: seja o *corpus* composto pelos três documentos (textos) :

D1: Este é um exemplo A.
D2: Este é um mostruário.
D3: Este é outro A, exemplo A

• De modo bem simples cada texto pode ser representado numericamente assim:

	A	**é**	**Este**	**exemplo**	**mostruário**	**outro**	**um**
D1:	1	1	1	1	0	0	1
D2:	0	1	1	0	1	0	1
D3:	1	1	1	1	0	1	0

Exemplo

• Se um termo é muito frequente no *corpus* inteiro ele deve ser pouco informativo para caracterizar textos individuais. Ex.: “é” e “este”.

Exemplo

- Problemas com esta representação?
    - Com textos de tamanhos muito diferentes ou com muitos termos distintos a maior parte dos valores serão iguais a zero!

- Obviamente, existem cálculos mais interessantes
    - Variações da frequência dos termos em um documento e no *corpus* como um todo...

Exemplo

- **TFxIDF (*Term Frequency* x *Inverse Document Frequency*)**

  - É uma forma de selecionar termos mais “importantes” ou menos importantes.

- **TF(i, D)**: número de vezes que o termo **i** aparece em relação ao total de termos do texto D.

- **IDF(i)**: *log* do número de textos no *corpus* dividido pelo número de textos que o termo **i** aparece.

Exemplo

A técnica estatística TF-IDF é utilizada no processo de mineração de texto e tem como principal utilidade descobrir palavras de importância em um texto não estruturado ou semi estruturado.

Atribui-se um valor a cada termo, que considera sua frequência no texto e em todos dos documentos da base, indicando sua importância.

Exemplo

- Voltando ao exemplo...

- Seja o *corpus* composto pelos três documentos:

D1: Este é um exemplo A.
D2: Este é um mostruário.
D3: Este é outro A, exemplo A.

Exemplo

- Para simplificar: tabela com a frequência de cada termo

	D1 (5 termos)	**D2** (4 termos)	**D3** (6 termos)
Termo	**Ocorrências**	**Ocorrências**	**Ocorrências**
A	1	0	2
é	1	1	1
Este	1	1	1
exemplo	1	0	1
mostruário	0	1	0
outro	0	0	1
um	1	1	0

D1:	Este é um exemplo A.
D2:	Este é um mostruário.
D3:	Este é outro A, exemplo A.

Exemplo

- **TFxIDF** de alguns termos
  - Um termo comum: “Este”

TF(“Este”, D1) = 1/5 = 0,2
TF(“Este”, D2) = 1/4 = 0,25
TF(“Este”, D3) = 1/6 = 0,17

IDF(“Este”) = log(3/3) = 0

TF(“Este”, D1) x **IDF**(“Este”) = 0,2 x 0 = 0
TF(“Este”, D2) x **IDF**(“Este”) = 0,25 x 0 = 0
TF(“Este”, D3) x **IDF**(“Este”) = 0,17 x 0 = 0

- **TF(i, D)**: número de vezes que o termo **i** aparece em relação ao total de termos do texto D.

- **IDF(i)**: *log* do número de textos no *corpus* dividido pelo número de textos que o termo **i** aparece.

D1: Este é um exemplo A.
D2: Este é um mostruário.
D3: Este é outro A, exemplo A.

Exemplo

- **TFxIDF** de alguns termos
  - Um termo mais raro: "**outro**"

TF("outro", D1) = 0/5 = 0
TF("outro", D2) = 0/4 = 0
TF("outro", D3) = 1/6 = 0,17

IDF("outro") = log(3/1) = 0,48

TF("outro", D1) x **IDF**("outro") = 0 x 0,48 = 0
TF("outro", D2) x **IDF**("outro") = 0 x 0,48 = 0
TF("outro", D3) x **IDF**("outro") = 0,17 x 0,48 = 0,08

- **TF(i, D)**: número de vezes que o termo **i** aparece em relação ao total de termos do texto D.

- **IDF(i)**: *log* do número de textos no *corpus* dividido pelo número de textos que o termo **i** aparece.

D1:	Este é um exemplo A.
D2:	Este é um mostruário.
D3:	Este é outro A, exemplo A.

Exemplo

- **TFxIDF** de alguns termos
  - Um termo de frequência mais variada: "**A**"

TF("A", D1) = 1/5 = 0,2
TF("A", D2) = 0/4 = 0
TF("A", D3) = 2/6 = 0,33

IDF("A") = log(3/2) = 0,18

TF("A", D1) x **IDF**("A") = 0,2 x 0,18 = 0,036
TF("A", D2) x **IDF**("A") = 0 x 0,18 = 0
TF("A", D3) x **IDF**("A") = 0,33 x 0,18 = 0,06

- **TF(i, D)**: número de vezes que o termo **i** aparece em relação ao total de termos do texto D.

- **IDF(i)**: *log* do número de textos no *corpus* dividido pelo número de textos que o termo **i** aparece.

D1:	Este é um exemplo A.
D2:	Este é um mostruário.
D3:	Este é outro A, exemplo A.

Exemplo

- Comparando o **TFxIDF** de alguns termos

Documento D1

	Ocorrências	TF	IDF	**TFxIDF**
“Este”	1	0,2	0	0
“outro”	0	0	0,48	0
“A”	1	0,2	0,18	0,036

Documento D2

	Ocorrências	TF	IDF	**TFxIDF**
“Este”	1	0,25	0	0
“outro”	0	0	0,48	0
“A”	0	0	0,18	0

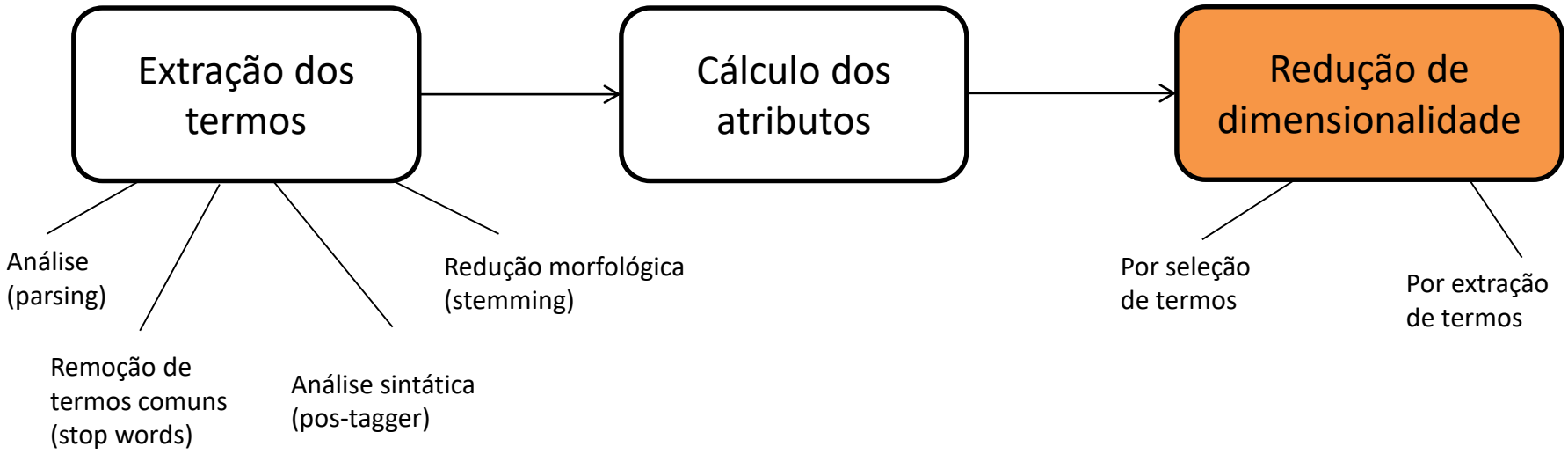
Documento D3

	Ocorrências	TF	IDF	**TFxIDF**
“Este”	1	0,17	0	0
“outro”	0	0,17	0,48	0,08
“A”	2	0,33	0,18	0,06

Exemplo

- Os textos são representados pelos valores **TFxIDF** de cada termo.
- **TFxIDF** igual a **zero** indica termo **não relevante**.
- **TFxIDF maior** indica termo **mais relevante**.
- Textos podem ser agrupados e categorizados com base no vetor de valores **TFxIDF**

Mineração de textos



Mineração de textos – redução de dimensionalidade

- O vetor de termos que representa os textos é bastante grande.

- Logo, é necessário reduzir o tamanho destes vetores.
  - Ou seja, diminuir o número de termos.

- Duas abordagens:
  - ✓ Seleção de termos
  - ✓ Extração ou remoção de termos

Mineração de textos

Softwares comerciais e abertos para Text Mining:

1. WEKA
2. Pacote NLTK (Natural Language Toolkit) do Python
3. SAS-Text Mining;
4. SPSS-Text Mining e Text Analysis para questionários;
5. STATISTICA Text Miner;
6. GATE - Natural Languagem Open Source;
7. GATE - Natural Languagem Open Source;
8. R-Language programming text mining;
9. Practical – text mining com Perl;
10. ODM – Oracle Data Mining;
11. Megaputeŕ s Text Analyst

Mineração de textos

Como fazer mineração de textos no WEKA?

Aplique o filtro não supervisionado **StringToWordVector** sobre os atributos (Expressão regular: [1-9]+.*)

Veja o vídeo:
https://www.youtube.com/watch?v=ycbGUfY8BzM

https://www.youtube.com/watch?v=8kH6C3CwxM0

Exercícios

Fazer os exercícios do **Tutorial 5: Text mining do link**

http://phdies.ing.unisi.it/corsi/matdid/26_WEKA.pdf

Veja também um tutorial explicando como se faz isso no WEKA, passo a passo:
https://medium.com/@karim_ouda/tutorial-document-classification-using-weka-aa98d5edb6fa

Referências

Slides baseados em:
http://professor.ufabc.edu.br/~ronaldo.prati/DataMining/Mineracao-Textos.pdf

https://www.researchgate.net/publication/289540770_Descoberta_de_Informacao_Atraves_da_Mineracao_de_Texto_-_Fundamentos_e_Aplicacoes

Links
https://www.youtube.com/watch?v=iIQax6NuRsg
https://www.youtube.com/watch?v=ycbGUfY8BzM&t=2s
https://www.youtube.com/watch?v=IY29uC4uem8
https://www.youtube.com/watch?v=HrixTPMOCD4

Artigo:
http://docplayer.com.br/40427433-Metodo-supervisionado-para-identificacao-de-duvidas-em-foruns-educacionais.html