

CLASSIFICAÇÃO MULTI-RÓTULO/MULTI-LABEL

Cristiane Neri Nobre

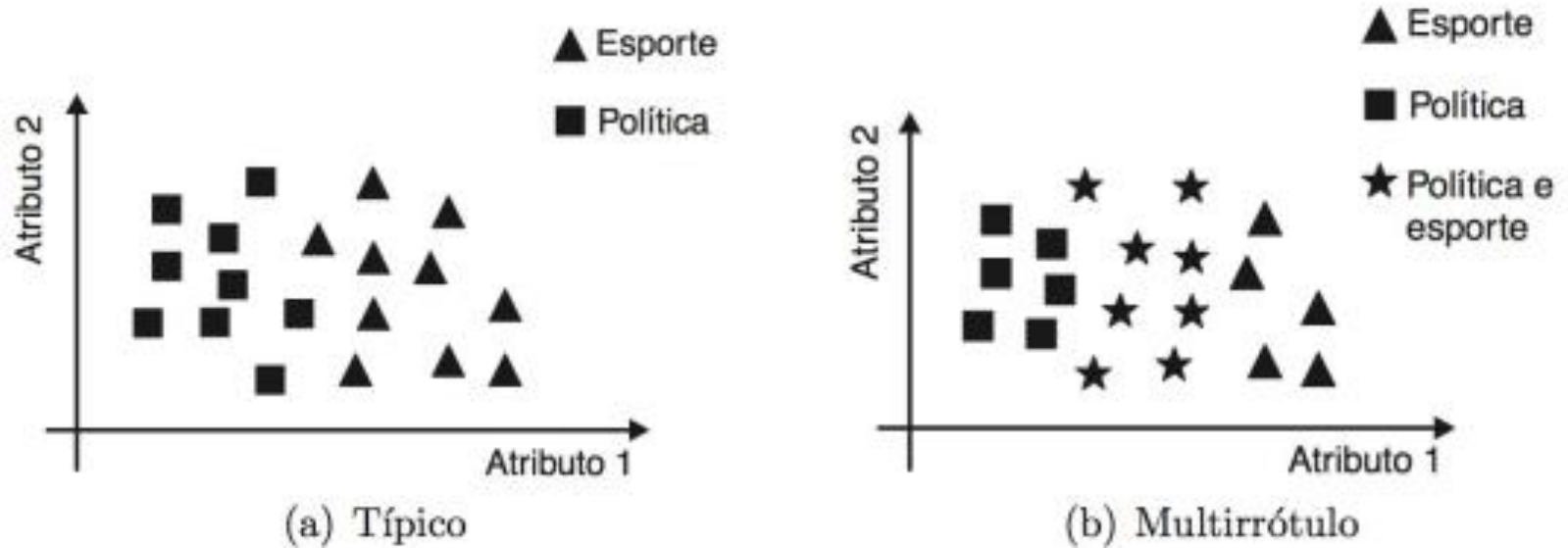
O que é classificação Multi-rótulo?

Problemas de **classificação mutirrótulo** é quando cada exemplo pode pertencer simultaneamente a mais de uma classe.

Algumas aplicações:

1. Classificação de texto: um documento pode pertencer área de CC e Física, por exemplo
2. Uma proteína pode executar uma ou mais funções
3. Um paciente pode estar com diabetes e gripe ao mesmo tempo

Ilustração de um problema de classificação Multi-rótulo



Fonte: Extraído de Facelli et al, 2015

Ilustração de um problema de classificação Multi-rótulo

X	Metal	Jazz	Bossa	Pop
X ₁	•			•
X ₂		•	•	
X ₃		•		
X ₄	•			
X ₅		•	•	•



X	Y
X ₁	Metal-Pop
X ₂	Jazz-Bossa
X ₃	Jazz
X ₄	Metal
X ₅	Jazz-Bossa-Pop

Competição no Kaggle na área de classificação multi-rótulo

A competição [Toxic Comment Classification](#) do Kaggle trata de um problema de **classificação de texto**, mais precisamente de classificação de comentários tóxicos.

Os participantes devem criar um modelo *multi-label* capaz de detectar diferentes tipos de toxicidade nos comentários, como ameaças, obscenidade, insultos e ódio baseado em identidade.

Base de dados

Veja o link da base:

<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data>

Características desta base de dados

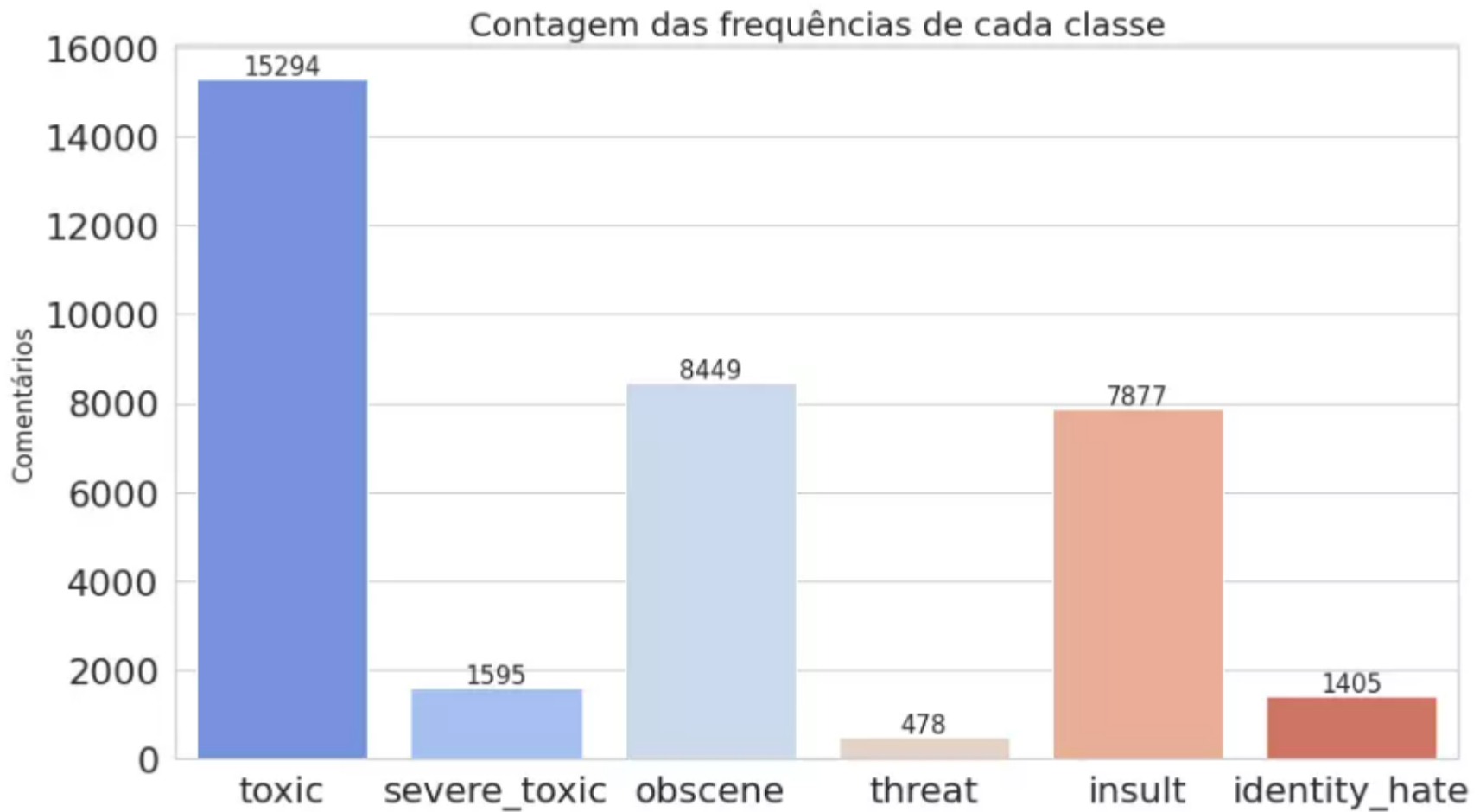
Número de instâncias: 159571

Número de atributos: 8

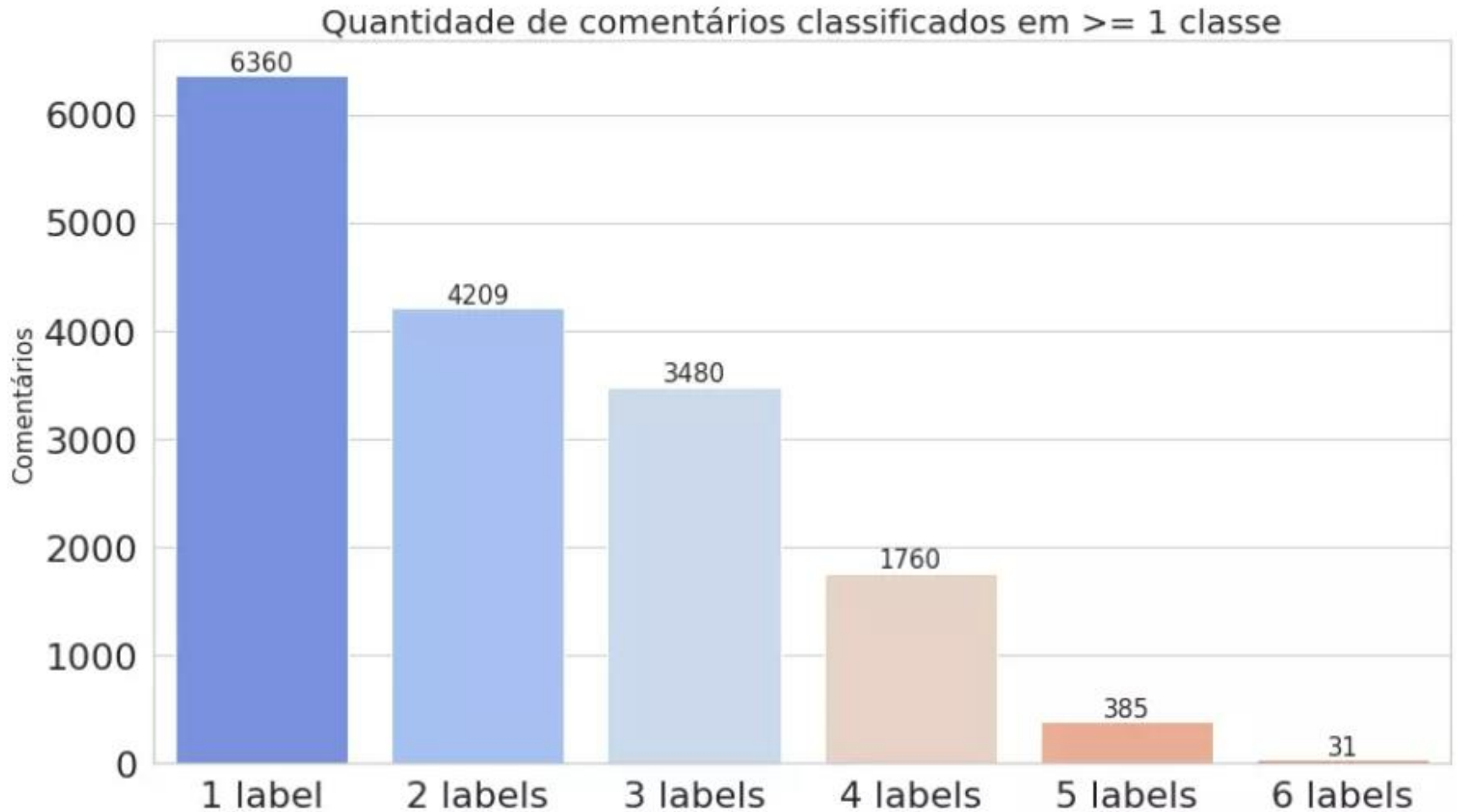
	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0
5	00025465d4725e87	"\n\nCongratulations from me as well, use the ...	0	0	0	0	0	0
6	0002bcb3da6cb337	AROUND ON MY WORK	1	1	1	0	1	0

Fonte: Extraído de <https://www.insightlab.ufc.br/aprenda-a-estratificar-dados-multi-label-com-scikit-multilearn/>

Características desta base de dados



Características desta base de dados



Etapas importantes de pré-processamento

1. Converter o texto apenas para letras minúscula
Remover as URLs
2. Remover pontuação
3. Remover números
4. Remoção das *stopwords*
5. Separar o texto em tokens – palavras ou grupos de palavras
6. etc

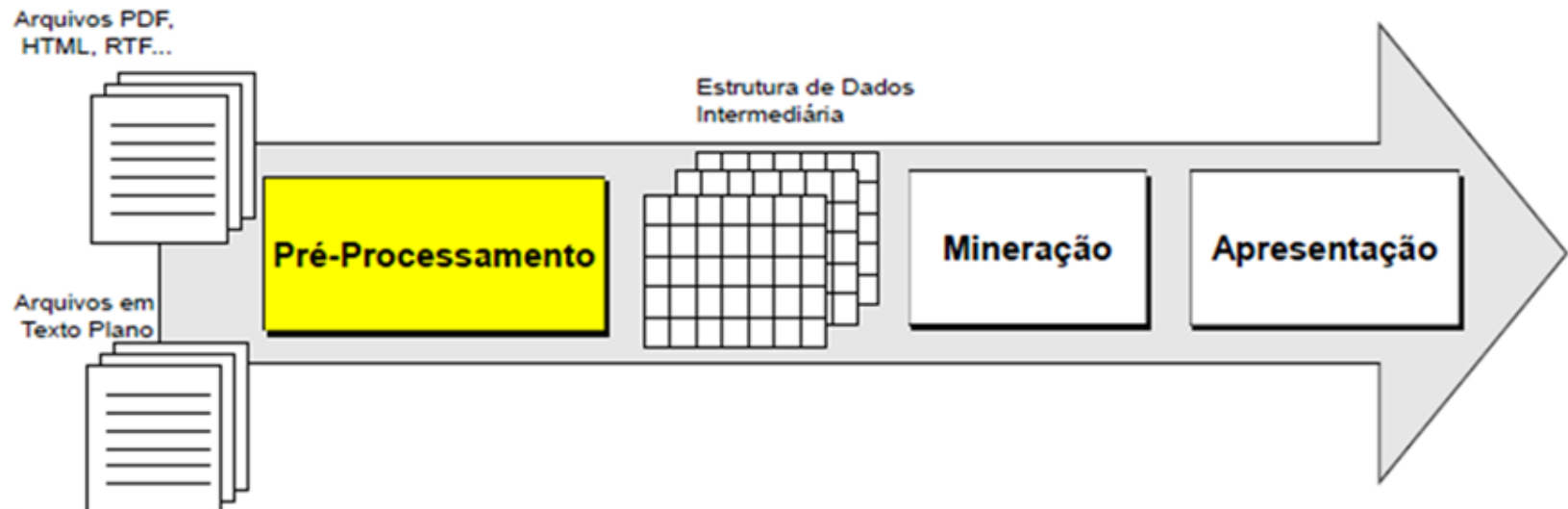
Etapas importantes de pré-processamento

	<code>comment_text</code>	<code>text_tokens</code>
0	Explanation\nWhy the edits made under my usern...	[explanation, edits, made, username, hardcore,...
1	D'aww! He matches this background colour I'm s...	[aww, matches, background, colour, seemingly, ...
2	Hey man, I'm really not trying to edit war. It...	[hey, man, really, trying, edit, war, guy, con...
3	"\nMore\nI can't make any real suggestions on ...	[make, real, suggestions, improvement, wondere...
4	You, sir, are my hero. Any chance you remember...	[sir, hero, chance, remember, page]

Breve introdução à mineração de texto

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**.

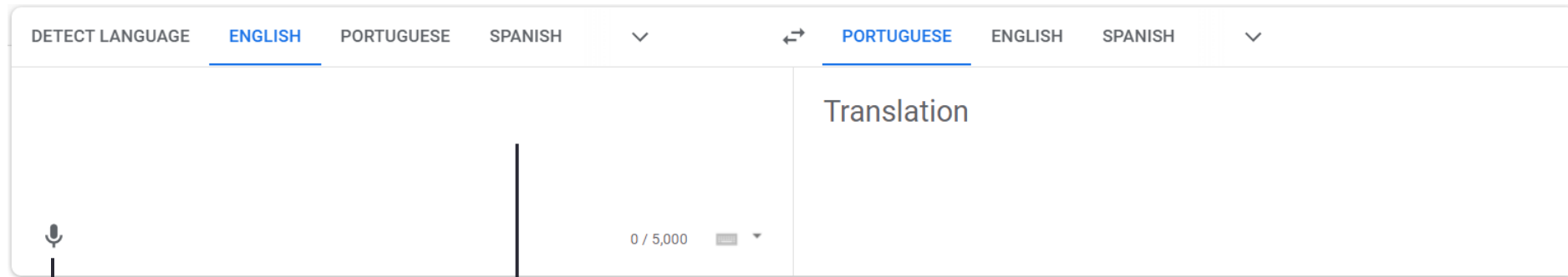


Mineração de textos

Algumas **aplicações típicas** para mineração de textos
(Processamento de Linguagem Natural - PLN):

- 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

Processamento de Linguagem Natural - PLN



The image shows a screenshot of the Google Translate web interface. At the top, there are tabs for 'DETECT LANGUAGE', 'ENGLISH', 'PORTUGUESE', and 'SPANISH'. The 'ENGLISH' tab is selected. To the right, there are tabs for 'PORTUGUESE', 'ENGLISH', and 'SPANISH'. The 'PORTUGUESE' tab is selected. Below the tabs, there is a large input field on the left and a large output field on the right. The output field is labeled 'Translation'. In the input field, there is a microphone icon on the left and a character count '0 / 5,000' on the right. Two arrows point from the input field to labels below: one from the microphone icon to 'Speech transcription' and one from the center of the input field to 'Tradução de texto'.

Speech
transcription

Tradução de
texto

Fonte: <https://translate.google.com.br/?hl=en>

Processamento de Linguagem Natural – PLN



Untitled document

Restrictions on the model interpretability generate numerous problems for those involved. For users, minority groups can be discriminated against, injustices may not be perceived, and the people harmed still have few resources to argue. Furthermore, in most cases, the entities that own the systems cannot explain how the decisions were made due to the opacity of systems. In high-risk scenarios, such as healthcare, the decision-maker feels insecure without explaining the model's results.

4 All suggestions

- systems · Correct article usage

• CLARITY

BETA

Consider adding a transition phrase to improve the flow of your paragraph.

... opacity of systems. ~~In~~ **Finally, in** high-risk scenarios, ...

Add transition phrase

☒ Highlight changes

? Learn more



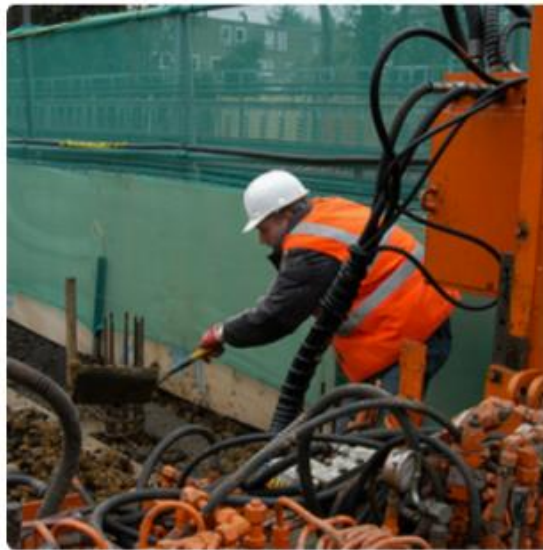
Fonte: <https://app.grammarly.com/>

Processamento de Linguagem Natural – PLN

Exemplo de Image Captioning



"man in black shirt is playing guitar."



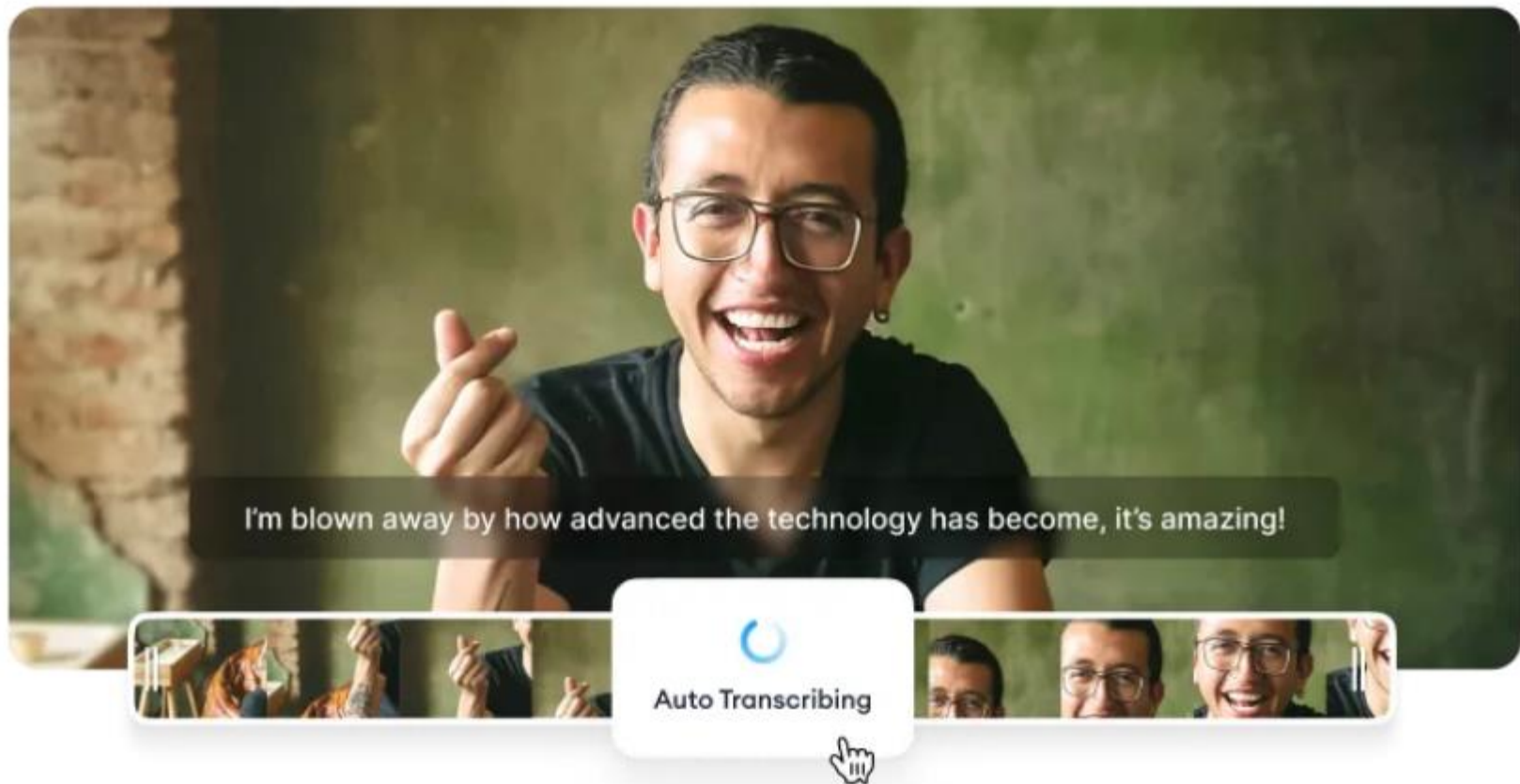
"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."

Processamento de Linguagem Natural – PLN

Exemplo de **Video Captioning**



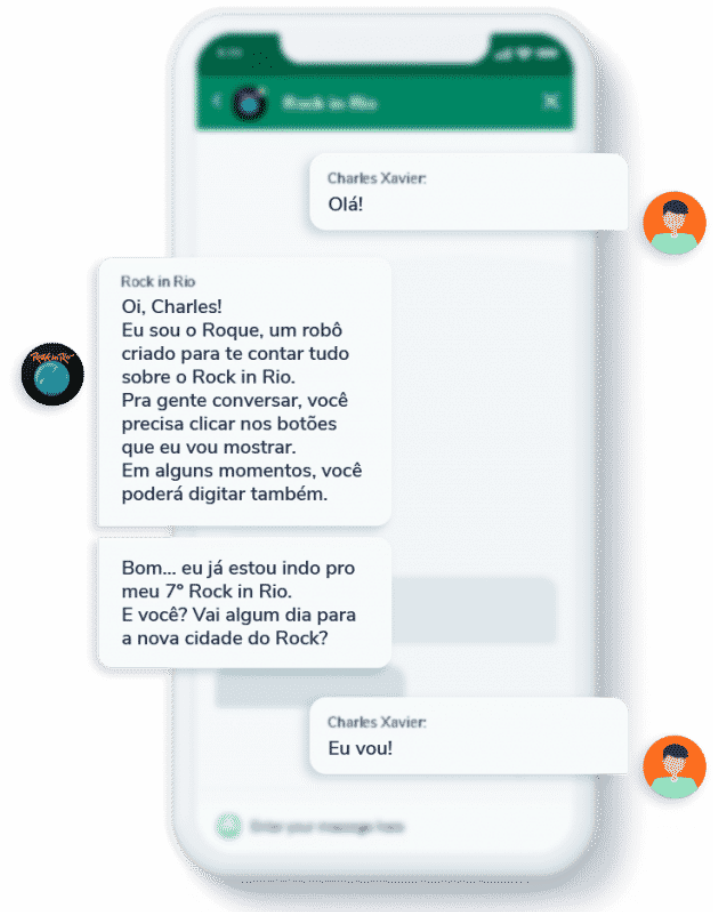
Fonte: <https://www.veed.io/tools/auto-subtitle-generator-online/video-caption-generator>

Processamento de Linguagem Natural – PLN

Exemplo de Chatbot



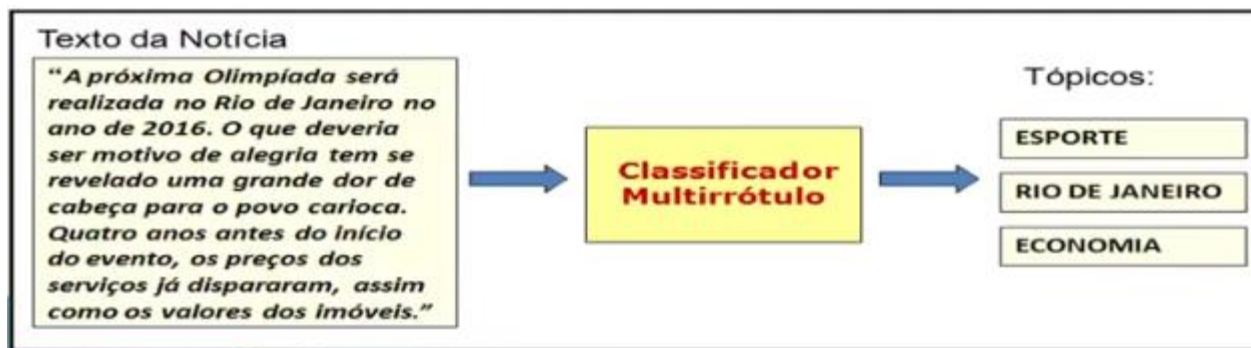
<https://tecnoblog.net/responde/o-que-e-a-alexa-ou-melhor-quem-e/>



Fonte: <https://www.take.net/blog/chatbots/chatbot/>

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

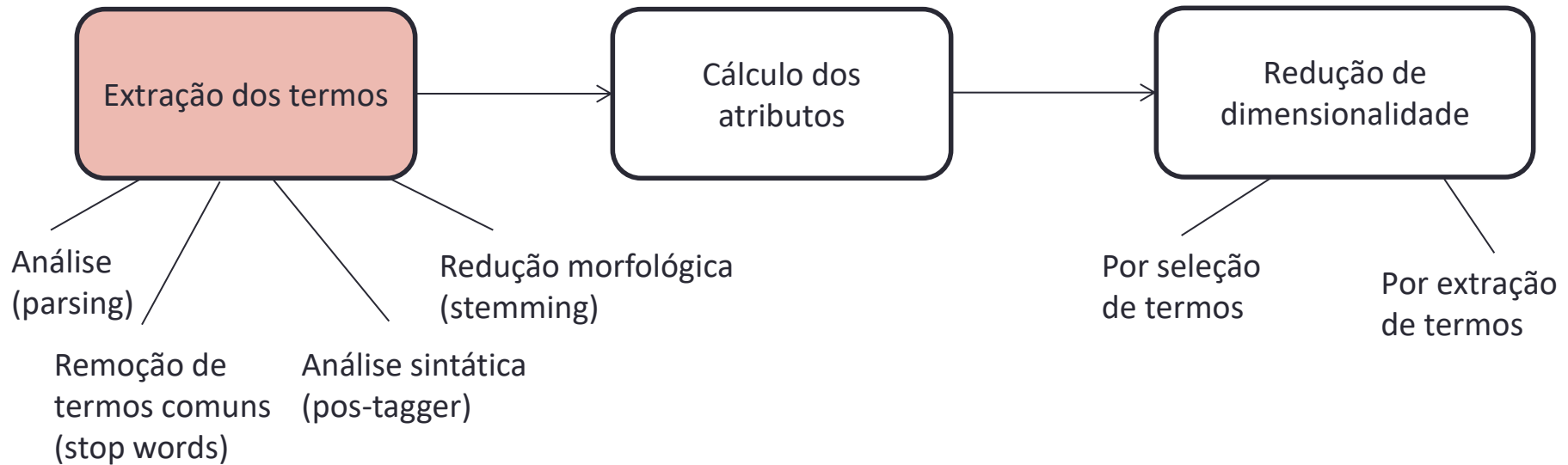
Frequência dos termos, ignorando informações semânticas

Processamento de linguagem natural.

Interpretação sintática e semântica das frases

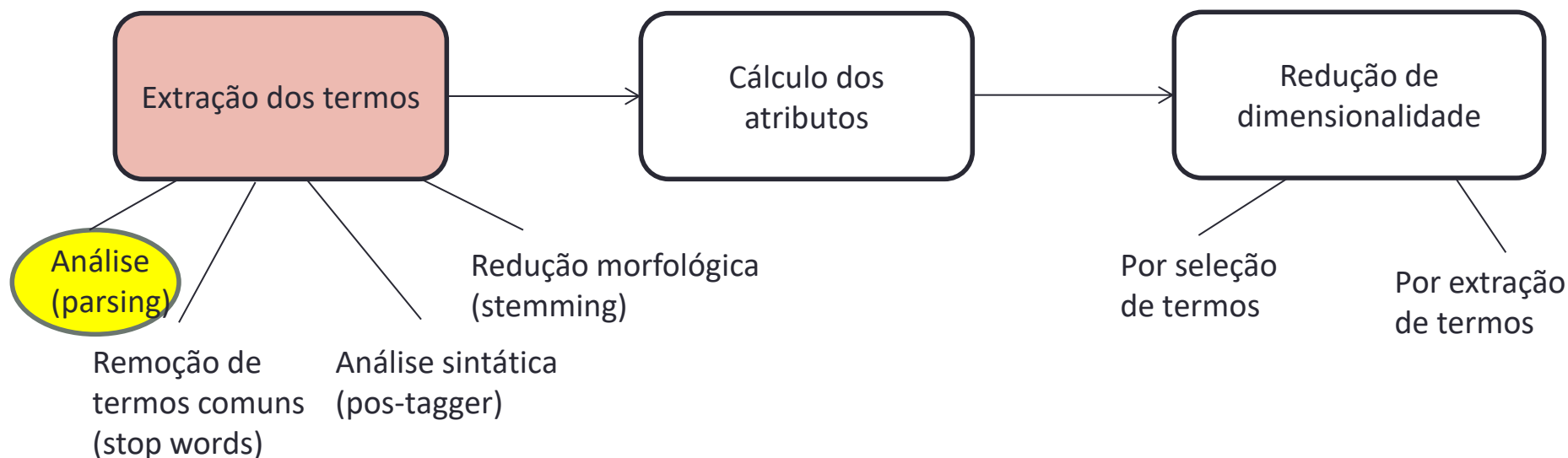
Mineração de textos

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?



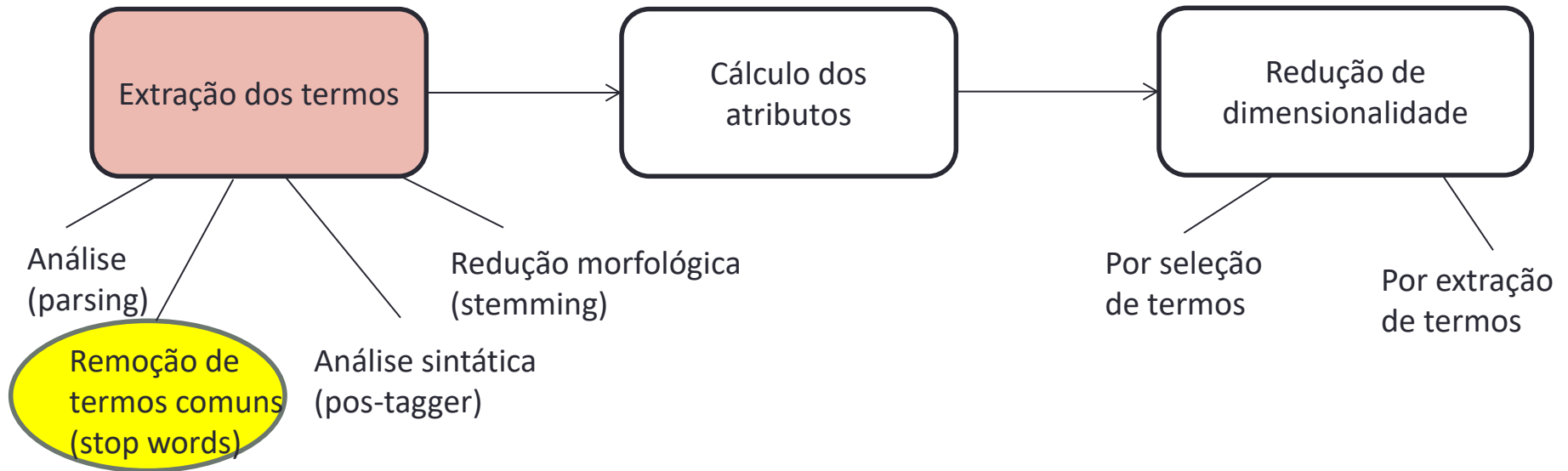
Mineração de textos – Análise de Parsing

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

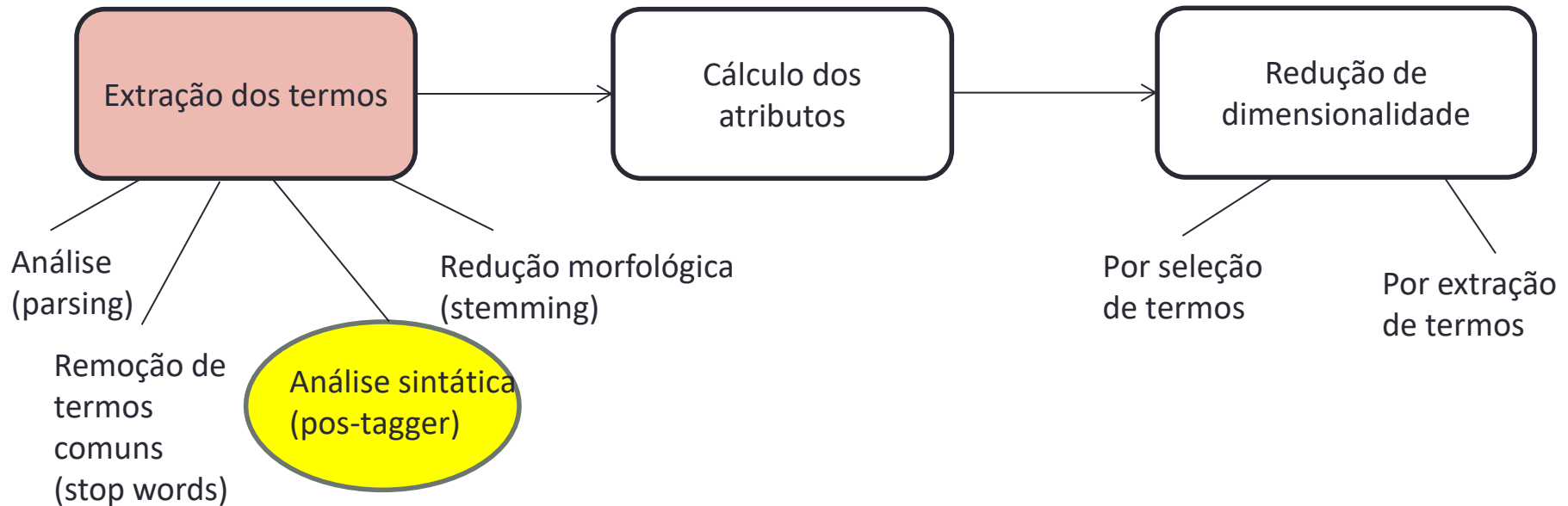
Um sistema de MT geralmente associa uma **stop list** com um conjunto de documentos

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 ambiguidade – 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

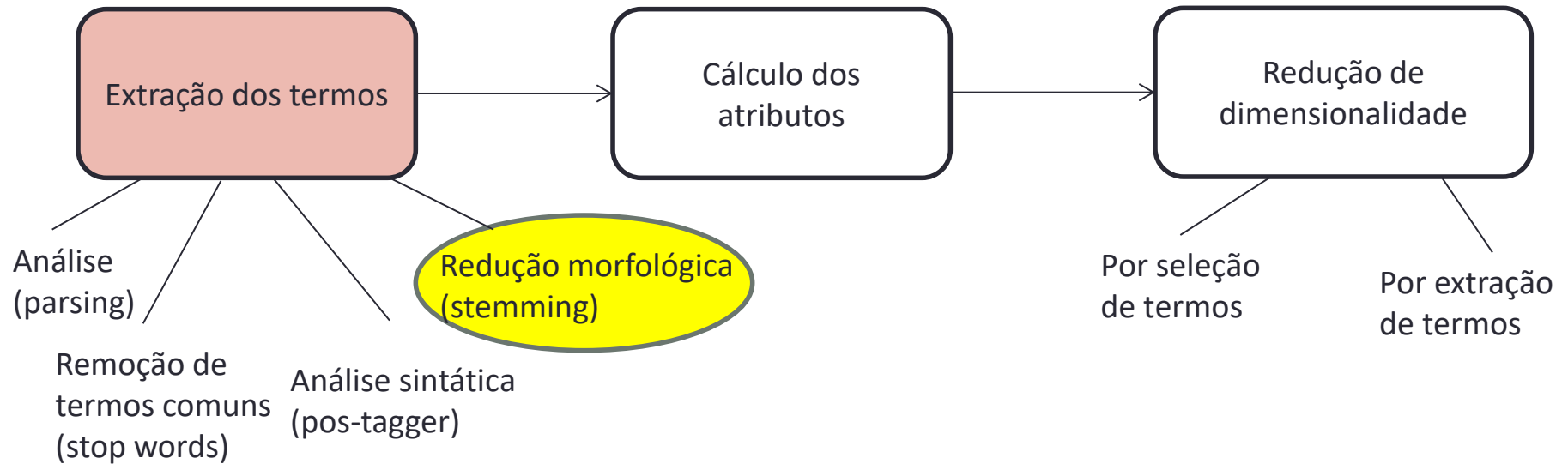
Sinonímia: uma palavra possui vários sinônimos

- 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

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 – Análise morfológica

Diferença entre **Lematização** e **Stemização**

- O processo de **stemização** consiste em reduzir uma palavra ao seu radical.
 - ✓ A palavra “meninas” se reduziria a “menin”, assim como “meninos” e “menininhos”.
 - ✓ As palavras “gato”, “gata”, “gatos” e “gatas” reduziriam-se para “gat”.
- Já a **lematização** reduz a palavra ao seu lema, que é a forma no **masculino** e **singular**. No caso de verbos, o lema é o **infinitivo**.
 - ✓ Por exemplo, as palavras “gato”, “gata”, “gatos” e “gatas” são todas formas do mesmo lema: “gato”.
 - ✓ As palavras “tiver”, “tenho”, “tinha”, “tem” são formas do mesmo lema “ter”.

Mineração de textos – Análise morfológica

Confira a diferença entre **Lematização** e **Stemização**

<https://iaexpert.academy/2020/02/06/lematizacao-x-stemizacao-processamento-de-linguagem-natural/>

<https://www.alura.com.br/artigos/lemmatization-vs-stemming-quando-usar-cada-uma>
Código: <https://colab.research.google.com/drive/1ldg-dolfy87KSi--kMRxcXn-3xTwMA8s?usp=sharing>

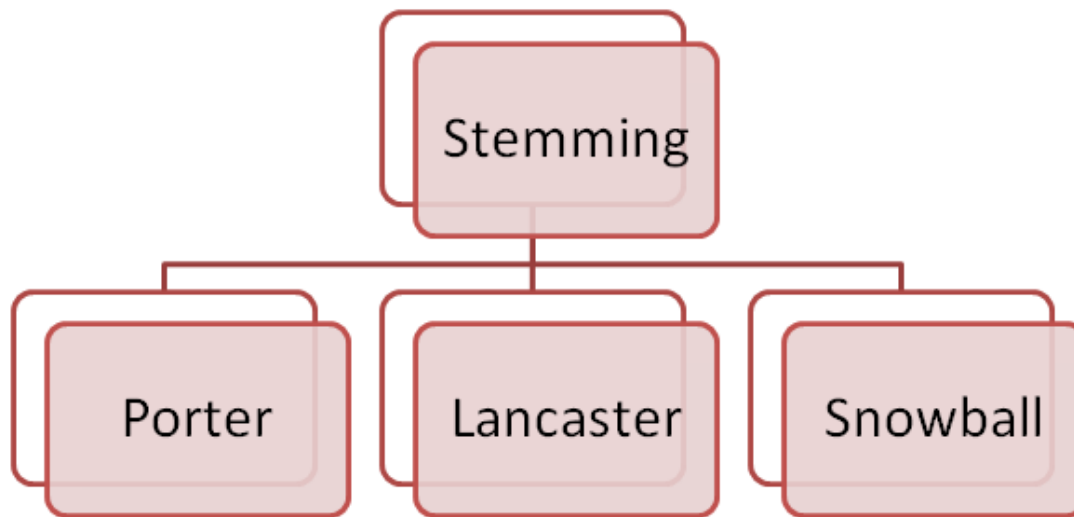
Veja aqui um **quadro comparativo** entre os dois termos:

<https://www.analyticssteps.com/blogs/what-stemming-and-lemmatization-nlp>

Mineração de textos – Análise morfológica

Stemização

Existem diversos **algoritmos de Stemização**: Porter, Lancaster e Snowball. Experimentem!



Fonte: <https://medium.com/fintechexplained/nlp-text-processing-via-stemming-and-lemmatisation-in-data-science-projects-ad4d5176060e>

Mineração de textos – Análise morfológica

Stemização

Existem diversos **algoritmos de Stemização**: Porter, Lancaster e Snowball. Experimentem!

```
from nltk import SnowballStemmer

# Function to apply stemming to a list of words
stemmer = SnowballStemmer()

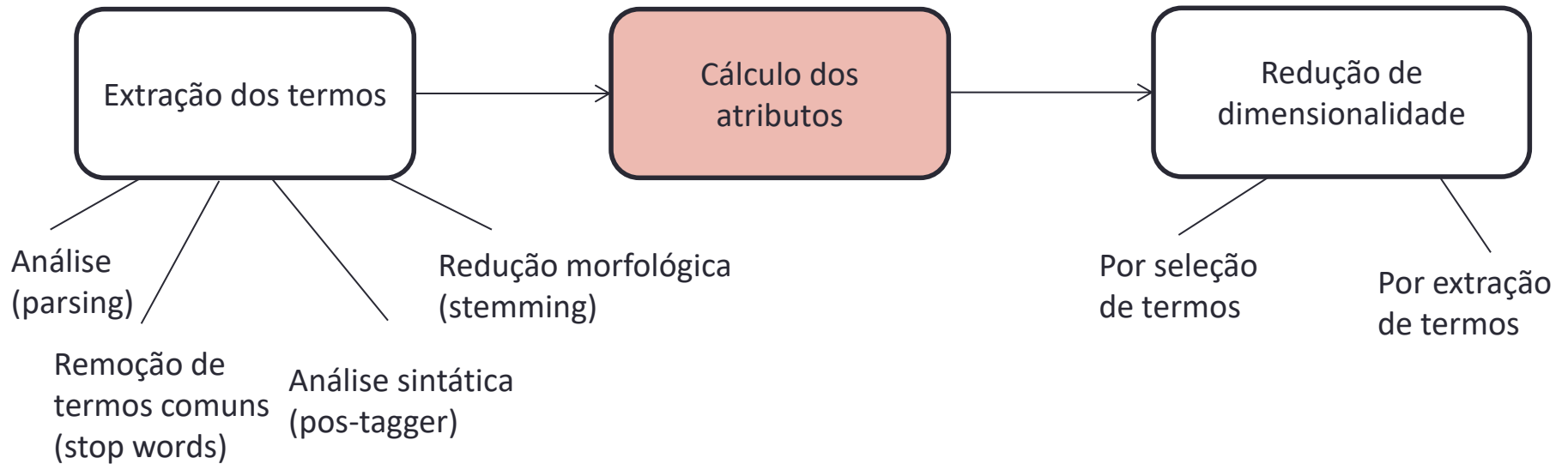
for word in ['blogging', 'blogged', 'blogs']:
    print(stemmer.stem(word))

#This will return blog, blog,
```

```
import nltk
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
print(lemmatizer.lemmatize("blogs"))

#Returns blog
```

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 \mathbf{v} no espaço t -dimensional \mathbb{R}^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”

$$\text{TF}(\text{“Este”}, D1) = 1/5 = 0,2$$

$$\text{TF}(\text{“Este”}, D2) = 1/4 = 0,25$$

$$\text{TF}(\text{“Este”}, D3) = 1/6 = 0,17$$

$$\text{IDF}(\text{“Este”}) = \log(3/3) = 0$$

$$\text{TF}(\text{“Este”}, D1) \times \text{IDF}(\text{“Este”}) = 0,2 \times 0 = 0$$

$$\text{TF}(\text{“Este”}, D2) \times \text{IDF}(\text{“Este”}) = 0,25 \times 0 = 0$$

$$\text{TF}(\text{“Este”}, D3) \times \text{IDF}(\text{“Este”}) = 0,17 \times 0 = 0$$

• **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**”

$$\text{TF}(\text{“outro”}, D1) = 0/5 = 0$$

$$\text{TF}(\text{“outro”}, D2) = 0/4 = 0$$

$$\text{TF}(\text{“outro”}, D3) = 1/6 = 0,17$$

$$\text{IDF}(\text{“outro”}) = \log(3/1) = 0,48$$

$$\text{TF}(\text{“outro”}, D1) \times \text{IDF}(\text{“outro”}) = 0 \times 0,48 = 0$$

$$\text{TF}(\text{“outro”}, D2) \times \text{IDF}(\text{“outro”}) = 0 \times 0,48 = 0$$

$$\text{TF}(\text{“outro”}, D3) \times \text{IDF}(\text{“outro”}) = 0,17 \times 0,48 = 0,08$$

• **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”

$$\text{TF}(\text{“A”}, D1) = 1/5 = 0,2$$

$$\text{TF}(\text{“A”}, D2) = 0/4 = 0$$

$$\text{TF}(\text{“A”}, D3) = 2/6 = 0,33$$

$$\text{IDF}(\text{“A”}) = \log(3/2) = 0,18$$

$$\text{TF}(\text{“A”}, D1) \times \text{IDF}(\text{“A”}) = 0,2 \times 0,18 = 0,036$$

$$\text{TF}(\text{“A”}, D2) \times \text{IDF}(\text{“A”}) = 0 \times 0,18 = 0$$

$$\text{TF}(\text{“A”}, D3) \times \text{IDF}(\text{“A”}) = 0,33 \times 0,18 = 0,06$$

• **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

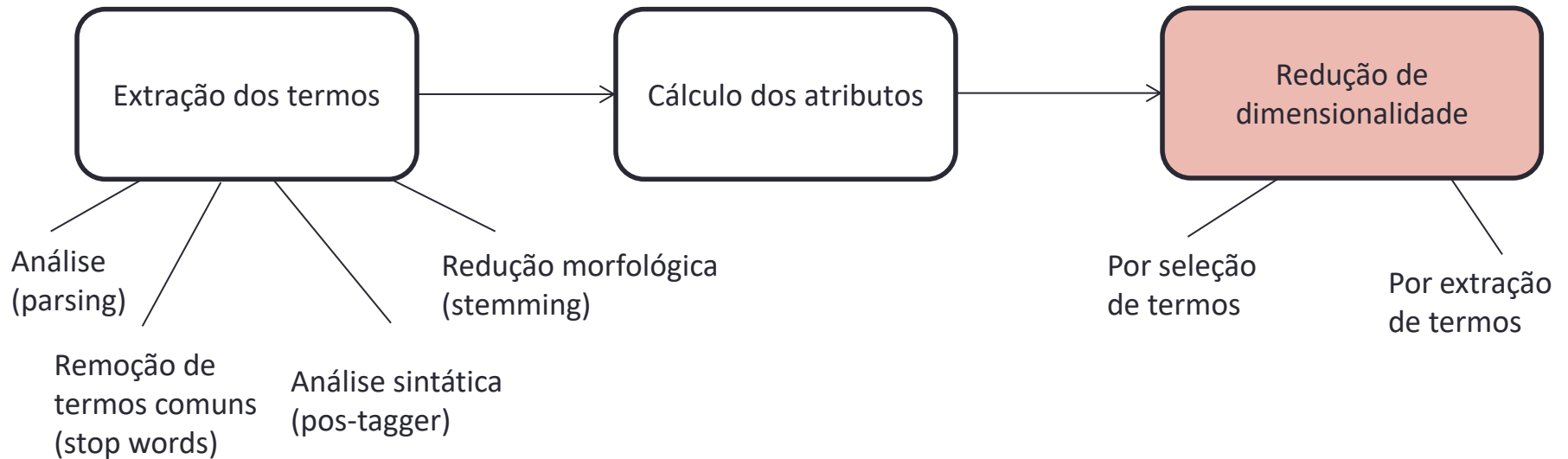
Documento D3

	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

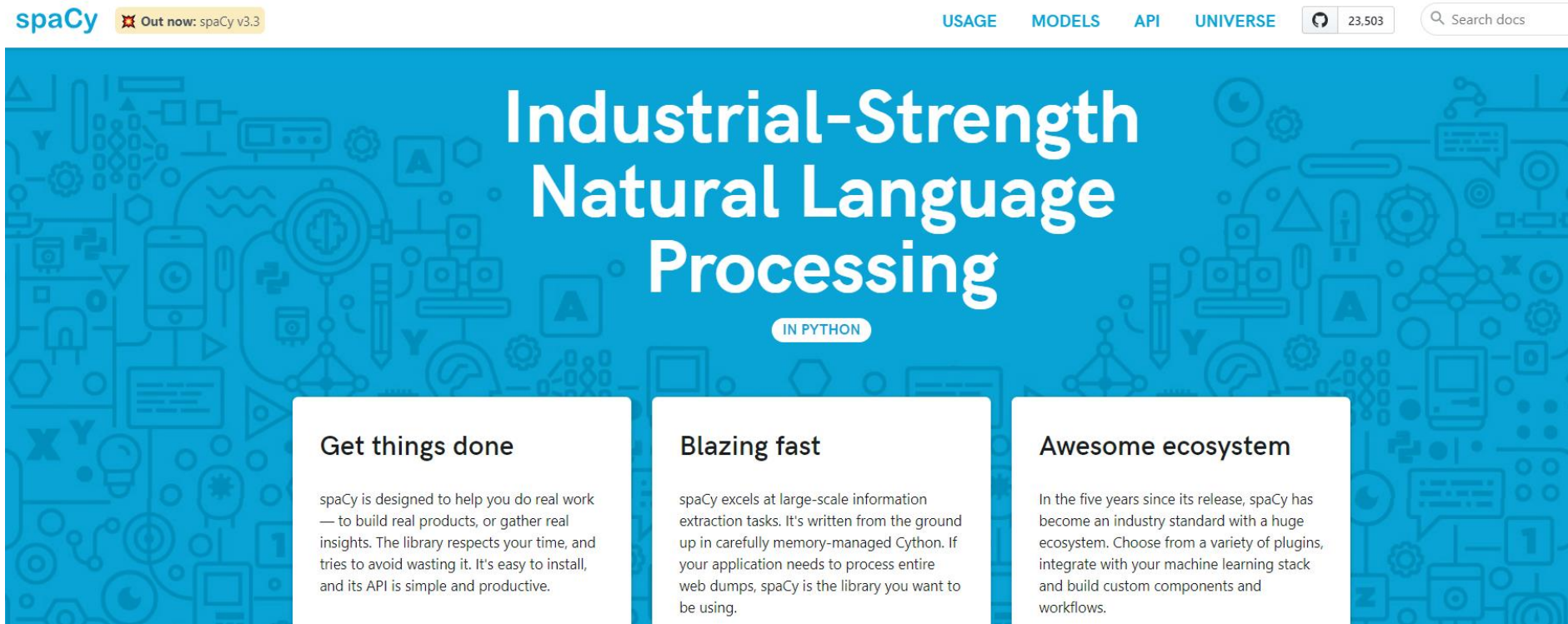
- O vetor de termos que representa os textos é bastante grande.
- Logo, é necessário reduzir o tamanho destes vetores.
 - Ou seja, diminuir o número de termos.
- Duas abordagens:
 - ✓ Seleção de termos
 - ✓ Extração ou remoção de termos

Mineração de textos

Softwares comerciais e abertos para Text Mining:

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

Manual do Spacy - <https://spacy.io/api/data-formats#pos-tagging>

The image shows the top portion of the spaCy website. At the top left is the 'spaCy' logo. Next to it is a yellow badge with a red 'X' icon and the text 'Out now: spaCy v3.3'. To the right are navigation links: 'USAGE', 'MODELS', 'API', and 'UNIVERSE'. Further right is a circular icon with a refresh symbol and the number '23,503'. At the far right is a search bar with the placeholder text 'Search docs'. The main hero section has a blue background with a pattern of white icons representing various technologies and data. The title 'Industrial-Strength Natural Language Processing' is written in large white text. Below the title is a small white pill-shaped button with the text 'IN PYTHON'. Below this are three white rectangular boxes, each with a title and a paragraph of text.

spaCy

Out now: spaCy v3.3

USAGE MODELS API UNIVERSE

23,503

Search docs

Industrial-Strength Natural Language Processing

IN PYTHON

Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

Como podemos resolver o problema de classificação multirrótulo?

1. Eliminação de Exemplos Multirrótulo

A estratégia mais simples que existe da transformação baseada em exemplos, mas também a mais ineficaz, é **eliminar** do conjunto de dados os exemplos que são multirrótulo.

A eliminação dos exemplos com mais de uma classe não resolve o problema multirrótulo original. Ela apenas muda o problema, transformando-o em outro mais simples e, provavelmente, não tão relevante quanto o original

X	Metal	Jazz	Bossa	Pop
X ₁	•			•
X ₂		•	•	
X ₃		•		
X ₄	•			
X ₅		•	•	•



X	Y
X ₃	Jazz
X ₄	Metal


Como podemos resolver o problema de classificação multirrótulo?

2. Criação de Novos rótulos para os Exemplos Multirrótulo Existentes

Nessa estratégia, para cada exemplo, todas as classes atribuídas àquele exemplo são combinadas em uma nova e única classe.

Com essa combinação, o número de classes envolvidas no problema pode aumentar consideravelmente, e algumas classes podem terminar com poucos exemplos que as representem

X	Metal	Jazz	Bossa	Pop
X ₁	•			•
X ₂		•	•	
X ₃		•		
X ₄	•			
X ₅		•	•	•



X	Y
X ₁	Metal-Pop
X ₂	Jazz-Bossa
X ₃	Jazz
X ₄	Metal
X ₅	Jazz-Bossa-Pop

Medidas de avaliação

A avaliação de classificadores multirótulo requer métricas diferentes das utilizadas em problemas de classificação simples-rótulo

Em classificação multirrótulo podemos ter:

1. Um classificador atribui corretamente a um exemplo pelo menos uma das classes a que ele pertence, mas também não atribui ao exemplo uma ou mais classes às quais ele pertence
2. Pode acontecer também de o classificador atribuir a um exemplo uma ou mais classes às quais ele não pertence

Medidas de avaliação

Cálculo da **precisão e recall**, por instância

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}}$$

$$\text{Recall} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}}$$

Medidas de avaliação

Vamos considerar dois exemplos de teste e sua predição:

1ª instância:

Classe verdadeira: [1, 1, 1]

Classe prevista: [0, 1, 1]

Comparando:

1º rotulo: verdadeiro = 1, previsto = 0 (FN)

2º rotulo: verdadeiro = 1, previsto = 1 (VP)

3º rotulo: verdadeiro = 1, previsto = 1 (VP)



- Verdadeiros Positivos (TP) = 2

- Falsos Positivos (FP) = 0

- Falsos Negativos (FN) = 1

$$\text{Precisão} = \frac{2}{2+0} = 1$$

$$\text{Recall} = \frac{2}{2+1} = \frac{2}{3} \approx 0,666$$

2ª instância:

Classe verdadeira: [1, 1, 0]

Classe prevista: [1, 1, 0]

Comparando:

1º rotulo: verdadeiro = 1, previsto = 1 (VP)

2º rotulo: verdadeiro = 1, previsto = 1 (VP)

3º rotulo: verdadeiro = 0, previsto = 0 (VN)



- Verdadeiros Positivos (TP) = 2

- Falsos Positivos (FP) = 0

- Falsos Negativos (FN) = 0

$$\text{Precisão} = \frac{2}{2+0} = 1$$

$$\text{Recall} = \frac{2}{2+0} = 1$$

Medidas de avaliação

Cálculo da **precisão e recall médios**

$$\text{Precisão} = \frac{2}{2+0} = 1$$

$$\text{Precisão} = \frac{2}{2+0} = 1$$

$$\text{Recall} = \frac{2}{2+1} = \frac{2}{3} \approx 0,666$$

$$\text{Recall} = \frac{2}{2+0} = 1$$

Precisão Média:

$$\text{Precisão Média} = \frac{\text{Precisão da Instância 1} + \text{Precisão da Instância 2}}{2} = \frac{1 + 1}{2} = 1$$

Recall Médio:

$$\text{Recall Médio} = \frac{\text{Recall da Instância 1} + \text{Recall da Instância 2}}{2} = \frac{1+0,6}{2} = 0,8$$

Medidas de avaliação

E como calcular estas métricas **por classe**?

Para cada rótulo:

- **Verdadeiros Positivos (TP)**: Número de vezes em que o rótulo foi previsto como positivo e era realmente positivo.
- **Falsos Positivos (FP)**: Número de vezes em que o rótulo foi previsto como positivo, mas era negativo.
- **Falsos Negativos (FN)**: Número de vezes em que o rótulo era positivo, mas foi previsto como negativo.

1ª instância:

Classe verdadeira: [1, 1, 1]

Classe prevista: [0, 1, 1]

2ª instância:

Classe verdadeira: [1, 1, 0]

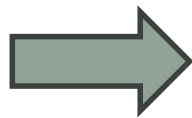
Classe prevista: [1, 1, 0]

Rótulo 1

VP = 1

FP = 0

FN = 1



$$\text{Precisão} = \frac{1}{1+0} = 1$$

$$\text{Recall} = \frac{1}{1+1} = 0,5$$

Medidas de avaliação

E como calcular estas métricas por classe?

Para cada rótulo:

- **Verdadeiros Positivos (TP)**: Número de vezes em que o rótulo foi previsto como positivo e era realmente positivo.
- **Falsos Positivos (FP)**: Número de vezes em que o rótulo foi previsto como positivo, mas era negativo.
- **Falsos Negativos (FN)**: Número de vezes em que o rótulo era positivo, mas foi previsto como negativo.

1ª instância:

Classe verdadeira: [1, 1, 1]

Classe prevista: [0, 1, 1]

2ª instância:

Classe verdadeira: [1, 1, 0]

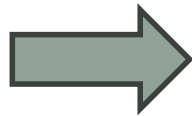
Classe prevista: [1, 1, 0]

Rótulo 2

VP = 2

FP = 0

FN = 0



$$\text{Precisão} = \frac{2}{2+0} = 1$$

$$\text{Recall} = \frac{2}{2+0} = 1$$

Medidas de avaliação

E como calcular estas métricas por classe?

Para cada rótulo:

- **Verdadeiros Positivos (TP)**: Número de vezes em que o rótulo foi previsto como positivo e era realmente positivo.
- **Falsos Positivos (FP)**: Número de vezes em que o rótulo foi previsto como positivo, mas era negativo.
- **Falsos Negativos (FN)**: Número de vezes em que o rótulo era positivo, mas foi previsto como negativo.

1ª instância:

Classe verdadeira: [1, 1, 1]

Classe prevista: [0, 1, 1]

2ª instância:

Classe verdadeira: [1, 1, 0]

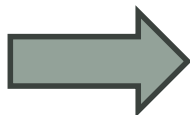
Classe prevista: [1, 1, 0]

Rótulo 3

VP = 1

FP = 0

FN = 0



$$\text{Precisão} = \frac{1}{1+0} = 1$$

$$\text{Recall} = \frac{1}{1+0} = 1$$

Medidas de avaliação

HammingLoss - Esta medida informa o número médio de predições binárias incorretas por objeto de teste

Seja \mathbf{X} um conjunto de dados multirrótulo com n exemplos $(\mathbf{x}_i, \mathbf{y}_i)$, com $i = 1, 2, \dots, n$ e $\text{sum}(\mathbf{y}_i) < k$, em que k é o conjunto de possíveis classes. Sejam ainda \hat{f} um classificador multirrótulo e $\mathbf{z}_i = \hat{f}(\mathbf{x}_i)$ um vetor binário com k elementos representando o conjunto de classes preditas por \hat{f} para um dado exemplo \mathbf{x}_i . Uma medida comumente utilizada para realizar uma avaliação baseada na classificação é o *Hamming Loss* (Schapire e Singer, 2000), definida por:

$$\text{Hamming Loss}(\hat{f}, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \frac{a(\mathbf{y}_i, \mathbf{z}_i)}{k}. \quad (19.3)$$

Nessa medida, $a(\mathbf{y}_i, \mathbf{z}_i)$ representa a distância de Hamming entre dois vetores e corresponde à operação *XOR* da lógica booleana (Tsoumakas e Katakis, 2007). Quanto menor for o valor do *Hamming Loss*, melhor é a classificação, sendo que a predição perfeita ocorre quando o seu valor é igual a zero

Medidas de avaliação – Distância de HammingLoss

Vamos considerar dois exemplos de teste e sua predição:

1ª instância:

Classe verdadeira: [1, 1, 1]

Classe prevista: [0, 1, 1]

Comparando:

1º rotulo: verdadeiro = 1, previsto = 0 (erro)

2º rotulo: verdadeiro = 1, previsto = 1 (acerto)

3º rotulo: verdadeiro = 1, previsto = 1 (acerto)



Número de erros = 1

2ª instância:

Classe verdadeira: [1, 1, 0]

Classe prevista: [1, 1, 0]

Comparando:

1º rotulo: verdadeiro = 1, previsto = 1 (acerto)

2º rotulo: verdadeiro = 1, previsto = 1 (acerto)

3º rotulo: verdadeiro = 0, previsto = 0 (acerto)



Número de erros = 0

Medidas de avaliação – Distância de HammingLoss

Vamos considerar dois exemplos de teste e sua predição:

Cálculo da Hamming Loss

Agora, somamos os erros e dividimos pelo total de rótulos (número de instâncias multiplicado pelo número de rótulos).

$$\text{Hamming Loss} = \frac{\text{número total de erros}}{\text{número de instâncias} \times \text{número de rótulos}} = \frac{1+0}{2*3} = \frac{1}{6} = 0,1666$$

A **Hamming Loss** para essas duas instâncias é **0,166**. Isso significa que, em média, um sexto dos rótulos foram previstos incorretamente

Biblioteca em Python para trabalhar com Multirótulo

Como trabalhar com classificação multi-rótulo em Python?

Biblioteca **Scikit-multilearn**

Veja: <http://scikit.ml/>

Investigue:

```
from skmultilearn.problem_transform import  
BinaryRelevance
```

Biblioteca em Python para dataset textual

Alguns datasets multi-rótulo:

- Reuters-21578 (Reuters News Dataset)
- MovieLens 20M Dataset
- Amazon Product Reviews
- European Union Emotions Dataset

Dataset para Fake News:

<https://www.kaggle.com/c/fake-news/data>

<https://github.com/jghm-f/FACTCK.BR>

Referência

- Capítulo 19 do livro
- Katti Faceli et al.
Inteligência Artificial, Uma abordagem de Aprendizado de Máquina, LTC, 2021.

