TÓPICOS AVANÇADOS EM INTELIGÊNCIA ARTIFICIAL

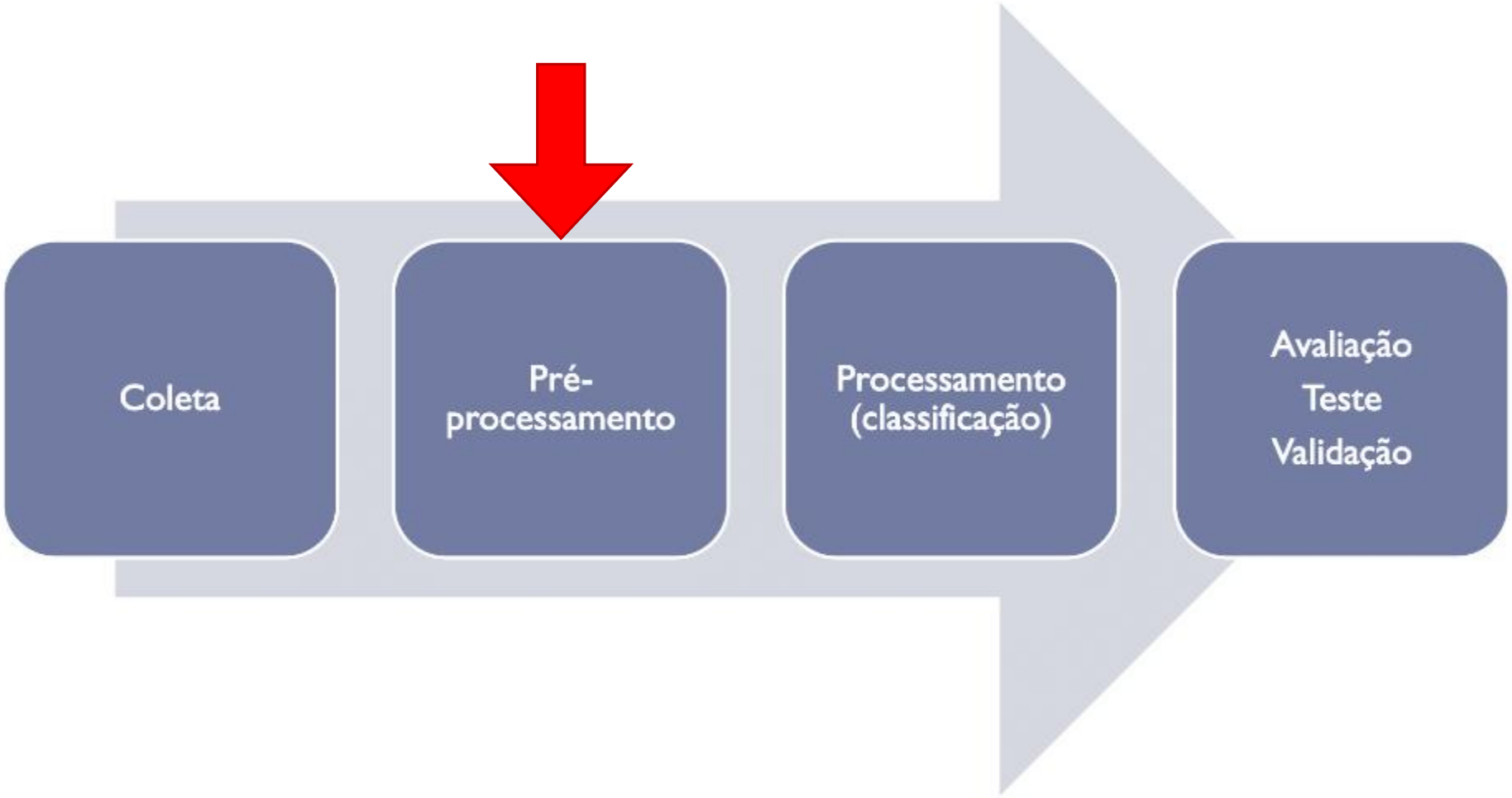
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA

PROFESSOR: DOUGLAS VITÓRIO (douglas.alisson17@gmail.com)

PRÉ-PROCESSAMENTO

AULA 04

PROCESSO DE MINERAÇÃO DE OPINIÃO



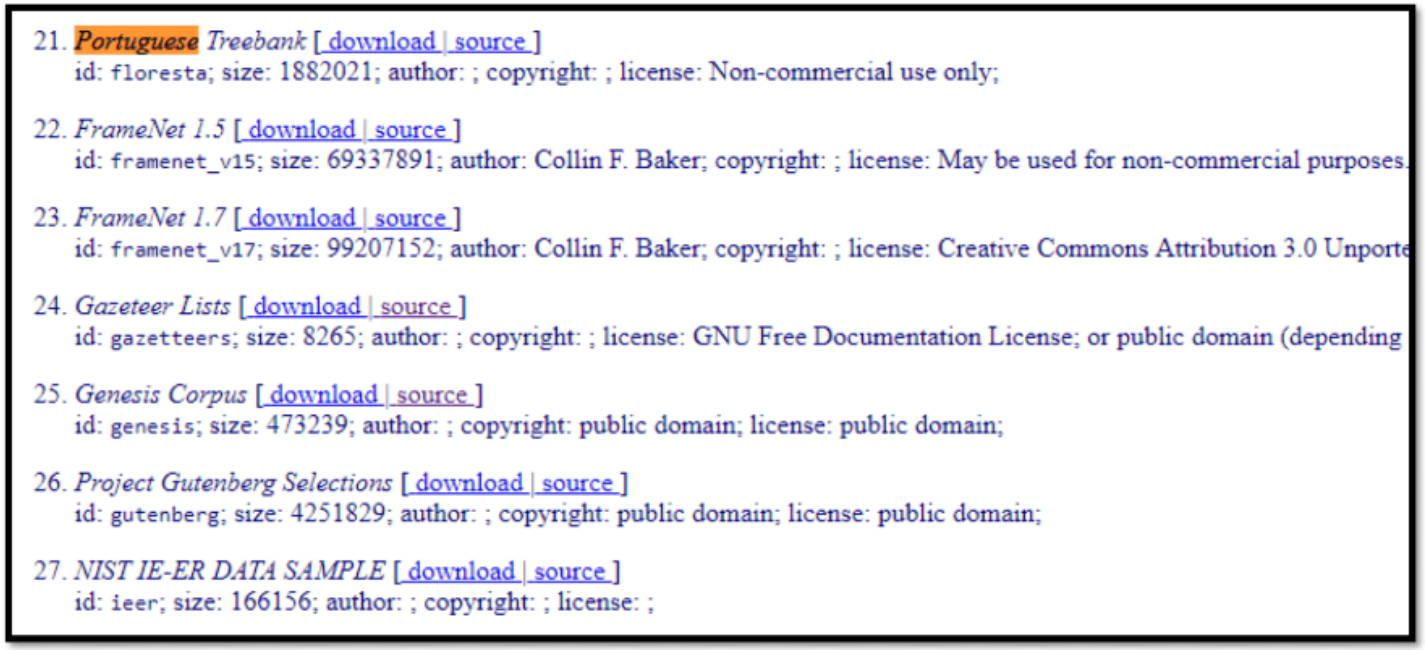
PRÉ-PROCESSAMENTO DE DADOS

Informação textual é, comumente, **não estruturada** e sem regras de padronização.

Para estruturar o texto e convertê-lo em um formato que seja entendível pelos classificadores, algumas técnicas de pré-processamento podem ser utilizadas.

NLTK

- http://www.nltk.org/nltk_data/
- http://www.nltk.org/howto/portuguese_en.html



21. *Portuguese Treebank* [download | source]
    id: floresta; size: 1882021; author: ; copyright: ; license: Non-commercial use only;

22. *FrameNet 1.5* [download | source]
    id: framenet_v15; size: 69337891; author: Collin F. Baker; copyright: ; license: May be used for non-commercial purposes.

23. *FrameNet 1.7* [download | source]
    id: framenet_v17; size: 99207152; author: Collin F. Baker; copyright: ; license: Creative Commons Attribution 3.0 Unporte

24. *Gazeteer Lists* [download | source]
    id: gazetteers; size: 8265; author: ; copyright: ; license: GNU Free Documentation License; or public domain (depending

25. *Genesis Corpus* [download | source]
    id: genesis; size: 473239; author: ; copyright: public domain; license: public domain;

26. *Project Gutenberg Selections* [download | source]
    id: gutenberg; size: 4251829; author: ; copyright: public domain; license: public domain;

27. *NIST IE-ER DATA SAMPLE* [download | source]
    id: ieer; size: 166156; author: ; copyright: ; license: ;

NLTK

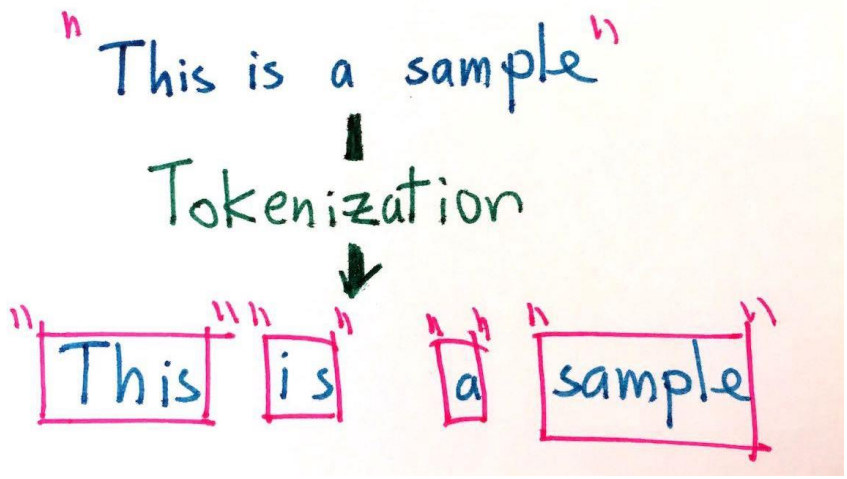

33. Lin's Dependency Thesaurus [download | source]
id: lin_thesaurus; size: 89154019; author: Dekang Lin; copyright: ; license: Distri
34. MAC-MORPHO: Brazilian Portuguese news text with part-of-speech tags [downl
id: mac_morpho; size: 3013904; author: ; copyright: ; license: Distributed with perm
Carlos, Universidade Federal de São Carlos (UFSCar), Universidade Estadual Paul
35. Machado de Assis -- Obra Completa [download | source]
id: machado; size: 6151774; author: Machado de Assis; copyright: ; license: Public
36. MASC Tagged Corpus [download
id: masc_tagged; size: 1602143; aut
development, including commercia
37. Sentiment Polarity Dataset Version
id: movie_reviews; size: 4004848; a
97. Word2Vec Sample [download | source]
id: word2vec_sample; size: 49396025; author: ; copyright: ; license: ;
98. VADER Sentiment Lexicon [download | source]
id: vader_lexicon; size: 90486; author: C.J. Hutto and Eric Gilbert; copyright: ; license: MI
99. Porter Stemmer Test Files [download | source]
id: porter_test; size: 200510; author: ; copyright: ; license: ;
100. RSLP Stemmer (Removedor de Sufixos da Lingua Portuguesa) [download | source]
id: rslp; size: 3805; author: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christia
101. Snowball Data [download | source]
id: snowball_data; size: 6785405; author: ; copyright: ; license: ;
102. Averaged Perceptron Tagger [download | source]
id: averaged_perceptron_tagger; size: 2526731; author: ; copyright: ; license: ;

TOKENIZAÇÃO

Tokenizadores são utilizados para dividir *strings* em listas de *substrings*.

***Sentence tokenizer*:** utilizado para dividir um texto em uma lista de **sentenças.**

***Word tokenizer*:** utilizado para dividir um texto em uma lista de **palavras.**



REMOÇÃO DE STOPWORDS

Stopwords são palavras comuns em uma língua.

Podem ser filtradas durante o pré-processamento do texto, já que não são muito úteis para identificação de sentimento, por exemplo.

Stopwords para o Português (NLTK): [‘a’, ‘ao’, ‘aos’, ‘aquela’, ‘aquelas’, ‘aquele’, ‘aqueles’, ‘aquilo’, ‘as’, ...]

STEMMING

Stemming é a técnica de reduzir palavras flexionadas (em gênero, número...) a seu tronco (***stem***).

É diferente da **lematização**, que converte os verbos para o infinitivo e nomes para o singular.

A biblioteca NLTK disponibiliza o **RSLPStemmer** para a Língua Portuguesa.

Palavra	Stem
deixa	deix
falar	fal
disse	diss
mentirosa	mentir
realmente	real

Palavra	Stem
piorada	pior
estressada	estress
apaixonou	apaixon
concretizar	concret
morto	mort

MARCAÇÃO DE PARTES DA FALA

Marcação de Partes da Fala (*Part-of-Speech tagging; POS tagging*) consiste na classificação sintática dos termos de um texto.

Auxilia a compreender o que está sendo dito em uma frase.

MARCAÇÃO DE PARTES DA FALA

Floresta Sinta(c)tica → *tags* gramaticais para o Português.

Symbols		Category
n		noun
prop		proper noun
adj		adjective
n-adj		between noun and adjective
v	v-fin	finite verb
	v-inf	infinitive verb
	v-pcp	participle verb
	v-ger	gerund verb
art		article
pro h	pron-pers	personal pronoun
	pron-det	determinative pronoun
	pron-indp	independent pronoun
adv		adverb
num		numeral
prp		preposition
intj		interjection
con j	conj-s	subordinating conjunction
	conj-c	coordinating conjunction

MARCAÇÃO DE PARTES DA FALA

Floresta Sinta(c)tica → *tags* gramaticais para o Português.

```
[('eu', 'pron-pers'), ('não', 'adv'), ('aguento', 'v-
fin'), ('mais', 'adv'), ('ter', 'v-inf'), ('problema',
'n'), ('com', 'prp'), ('minha', 'pron-det'), ('volta',
'n'), ('da', 'n'), ('australiaaaaa', 'n'), ('!', '!'),
('!', '!'), ('para', 'prp'), ('de', 'prp'), ('cancelar',
'v-inf'), ('meus', 'pron-det'), ('voos', 'n'), ('@', 'n'),
('LATAM_BRA', 'n')]
```

VSM

Para transformar a lista de tokens em um vetor de números, nós utilizamos o **Modelo de Espaço Vetorial (VSM)**.

No **VSM**, cada **token** (palavra, termo) que apareceu no nosso conjunto de dados se torna uma **característica**.

Pode-se notar que a dimensionalidade se torna ENORME (técnicas de redução são utilizadas).

VSM

Imaginemos uma lista de documentos:

1	ele vai comer
2	ele vai sair e eu vou sair
3	eu vou beber e sair

Dessa forma, nossa lista de palavras é:

“ele”, “vai”, “comer”, “sair”, “e”, “eu”, “vou”, “beber”

VSM

Há três formas básicas de se transformar as palavras em um vetor:

- **Com um valor binário:**

Documento	ele	vai	comer	sair	e	eu	vou	beber
1	1	1	1	0	0	0	0	0
2	1	1	0	1	1	1	1	0
3	0	0	0	1	1	1	1	1

VSM

Há três formas básicas de se transformar as palavras em um vetor:

- **Contando a quantidade de vezes (frequência):**

Documento	ele	vai	comer	sair	e	eu	vou	beber
1	1	1	1	0	0	0	0	0
2	1	1	0	2	1	1	1	0
3	0	0	0	1	1	1	1	1

VSM

Há três formas básicas de se transformar as palavras em um vetor:

- **TF-IDF:**

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

VSM

Há três formas básicas de se transformar as palavras em um vetor:

- **TF-IDF:**

Documento	ele	vai	comer	sair	e	eu	vou	beber
1	0.058	0.058	0.159	0.000	0.000	0.000	0.000	0.000
2	0.025	0.025	0.000	0.050	0.025	0.025	0.025	0.000
3	0.000	0.000	0.000	0.035	0.035	0.035	0.035	0.095

N-GRAM

O modelo **n-gram** é um modelo de linguagem probabilístico amplamente utilizado na classificação de texto.

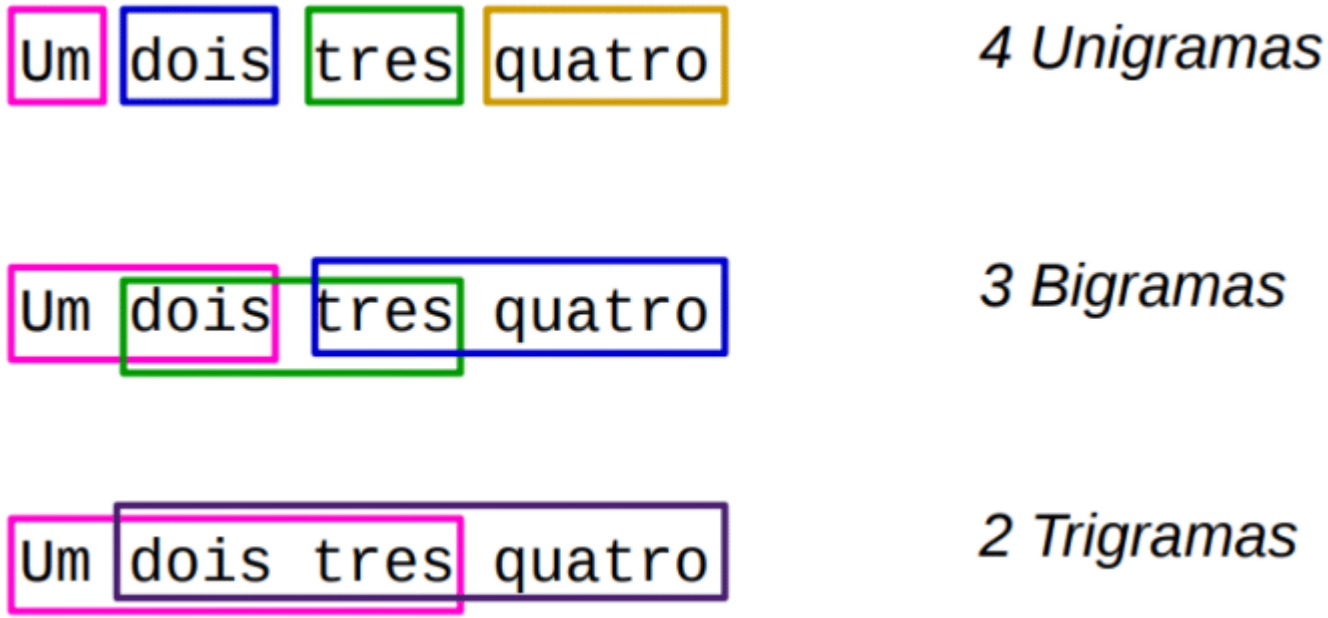
Um n-gram é uma sequência contínua de *n* itens de um dado (fonemas, personagens, **palavras**, **caracteres**, etc.) obtida através de um corpus.

Número de elementos	Nome
1	Unigram
2	Bigram
3	Trigram
4	4-gram
5	5-gram

N-GRAM

O modelo **n-gram** pode ser utilizado a **nível de palavra** ou a **nível de caractere.**

Nível de palavra:



N-GRAM

Nível de palavra:

"sendo muito bem atendido pelo pessoal do @alobradesco"

'sendo muito', 'muito bem', 'bem atendido', 'atendido pelo', 'pelo pessoal', 'pessoal do', 'do @', '@ alobradesco'