Processamento de linguagem natural com Python

Coordenador
Fernando Ferreira

INSTITUTO INFNET 1994

Institucional

Reitor

Prof. Eduardo Ramos
Eduardo.ramos@Infnet.edu.br

Coordenação do Curso

Prof. Fernando Ferreira
Fernando.gFerreira@infnet.edu.br

Gerente Acadêmica

Ana Curi
ana.curi@Infnet.edu.br

Coordenador/Professor

Prof. Fernando Ferreira

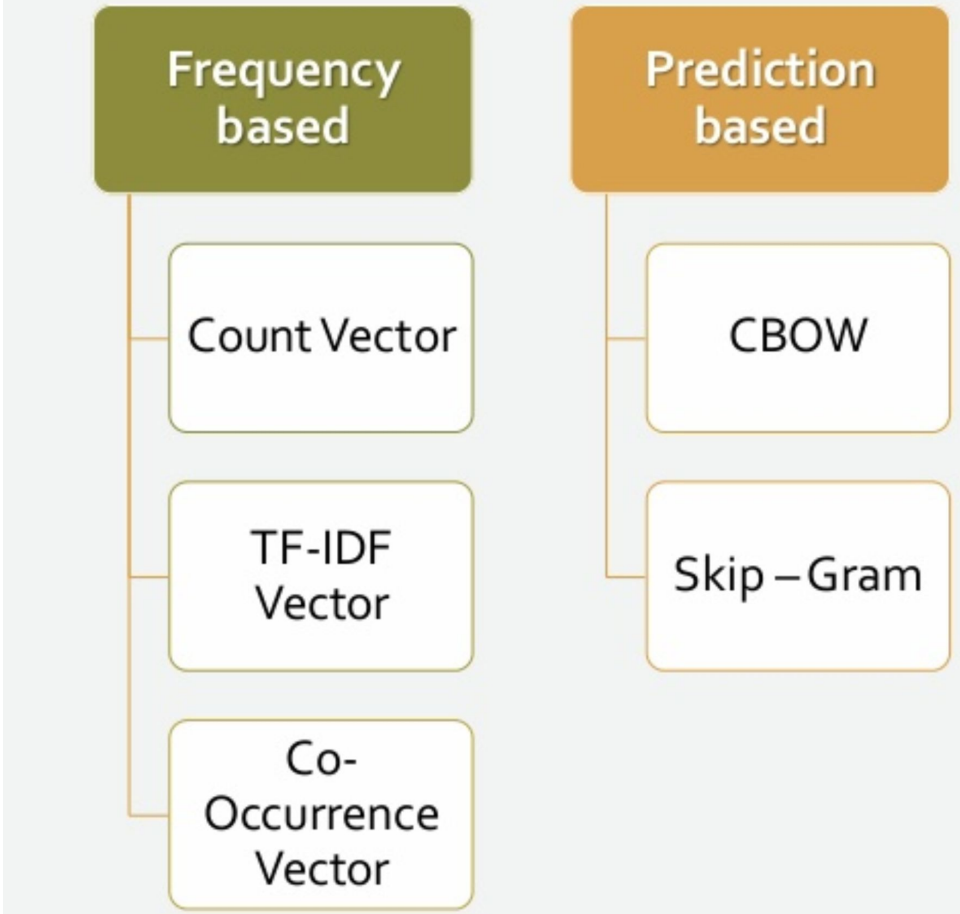
fernando.gferreira@infnet.edu.br

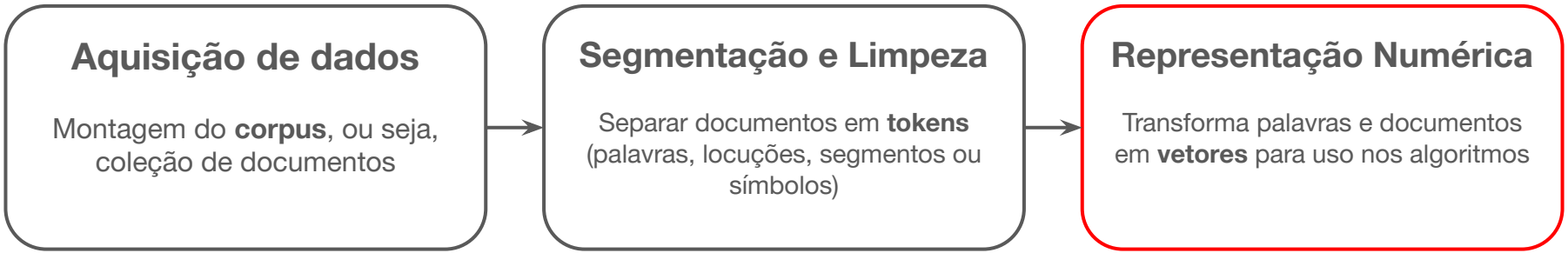
https://www.linkedin.com/in/nandoferreira/

Mini-bio

- Doutor pelo Programa de Engenharia Elétrica da Universidade Federal do Rio de Janeiro (Coppe-UFRJ) - ênfase em Inteligência Computacional.
- Sócio-fundador da empresa Twist (https://twist.systems/) - residente do Parque Tecnológico da UFRJ.
- Coordenador do MBA de Data Science e do MIT de Inteligência Artificial do Instituto Infnet.
- Professor de disciplinas de desenvolvimento ágil de software na graduação da Infnet. Professor do MBA de Business Analytics e Big Data da Fundação Getúlio Vargas.
- Membro da colaboração internacional CERN/UFRJ.
- [Twist] Empresa membro-nato do CDIA.Rio - Rede de Ciência de Dados & Inteligência Artificial do Rio de Janeiro

Representação de documentos em vetores




Aquisição de dados
Montagem do **corpus**, ou seja, coleção de documentos
Segmentação e Limpeza
Separar documentos em **tokens** (palavras, locuções, segmentos ou símbolos)
Representação Numérica
Transforma palavras e documentos em **vetores** para uso nos algoritmos

Segmentação em Frases e Palavras

It only outputs 1.5 amperes. That's the main weakness.

Punkt Tokenizer (Segmentação em Frases)

It only outputs 1.5 amperes.	That's the main weakness.

Penn Treebank (Segmentação em Palavras)

It	only	outputs	1.5	amperes	.	That	's	the	main	weakness	.

Segmentação em Frases e Palavras

It only outputs 1.5 amperes. That's the main weakness.

Punkt Tokenizer (Segmentação em Frases)

It only outputs 1.5 amperes. | That's the main weakness.

Penn Treebank (Segmentação em Palavras)

It	only	outputs	1.5	amperes	.	That	's	the	main	weakness	.

`tokenizers::tokenize_ptb()`

Representação Vetorial de Documentos

Objetivo

- Ter um vetor **numérico** que representa todo um documento
- Usar esse vetor para aplicar a algoritmos com entrada numérica (árvores de decisão, k-Means, SVM, etc)

Obs

Naïve Bayes não precisa de uma representação vetorial completa. Basta a etapa de segmentação de palavras e definição de vocabulário.

Vector Space Model

- Criação de um **espaço vetorial** no qual cada ponto é um **documento**.
- Cada dimensão do vetor corresponde a um **token**.
- Um **vocabulário** precisa ser definido a priori.
- Representação "bag of words" (multiconjunto de palavras)

Vector Space Model

	outputs	amperes	main	weakness	obama	speaks	media	illinois
It only outputs 1.5 amperes. That's the main weakness.	1	1	1	1	0	0	0	0
Obama speaks to the media in Illinois	0	0	0	0	1	1	1	1

- O Vocabulário pode conter centenas de milhares de termos: **representação esparsa**.
- Descarta a **sequência** e relação **semântica** entre termos
- Diferentes formas de preencher os valores:
  - **One-Hot**: **1** se a palavra estiver presente, **0** se não estiver.
  - **TF** (Terms Frequency): Número de vezes que a palavra aparece no documento.
  - **TF-IDF** (TF, Inverse Document Frequency): Ponderar TF pelo inverso da frequência em da palavra em todos os documentos (corpus)

Vector Space Model

	outputs	amperes	main	weakness	obama	speaks	media	illinois
It only outputs 1.5 amperes. That's the main weakness.	1	1	1	1	0	0	0	0
Obama speaks to the media in Illinois	0	0	0	0	1	1	1	1

- O Vocabulário pode conter centenas de milhares de termos: **representação esparsa**.
- Descarta a **sequência** e relação **semântica** entre termos
- Diferentes formas de preencher os valores:
  - One-Hot: **1** se a palavra estiver presente, **0** se não estiver.
  - TF (Terms Frequency): Número de vezes que a palavra aparece no documento.
  - $$\text{TF-IDF}_{i,j} = \frac{\text{no. de vezes que a palavra } j \text{ aparece no documento } i}{\text{proporcao dos documentos nos quais a palavra } j \text{ aparece}}$$

Vector Space Model (TF-IDF)

	outputs	amperes	main	weakness	obama	speaks	media	health	illinois
It only outputs 1.5 amperes. That's the main weakness.	**1**	**1**	.5	.5	0	0	0	0	0
Obama speaks to the media in Illinois. Obama speaks very well.	0	0	0	0	.67	**1**	**1**	0	.5
The main weakness of Obama is the Illinois health system	0	0	.5	.5	.33	0	0	**1**	.5