FIAP GRADUAÇÃO

Processamento de Linguagem Natural, chatbots & virtual agents

Prof . : Guilherme Cardim Mattos
Email:profguilherme.mattos@fiap.com.br

2023

Objetivo geral da aula de hoje

Visão geral sobre métodos de tokenização e similaridade

Ementa – visão geral

Processamento de Linguagem Natural, chatbots & virtual agents

Elementos de linguagem. Interpretação de linguagem natural e semântica.

Tokenização. IA e Machine Learning aplicados à análise de linguagem.

Técnicas para análise e organização de dados não estruturados (voz e texto).

Análise de sentimentos. Classificação, clusterização.

Construção de chatbots usando dialogFlow e outros frameworks. Implementação, testes e verificação.

Bibliografia

L. T. Cruz, a. J. Alencar, E. A. Schmitz, Assistentes Virtuais Inteligentes e Chatbots, Editora Brasport, 2019

A. B. Valdati, Inteligência artificial – IA, Ed Contentus, 2020

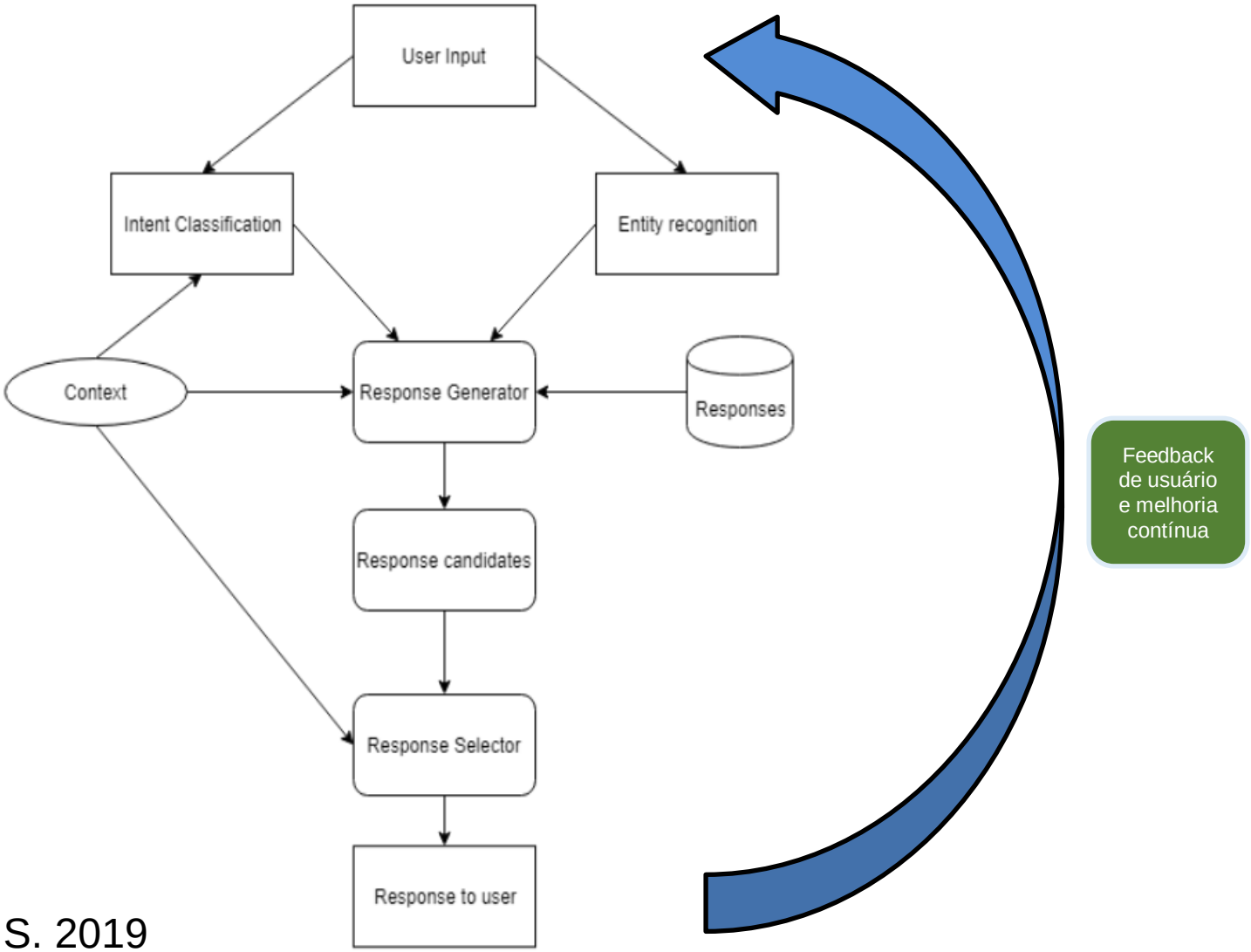
A. U. Rech, Artificial intelligence, environment and smart cities, Editora Educs, 2021

G. C. L. Leal, Linguagem, programação e banco de dados: guia prático de aprendizagem - 1° Edição, Editora Intersaberes, 2015

T. C. Guimarães, Comunicação e linguagem – 2ª edição, Editora Pearson, 2020

A. M. T. Ibanos, et al., Fundamentos linguísticos e computação, Editora EdiPUC-RS, 2015

Arquitetura tradicional Chatbot



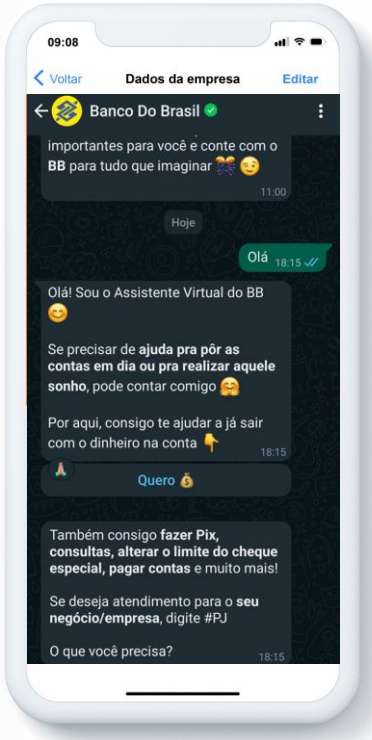
Shah, V., Shah, S. 2019

Chatbots na prática

Bacio di Latte

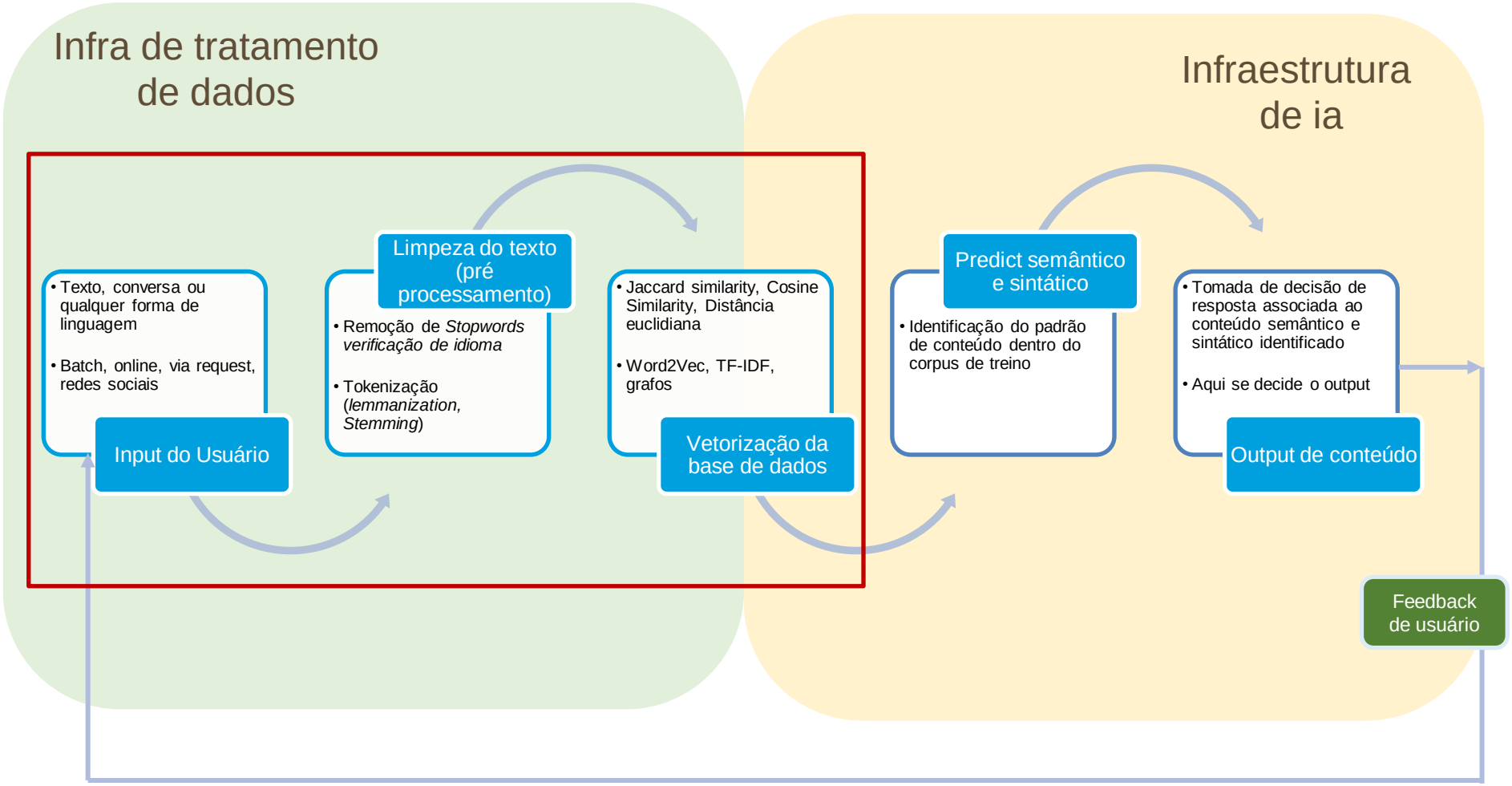
Banco itaú

Banco do brasil



Xp investimentos

Pipeline de um projeto de NLP



Approuch pré processamento de texto

NLTK Stemmers

Interfaces used to remove morphological affixes from words, leaving only the word stem. Stemming algorithms aim to remove those affixes required for eg. grammatical role, tense, derivational morphology leaving only the stem of the word. This is a difficult problem due to irregular words (eg. common verbs in English), complicated morphological rules, and part-of-speech and sense ambiguities (eg. `ceil-` is not the stem of `ceiling`).

StemmerI defines a standard interface for stemmers.

Stem: the main long, thin part of a plant above the **ground from which the leaves or flowers grow**; a smaller part that grows from this and supports flowers or leaves

Oxford Dictionary

NLTK :: nltk.stem package

Approuch pré processamento de texto

Component for assigning base forms to tokens using rules based on part-of-speech tags, or lookup tables. Different `Language` subclasses can implement their own lemmatizer components via language-specific factories. The default data used is provided by the `spacy-lookups-data` extension package.

For a trainable lemmatizer, see `EditTreeLemmatizer` .

New in v3.0

As of v3.0, the `Lemmatizer` is a **standalone pipeline component** that can be added to your pipeline, and not a hidden part of the vocab that runs behind the scenes. This makes it easier to customize how lemmas should be assigned in your pipeline.

If the lemmatization mode is set to `"rule"`, which requires coarse-grained POS (`Token.pos`) to be assigned, make sure a `Tagger` , `Morphologizer` or another component assigning POS is available in the pipeline and runs *before* the lemmatizer.

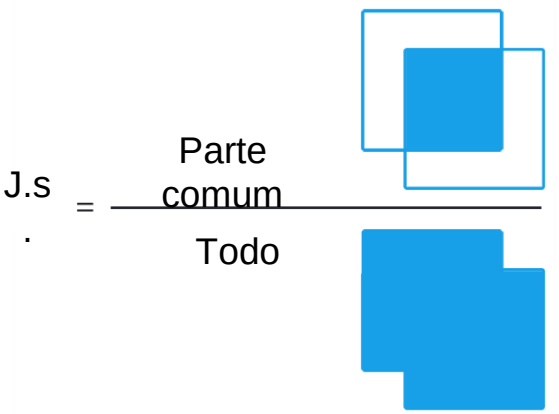
“Lemmatization is the task of finding the base form – the lemma – of a given word form.”
Ingason et. al. (2008).

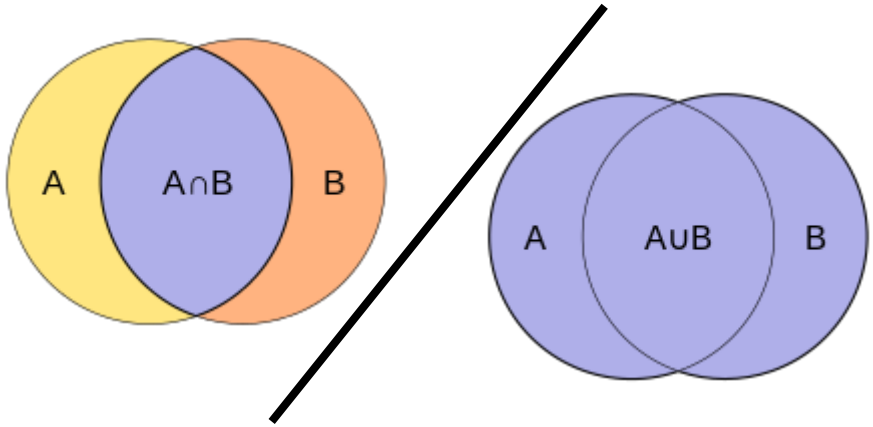
https://spacy.io/api/lemmatizer

Critérios de semelhança

Jaccard Similarity:

J.s = Parte comum / Todo = %





$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Prática

- https://colab.research.google.com/drive/1uCWMw3F5z4DDF-pDQsaRVmg0a7aFsXbs?usp=sharing

Curiosidades

- https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

FIAP GRADUAÇÃO