Processamento de Linguagem Natural (PLN) – aula I

por: Rafael Stoffalette João

Data: 05/09/2020

UNIversidade Paulista (UNIP) – Araçatuba
Pós-graduação em Data Science e Machine Learning

Objetivos da aula de hoje

A disciplina visa apresentar conceitos e aplicações com campo de processamento de linguagem natural.

Serão apresentados técnicas inovadoras de processamento de linguagem natural para processar falas e analisar textos baseados em modelos probabilísticos de deep learning, como os modelos ocultos de Markov e redes neurais recorrentes.

Apresentará também extração de informações, marcação morfossintática, sintaxe e semântica, modelos estatísticos e modelos baseados em regras, modelagem linguística, clustering e etc.

Agenda da disciplina

Módulo 01 - Introdução ao Processamento de Linguagem Natural

Módulo 02 - Análise Semântica e Morfológica

Módulo 03 - Processo de Mineração de Texto

Módulo 04 - Modelagem Estatística da Linguagem

Módulo 05 - Word Embeddings

Módulo 06 - Classificação de Texto

Módulo 07 - Extração de Informação

Módulo 08 - Geração de Resumo

Módulo 09 - Análise de Sentimentos

Módulo 10 - Deep Learning aplicado ao Processamento de Linguagem Natural

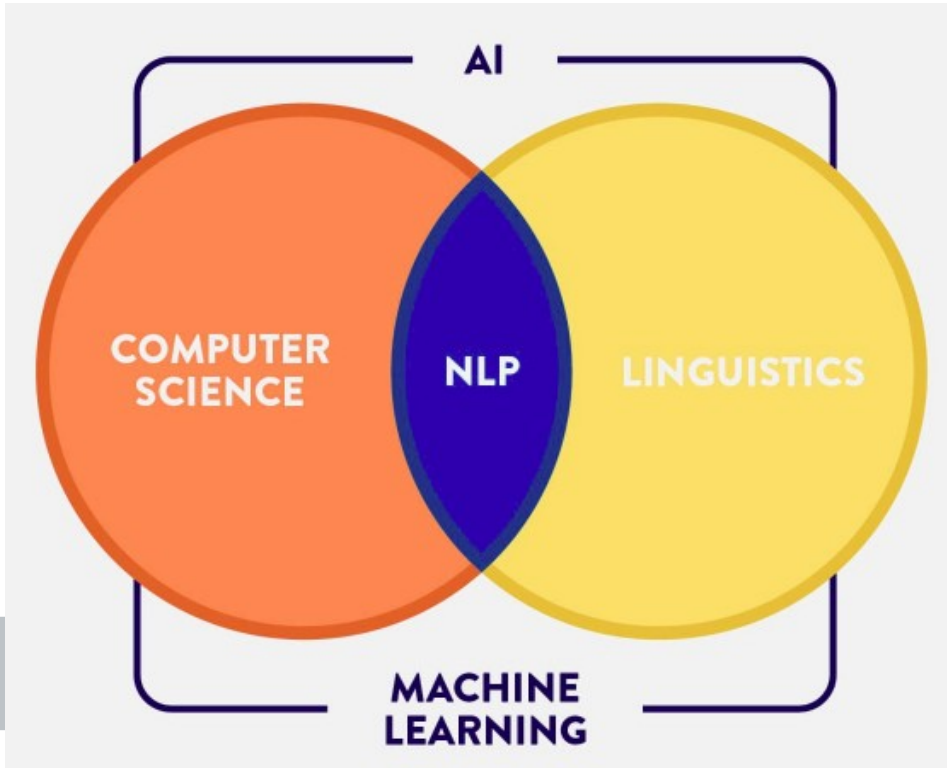
Processamento de Linguagem Natural (PLN)

"... é uma área da computação que tem como objetivo extrair representações e significados mais completos de textos livres escritos em linguagem natural" (INDURKHYA; DAMERAU, 2010)

Facilitar a interação entre humanos e máquinas

linguagem natural

uma linguagem que é usada para comunicações do dia-a-dia feitas por humanos;



Processamento de Linguagem Natural (PLN)

PLN: dois campos

Oral:

– Speech-to-Text e Text-to-Speech

Escrita:

– Assistentes de escrita, motores de busca, tradutores automáticos…

Processamento de Linguagem Natural (PLN)

Linguagens de programação são EXATAS e estruturadas.

Uma vírgula, letra trocada, espaço em branco ou simples quebra de linha implicam em falha.

Compreensão binária. Ou funciona ou não.

```
while  True:
    if V == 0:
        break
    contadorDeTestes += 1
    S = 0
    i = 0
    notas = [0, 0, 0, 0]

    while S < V:
        if caixa[i] + S <= V:
            S = S + caixa[i]
            notas[i] += 1
        else:
            i += 1
```

Um gaúcho entende completamente um Alagoano falando?...

Processamento de Linguagem Natural (PLN)

Tarefas mais comuns:

- Análise morfológica;
- Sumarização;
- Tradução;
- Reconhecimento de fala.
- etc...

Processamento de Linguagem Natural (PLN)

Análise morfológica:

Forma das palavras

- É um nome, pronome, verbos, advérbios, adjetivo, preposições, ...
- Plural <-> Singular

Por meio da análise morfológica, correções automáticas podem ser conduzidas.

E os erros ortográficos, prejudicam o desempenho?

Processamento de Linguagem Natural (PLN)

Reconhecimento de voz:

Assistentes virtuais, como Siri, Alexa, Cortana, dentre outras, tem ganhado muita importância em nosso dia a dia.

Celulares smartfones que “ouvem conversas” mesmo desligados.

```
import speech_recognition as sr
```

Processamento de Linguagem Natural (PLN)

Chatbots:

Hoje focados no atendimento e marketing digital.

horários que um funcionário não poderia estar atendendo.

Mas o aprendizado que eles possuem pode levar a lideranças

Inbot, Tinbot, etc..

O Processamento de Linguagem Natural...

Chatbots deixam de atender clientes para interagir de forma ativa

https://www.youtube.com/watch?v=t5bYtcjWcPQ

https://www.youtube.com/watch?v=VZuSoAD1LtQ

https://www.youtube.com/watch?v=mztWx0GuAEA

https://www.youtube.com/watch?v=HK17k65CAno

https://www.youtube.com/watch?v=_IVq9TBpXzY

O Processamento de Linguagem Natural...

chatbot @TayandYou (2016 - Microsoft)

Aprender e manter conversa “natural” com usuários do Twitter

O Processamento de Linguagem Natural...

chatbot @TayandYou (2016 - Microsoft)

Aprender e manter conversa “natural” com usuários do Twitter

Em menos de 24 horas:

Tay aprendeu discursos racistas, homofóbicos e de ódio.

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

O Processamento de Linguagem Natural...

Linguagem natural pode ser falada ou escrita.

Ambas tem suas peculiaridades e dificuldades para compreensão.

- escrita: regras gramaticais ajudam na construção de frases - difícil processo
- falada: a falta de regras torna a compreensão difícil – facilidade na construção

O Processamento de Linguagem Natural...

O grande desafio do PLN é:

transformar textos e falas em conjuntos de dados para algoritmos de machine learning.

Linguagem natural é não estruturada,

Algoritmos tentam construir informação

O Processamento de Linguagem Natural...

Desafio:

Você tem 10 segundos para ler um texto de 50 linhas.

Qual estratégia utilizar?

A comida que vai à mesa do araçatubense está bem mais salgada do que de costume. Itens básicos tiveram sucessivos aumentos de preços, como o arroz de cada dia, que teve uma alta superior a 50%. A inflação é motivada pela alta do dólar, que tornou mais competitivos os produtos brasileiros lá fora, estimulando as exportações, e também pelo aumento do consumo nas residências com o isolamento social por causa da pandemia de Covid-19.

Os itens que tiveram altas mais expressivas foram o óleo de soja, o arroz e as carnes suínas e bovinas, mas houve aumentos pontuais no preço do feijão e do leite, por causa da estiagem, que reduz a produção no campo. Os demais itens sofreram o impacto das vendas para o mercado externo – com o dólar acima dos R$ 5,00, os produtores optaram pela exportação, em detrimento do abastecimento interno.

Para se ter uma ideia, o custo de um pacote de cinco quilos de arroz para os supermercados era de R$ 12,00 em fevereiro e, hoje, é de R$ 23,00, alta de 91,66%. A projeção é de que chegue aos R$ 25,00 até o mês que vem. Para o consumidor final, o mesmo produto era vendido a R$ 17,90 no início do ano e, hoje, é comercializado a R$ 25,00, aumento de 39,66%.

As carnes suína e bovina também tiveram alta nos preços, motivada pelos embarques ao exterior. Conforme relatório da Secex (Secretaria de Comércio Exterior), foram exportadas 87,7 mil toneladas de carne suína in natura em agosto, leve queda de 2,8% na comparação com o total de julho, mas 87,5% acima do volume embarcado em agosto de 2019.

Segundo o Cepea (Centro de Estudos de Avançados em Economia Aplicada), a expressiva demanda chinesa por carnes, especialmente bovina e suína, tem sido o principal motivo de elevação nos embarques brasileiros.

Com a forte demanda no exterior e a consequente queda nos estoques internos, o preço subiu para o consumidor brasileiro. A carne suína teve alta superior a 54%, levando-se em conta que o quilo da costela suína, antes vendido a R$ 11,00, hoje é comercializado a R$ 17,00.

O mesmo ocorreu com a carne bovina. O acém, que custava R$ 16,00 o quilo, agora é vendido a R$ 25,00, 56% a mais. Neste caso, além das exportações, há também a entressafra que impacta nos preços. O leite também teve o preço reajustado por causa da estiagem, com aumento médio de 15%.

Outro produto em alta é o óleo de soja. Antes, um litro do produto custava R$ 3,20 no atacado; agora, custa R$ 5,75, ou seja, 79,68% a mais. Nas gôndolas, o consumidor que pagava R$ 4,50, hoje chega a pagar R$ 6,00. Os que supermercados que conseguem preços menores têm estoques do produto e ainda não repassaram os aumentos.

Neste caso, no entanto, a tendência, agora, é de queda, segundo o Cepea (Centro de Estudos Avançados em Economia Aplicada) da Esalq/USP, porque, com a demanda por farelo de soja, há excedente de óleo – ao ser processada, a soja resulta, em geral, em 70% de farelo e entre 19% e 20% de óleo.

Para o supermercadista Carlos Fernando Felipe, dono do Rosa Felipe e presidente da Rede Pas, que reúne 24 supermercados entre Castilho e Cafelândia, a preocupação é de que a tendência ainda é de alta para alguns produtos. “É muito difícil porque impacta demais no orçamento da população”, afirmou.

Segundo ele, os supermercados que possuem estoques estão tentando segurar o aumento para não repassar aos consumidores. “As altas foram sucessivas e estamos tentando segurar ao máximo, mas uma hora os estoques vão acabar e o dólar não tem tendência de queda, o que significa que as exportações vão continuar”, observou.

Outro fator preocupante, conforme Felipe, é que, com a pandemia de Covid-19, as indústrias de embalagem deixaram de produzir e as empresas já enfrentam falta de matéria-prima para embalagens de plásticos. “A indústria já sinalizou 30% de alta neste mês e os preços devem subir mais 20% em outubro, o que também vai impactar no valor final dos produtos”, antevê.

Para ele, a situação atual é uma novidade para toda a sociedade e é difícil prever o que irá acontecer daqui para frente. No entanto, ele calcula que a volta à normalidade deve levar pelo menos seis meses.

O Processamento de Linguagem Natural...

Desafio:

Você tem 10 segundos para ler um texto de 50 linhas.

Qual estratégia utilizar?

Leitura dinâmica?

Ler uma palavra por parágrafo?

Ler só as palavras do meio do parágrafo?

O Processamento de Linguagem Natural...

Sumarização de textos

Uma das tarefas mais utilizadas do PLN. Consiste em reduzir uma quantidade de texto sem perder a semântica.

Por meio de sumários nós podemos escolher filmes, ler artigos, aprender novos assuntos,...

...Enfim, realizar mais de uma tarefa ao mesmo tempo.

O Processamento de Linguagem Natural...

http://textsummarization.net/text-summarizer

Ver também a API

O Processamento de Linguagem Natural...

Sumarização, condensação e resumo de textos

O grande desafio é extrair termos representativos do conteúdo dos documentos

Termos: palavras ou frases.

Tarefas comumente realizadas:

1) Contar palavras num texto;

2) Comparar a uma lista de palavras proibidas;

3) Eliminar palavras não significativas (artigos, preposições, conjunções, etc.) e

4) Ordenar as palavras de acordo com sua frequência.

O Processamento de Linguagem Natural...

Sumarização, condensação e resumo de textos

O grande desafio é extrair termos representativos do conteúdo dos documentos

Termos: palavras ou frases.

http://localhost:8888/notebooks/contadorDepalavras.ipynb

Tarefas comumente realizadas:

1) Contar palavras num texto;

2) Comparar a uma lista de palavras proibidas;

3) Eliminar palavras não significativas (artigos, preposições, conjunções, etc.) e

4) Ordenar as palavras de acordo com sua frequência.

O Processamento de Linguagem Natural...

Sumarizadores automáticos nem sempre têm muito sucesso

As frases podem não fizessem sentido reuni-las.

Uma estratégia é:

Detecção do tema mais relevante do texto

O Processamento de Linguagem Natural...

Mas para isso, alguns conceitos devem ser compreendidos.

O Processamento de Linguagem Natural...

Normalização da grafia

Substituição das palavras por semelhantes apenas com caracteres maiúsculos ou minúsculos

Data Mining == datamining?

Para um hamano?

Para um código em C?

	00	16	32	48	64	80	96	112
0	NUL	DLE		0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
10	LF	SUB	*	:	J	Z	j	z
11	VT	ESC	+	;	K	[	k	{
12	FF	FS	,	<	L	\	l	|
13	CR	GS	-	=	M	]	m	}
14	SO	RS	.	>	N	^	n	~
15	SI	US	/	?	O	_	o	DEL

O Processamento de Linguagem Natural...

Normalização da grafia

Remoção de numerais

Remova também os símbolos, como “R$”, “$”, “US$”, “kg”, “km”, “milhões”, “bilhões” dentre outros.

Assim como as pontuações “?”, “!”, ...

O Processamento de Linguagem Natural...

Corpus e corpora

Corpus é um conjunto de frases/parágrafos/textos que descrevem um determinado assunto.

- Livros de um autor;
- Textos sobre energia solar;
- Notícias sobre coronavírus;
- Comentários em uma rede social ...

Conjunto de textos escritos e registros orais em uma determinada língua e que serve como base de análise.

Corpora

Conjunto de corpus (plural)

O Processamento de Linguagem Natural...

Termo/token

É a menor parte, ou combinação, de um corpus.

Elementos entre dois delimitantes (espaço, pontos, etc)

“O abacate é uma fruta muito boa e versátil”.

Todas as palavras são tokens/termos

Stop words

Processamento automático.

Estratégia é a limpeza dos conceitos menos importantes.

1 – Reduzir os predicados e conjunções;

de, o, a, da, para, em, que, … ?

Palavras com menos de 4 caracteres? Faz, ter, ver não são importantes?

Mas qual a melhor estratégia?

Mas quais são os predicados, artigos e conjunções que devem ser removidos?

Bag of words

O Processamento de Linguagem Natural...

Bag of word

Um vetor de palavras de um documento/assunto.

Por exemplo:

Bow = {neymar, bola, gol, Brasil, corinthians, placar}

Um texto pode ser identificado se é relacionado à futebol quando tem muitos correspondentes na matriz de similaridade

	neymar	bola	gol	Brasil	corinthians	placar
texto1	0	31	0	3	3	1
texto2	0	2	0	45	12	3
texto3	4	12	1	1	0	6
...						

O Processamento de Linguagem Natural...

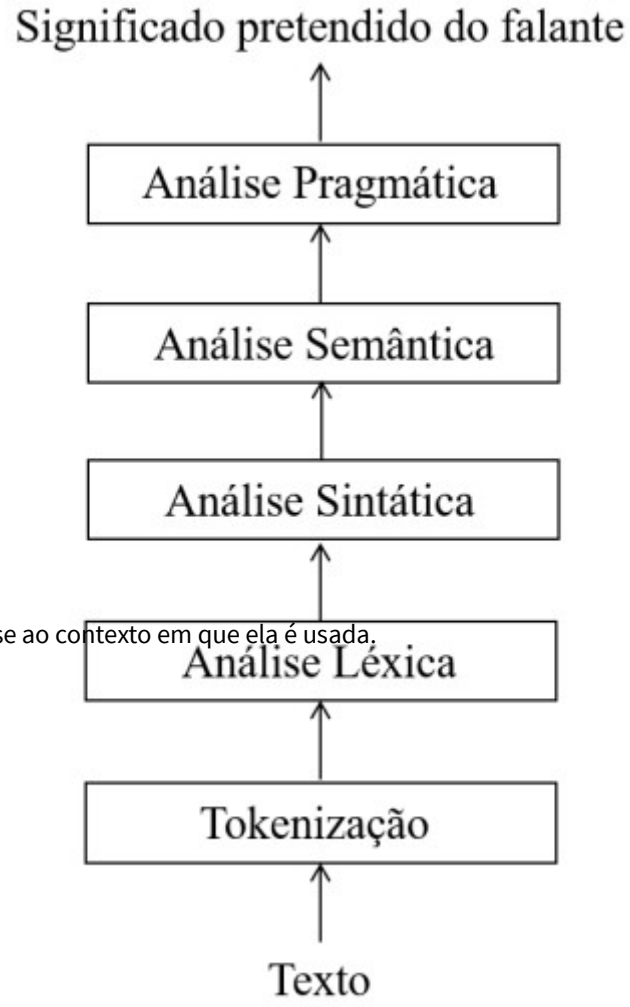
Identificação de um contexto por Bag of words...

http://localhost:8888/notebooks/bagOfWords.ipynb

Processamento de Linguagem Natural (PLN)

Sistematicamente:

5 fases de análise de texto



Pragmática: adequa o significado de uma frase ao contexto em que ela é usada.

Sumarização

Como um processo de sumarização pode ser conduzido?

Sumarização

Como um processo de sumarização pode ser conduzido?

A primeira etapa é a normalização da grafia;

Após isso, vem a identificação de tokens;

...

Processamento de Linguagem Natural (PLN)

Tokenização

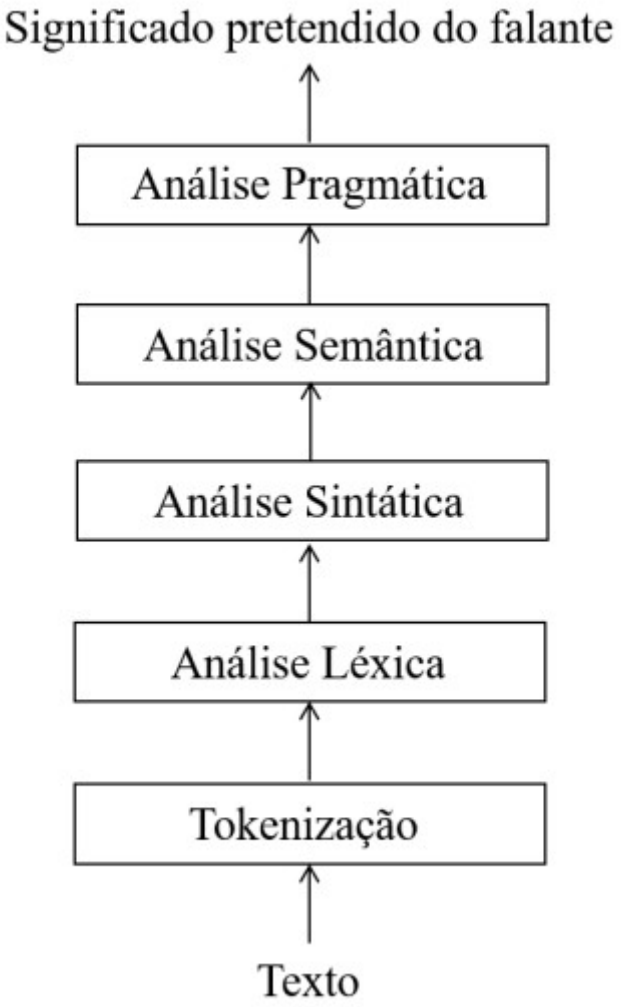
Uma das tarefas mais essenciais do PLN - identificar sequências de caracteres que separam textos de forma qu cada token possa ser comparado.

Ambiguidades são um grande problema.

10 mil == 10.000,00 ?

Sr. Rubens == Senhor Rubens?

Av. Brasil == 1.33 ?



Processamento de Linguagem Natural (PLN)

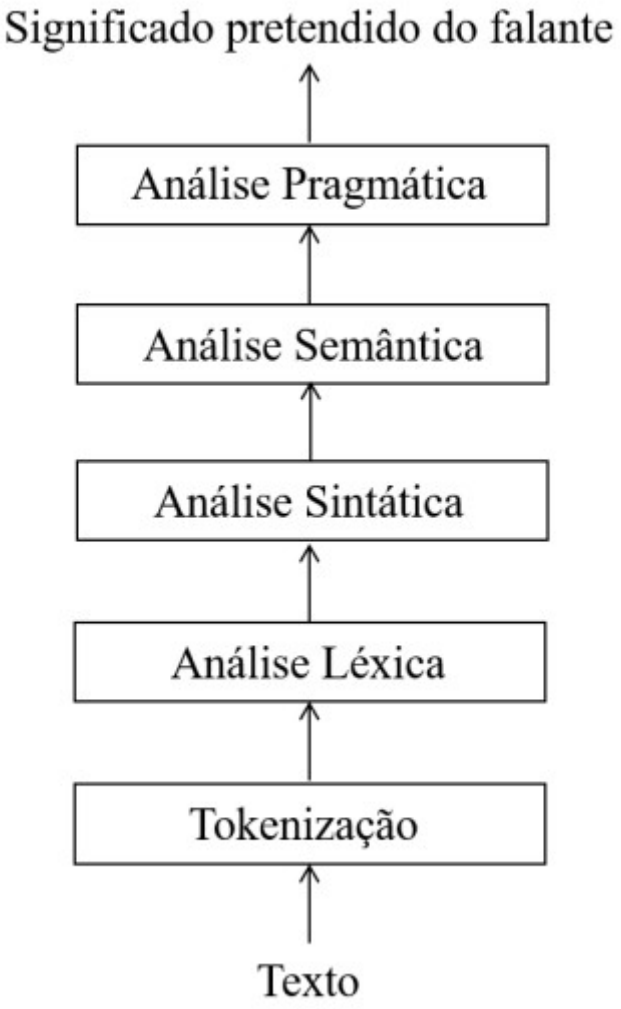
Análise Léxica

Após a tokenização, relaciona variantes morfológicas

Identificação de Lemmas (que originam palavras)

cantar, cantor, cantoria, cantando, etc…

entrega, entregador, entregar, entregando, etc...



Processamento de Linguagem Natural (PLN)

Análise Léxica

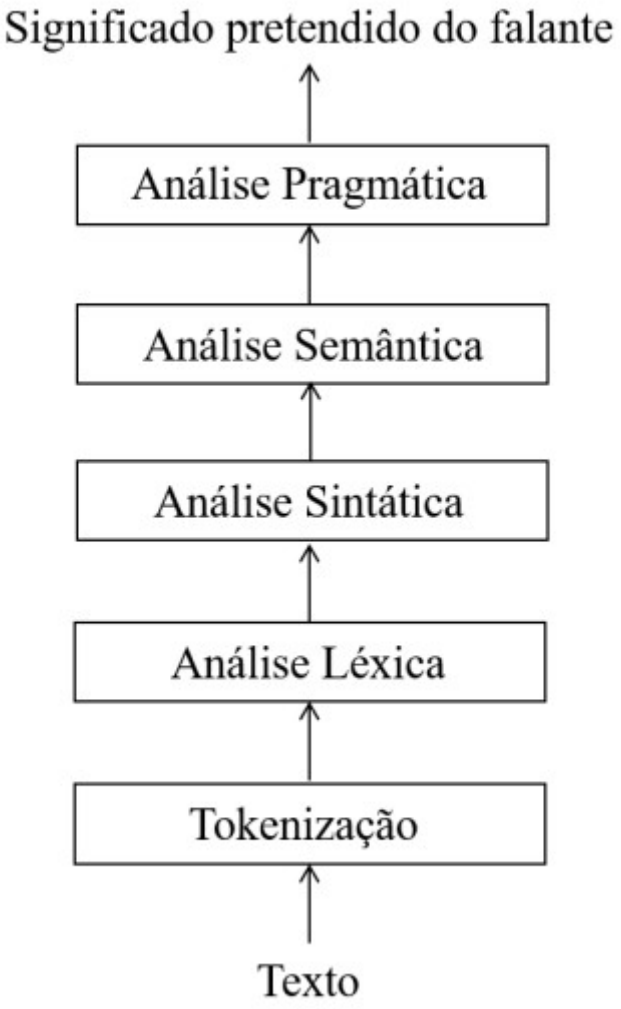
Nesta etapa é preciso considerar um dicionário de lemmas

- parsing side:

  mapeamento da palavra para seu lemma,

- geração morfológica:

  do lemma para a palavra, chamado de



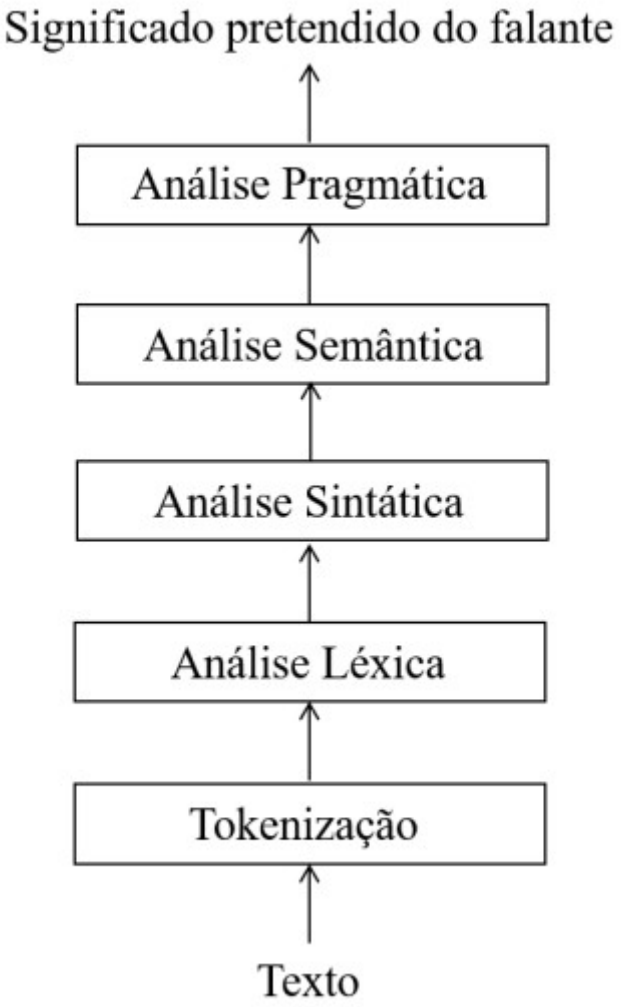
Processamento de Linguagem Natural (PLN)

Análise Léxica

Stemming: identificar o radical das variações morfológicas

cantar, cantor, cantoria, cantando, etc…

entrega, entregador, entregar, entregando, etc...



Processamento de Linguagem Natural (PLN)

O stemming consiste em reduzir a palavra à sua raiz (sem levar em conta a classe gramatical)

amig : amigo, amiga, amigão

gat : gato, gata, gatos, gatas

Lemmatization (a ação de reduzir em Lemmas)

Lemma: Forma básica da palavra

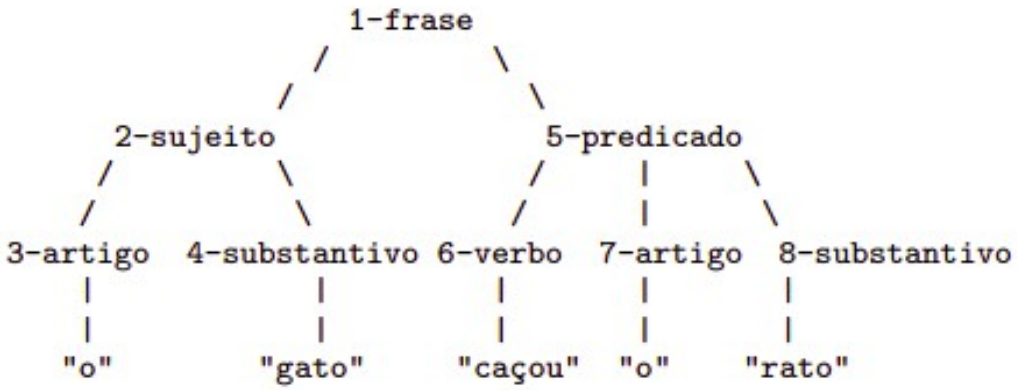
Lemmatizer: O artefato (programa)

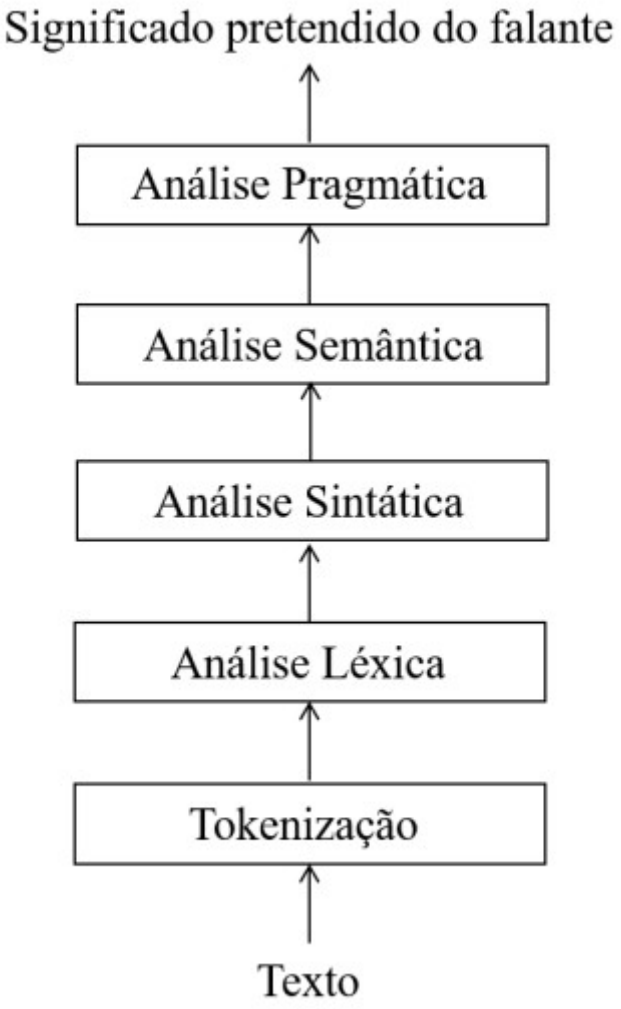
Algorithm for lemmatization

Processamento de Linguagem Natural (PLN)

Análise Sintática

syntax tree ou parse tree



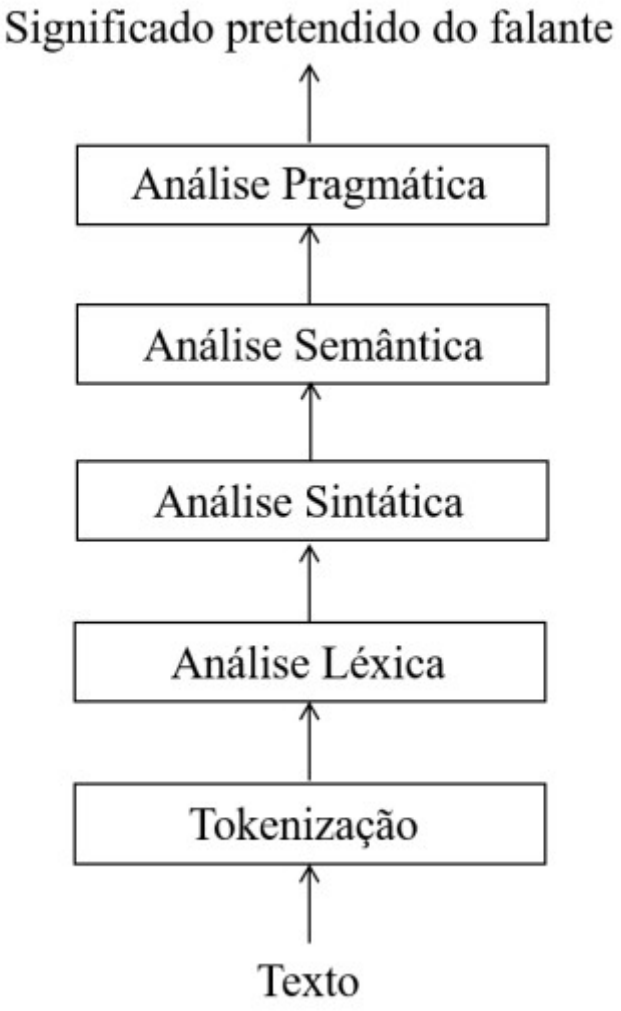


Processamento de Linguagem Natural (PLN)

Análise Pragmática

É a etapa que analisa a construção das frases do resumo.

Ou seja, qual a relação de cada toquem importante com os demais tokens do documento.



Sumarização

Como um processo de sumarização pode ser conduzido?

A primeira etapa é a normalização da grafia;

Após isso, vem a identificação de tokens;

Logo em seguida, a etapa de mensura de importância;

...

O Processamento de Linguagem Natural...

Após as ações básicas de pré-processamento

A relevância/importância de cada token/termo deve ser classificada.

Text Frequency-Inverse Document Frequency) (TF-IDF).

de natureza estatística, avalia o quanto um determinado termo, ou palavra, é importante em um documento

O Processamento de Linguagem Natural...

Text Frequency-Inverse Document Frequency) (TF-IDF).

Avalia o quanto um determinado termo, ou palavra, é importante em um documento

Termos que ocorrem com maior frequência isolados em geral não representam bem o documento, de fato, aumentam o nível de ruído nas respostas

Deve, ser avaliados quanto à todos os documentos do corpora.

Ideia → remover as palavras e ou frases de menor relevância.

O texto restante é considerado de maior relevância

Consequentemente, o texto já é reduzido

O Processamento de Linguagem Natural...

1. Identifica cada termo
2. Para cada termo, calcular TF, IDF e TF-IDF

$$TF = \frac{t}{tp} \qquad IDF = log\frac{N}{n} \qquad TF\text{-}IDF = TF * IDF$$

t é a quantidade de vezes que um termo ocorreu no texto;

tp é o total de termos existente no documento;

N é o total de documentos no corpus;

n = t

http://localhost:8888/notebooks/tf-idf.ipynb

O Processamento de Linguagem Natural...

1. Identifica cada termo
2. Para cada termo, calcular TF, IDF e TF-IDF
3. Para cada palavra importante calcula o TF-IDF das palavras a sua direita e a sua esquerda.

#	Texto original	Texto sem *stopword*	Texto apenas com palavras relevantes
1	A busca por fontes de energia limpa e renováveis é um dos grandes desafios da população mundial, o que faz com que pesquisadores procurem por soluções cada vez mais incomuns e inovadoras.	busca fontes energia limpa renováveis desafio população mundial faz pesquisadores procurem soluções mais incomuns inovadoras	fontes energia limpa renováveis desafios pesquisadores procurem soluções incomuns inovadoras
2	Pesquisadores da Universidade Brigham Young, de Washington (EUA), criaram uma célula de combustível que retira energia elétrica a partir da glicose e de outros açúcares, também conhecidos como carboidratos.	pesquisadores universidade brigham young, washington (eua) criaram célula combustível retira energia elétrica partir glicose outros açúcares conhecidos carboidratos	brigham young, washington (eua) criaram célula combustível retira energia elétrica glicose conhecidos carboidratos
3	Isso mesmo, a fonte de energia preferida do corpo humano pode, em um futuro próximo, alimentar desde o celular até um carro.	isso mesmo fonte energia preferida corpo humano futuro próximo alimentar desde celular carro	fonte energia preferida humano próximo alimentar celular

O Processamento de Linguagem Natural...

Cada palavra do resumo de primeira ordem é analisada quanto ao texto original:

Palavras	Palavras de maior frequência de associação						
	Palavras à esquerda				Palavras à direita		
fontes	as	das	de		de	do	e
	0,9%	0,81%	0,64%		0,18%	0,6%	0,02%
energia	de	a	e		a	e	solar
	0,45%	0,09%	0,09		0,02	0,07	0,03

Qual a palavra mais frequente à sua direita/esquerda?

O resumo é refeito com elas.

Vamos fazer com 1 palavra?

Sumário produzido automaticamente
As fontes de energia limpa e renováveis são desafios a pesquisadores que procurem soluções incomuns inovadoras. brigham young washington (eua) criaram uma célula de combustível que retira energia elétrica da glicose conhecidos por carboidratos. a fonte de energia preferida do humano é próximo de alimentar um celular

https://spacy.io/usage/models

O spaCy é uma biblioteca de processamento de linguagem natural (para Python) que tem desde funcionalidades “básicas” como Tokenização, que consiste em um pré-processamento do texto, até mais complexas que permitem treinar modelos estatísticos para classificação de textos.

Nltk é outra possibilidade

http://localhost:8888/notebooks/nltkSumariza%C3%A7%C3%A3o.ipynb

http://localhost:8888/notebooks/sumariza%C3%A7%C3%A3o%20com%20spacy.ipynb

Entidades

Reconhecimento de entidades

http://localhost:8888/notebooks/Entidades.ipynb

No próximo encontro...

Iremos trabalhar com Classificação e análise de sentimentos...

Mas, quando vai ser?