DEEP LEARNING CIn.AI Centro de Informática UFPE Universidade Federal de Pernambuco

Processamento de Linguagem Natural

Conceitos Introdutórios e Processamento de Texto

Prof. Luciano Barbosa &
Prof. Johny Moreira
{luciano, jms5}@cin.ufpe.br

O que é PLN?

Processamento de Linguagem Natural

É um conjunto de algoritmos, estratégias, tarefas e problemas que utilizam textos produzidos por humanos como entrada para a produção, extração e manipulação de informações úteis que antes não poderiam ser facilmente entendíveis por máquinas.

O que é PLN?

Processamento de Linguagem Natural

É um conjunto de algoritmos, estratégias, tarefas e problemas que utilizam textos produzidos por humanos como entrada para a produção, extração e manipulação de informações úteis que antes não poderiam ser facilmente entendíveis por máquinas.

- Estruturar Informação
- Obter Rótulos
- Geração de Representações Semânticas
- Extração de Relações Semânticas
- Tradução
- Sumarização
- Geração Automática de Texto
- Reconhecimento de Entidades
- Classificação de Texto
- Análise de Sentimento
- Classificação gramatical de palavras
- Parsing Sintático
- Mecanismos de Busca
- Sistemas de Diálogo (Chatbots)
- Perguntas e respostas
- Correção de erros de escrita e gramaticais

Corpus & Corpora

- Um único documento ou uma coleção de documentos
- Corpora: plural de corpus
- Types (Vocabulário): número de palavras distintas no corpus

Herdan's Law (1960) ou Heaps' Law (1978)

$$|V| = kN^{\beta}, \text{ geralmente... } 30 \leq k \leq 100,\ 67 < \beta < .75$$

N: número de tokens
V: vocabulário

Corpus	Tokens = *N*	Types = |*V*|
Shakespeare	884 mil	31 mil
Brown corpus	1 milhões	38 mil
Switchboard telephone conversations	2.4 milhões	20 mil
COCA	440 milhões	2 milhões
Google n-grams	1 trilhões	13 milhões

O **tamanho do vocabulário** cresce na razão do **número de tokens no corpus**

Diversos fatores influenciam a construção de um Corpora e o desenvolvimento das abordagens de PLN

- Quem escreveu (idade, gênero, classe social)
- Quando foi escrito
- Finalidade:
  - notícia, artigos científicos, livros de romance, tweets ou postagem na internet
- Língua: 7097 línguas no mundo
- Múltiplas línguas e dialetos
  - Mais populares: Chinese, Spanish, Japanese, German
  - Dialetos regionais
  - “Code switching”

African American English (AAF)	iont	talmbout
Mainstream American English (MAE)	I don’t	talking about

Por primera vez veo a @username actually being hateful! it was beautiful:)
[*Pela primeira vez vejo a @nomedeusuario sendo realmente odiosa! Foi lindo:)*]

dost tha or ra- hega ... dont wory ... but dherya rakhe
[*Ele foi e continuará sendo um amigo… não se preocupe… mas tenha fé*]

Palavras

“...se uma máquina deve ser infalível, ela não pode ser inteligente”.

O texto acima apresenta quantas palavras?

Palavras

“...se uma máquina deve ser infalível, ela não pode ser inteligente”.

O texto acima apresenta quantas palavras?

N: 16
V: [se, uma, máquina, deve, ser, infalível, ela, não, pode, inteligente, ‘.’, ‘...’, ‘,’, “‘“, “’”]

Palavras

“...se uma máquina deve ser infalível, ela não pode ser inteligente”.

O texto acima apresenta quantas palavras?

N: 16
V: [se, uma, máquina, deve, ser, infalível, ela, não, pode, inteligente, ‘.’, ‘...’, ‘,’, “‘“, “’”]

N: 11
V: [se, uma, máquina, deve, ser, infalível, ela, não, pode, inteligente]

Palavras

“...se uma máquina deve ser infalível, ela não pode ser inteligente”.

O texto acima apresenta quantas palavras?

N: 16
V: [se, uma, máquina, deve, ser, infalível, ela, não, pode, inteligente, ‘.’, ‘...’, ‘,’, “‘“, “’”]

N: 11
V: [se, uma, máquina, deve, ser, infalível, ela, não, pode, inteligente]

A pontuação é um fator crítico em alguns casos…

- Encontrar limites
- Identificar aspectos de significado
- Perguntas
- Marcas de exclamação
- Citações
- Identificar classes gramaticais
- Análise sintática

N: número de tokens
V: vocabulário

n-gramas

- Uma sequência de N itens (sejam eles palavras, caracteres, sílabas, fonemas, números…)
- Também chamados de sintagmas

“...se uma máquina deve ser infalível, ela não pode ser inteligente”.

Bi-grams: {..., “se uma”, “uma máquina”, “máquina deve”, …, “pode ser”, “ser inteligente”}
Tri-grams: {..., “se uma máquina”, “uma máquina deve”, “máquina deve ser”, …}

Alguns outros conceitos…

- Léxico (Dicionário)
  - Lista de palavras usadas pelo sistema
  - Contém outras informações como classe gramatical, gênero da palavra…
- Sintaxe
  - Maneira como as palavras são dispostas no texto e das relações entre elas, modelando o conhecimento inconsciente que as pessoas têm da gramática
- Gramática
  - Regras que definem as cadeias de palavras válidas em uma língua
- Semântica
  - É o significado usado pelos humanos para se expressar através da linguagem

Etapas do Processamento de Linguagem Natural

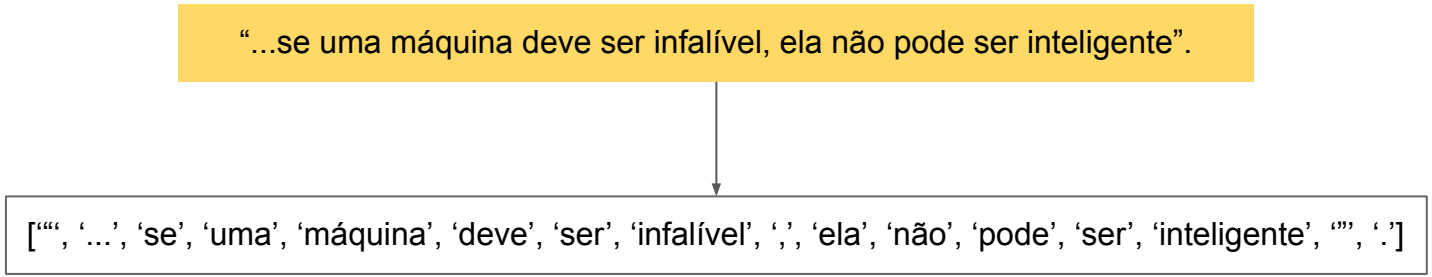
- Processamento morfológico
  - Corrige erros de digitação e trata variações de escrita
  - Utiliza o Léxico
- Processamento morfossintático
  - POS-tagging
- Processamento sintático
  - Utiliza a Gramática para determinar a estrutura das frases
  - Parsing sintático
- Análise semântica
  - Determinar o significado das palavras, frases e textos
- Análise do discurso
  - A produção de sequências estruturadas de frases
- Processamento pragmático
  - O uso da língua na interação social

Preparação: Pré-processamento

1. Tokenização de Palavras
2. Normalização do Formato de Palavras
3. Segmentação de Sentenças

Tokenização

- Quebrar sequência de caracteres em palavras
- Exemplo de abordagens simples:
  - Qualquer sequência de caracteres alfanuméricos de tamanho mínimo 3
  - Terminado em espaço
  - Terminado em algum caracter especial



Tokenização: Dificuldades

- Palavras pequenas podem ser importantes em algumas consultas
  - Ex: am, pm, el (paso), (world war) II, se
- Hífens algumas vezes são necessários
  - Ex: e-bay, cd-rom, t-shirts, guarda-chuva, bem-te-vi
- Caracteres especiais são importantes para URL, tags, preços e código em documentos
  - AT&T, R$45.55, 01/02/06, #NLPRules, usuario@gmail.com
- Apóstrofo pode ser parte de uma palavra, parte de um possessivo (inglês), ou erro

TEXT: Visite o nosso site: https://www.itv.org/
TOKENS: [“visite”, “o”, “nosso”, “site”, “https”, “www”, “itv”, “org”]

TEXT: “Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing”
TOKENS: [“mr”, “neill”, “thinks”, “that”, “the”, “boys”, “stories”, “about”, “chile”, “capital”, “aren”, “amusing”]

Tokenização: Dificuldades

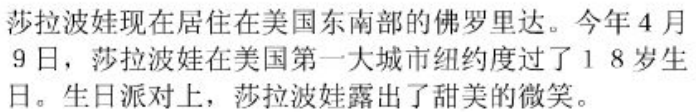
- Números podem ser importantes
  Ex: nokia 3250, united 93, quicktime 6.5 pro
- Pontos podem estar em números, abreviações, URLs, fim de sentenças etc
  Ex: I.B.M., Ph.D., U.S.A., N.S.A

TEXT: “Bigcorp’s 2007 bi-annual report showed profits rose 10%”
TOKENS: [“bigcorp”, “2007”, “annual”, “report”, “showed”, “profits”, “rose”]

TEXT: “She bought the newly announced iphone 14 pro”
TOKENS: [“she”, “bought”, “the”, “newly”, “announced”, “iphone”, “pro”]

Tokenização: Dificuldades

- Linguagens como Chinês, Japonês e Tailandês que não usam espaços para marcar limites de palavras
- Cada caractere representa uma unidade de significado (morfema)
- Decidir o que é uma palavra em linguagens do tipo, é mais complexo



► **Figure 2.3** The standard unsegmented form of Chinese text using the simplified characters of mainland China. There is no whitespace between words, not even between sentences – the apparent space after the Chinese period (。) is just a typographical illusion caused by placing the character on the left side of its square box. The first sentence is just words in Chinese characters with no spaces between them. The second and third sentences include Arabic numerals and punctuation breaking up the Chinese characters.

Fonte: https://nlp.stanford.edu/IR-book/ - Capítulo 2 - The term vocabulary & postings lists

Tokenização: Dificuldades

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

Japonês - Múltiplos sistemas de escrita: Caracteres chineses com sílabas hiragana e letras em latin formanco uma expressão japonesa.

Mais detalhes consultar fonte: https://nlp.stanford.edu/IR-book/ - Capítulo 2 - Página 31.

Tokenização: Dificuldades

姚明进入总决赛
(“Yao Ming reaches the finals”)

1 - Segmentação Chinese Treebank

姚明	进入	总决赛
YaoMing	reaches	finals

2 - Segmentação Peking University

姚	明	进入	总	决赛
Yao	Ming	reaches	overall	finals

3 - Segmentação por caracteres

姚	明	进	入	总	决	赛
Yao	Ming	enter	enter	overall	decision	game

Tokenização: Ambiguidade

和尚

► **Figure 2.4** Ambiguities in Chinese word segmentation. The two characters can be treated as one word meaning 'monk' or as a sequence of two words meaning 'and' and 'still'.

Fonte: https://nlp.stanford.edu/IR-book/ - Capítulo 2 - The term vocabulary & postings lists

Tokenização: Substantivos compostos sem espaços

Exemplos em Alemão:

computerlinguistik
computer + linguistik
computational linguistics

lebensversicherungsgesellschaftsangestellter
leben + versicherung + gesellschaft + angestellter
life insurance company employee

Tokenização: Morfologia complexa

Palavra Turca:

Uygarlastiramadiklarimizdanmissinizcasina

“(behaving) as if you are among those whom we could not civilize”

Uygar “civilized” + las “become” + tir “cause” + ama “not able”

+ dik “past” + lar “plural” + imiz “p1pl” +

dan “abl” + mis “past” + siniz “2pl” + casina “as if”

Normalização

- É a tarefa de colocar as palavras/tokens em um formato padrão
- Existe perda de informação ortográfica
- Forma mais simples: **transformação de caixa (maiúsculas > minúsculas)**
  - Não recomendável em alguns casos:
    - Extração de Informação
    - Tradução
    - Tarefas de classificação de texto ou análise de sentimento…

Sem Transformação	Com Transformação
Peru (País)	peru (ave)
UNE (organização)	une (unir)
US (País)	us (verbo nós em inglês)
Apple (empresa)	apple (fruta)

Normalização: Stopwords

- Palavras que aparecem muito na coleção
- Não possuem muito significado
- Usualmente, não são boas para diferenciar
- Artigos, preposições, conjunções etc
- Criadas a partir de palavras com alta frequência ou de listas existentes
  - Ex: Stopwords em Português (https://gist.github.com/alopes/5358189)

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Figure 2.5 A stop list of 25 semantically non-selective words which are common in Reuters-RCV1.

Fonte: https://nlp.stanford.edu/IR-book/ - Capítulo 2 - The term vocabulary & postings lists

Normalização: Lematização

- Agrupar palavras com mesma raiz (lemma)
- Verbos no passado são colocados no presente
- sinônimos são unificados

foi → vai
melhor → bom
correr, corre, correu → correr
quero, queres, quereres → querer
am, are, is → be
car, cars, car's, cars' → car
He is reading detective stories → He be read detective story

Lematização: Parsing Morfológico

- Estrutura das palavras: quebra a palavra em morfemas
- Morfemas
  - Menor unidade com “significado” que compõe a palavra
  - Stem (radical): unidade com o significado principal
  - Affixes (afixos): adiciona ao stem com uma função gramatical

Exemplo	Palavra:	unlikeliest
	Stem (radical):	likely
	Affixes (afixos):	un-, -est

- Essencial para linguagens morfologicamente complexas como árabe e turco

Normalização: Stemming

- Reduz variações morfológicas das palavras para um stem em comum
- Remove prefixo ou sufixo -> stem

  Exemplo

  Palavras: connected, connecting, connection, connections
  Stem: connect

- Não há um consenso sobre benefícios (depende da língua)

Stemming: Porter Stemmer

- Um dos mais utilizados
- Consiste de uma série de regras executadas em cascata
- Comete erros e difícil de modificar

ATIONAL → ATE (e.g., relational → relate)
ING → ϵ if stem contains vowel (e.g., motoring → motor)
SSES → SS (e.g., grasses → grass)

This was not the map we found in Billy Bones's chest, but an accurate copy, complete in all things-names and heights and soundings-with the single exception of the red crosses and the written notes.

Thi wa not the map we found in Billi Bone s chest but an accur copi complet in all thing name and height and sound with the singl except of the red cross and the written note

Segmentação de Sentenças

- Caracters ! e ? são precisos para separar
- Caracter “.” nem tanto
  - Abreviação: Dr.
  - Número: 7.5%
- Estratégia mais usada:
  - Toqueniza primeiro
  - Usa regras ou Machine Learning para classificar o ponto
  - Uso de dicionário de abreviação

Ferramentas e Bibliotecas

- Stanford tokenizer para inglês (http://nlp.stanford.edu/software/tokenizer.shtml)
- CoreNLP (https://stanfordnlp.github.io/CoreNLP/)
- **NLTK** (https://www.nltk.org/)
- **Spacy** (https://spacy.io/)

Aula Prática

Google Colab

Resumo da Aula

- O que é PLN?
- Conceitos Básicos
  - Corpus
  - Corpora
  - Vocabulário
  - Palavras
  - N-gramas
  - Léxico
  - Sintaxe
  - Gramática
- Etapas do Processamento de Linguagem Natural
- Etapa de Preparação dos dados: Pré-processamento
  - Tokenização
  - Normalização
  - Lematização
  - Stemming
  - Segmentação de Sentenças
- Ferramentas e Bibliotecas

Exercício Proposto

Google Colab

Referências

Dan Jurafsky, James H. Martin. Speech and Language Processing. (3rd ed. Draft). 2021. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/2.pdf>. Capítulo 2. Acesso em: 01 Setembro de 2022.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. Disponível em: <https://nlp.stanford.edu/IR-book/>. Capítulo 2. Acesso em: 01 Setembro de 2022.

Recursos e Ferramentas para a Lingua Portuguesa
http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources