

Processamento de linguagem natural e processamento de imagens

Aula 1

James Moraes de Almeida



Conteúdo do curso

- Qual o plano?
 - Processamento de linguagem natural.
 - Processamento de imagens.

Conteúdo do curso

- Qual o plano?
 - Processamento de linguagem natural (PNL ou NLP do inglês).
 - Mineração de dados.
 - Histórico.
 - Tokenização.
 - Pré-processamento.
 - Similaridade de sentenças.
 - Bag-Of-Words.
 - TF-IDF.
 - Modelagem de tópicos.
 - Grandes Modelos de Linguagem para extração de dados.
- Material principal: <https://www.youtube.com/@futuremojo>

Conteúdo do curso

- Qual o plano?
 - Processamento de imagens.
 - O que são imagens.
 - Filtros de imagens.
 - Transformadas em imagens.
 - Segmentação de imagens.
 - Detecção de objetos.
 - Aprendizado de máquina para análise de imagens.
 - Aprendizado de máquina profundo para análise de imagens.
- Material principal: <https://www.youtube.com/watch?v=Q5yVpymvjxU>

Avaliação

- Somativa:
 - Relatórios em dupla.
 - 1 relatório de processamento de linguagem natural.
 - 1 relatório de processamento de imagens.
 - Irei acompanhar o progresso dos relatórios.
 - Estejam prontos para me mostrar em visitas as mesas.
 - Em formato de artigo.
 - No máximo quatro páginas!
 - Estilo *letter*.
 - Testes pontuais.



O que é?

- Processamento de linguagem natural?
- É algo tão antigo quanto a computação!

A. M. Turing (1950) *Computing Machinery and Intelligence*. *Mind* 49: 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words in question. It is dangerous, if the meaning of the words "machine" and "think" are not defined, to examine how they are commonly used it is difficult to escape the conclusion that the answer to the question, "Can machines think?" is "no." In a statistical survey such as a Gallup poll. But this is absurd. Instead of the original question I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

ChatGPT broke the Turing test – the race is on for new ways to assess AI

Large language models mimic human chatter, but scientists disagree on their ability to reason.

By [Celeste Biever](#)

Processamento de Linguagem Natural

- PLN simbólico ou baseado em regras (1950).
 - Regras são usadas para processar o texto.
 - Regras programadas por pessoas com domínio na língua e especialistas no assunto.
 - Podem ser usadas regras linguísticas.
- Vantagens:
 - Altamente preciso para extração de dados.
 - Preciso para reconhecimento do texto.
- Desvantagens:
 - Muitas regras precisam ser programadas.
 - Pouco extrapolável.

Processamento de Linguagem Natural

- PNL estatístico ou aprendizado de máquina (1980).
 - Matemática empregada para entender a linguagem.
 - Quais são os padrões comuns encontrados na linguagem?
 - Com que frequência as palavras são usadas?
 - Há palavras que são utilizadas conjuntamente com grande frequência?
 - Vantagens:
 - Não é necessário programar regras.
 - Eficiente para análise do texto e estabelecer relações entre as palavras.
 - Desvantagens:
 - Como é estatístico, perde-se precisão nas extrações.
 - Pode variar bastante com os textos fornecidos.

Processamento de Linguagem Natural

- PNL de aprendizado de máquina profundo (2010).
 - Passa-se a utilizar redes neurais profundas.
 - Facilita-se mais a implementação.
 - Porém o custo computacional é maior.
- Vantagens:
 - A programação é "simples".
 - Muito eficaz em gerar textos.
- Desvantagens:
 - Custo computacional.
 - Pouco valor para dados quantitativos.

Conceitos básicos

- **Corpora:**
 - Se refere ao conjunto de textos coletados para sua análise com de PLN.
 - Comumente referido como **corpus**, que é o singular de corpora.
- **Token:**
 - São as palavras do texto, mas não só.
 - Podem ser números.
 - Também conjuntos de palavras (**N-gramas**).
 - "mesa redonda"
 - "guarda-chuva"
- **Stop Words:**
 - Conectivos, preposições, ...
 - Que, a, o, de, e, ...

Precisamos de um corpus!



Tokenização

- Precisamos estruturar um *corpus*.
 - O primeiro passo é separar os documentos em sentenças.
 - Depois as sentenças em palavras.
 - Mas textos não tem apenas palavras.
 - Pontuação.
 - Números.
 - Unidades.

Tokenização

- Poderíamos apenas utiliza a função split de processamento de string.
 - "A assinatura de inteligências artificiais de P.L.N. custa ~R\$100,00 por mês."

```
In [1]: 1 string="A assinatura de inteligências artificiais de P.L.N. custa ~R$100,00 por mês."
```

```
In [2]: 1 print(string.split(" "))  
['A', 'assinatura', 'de', 'inteligências', 'artificiais', 'de', 'P.L.N.', 'custa', '~R$100,00', 'por', 'mês.']
```

- Vejam que temos alguns problemas nesta tokenização.
 - Pontuação e unidades junto com as palavras.
 - Não seria interessante representar 'inteligências artificiais' como um token?
 - Como lidar com abreviações, como a 'P.N.L.'?
- Em geral há vários problemas e exceções.
 - Cada *corpus* teriam diferentes abordagens.

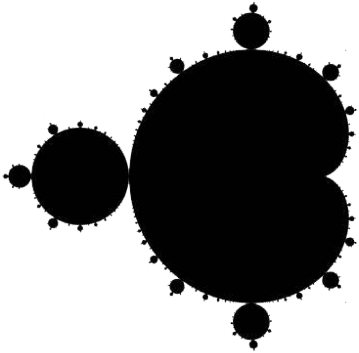
Tokenização

- Alguém já passou por este problema antes.
 - Felizmente já há soluções prontas e disponíveis em bibliotecas Python.

spaCy



Tokenizers



TextBlob



topic modelling for humans



NLTK



ChemDataExtractor 2

Tokenização

- Nosso material de apoio utilizará o spaCy
 - Há exemplos para o NLTK também.

Pré-Processamento

- O pré-processamento pode ser feito antes ou depois da tokenização.
- Consiste em geral nas seguintes estratégias.
 - Remover letras maiúsculas.
 - Diminui o vocabulário.
 - Pode ajudar, mas também pode levar a perda de informação:
 - Terra, terra.
 - Remover *stop words*.
 - Artigos, conjunções, pronomes, ...
 - *a, agora, ainda, alguém, algum, alguma, algumas, alguns, ampla, amplas, amplo, amplos, ante, antes, ao, aos, após, aquela, aquelas, aquele, aqueles, aquilo, as, até, através, cada, coisa, coisas, com, como, contra, contudo, da, daquele, daqueles, das, de, dela, delas, dele, deles, depois, dessa, dessas, desse, desses, desta, destas, deste, deste, destes, deve, devem, devendo, dever, deverá, deverão, deveria, deveriam, devia, deviam, disse, disso, disto, dito, diz, dizem, do, dos, e, é, ela, elas, ele, eles, em, enquanto, entre, era, essa, essas, esse, esses, esta, está, estamos, estão, estas, estava, estavam, estávamos, este, estes, estou, eu, fazendo, fazer, feita, feitas, feito, feitos, foi, for, foram, fosse, fossem, grande, grandes, há, isso, isto, já, la, lá, lhe, lhes, lo, mas, me, mesma, mesmas, mesmo, mesmos, meu, meus, minha, minhas, muita, muitas, muito, muitos, na, não, ...*
 - Pode ser muito bom para quando não tratamos contexto.
 - Diminui o vocabulário.

Pré-Processamento

- O pré-processamento pode ser feito antes ou depois da tokenização.
- Consiste em geral nas seguintes estratégias.
 - Radicalização das palavras.
 - Do inglês, *stemming*.
 - Levar as palavras a seu radical.
 - Diminui o vocabulário.

Radical	Exemplo
Mar:	Marinho
	Marítimo
	Maresia
Sol:	Solar
	Insolar
	Solecismo
Terr (relacionado a terra):	Território
	Terrestre
	Subterrâneo
Verb (relacionado a palavra ou falar):	Verbal
	Verbo
	Verborragia
Hidr (relacionado a água):	Hidrante
	Hidrografia
	Desidratação
Psic (relacionado a mente):	Psicologia
	Psiquiatra
	Psicossocial
Vid (relacionado a vida):	Vívido
	Reviver
	Sobrevivência

Pré-Processamento

- O pré-processamento pode ser feito antes ou depois da tokenização.
- Consiste em geral nas seguintes estratégias.
 - Lematização das palavras:
 - Leva os tokens a palavras presentes no dicionário.
 - Diminui o vocabulário.
 - Também diminui a quantidade de informação.

	Original	nlTK_stemmer	spacy_lemma
0	amigos	amig	amigo
1	amigas	amig	amigo
2	amizade	amizad	amizade
3	carreira	carr	carreira
4	carreiras	carr	carreira

Pré-Processamento

- *Named Entity Recognition* (NER):
 - Em português: Reconhecimento de Entidades Nomeadas.
 - Mais precisa que a POS.
 - Pode ser muito útil para extração de dados.
 - Estabelecendo relações entre as entidades das sentenças.

Categoria	Entidades Nomeadas	Proporção
Pessoa	2035	27.11%
Local	1250	18.15%
Organização	960	14.02%
Acontecimento	302	4.21%
Obra	437	6.31%
Abstração	278	4.50%
Coisa	304	4.38%
Tempo	1189	15.21%
Valor	352	4.53%
Outro	79	1.2%

Pré-Processamento

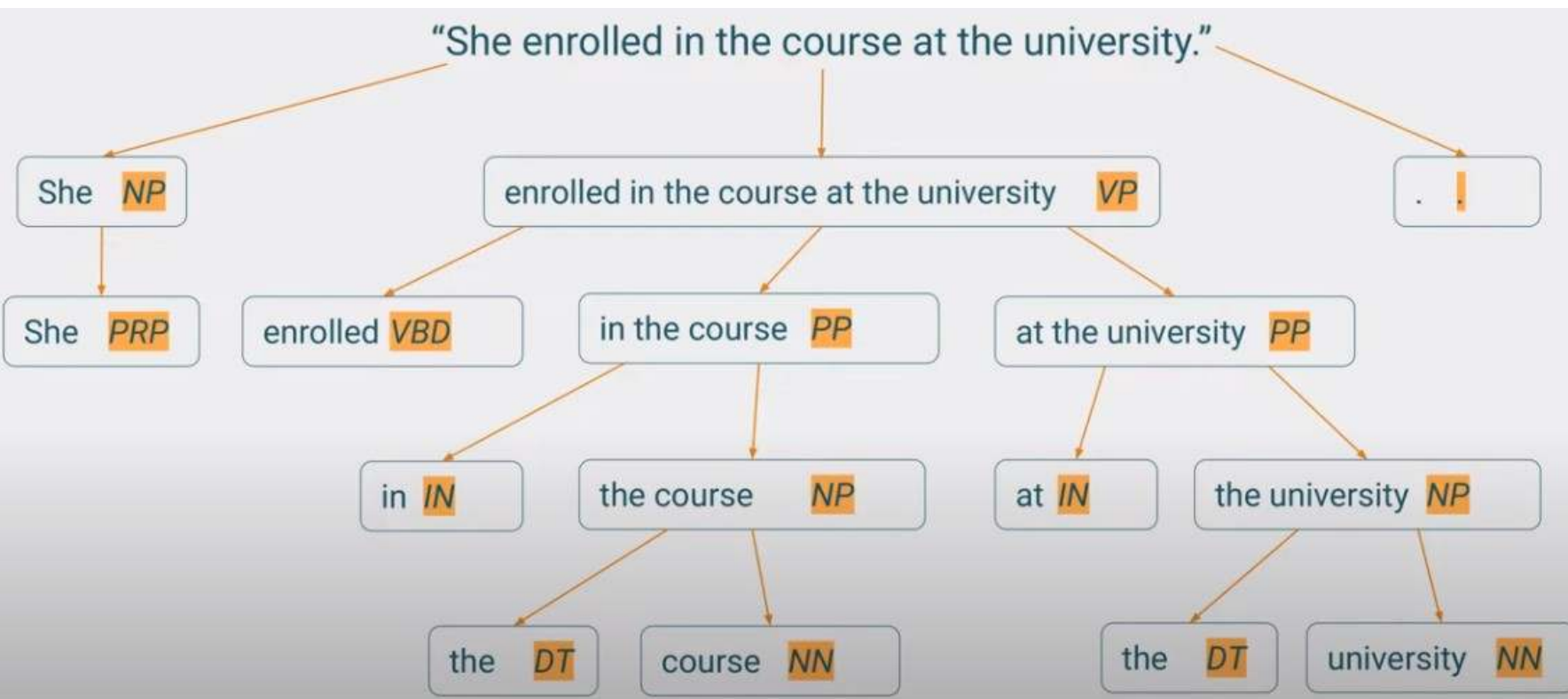
- Part-of-Speech (POS) tagging.
 - Em português: Etiquetagem morfosintática.
 - Classifica como as palavras estão sendo utilizadas no texto.

Palavras	O	cavalo	de	Napoleão	é	branco
Classes	ART	SUBS	ADP	SUBS	V	ADJ

- Ótimo para contexto:
 - "Eu gosto de comer **doce**."
 - "O pudim estava muito **doce**."

Pré-Processamento

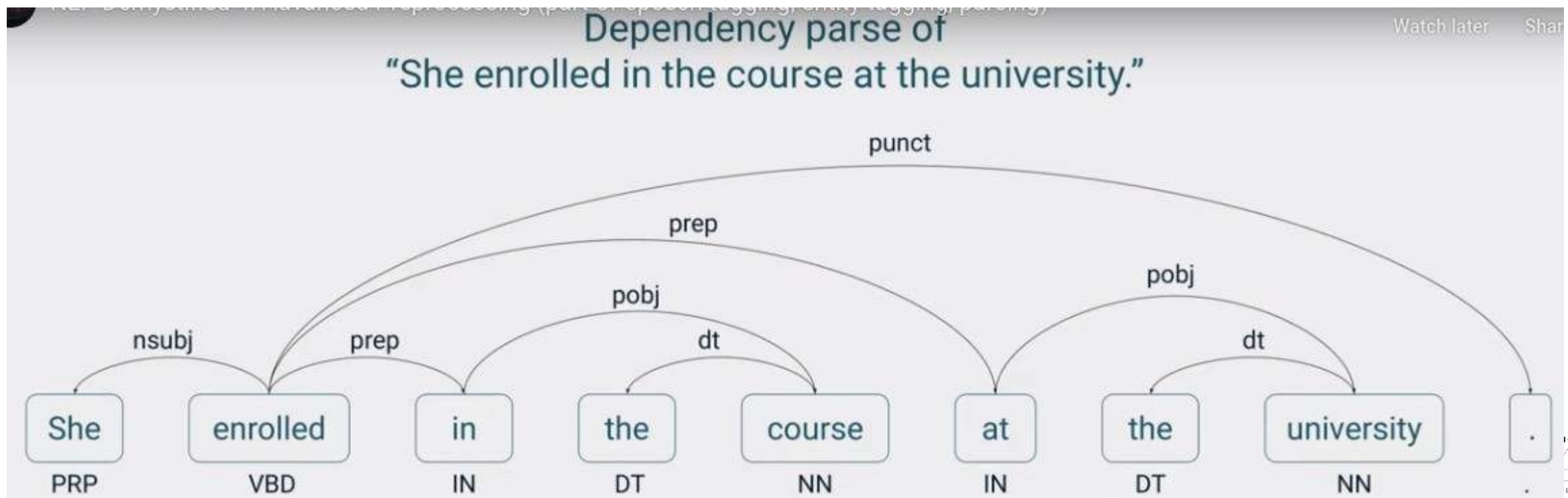
- *Parsing*:
 - Em português: Análise.
 - *Constituency parsing*: Análise por constituição.
 - Divide a sentença em diferentes partes.
 - Uma árvore é criada.



NP	Noun Phrase
VP	Verb Phrase
PP	Prepositional Phrase
NN	Noun
PRP	Personal Pronoun
NNP	Proper Noun
VB	Verb (base form)
DT	Determiner
IN	Preposition

Pré-Processamento

- *Parsing*:
 - Em português: Análise.
 - *Dependency parsing*: Análise por dependência.
 - Estabelece as dependências entre os tokens.
 - Pode ser poderosa para extração, usada em conjunto com NER.



Precisamos de um corpus!

- Vamos obter resumos do Web of Science.
- Escolham um tema que querem estudar.
 - Pensem numa propriedade quantitativa que querem obter.
 - Ex: Módulos de Young de metais.
- Após escolherem um tema, entrem no www.periodicos.capes.gov.br

OBRIGADO

James Moraes de Almeida
james.almeida@ilum.cnpem.br

ilum.cnpem.br