Formação Cientista de Dados

NLP

Corpus

- Conjunto de dados (texto não estruturado) em linguagem natural

```
1;Na era da informação e conhecimento, analisar dados não é uma atividade qualquer, para empresas e governos, é uma questão de sobrevivência. Num mundo globalizado e cada ve
2;E como se extrai informação e conhecimento de dados? Implementando projetos de Big Data. Um projeto, segundo definição do PMBOK (2013) é um esforço temporário para produzi
mação e conhecimento relevantes para a tomada de decisão.
3;Tudo isso para dizer que, de uma maneira geral, um projeto de Big Data nada mais é do que um projeto qualquer, com início e fim, escopo, orçamento, cronograma entre outros
4;Nesta obra vamos usar alguns termos que precisamos definir antes, de forma há não haver um entendimento ambíguo do seu significado:
5;Projeto de análise de dados ou de Big Data: Nesta obra vamos usar projetos de análise de dados ou de Big Data como uma definição genérica para qualquer tipo de projeto que
6;Dados de origem: analisar dados requer coletar dados de algum lugar. Vamos usar dados de origem para nos referirmos a dados ou conjuntos de dados que são utilizados para p
7;Dados de Staging: Muitos processos de analises de dados possuem uma etapa intermediária, entre os dados de origem e o resultado dos dados já consolidados. Esta etapa inter
8;Dados de destino: aqui estaremos sempre nos referindo ao resultado: o cubo, a previsão de uma análise preditiva, o relatório, o arquivo gerado, entre outros.
9;O PMBOK nos ensina que nem todos os seus processos são obrigatórios. Tampouco temos que aplicar os processos sempre com a mesma intensidade. Quais processos usar e em qual
10;Se você está lendo esta obra provavelmente já ouviu e leu muito sobre Big Data, e não será nosso objetivo aqui apresentar este conceito. Porém, muitos entendimentos são a
11;Falamos em seção anterior mas vamos repetir: cabe ao gerente de projeto decidir quais processos e com qual intensidade eles serão aplicados, conforme a complexidade do pr
12;Desde a pré-história o homem analisa dados. A análise de dados eletrônicos é um evento mais recente, com pouco mais de 70 anos. A capacidade de usar um computador para an
13;Mas a análise de dados só começou a tomar força na década de 90, foi quando os grandes armazéns de dados, ou data warehouse começaram a se tornar mais comuns para apoiar
14;Mas o que diferencia um projeto de análise de dados tradicional, como os popularizados nos anos 90, de um projeto de Big Data? Primeiro, os "Vs" que falamos na seção ante
15;Velocidade: a velocidade diz respeito não somente a da produção do dado em si, mas a velocidade do processamento e produção de informação e conhecimento, visto que o valo
e reais. O gráfico da Figura 1-1 abaixo mostra a relação inversa entre tempo e valor da informação. Toda a informação chegará a um tempo em que seu valor é zero, isso pode oc
16;Volume: projetos tradicionais eram construídos em armazéns de dados contendo por volta de terabytes de dados. Projetos de Big Data são de petabytes ou mais.
17;Variedade: projetos tracionais carregavam dados estruturados de sistemas de operação tradicionais, em modelos relacionais, hierárquicos ou de rede. Projetos de Big Data p
18;Mas além dos "Vs" existem outras diferenças significativas que devem ser consideradas em projetos de B&A. Vamos ver alguns:
19;Primeiro, do ponto de vista de arquitetura: projetos tradicionais tem uma arquitetura centralizada, enquanto Big Data é distribuída. Se o projeto tradicional precisar cre
20;Em projetos tradicionais, existe uma grande preocupação em só carregar dados em que, a partir de uma análise prévia, se vê valor. Estes dados são tratados e carregados em
21;Outra forma que podemos olhar uma solução de Big Data é sob sua arquitetura básica. Neste contexto, temos quatro elementos: fontes de dados, carga, armazenamento, analise
22;Quantos as fontes de dados, podemos ter nos dois casos os mesmos elementos: dados estruturados ou não estruturados. Porém, projetos de Big Data tem mais presentes fontes
```

Annotations

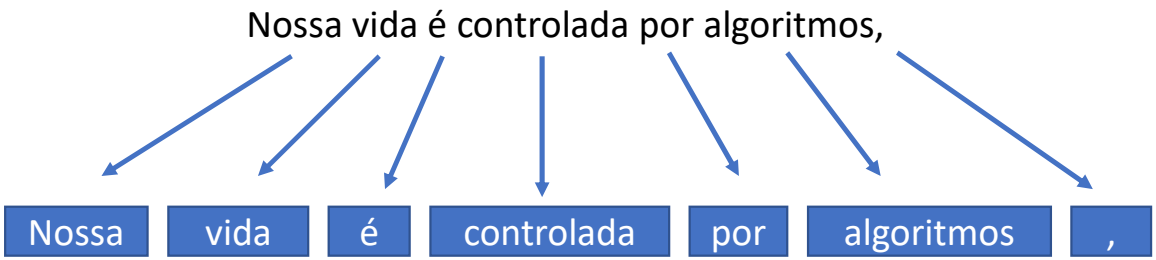
- Processo mais importante em NLP

- Coloca anotações no texto, como flexões, dependências etc.

1\n# newpar\n# sent_id = 1\n# text = Nossa vida é controlada por algoritmos, disse artista e professor de artes digitais de uma universidade americana\n1\tNossa\t_\tDET\tDET\t_\t2\tdet:poss\t_\t_\n2\tvida\t_\tNOUN\tNOUN\t_\t4\tnsubj:pass\t_\t_\n3\té\t_\tAUX\tAUX\t_\t4\taux:pass\t_\t_\n4\tcontrolada\t_\tVERB\tVERB\t_\t8\tccomp\t_\t_\n5\tpor\t_\tADP\tADP\t_\t6\tcase\t_\t_\n6\talgoritmos\t_\tNOUN\tNOUN\t_\t4\tnmod\t_\tSpaceAfter=No\n7\t,\t_\tPUNCT\t.\t_\t8\tpunct\t_\t_\n8\tdisse\t_\tVERB\tVERB\t_\t0\troot\t_\t_\n9\tartista\t_\tNOUN\tNOUN\t_\t8\tnsubj\t_\t_\n10\te\t_\tCCONJ\tCONJ\t_\t11\tcc\t_\t_\n11\tprofessor\t_\tNOUN\tNOUN\t_\t9\tconj\t_\t_\n12\tde\t_\tADP\tADP\t_\t13\tcase\t_\t_\n13\tartes\t_\tNOUN\tNOUN\t_\t17\tnmod\t_\t_\n14\tdigitais\t_\tADJ\tADJ\t_\t13\tamod\t_\t_\n15\tde\t_\tADP\tADP\t_\t17\tcase\t_\t_\n16\tuma\t_\tDET\tDET\t_\t17\tdet\t_\t_\n17\tuniversidade\t_\tNOUN\tNOUN\t_\t9\tnmod\t_\t_\n18\tamericana\t_\tADJ\tADJ\t_\t17\tamod\t_\tSpacesAfter=\\n\n\n"

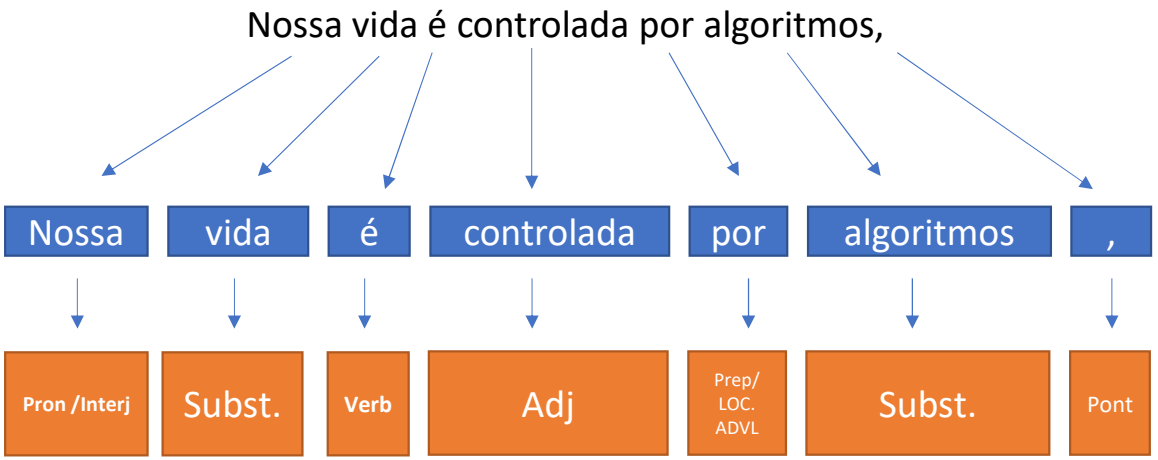
Tokenization

- Processo de separar a sentença em suas partes: palavras, pontos, símbolos etc.



Parts-of-Speech Tagging (POS)

Adiciona tags a cada token, como por exemplo, se é verbo, substantivo, adjetivo etc.



POS Tagging

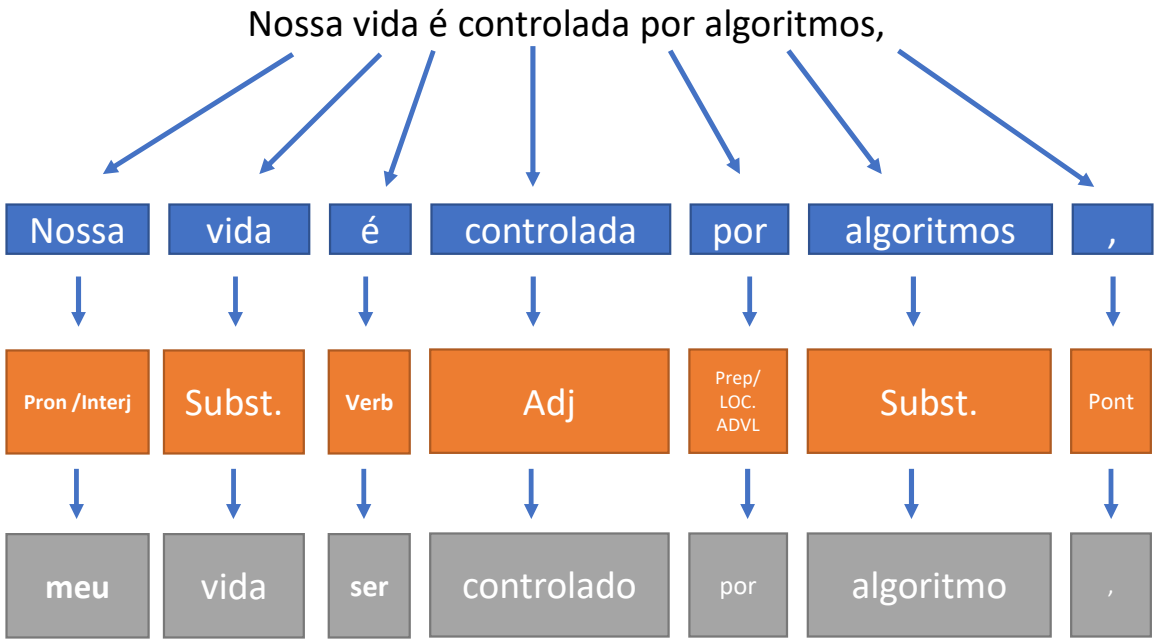
ABREVIAÇÃO	SIGNIFICADO	EXEMPLO
PROPN	Nome Próprio	José, Maria
VERB	verbo	Andar, Dirigir
ADP	Adposição	De, em, durante
DET	Determinante	A, Aquela, muitas
NOUN	Substantivo	Casa, carro
PUNCT	Pontuação	,.;
ADJ	Adjetivo	Infeliz, apavorado, brasileiro
CCONJ	Conjunções Coordenativas	E, nem, mas, entretanto
SCONJ	Conjunções Subordinativas	Embora, mesmo que, uma vez que
AUX	Verbos Auxiliares	Ser, estar, ter
PART	Funções de Partícula	Se, que
PRON	Pronomes	Meu, minha, meus, os quais
NUM	Números	10, vinte
ADV	Advérbios	Tarde, aqui, mal
SYM	Sinais Gráficos	~, ¨, ‘
INTJ	Interjeição	Ah, droga, psiu, hum
X	Outros	

Formação Cientista de Dados

PROF. FERNANDO AMARAL
www.datascientist.com.br

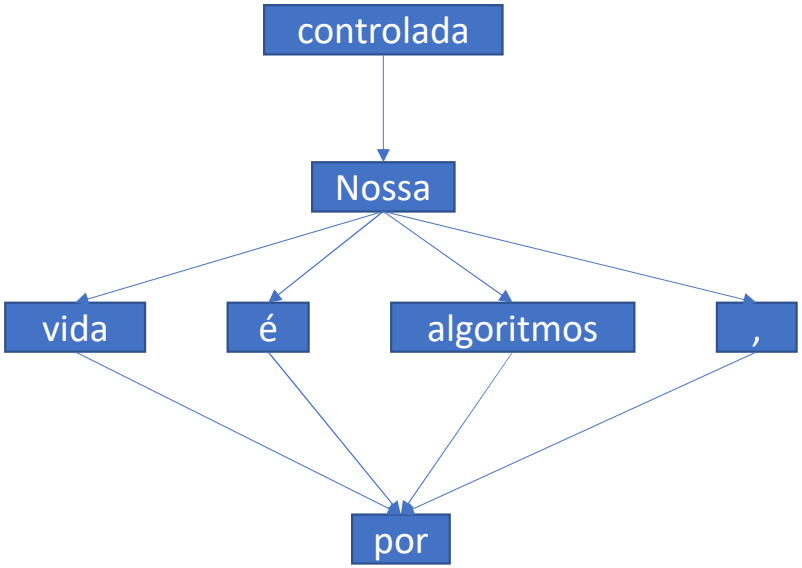
Lemmatizing (Lemma)

- Traz a palavra na sua flexão, de modo que possam ser analisadas juntas



Dependency Parsing

Encontra relação entre palavras “pais” e “filhos”

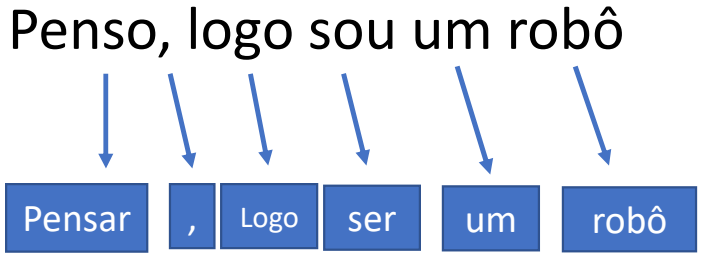


Modelo

- Análise
  - Verbo? Substantivo? Quais são as flexões? Quais as dependências?

- Um modelo é um banco de dados linguístico
- Específico de cada idioma
- Maioria das plataformas de NLP tem seus próprios modelos (ou usam de terceiros)
- Você pode criar o seu!

Pergunta: defina o processo abaixo:



- ❑Tokenization
- ❑Parts-of-Speech Tagging
- ☑Lemmatizing
- ❑Dependency Parsing