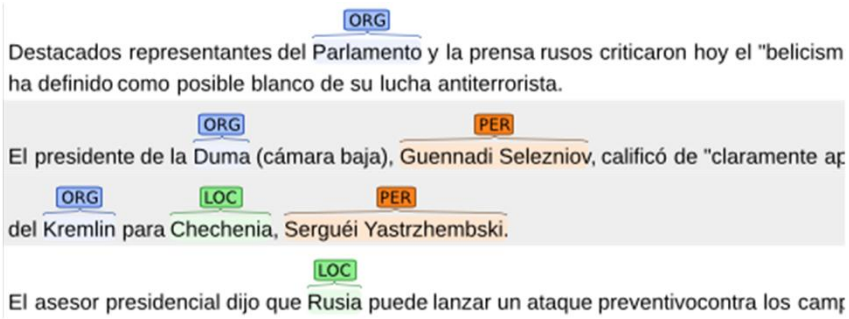
Corpus

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

```
1;Na era da informação e conhecimento, analisar dados não é uma atividade
2;E como se extrai informação e conhecimento de dados? Implementando proje
mação e conhecimento relevantes para a tomada de decisão.
3;Tudo isso para dizer que, de uma maneira geral, um projeto de Big Data n
4;Nesta obra vamos usar alguns termos que precisamos definir antes, de for
5;Projeto de análise de dados ou de Big Data: Nesta obra vamos usar projet
6;Dados de origem: analisar dados requer coletar dados de algum lugar. Vam
7;Dados de Staging: Muitos processos de analises de dados possuem uma etap
8;Dados de destino: aqui estaremos sempre nos referindo ao resultado: o cu
9;O PMBOK nos ensina que nem todos os seus processos são obrigatórios. Tam
10;Se você está lendo esta obra provavelmente já ouviu e leu muito sobre B
11;Falamos em seção anterior mas vamos repetir: cabe ao gerente de projeto
12;Desde a pré-história o homem analisa dados. A análise de dados eletrôni
13;Mas a análise de dados só começou a tomar força na década de 90, foi qu
14;Mas o que diferencia um projeto de análise de dados tradicional, como o
15;Velocidade: a velocidade diz respeito não somente a da produção do dado
e reais. O gráfico da Figura 1-1 abaixo mostra a relação inversa entre temp
16;Volume: projetos tradicionais eram construídos em armazéns de dados con
17;Variedade: projetos tracionais carregavam dados estruturados de sistema
18;Mas além dos “Vs” existem outras diferenças significativas que devem se
19;Primeiro, do ponto de vista de arquitetura: projetos tradicionais tem u
20;Em projetos tradicionais, existe uma grande preocupação em só carregar
21;Outra forma que podemos olhar uma solução de Big Data é sob sua arquite
22;Quantos as fontes de dados, podemos ter nos dois casos os mesmos elemen
```

Anotações: Annotations

- Localizar e classificar elementos específicos no texto
- Exemplos:
  - Anotar sentimentos para treinar um modelo de IA
- Pode ser específico do domínio: Ex: medicina
- Existem empresas especializadas em anotar
- Existem ferramentas especializadas: Doccano, brat etc.
- Alguns tipos podem ser feitos por máquina



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_

Tokenization

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

Nossa vida é controlada por algoritmos,						
Nossa	vida	é	controlada	por	algoritmos	,

Parts-of-Speech Tagging (POS)

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

Nossa vida é controlada por algoritmos,						
Nossa	vida	é	controlada	por	algoritmos	,
Pron /Interj	**Subst.**	**Verb**	**Adj**	**Prep/ LOC. ADVL**	**Subst.**	**Pont**

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**	

Parts-of-Speech Tagging

- Vamos comer porco (SUBS)
- Ele não toma banho, é um porco! (ADJ)

Lemmatizing (Lemma)

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

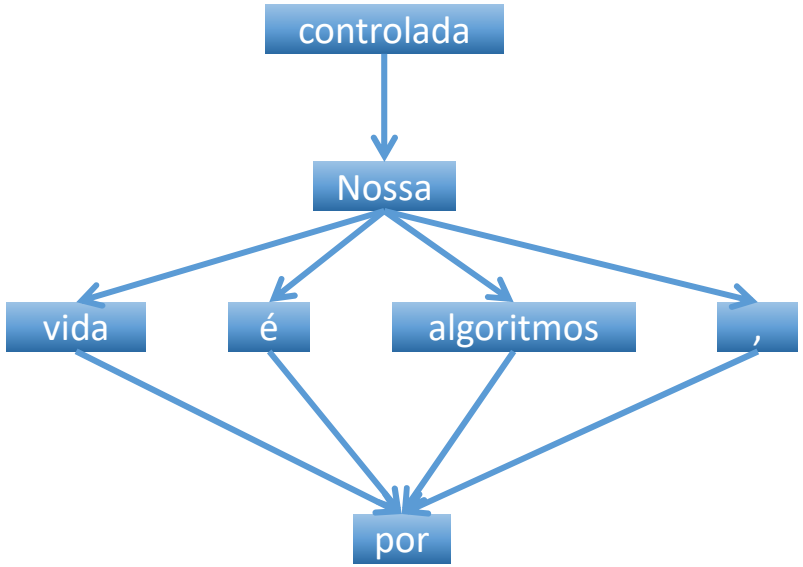
Nossa vida é controlada por algoritmos,						
Nossa	vida	é	controlada	por	algoritmos	,
Pron /Interj	**Subst.**	**Verb**	**Adj**	**Prep/ LOC. ADVL**	**Subst.**	**Pont**
meu	vida	ser	controlado	por	algoritmo	,

Stemming

- Cortas palavras, buscando ter uma representação raiz e única
- Diferentes técnicas
- Lemmatization é mais sofisticado
- Amigo, amigos, amiga, amigas => amig

Dependency Parsing

Encontra relação entre palavras “pais” e “filhos”



NGRAM

- N palavras consecutivas
- Bigrams e trigrams
- 4 ou mais não usado para palavras devido a esparsidade
- Pode também ser aplicado a letras

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!