Insper

Análise de texto de fontes desestruturadas e Web

Maciel Calebe Vidal
macielcv@insper.edu.br

Separação em Treino e Teste

- Train Test Split
- O que é Machine Learning?

Separação em Treino e Teste

- Train Test Split
- O que é Machine Learning?
- O modelo generaliza o que aprendeu?
- Irá ter boa perfomance em amostras nunca vistas?

Separação em Treino e Teste

- Aleatório

	Secao	Titulo
0	politica	Novas parcelas do auxílio emergencial ganham f...
1	economia	Café arábica recua de máxima de 4 anos na ICE;...
2	economia	Milho e soja tocam máximas de vários anos em C...
3	tecnologia	Brasil termina abril com 82.000 mortes por cov...
4	tecnologia	Vai comprar como?
5	politica	'Há uma grande operação abafa em curso', diz e...
6	politica	Conheça os projetos que propõem prorrogar o au...
7	tecnologia	A reestruturação do SAS
8	politica	TSE mantém protocolos sanitários no segundo turno
9	tecnologia	Laboratório Moderna, pioneiro da vacina antico...

Insper

Separação em Treino e Teste

- Base out-of-time

Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago

	Secao	Titulo
0	politica	Novas parcelas do auxílio emergencial ganham f...
1	economia	Café arábica recua de máxima de 4 anos na ICE;...
2	economia	Milho e soja tocam máximas de vários anos em C...
3	tecnologia	Brasil termina abril com 82.000 mortes por cov...
4	tecnologia	Vai comprar como?
5	politica	'Há uma grande operação abafa em curso', diz e...
6	politica	Conheça os projetos que propõem prorrogar o au...
7	tecnologia	A reestruturação do SAS
8	politica	TSE mantém protocolos sanitários no segundo turno
9	tecnologia	Laboratório Moderna, pioneiro da vacina antico...

Text Vectorization

- Word Embeddings
- O que queremos?
  - Treinar um modelo para classificar notícias
- Como o modelo vai aprender a partir do texto?

Google estende ações para proibição de anúncios políticos

CSN estende venda de suas ações da Usiminas

Text Vectorization

Converter texto em um conjunto de números (vetores)

Bag of Words

Notícia 1

CSN estende venda de suas ações da Usiminas

Notícia 2

Google estende ações para proibição de anúncios políticos

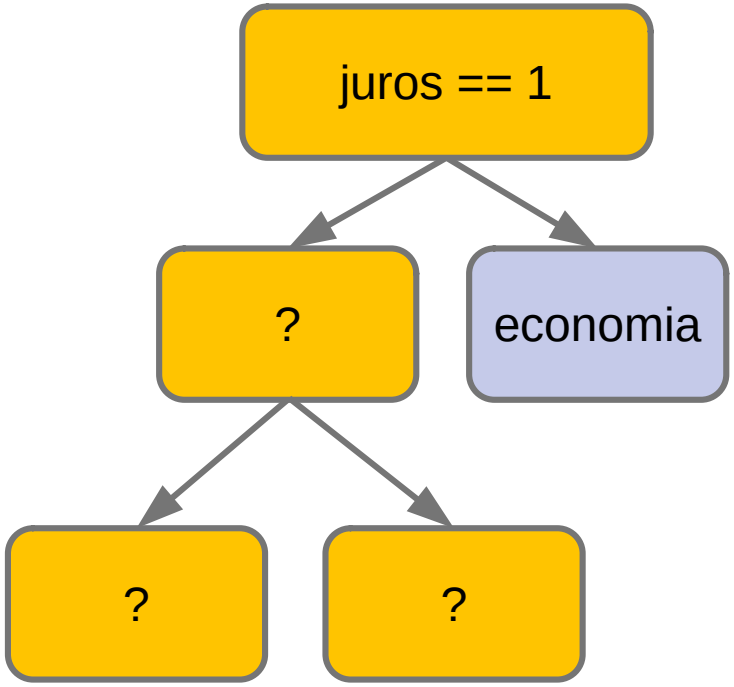
Palavra	Notícia 1	Notícia 2
CSN		
Ações		
Anúncios		
Usiminas		
proibição		
estende		
Google		
políticos		
venda		

Text Vectorization: Resultado

Notícia	governo	dólar	STF	COVID	...	juros	Categoria
1	1	1	1	1		0	política
2	0	0	0	1		0	tecnologia
3	1	1	0	1		1	economia
4	1	1	0	0		1	economia
5	0	0	1	1		0	política
6	0	1	1	1		0	política
7	0	1	1	1		1	economia
...	...	...	...	...	...	...	...
n	1	0	0	0		1	política

Exemplo Modelo

juros	Categoria
0	política
0	tecnologia
1	economia
1	economia
0	política
0	política
1	economia
...	...
1	política



Insper

Dúvidas?

Obrigado pela participação!