

SHIFT
FIAP

Ana Raquel



Carreira

- Tecnólogo em banco de dados pela faculdade FIAP.
- MBA em inteligência artificial pela FIAP.
- Mais de 8 anos de experiência como profissional na área de dados tendo atuado em diversos projetos de Banco de Dados, BI, Analytics e Data Science.
- Cientista de dados na FIAP e professora de Machine Learning , Deep Learning, Processamento de Linguagem Natural e Data Viz na FIAP.

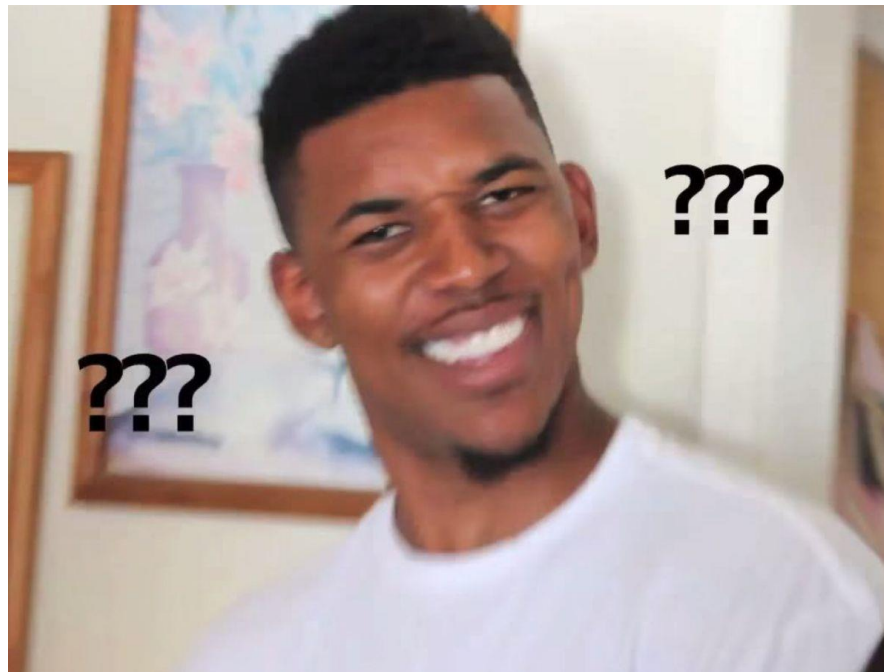


Case end to end - **Análise de sentimento** **em filmes**

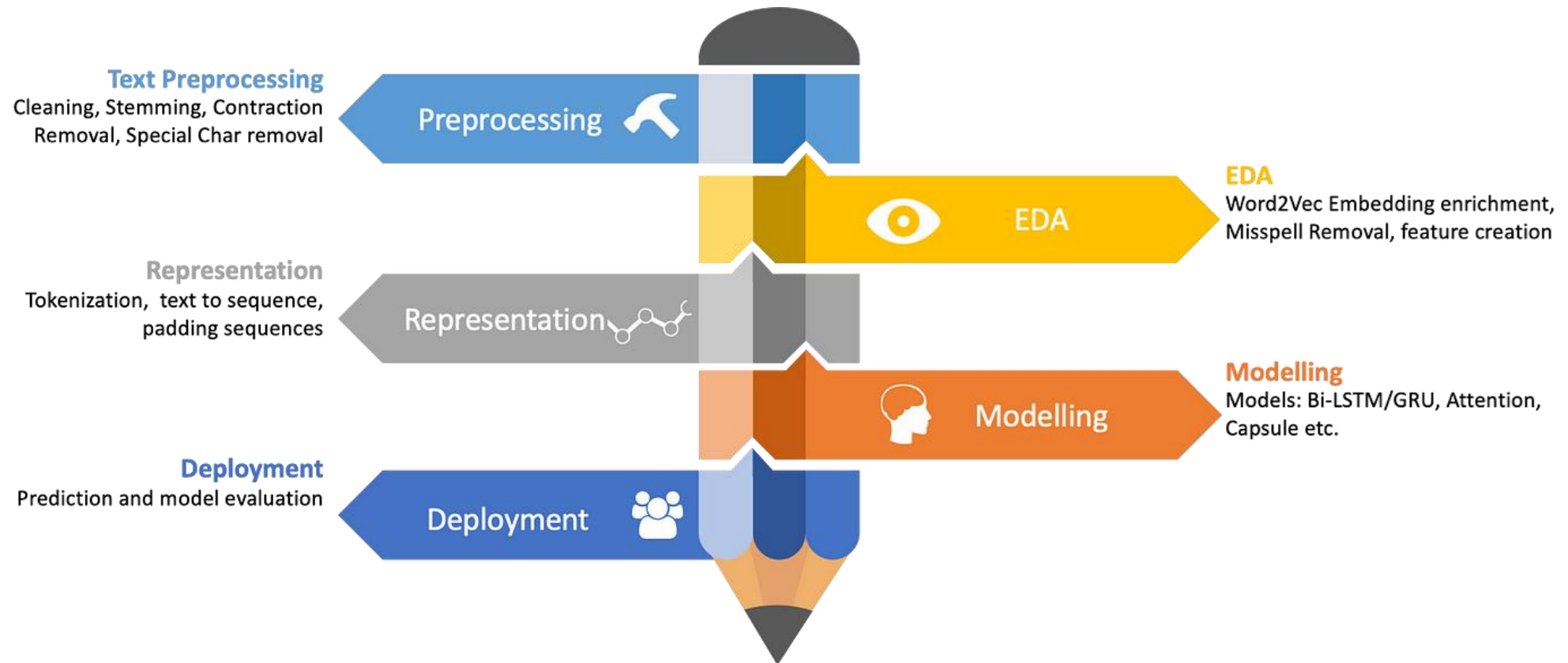
Como eu inicio um projeto ponta a ponta de NLP?


Aqui está um guia para seus próximos projetos de processamento de linguagem natural!

Visto as técnicas aprendidas em aula, como eu posso colocar em prática todos os conhecimentos? Por onde eu começo?



Como eu inicio um projeto ponta a ponta de NLP?





Base de dados representativa

Base de dados representativa

Com as bases de dados de estudo que utilizamos em aula, como por exemplo as bases de dados do kaggle, a coleta de dados parece ser uma tarefa muito simples! Onde, se eu tenho textos, eu posso analisar! Simples assim! Não é? Não :)

Extrair dados representativos se torna uma tarefa minuciosa, podendo se levar em conta alguns pontos importantes como:

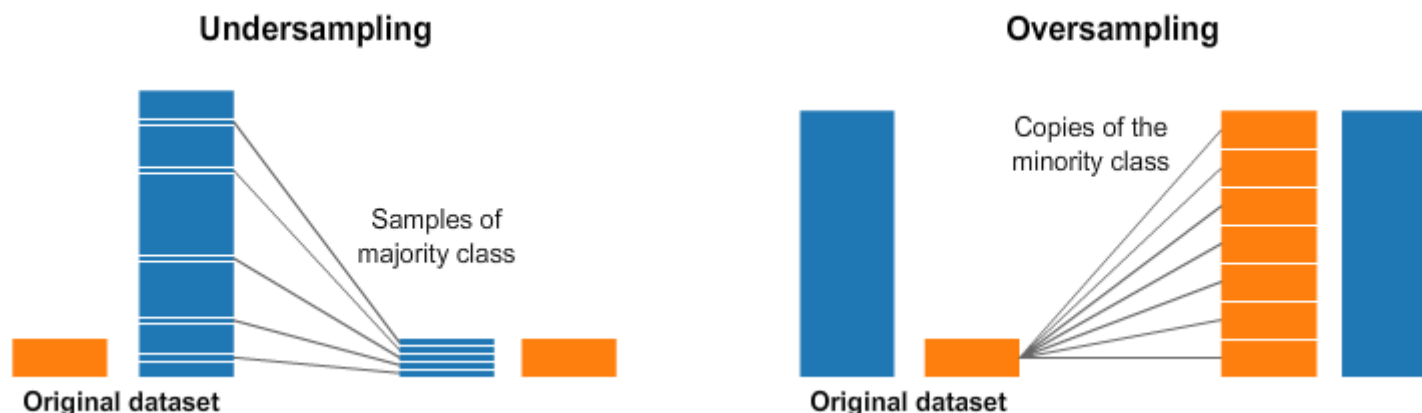
- Volume de dados (sua base é grande o suficiente para treinar um algoritmo? Cuidado com o underfitting!)
- Seus dados são representativos? (ou seja, seus dados de fato representam algo consistente?)
- Se você pensa em fazer um modelo de classificação, sua base está classificada?



Dados rotulados

Pensando aqui em um classificador de texto (como nosso case, um classificador de sentimentos), você já tem sua base pré-rotulada?

Aqui é importante manter a **base de dados equilibrada**, ou seja, **manter aproximadamente o mesmo volume de dados para todas as classes que deseja classificar**. O desequilíbrio da base de dados pode gerar uma amostra não representativa suficiente dos dados para o algoritmo tomar conhecimento dos comportamentos. Aqui você pode pensar em algumas técnicas tais como “undersampling” e “oversampling”, ou até mesmo, coletar mais dados :)



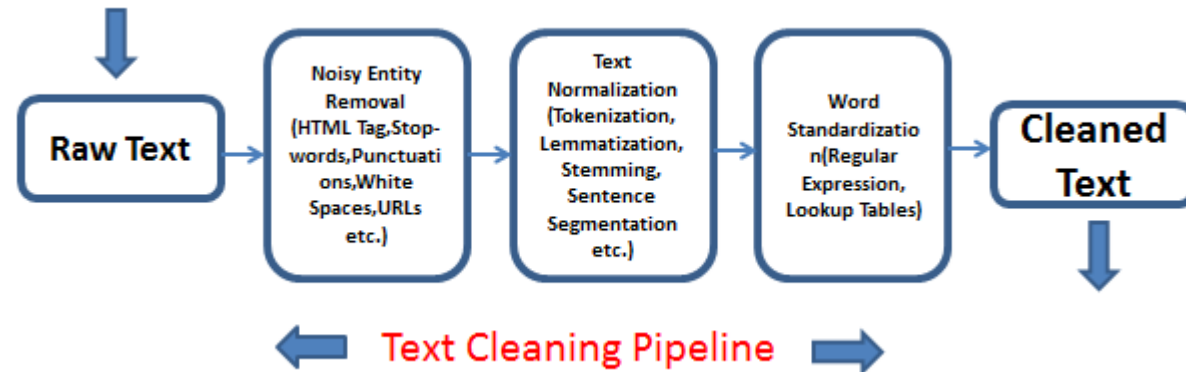


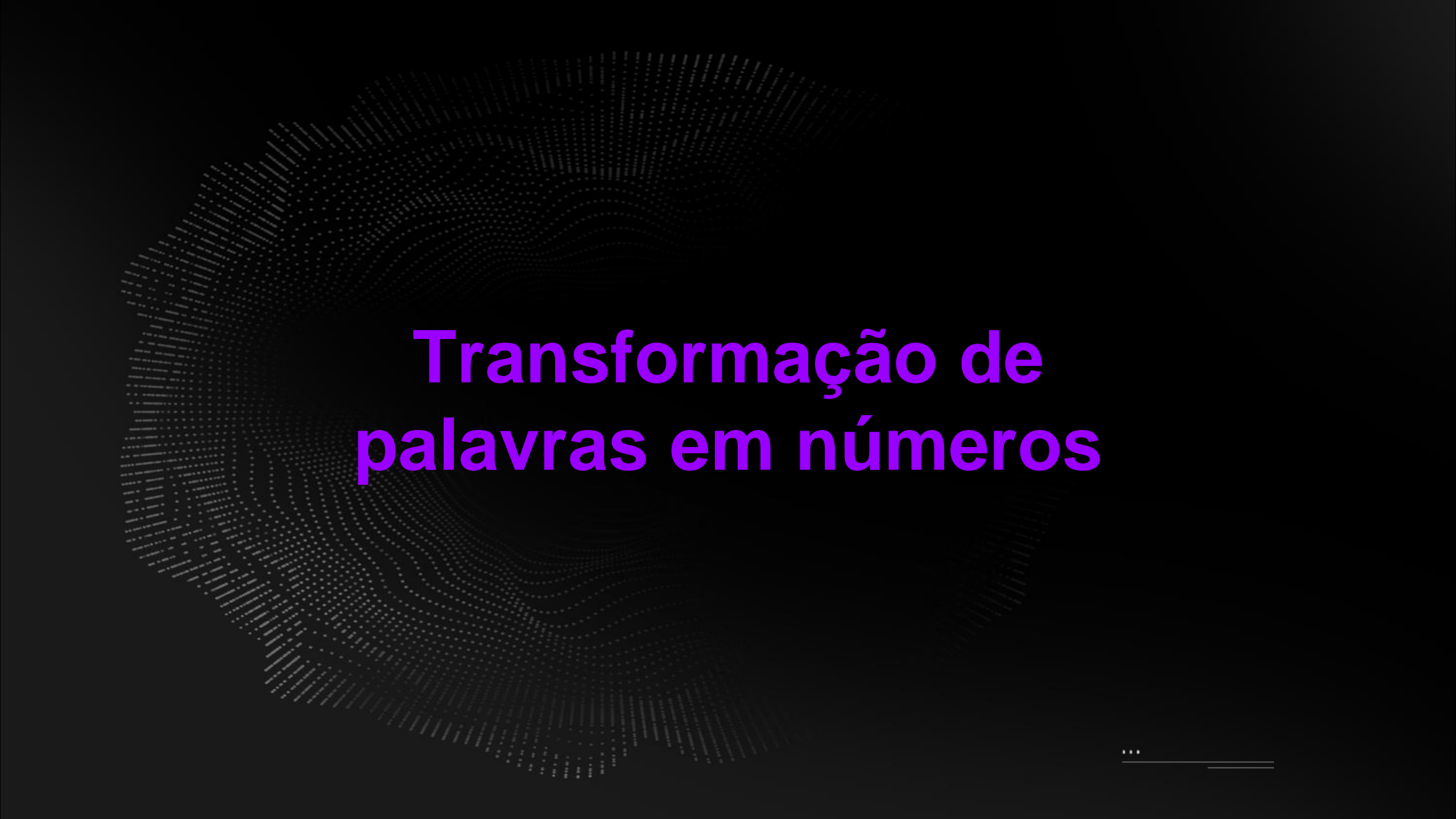
Dados “limpos”

Dados “limpos”

Um próximo passo para garantir um modelo de sucesso é a etapa de normalização dos dados:

- Remoção de dados nulos;
- Remoção de stop words;
- Lematização das palavras;
- Regex para limpar texto;

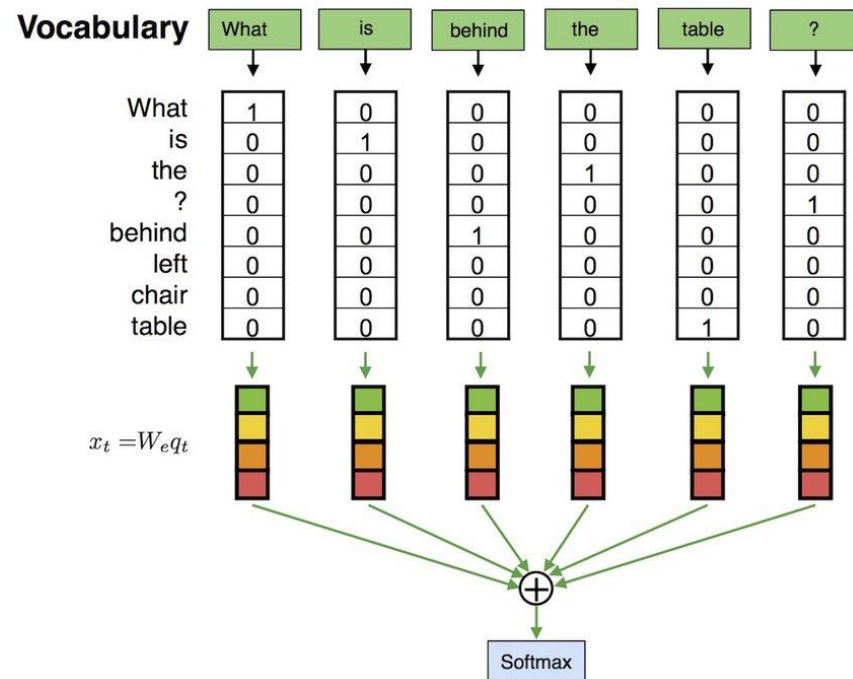




Transformação de palavras em números

Vetorização das palavras

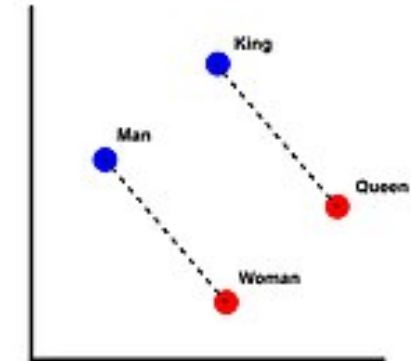
Técnica para transformar palavras em vetores numéricos para que possam ser processadas por algoritmos de aprendizado de máquina. Exemplos:



BOW

term	query			
	tf	df	idf	tf-idf
auto	0	5000	2.3	0
best	1	50000	1.3	1.3
car	1	10000	2.0	2.0
insurance	1	1000	3.0	3.0

TF-IDF



Word2vec



Próximos passos...

Obrigada!

Ana Raquel



[linkedin.com/ana-raquel-fernandes-cunha](https://www.linkedin.com/ana-raquel-fernandes-cunha)

Copyright © 2023 | Ana Raquel Fernandes Cunha

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem consentimento formal, por escrito, do professor/autor.