

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.

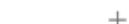


Text Mining

•

•

+



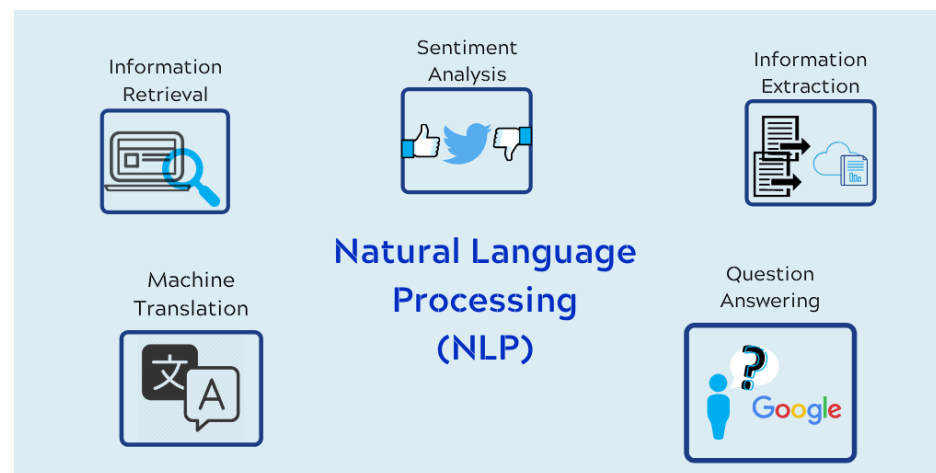
Algumas técnicas de data mining:

Recuperação de informações (Search Engine), como por exemplo o Google.



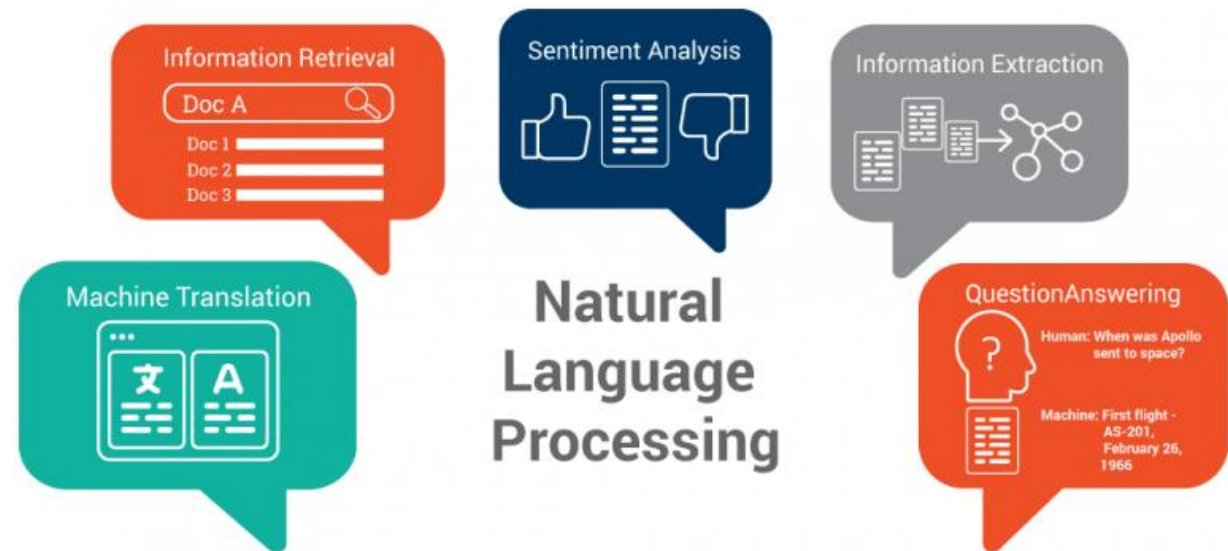
🔍 Pesquise no Google ou digite um URL

Análise de NLP, assim como estamos aprendendo aqui no curso 😊



Aonde podemos utilizar as técnicas de text mining?

- Chatbots
- Reviews de produtos
- Análise de sentimento
- Filtro de spam em e-mails



The background features a complex, abstract pattern of concentric circles and lines, creating a sense of depth and movement. The pattern is composed of numerous small dots and lines, which are arranged in a way that suggests a spiral or a series of overlapping waves. The colors are primarily dark, with some lighter, golden-brown highlights that add to the visual complexity.

Latent Dirichlet Allocation (LDA)

Introdução ao LDA

Para qual grupo você atribuiria esse documento?

Embora a indústria do turismo sempre tenha investido em Inteligência Artificial, nunca este processo teve tanta agilidade e implementação. Esta aceleração do sistema, gerada pela pandemia, redefine todo o mercado trazendo vantagens para o viajante e para todo o setor.

Ferramentas de IA, nuvens digitais, big data, RV e RA agilizam os processos, atendem as demandas, otimizam as receitas da empresa e possibilitam uma experiência única para os viajantes. Com a utilização de dados pessoais, de forma segura e perante aprovação prévia, é possível garantir a personalização do início ao final de cada viagem.

A satisfação do cliente também está na escala do atendimento, nada pior que aguardar atendimento quando hoje em dia o imediatismo faz parte do nosso dia a dia. Com o uso de IA para o turismo, o atendimento em grande escala, através de Chatbots, é imprescindível. Com o cliente sempre conectado, em poucos cliques, decisões e resolução rápida de problemas fidelizam o viajante.



Inteligência Artificial



Jurídico



Turismo

Introdução ao LDA

Para qual grupo você atribuiria esse documento?

Aparentemente, poderíamos classificar esse documento sobre o tema “**Turismo**” ou talvez “**Inteligência Artificial**”?



Introdução ao LDA

No documento temos algumas palavras chaves:

Embora a indústria do turismo sempre tenha investido em Inteligência Artificial, nunca este processo teve tanta agilidade e implementação. Esta aceleração do sistema gerada pela pandemia, redefine todo o mercado trazendo vantagens para o viajante e para todo o setor.

Ferramentas de IA, nuvens digitais, big data, RV e RA agilizam os processos, atendem as demandas, otimizam as receitas da empresa e possibilitam uma experiência única para os viajantes. Com a utilização de dados pessoais, de forma segura e perante aprovação prévia, é possível garantir a personalização do início ao final de cada viagem.

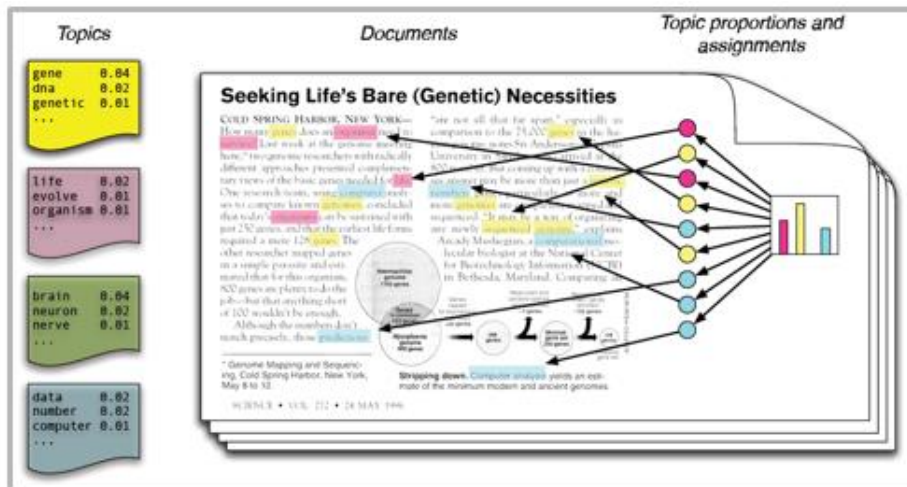
A satisfação do cliente também está na escala do atendimento, nada pior que aguardar atendimento quando hoje em dia o imediatismo faz parte do nosso dia a dia. Com o uso de IA para o turismo, o atendimento em grande escala, através de Chatbots, é imprescindível. Com o cliente sempre conectado, em poucos cliques, decisões e resolução rápida de problemas fidelizam o viajante.

Introdução ao LDA

Perceba que ambos os grupos fazem parte do mesmo documento, ou seja, **em um mesmo documento podem existir vários tópicos diferentes**.

Isso pode ser um problema para os **paradigmas da classificação e agrupamento**, pois o objetivo dessas técnicas é **definir uma classe para categorizar os dados**.

Por isso, temos como uma possível solução para esse problema o **Topic Modelling**, onde a técnica atribui para um determinado documento **probabilidades de pertencimento a um número “k” de tópicos**.



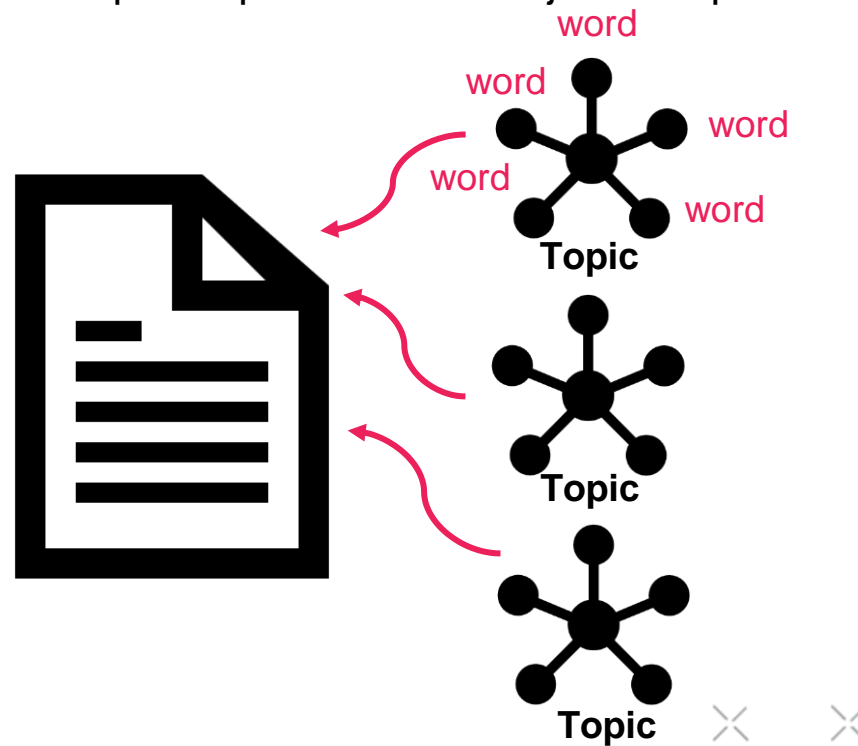
Encontrar tópicos que melhor descrevem os documentos

Tarefa não supervisionada

Conceito do LDA

“Cada documento pode ser descrito por **uma distribuição de tópicos** e **cada tópico pode ser descrito por uma distribuição de palavras**”.

Assim, podemos dizer que um documento foi gerado a partir de um conjunto de tópicos e cada tópico a partir de um conjunto de palavras.



Como descobrir os tópicos?

Imagine um conjunto de 1000 palavras e 1000 documentos. Assuma que cada documento possui, em média, 500 dessas palavras. Como descobrir a categoria a que cada documento pertence?

- Uma maneira é conectar cada documento a cada **palavra baseado em sua aparição no documento.**

		Palavras
Doc 1	Tomei um remédio por meio de injeção no hospital ontem. Doeu muito...	
		Remédio
		Injeção
		Governo
Doc 2	O fundo do governo aumentou o fornecimento de remédios.	Flores
		Árvores
Doc 3	Eu adoro a natureza! O mar, as montanhas, o céu, as árvores	Médico
		Montanhas
		Mar
		Céu
		Hospital

Como descobrir os tópicos?

Entretanto, essa abordagem **não é escalável**, pois ficaria impossível, visualmente falando, conseguir identificar todas as relações.

Doc	Palavras
Doc 1	
Tomei um remédio por meio de injeção no hospital ontem. Doeu muito...	
	Remédio
	Injeção
Doc 2	
O fundo do governo aumentou o fornecimento de remédios.	Governo
	Flores
	Árvores
Doc 3	
Eu adoro a natureza! O mar, as montanhas, o céu, as árvores	Médico
	Montanhas
	Mar
	Céu
	Hospital

Então como resolver?

Como descobrir os tópicos?

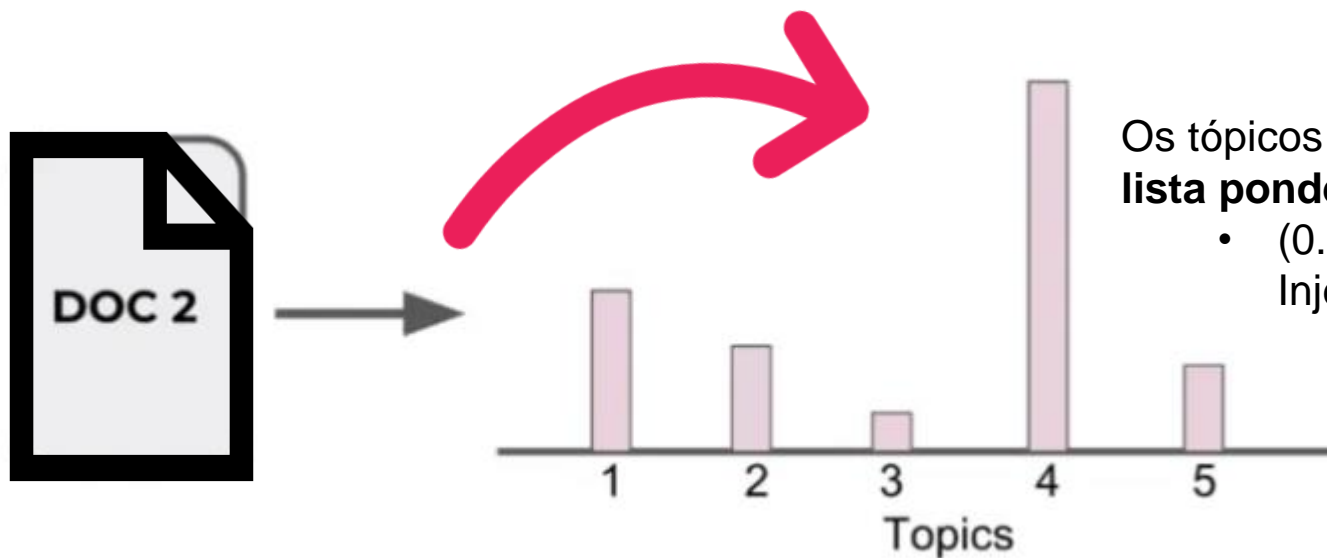
Utilizamos uma camada oculta nomeada como “latent” , supondo que **existam k tópicos que apareçam em todos os documentos.**

Documento 1					Palavras
Tomei um remédio por meio de injeção no hospital ontem. Doeu muito...					
Documento 2					
O fundo do governo aumentou o fornecimento de remédios.					
Documento 3					
Eu adoro a natureza! O mar, as montanhas, o céu, as árvores					

The diagram illustrates the process of discovering topics from a set of documents. Three documents are shown on the left, each with a highlighted title and a sentence. On the right, three topics are identified: 'Saúde' (Health) in blue, 'Política' (Politics) in red, and 'Natureza' (Nature) in green. Blue lines connect words in the documents to these topics. For example, 'Remédio', 'Injeção', and 'Hospital' connect to 'Saúde'; 'Governo' connects to 'Política'; and 'Flores', 'Árvores', 'Montanhas', 'Mar', and 'Céu' connect to 'Natureza'. Red curved arrows point from each document to its corresponding topic.

Como descobrir os tópicos?

Assim, posso usar essa informação conectando palavras a tópicos, dependendo quão bem essa palavra se ajuste a esse tópico, e então conectar os tópicos aos documentos com base nos tópicos abordados em cada documento.



Os tópicos são representados como uma **lista ponderada de palavras**. Exemplo:

- $(0.1 * \text{Hostpita}, 0.1 * \text{Remédio}, 0.1 * \text{Injeção})$ representando **“Saúde”**

Workflow do Topic Modeling

Cada documento será atribuído a um determinado tópico de acordo com a probabilidade de pertencimento, quanto maior a probabilidade de pertencer a um determinado tópico x, esse será considerado o tema que representa esse documento.

Documents

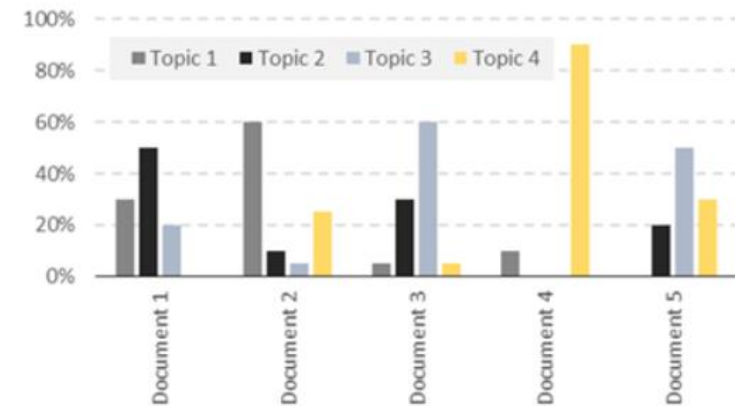


LDA

Creation of topics

	weight	words
Topic 1	3%	flower
	2%	rose
	1%	plant
...		
Topic 2	2%	company
	1%	wage
	1%	employee

Topics allocation to documents



Workflow do Topic Modeling



Documento 1

The Beatles foi uma banda de [rock](#) britânica formada em 1960 na cidade de [Liverpool](#). Formada por [John Lennon](#), [Paul McCartney](#), [George Harrison](#) e [Ringo Starr](#), é considerada a banda mais influente de todos os tempos. O grupo fez parte do desenvolvimento da [contracultura da década de 1960](#) e do reconhecimento da música popular como forma de arte.



LDA
Algorithm

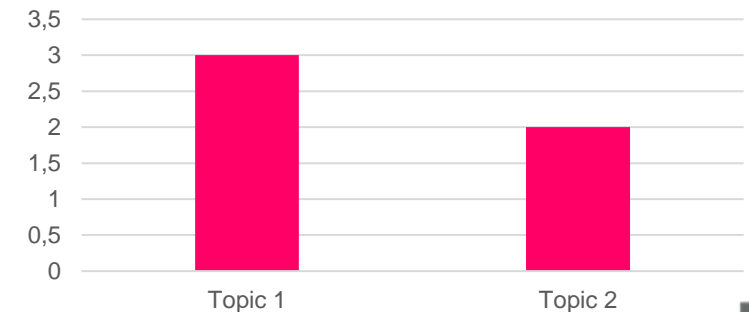


Topic 1	
Peso	Palavra
3%	banda
2%	música
1%	arte
Topic 2	
Peso	Palavra
1%	decada
1%	1960
0%	história

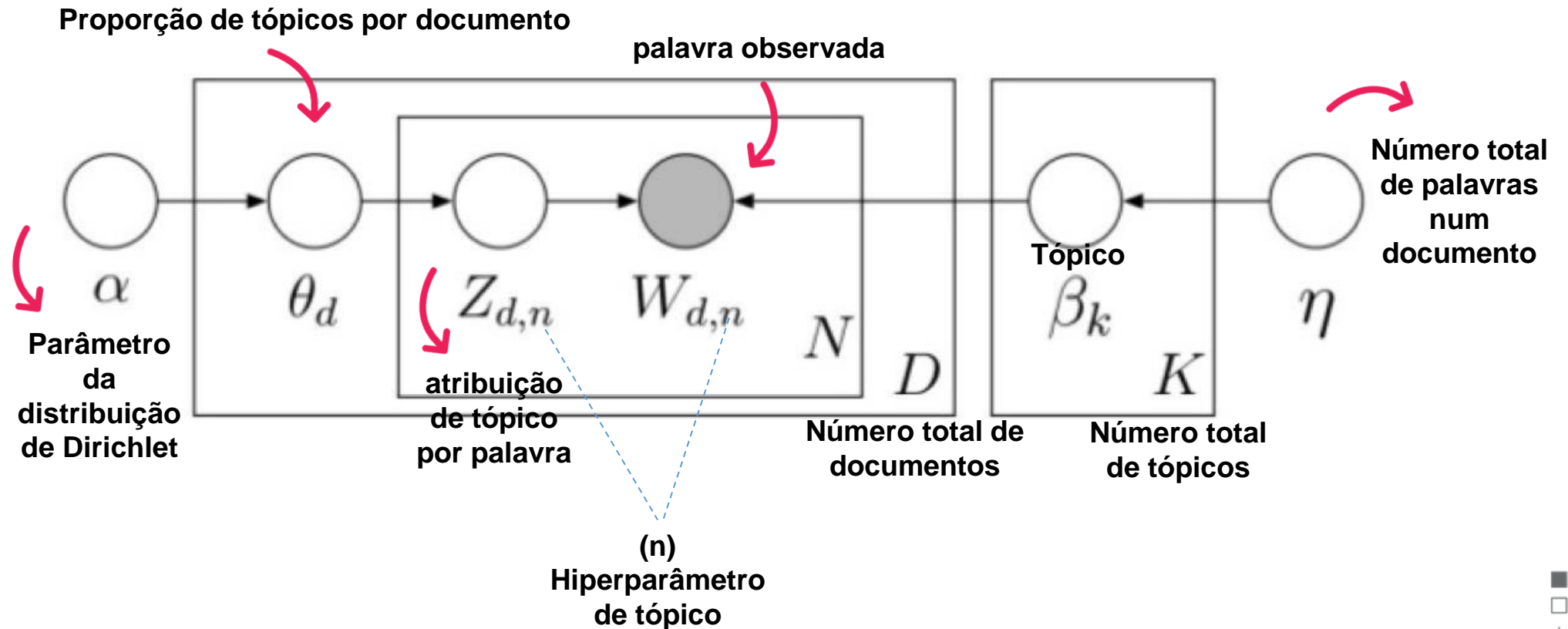


O documento 1 **tem maior probabilidade de pertencer ao Topic 1** por conter mais palavras contidas no topic 1 do que no Topic 2.

Topics allocations to documento 1



Explicação matemática por trás do LDA



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.