

# Classificação não- supervisionada hierárquica de artigos jornalísticos

Cirillo Ferreira

MAC0449 – Trabalho de Formatura Supervisionado

IME/USP

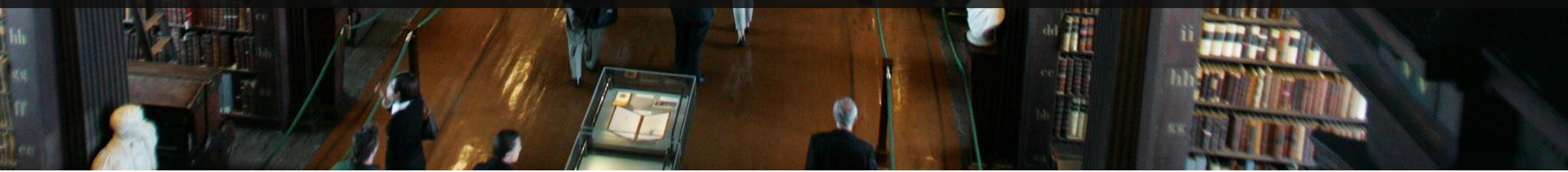
# Agenda

1. Introdução
2. Problema
3. Objetivo
4. Solução
  1. A biblioteca
  2. O sistema hVINA
5. Conclusão

# **1 INTRODUÇÃO**



A tarefa de classificar e agrupar documentos textuais remonta desde a antiguidade.



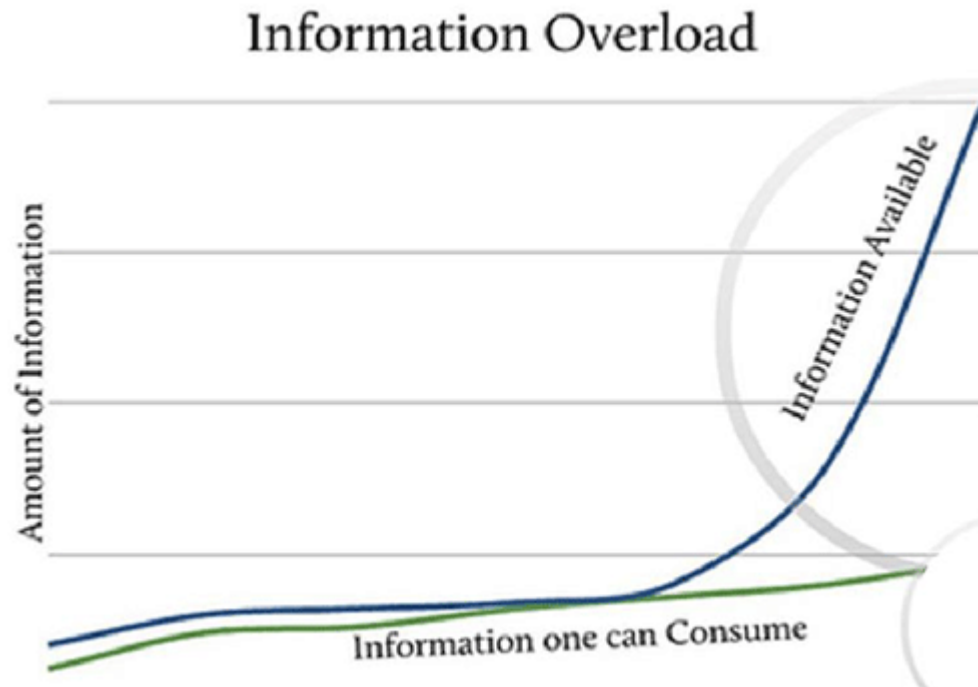
**2**

# **PROBLEMA**

Com o advento da internet houve uma **explosão de informação** que tornou muito difícil a classificação manual desses novos documentos.



Nunca se produziu tanta informação como nos tempos atuais ...



# INFORM ATION

# ANXIETY ANXIETY

What to do when information doesn't  
tell you what you need to know

**RICHARD SAUL WURMAN**

Introduction by John Naisbitt, author of *Megatrends 2000*



(...) O problema está  
em não saber filtrar  
o que é importante.

Clay Shirky (NYU)



Quanto maior a produção de informação, maior a necessidade de mecanismos automáticos para armazenar, **organizar** e recuperá-la.

**3**

# OBJETIVO DO TRABALHO

## Principal objetivo

Criação de uma biblioteca para agrupamento hierárquico de artigos jornalísticos.

Mas que seja:

- **Modular**

Implementada de forma que permita a extensão para novos algoritmos de agrupamento de forma simples.

- **Flexível**

Possa trabalhar com diversas coleções de artigos jornalísticos, em diversos idiomas.



# Criação de um sistema para visualização dos agrupamentos (*hVINA*).





## O que é um agrupamento?

É uma classificação que tem como objetivo o particionamento de objetos em grupos cujo membros sejam **similares entre si** e diferentes dos membros de outros grupos.

**Agrupamento**  
**=**  
**Classificação não-supervisionada**

**Mas agrupamento não é a mesma coisa que classificação?**

**NÃO**

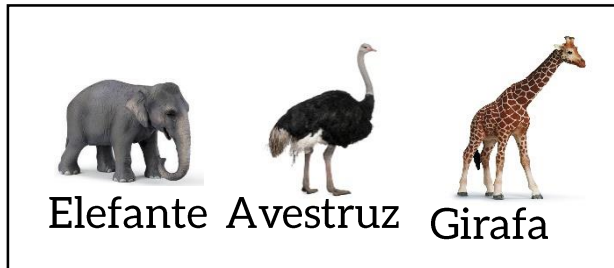
# Agrupamento x Classificação

	Agrupamento	Classificação
Tipo de aprendizado	Não-supervisionado	Supervisionado
Requer dados de treinamento	Não	Sim
Conjunto de classes	Inicialmente desconhecida	Predefinida
Abordagem	Agrupa objetos baseado em uma medida de similaridade	Utiliza “regras” para atribuir rótulos aos novos objetos

# Classificação (supervisionada)

Elefante

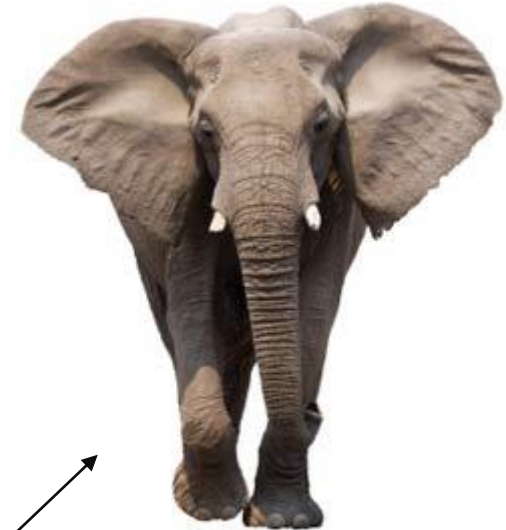
Conjunto de treinamento



Entrada

Resultado

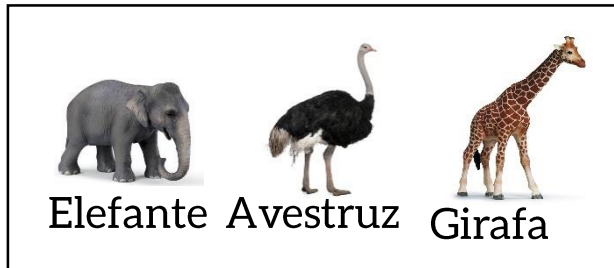
**Classificador**





# Classificação (supervisionada)

Conjunto de treinamento



????

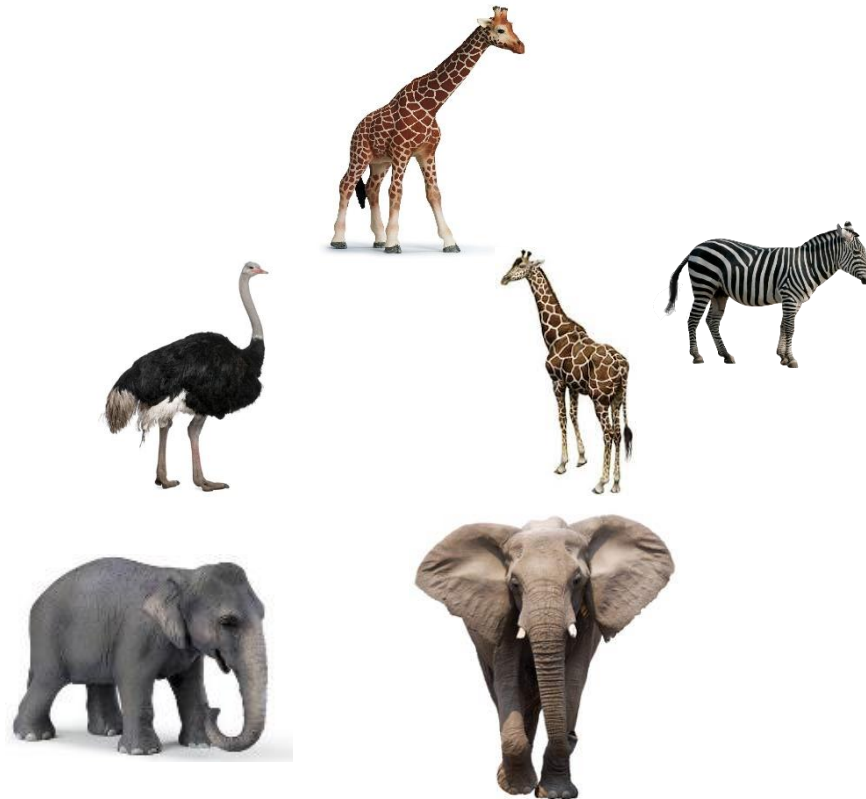


Entrada

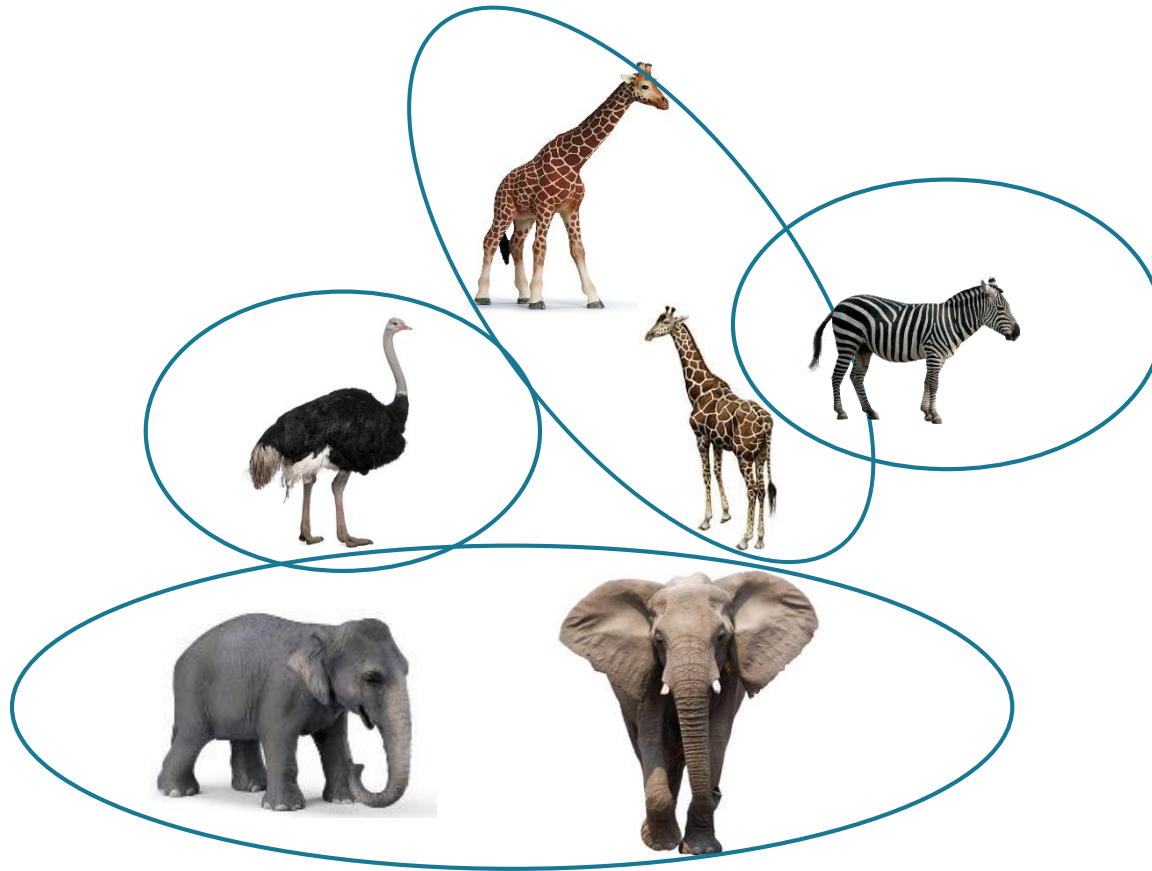
Resultado

**Classificador**

# Agrupamento

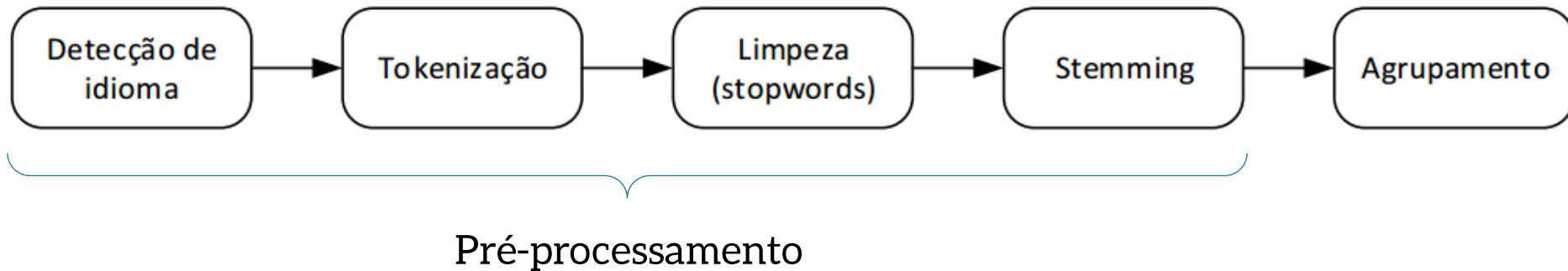


# Agrupamento



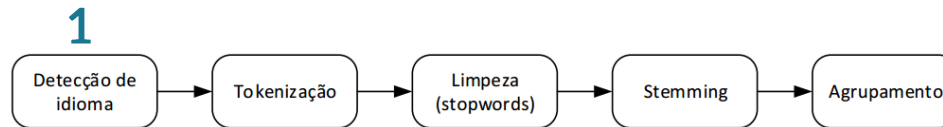
# **4 SOLUÇÃO**

# Arquitetura da biblioteca



O primeiro passo é determinar o idioma utilizado na coleção de artigos.

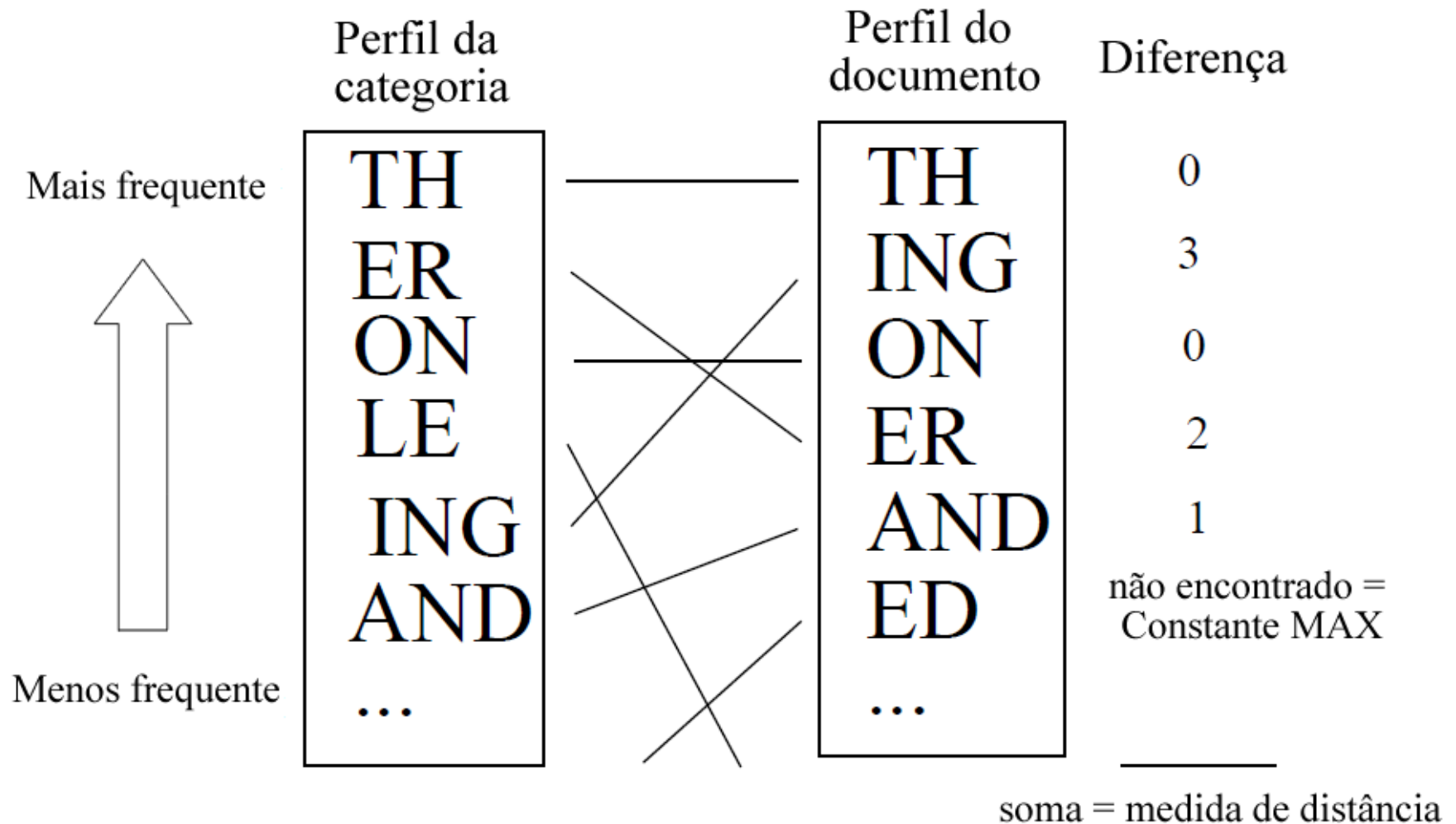


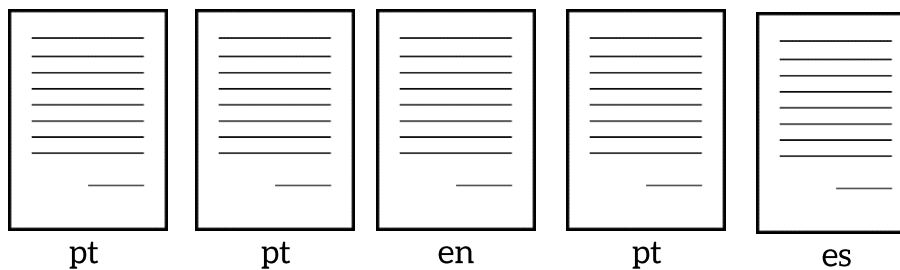
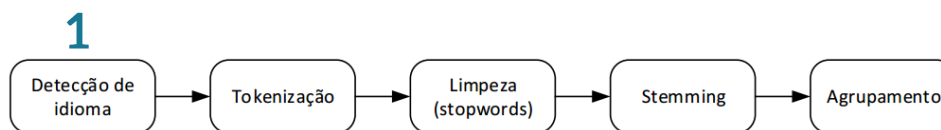


## Detecção de idioma

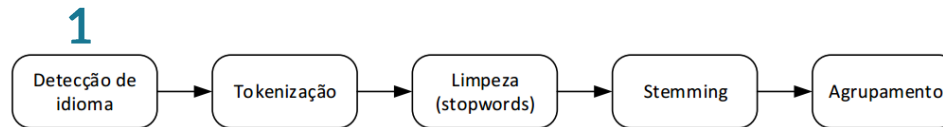
Em si, é um problema de classificação de padrão, onde as classes são os idiomas existentes.

# Método de n-grama



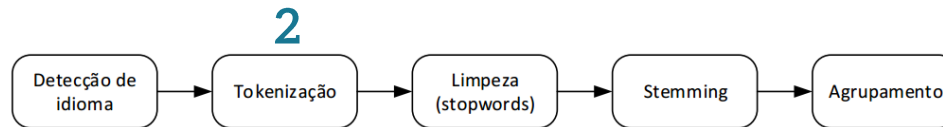


pt



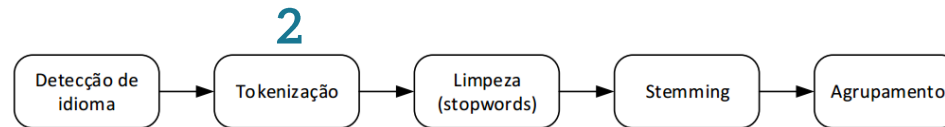
## Atualmente:

1. Português (pt)
2. Inglês (en)
3. Espanhol (es)

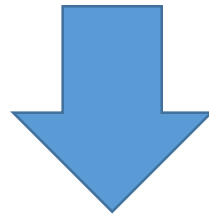


## Tokenização

Segmenta os textos em palavras.



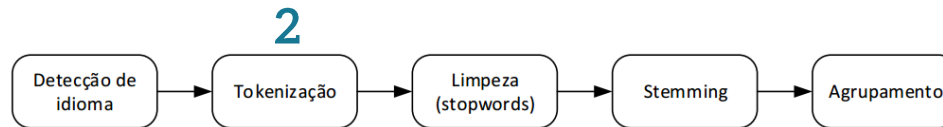
O crescimento da economia não abrange só os mercados do Médio Oriente – mas também os mercados globais, sobretudo os emergentes...



O / crescimento / da / economia / não / abrange / só / os / mercados / do / Médio / Oriente / – / mas / também / os / mercados / globais / , / sobretudo / os / emergentes / ... /

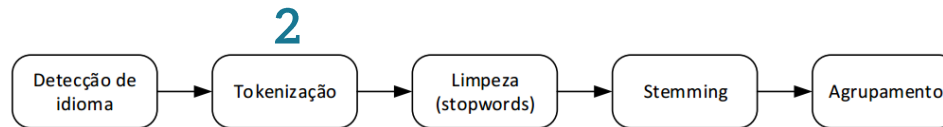


Porém não é tão trivial assim...



## Como lidar com palavras compostas?

(...) O **vice-presidente** da Intel disse que o futuro está na interação **homem-computador** (...)



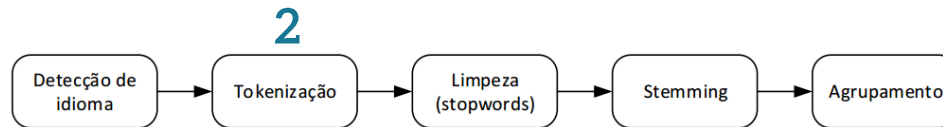
## Como lidar com palavras compostas?

(...) O **vice-presidente** da Intel disse que o futuro está na interação **homem-computador** (...)

“vice-presidente”

ou

“vice” / “-” / “presidente”



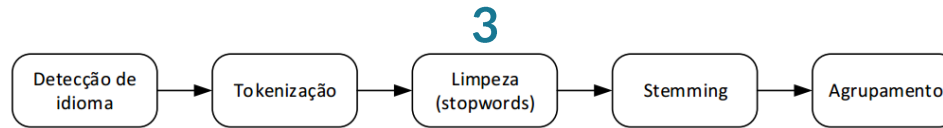
## Como lidar com palavras compostas?

(...) O **vice-presidente** da Intel disse que o futuro está na interação **homem-computador** (...)

“homem-computador”

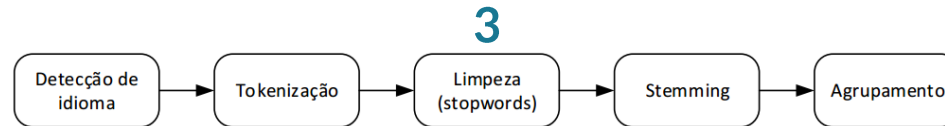
ou

“homem” / “-” / “computador”

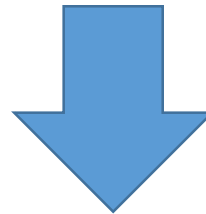


## Limpeza

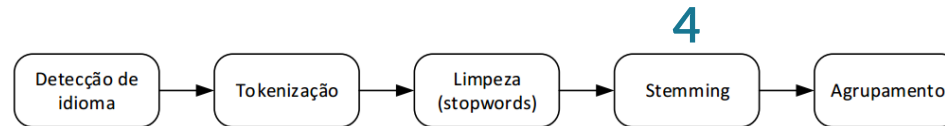
Remove as palavras que possuem pouca relevância no texto.



O / crescimento / da / economia / não / abrange / só / os / mercados / do /  
 Médio / Oriente / – / mas / também / os / mercados / globais / , / sobretudo /  
 os / emergentes / ... /

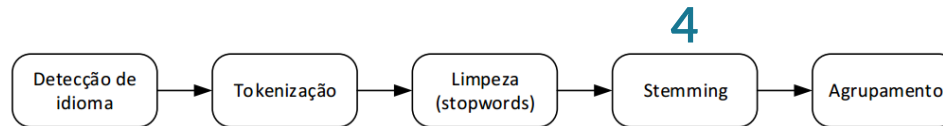


crescimento / economia / abrange / mercados / Médio / Oriente / mercados /  
 globais/ emergentes /



## Stemming

Unifica formas variantes de palavras que possuem o mesmo significado.

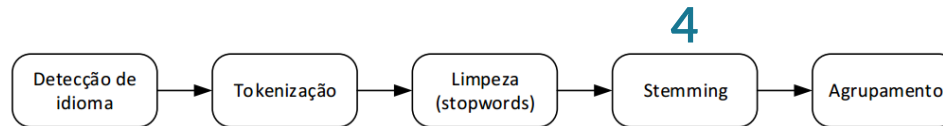


econômico => econom

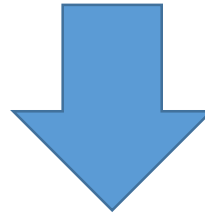
economia => econom

economias => econom

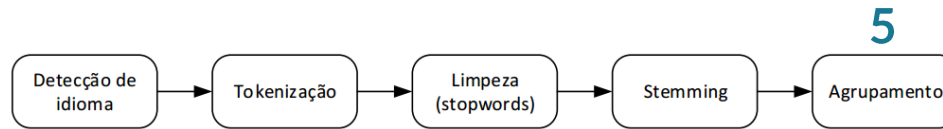




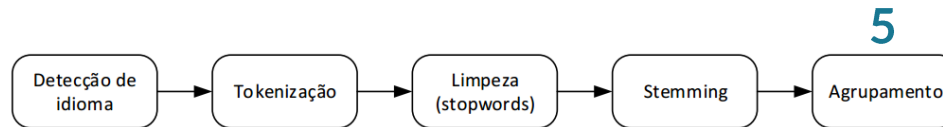
crescimento / economia / abrange / mercados / Médio / Oriente / mercados /  
globais/ emergentes /



cresc / econom / abrang / merc / Médi / Orient / merc / glob / emerg



## Algoritmo de agrupamento em si



**Inicialmente foi implementado o FIHC.**

Agrupamento hierárquico baseado em conjunto de itens frequentes.

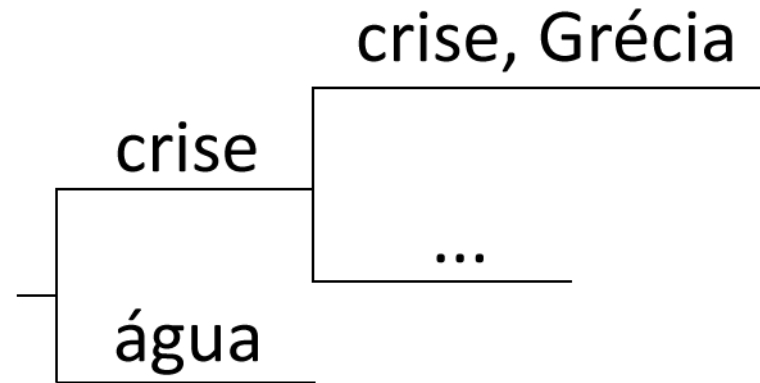
**Por que usar hierárquico?**

Evidencia relações implícitas entre os grupos (tópicos)

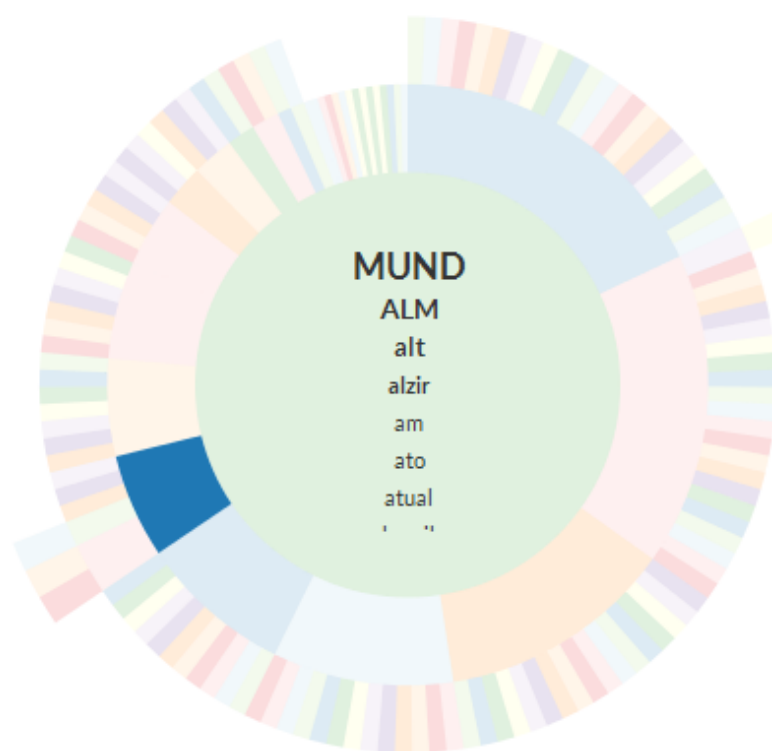
# Agrupamento plano

# x Hierárquico

- crise
- água
- Grécia



Permite a navegação entre os grupos



1. A Mensagem da Cruz
2. Independência e Novo Mandamento
3. Estrutura de um mundo novo
4. Deus tem muitos sinônimos
5. As crianças e a Mãe de Jesus
6. Morte e Ressurreição
7. Vencer o sofrimento do corpo e da Alma
8. Paiva Netto escreve: "Vencer o sofrimento do corpo e da Alma"
9. Ano-Novo! Ano-Bom?
10. Novas conquistas
11. Que falta ao mundo para que haja Paz?
12. Jesus ressuscitou. E nós com Ele
13. Paz em 2014
14. O perigo é real
15. Hiroshima
16. A Ideologia das ideologias
17. Tesouro precioso
18. Natal Permanente e renovação planetária
19. Desumanidade gera desumanidade
20. Em louvor à Paz
21. Paz em 2010
22. Sustentabilidade e reeducação
23. A grande família Humanidade
24. Não perder a fé
25. Francisco



# DEMO

# **5 CONCLUSÃO**

O problema não é a quantidade de informação, mas sim a maneira como a organizamos.

É viável um sistema para **classificação não-supervisionada** de artigos jornalísticos que seja acessível a qualquer usuário.

# Imagens

- [http://www.meetinireland.com/BusinessTourism/media/main\\_site/Blog/EUCHARISTI C-CONGRESS---Fam-trip-in-Trinity.jpg](http://www.meetinireland.com/BusinessTourism/media/main_site/Blog/EUCHARISTI C-CONGRESS---Fam-trip-in-Trinity.jpg)
- <http://cfile30.uf.tistory.com/image/156899344FEA735C2ADC9A>
- [http://www.athomearkansas.com/sites/athomearkansas.com/files/images/2/gallery\\_images/pug-n-shelves.jpg](http://www.athomearkansas.com/sites/athomearkansas.com/files/images/2/gallery_images/pug-n-shelves.jpg)
- [http://digital.coolspringspress.com/rp\\_columns\\_images/images/666.jpg](http://digital.coolspringspress.com/rp_columns_images/images/666.jpg)

# Referências

- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2009
- <http://www.forbes.com/sites/johnnosta/2013/06/13/information-overload-the-big-challenge-for-digital-health/>
- Richard S Wurman. Information Anxiety. 1989
- William B. Cavnar, John M. Trenkle. N-Gram-Based Text Categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994

**Obrigado!!**