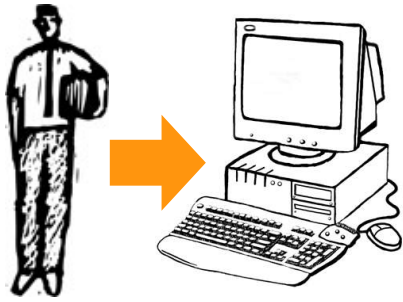


Identificação de Contexto Geográfico Implícito em Texto Auxiliada pela *Wikipedia*

Mestrando: Rafael Odon de Alencar
(odon@dcc.ufmg.br)

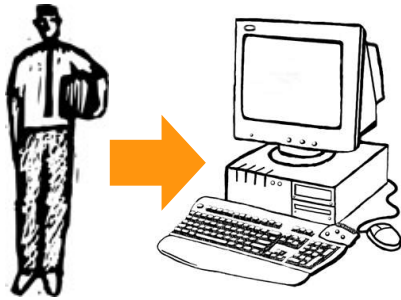
Orientador: Clodoveu Augusto Davis Jr.
(clodoveu@dcc.ufmg.br)



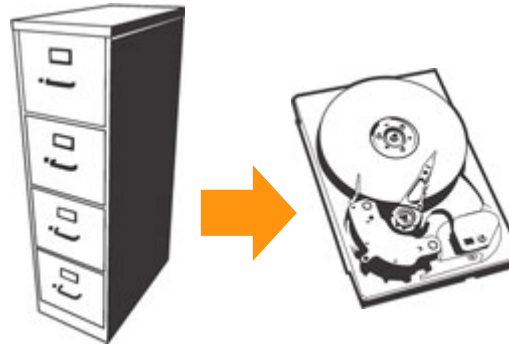


Sistemas de
Informação



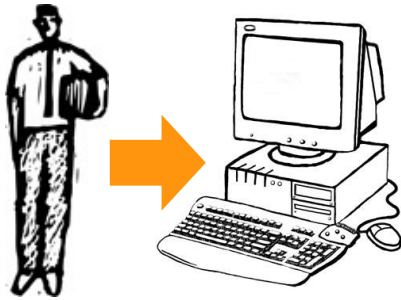


Sistemas de
Informação

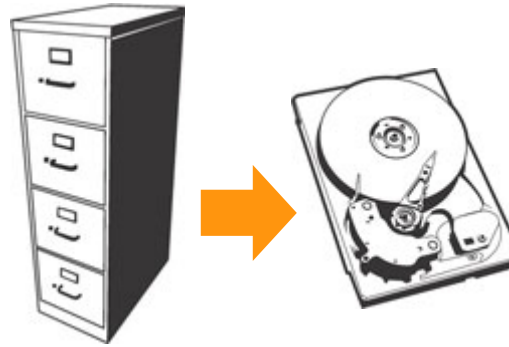


Dispositivos de
Armazenamento em
Massa

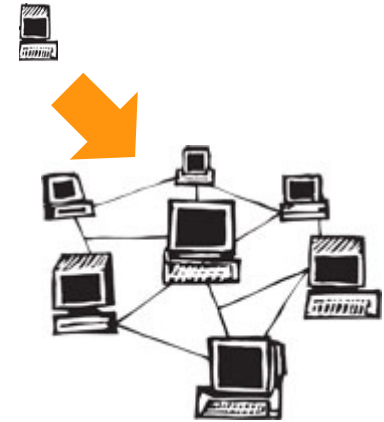




Sistemas de
Informação

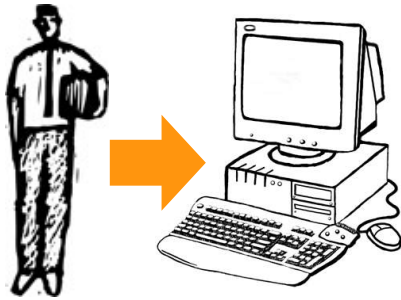


Dispositivos de
Armazenamento em
Massa

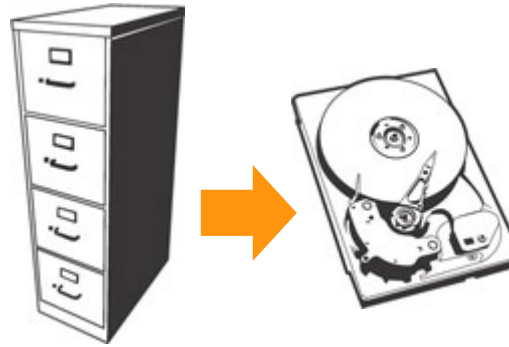


Advento da
Internet

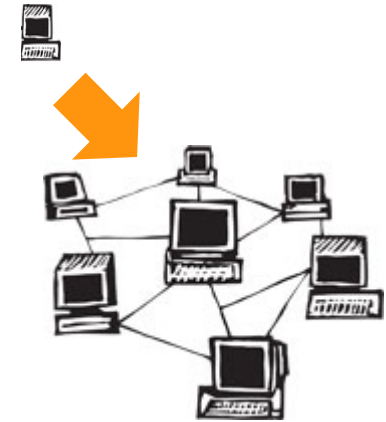
Recuperação de Informação



Sistemas de
Informação



Dispositivos de
Armazenamento em
Massa



Advento da
Internet



Google™

Google Search

I'm Feeling Lucky

Quantidade
considerável de
consultas com
intenção geográfica

...no Brasil
Onde fica... ...leste
...perto da UFMG
...Minas Gerais

Quantidade
considerável de
consultas com
intenção geográfica

...no Brasil
Onde fica... ...leste
...perto da UFMG
...Minas Gerais

Documentos
podem não
conter anotação
geográfica
formal



Quantidade
considerável de
consultas com
intenção geográfica

...no Brasil
Onde fica... ...leste
...perto da UFMG
...Minas Gerais

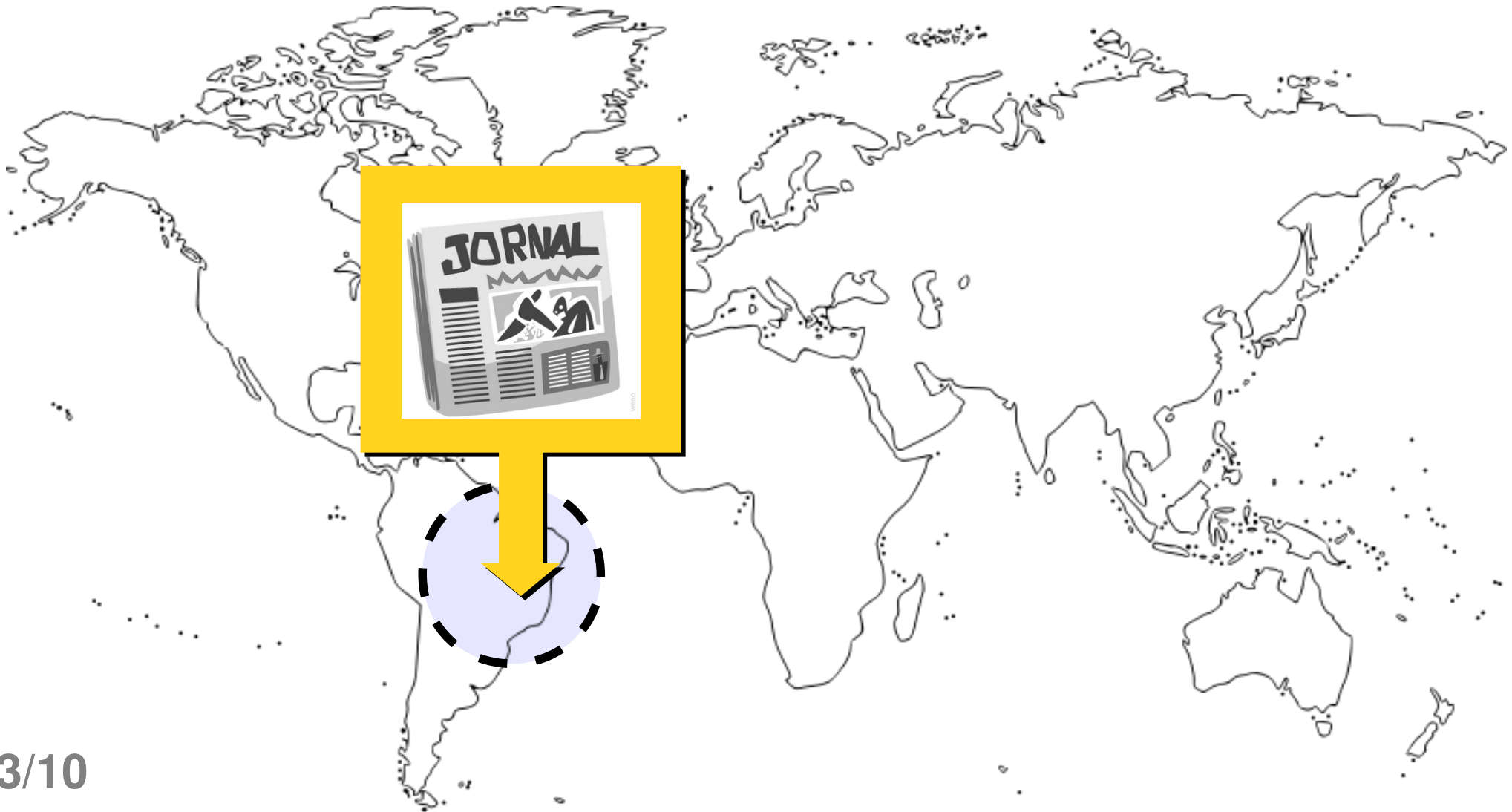
Documentos
podem não
conter anotação
geográfica
formal



Problema → Identificação de contexto geográfico
(identificação de escopo geográfico ou geo-tagging)

Desafio da Identificação de Contexto Geográfico:

Tenta descobrir quais **locais** no globo estão relacionados com uma determinado recurso, dado o conjunto de suas características.



Desafio da Identificação de Contexto Geográfico:

Tenta descobrir quais **locais** no globo estão relacionados com uma determinado recurso, dado o conjunto de suas características.



mecanismos
de busca local

promoção de
publicidade adequada

City Room



April 5, 2010, 6:09 PM

Searching for Sept. 11 Victims, Again

By [ANDY NEWMAN](#)

The search for the remains of the victims of the [World Trade Center attacks](#) has begun anew.

As the morning sun climbed over what was once the Fresh Kills landfill on Staten Island Monday, workers under the supervision of the medical examiner's office began sifting through 844 cubic yards of dirt and debris excavated from the area around the World Trade Center site since 2007.

Officials are hoping to turn up traces of some of the 1,123 victims of the terrorist attacks whose remains were never found.

"There is soil that in theory could — nobody knows whether it does or not — contain remains from the terrible tragedy that occurred on 9/11," Mayor [Michael R. Bloomberg](#) said.

"We are not going to walk away from our obligation to do anything that is reasonable and practical and affordable to recover and identify any remains that might be there."

The previous sifting effort, which ended in 2007, sorted through than 15,000 yards of debris and turned up 1,772 objects that were deemed potential human remains.

Most of the debris being analyzed in the new search — 630 of the 844 cubic yards — was dug up beneath West Street, between Albany Street and Vesey Street.

Another 105 cubic yards came from the rooftop of [Fiterman Hall](#) at 30 West Broadway, which was damaged on 9/11.

City Room

April 5, 2010, 6:09 PM

Searching for Sept. 11 Victims, Again

By [ANDY NEWMAN](#)

The search for the remains of the victims of the [World Trade Center attacks](#) has begun anew.

As the morning sun climbed over what was once the Fresh Kills landfill on Staten Island Monday, workers under the supervision of the medical examiner's office began sifting through 844 cubic yards of dirt and debris excavated from the area around the World Trade Center site since 2007.

Officials are hoping to turn up traces of some of the 1,123 victims of the terrorist attacks whose remains were never found.

"There is soil that in theory could — nobody knows whether it does or not — contain remains from the terrible tragedy that occurred on 9/11," Mayor [Michael R. Bloomberg](#) said.

"We are not going to walk away from our obligation to do anything that is reasonable and practical and affordable to recover and identify any remains that might be there."

The previous sifting effort, which ended in 2007, sorted through than 15,000 yards of debris and turned up 1,772 objects that were deemed potential human remains.

Most of the debris being analyzed in the new search — 630 of the 844 cubic yards — was dug up beneath West Street, between Albany Street and Vesey Street.

Another 105 cubic yards came from the rooftop of [Fiterman Hall](#) at 30 West Broadway, which was damaged on 9/11.

Staten Island (bairro)

West Street (rua)

Albany Street (rua)

Vesey Street (rua)

West Broadway (rua)

City Room

April 5, 2010, 6:09 PM

Searching for Sept. 11 Victims, Again

By [ANDY NEWMAN](#)

The search for the remains of the victims of the [World Trade Center attacks](#) has begun anew.

As the morning sun climbed over what was once the Fresh Kills landfill on Staten Island Monday, workers under the supervision of the medical examiner's office began sifting through 844 cubic yards of dirt and debris excavated from the area around the World Trade Center site since 2007.

Officials are hoping to turn up traces of some of the 1,123 victims of the terrorist attacks whose remains were never found.

"There is soil that in theory could — nobody knows whether it does or not — contain remains from the terrible tragedy that occurred on 9/11," Mayor [Michael R. Bloomberg](#) said.

"We are not going to walk away from our obligation to do anything that is reasonable and practical and affordable to recover and identify any remains that might be there."

The previous sifting effort, which ended in 2007, sorted through than 15,000 yards of debris and turned up 1,772 objects that were deemed potential human remains.

Most of the debris being analyzed in the new search — 630 of the 844 cubic yards — was dug up beneath West Street, between Albany Street and Vesey Street.

Another 105 cubic yards came from the rooftop of [Fiterman Hall](#) at 30 West Broadway, which was damaged on 9/11.

Staten Island (bairro)

West Street (rua)

Albany Street (rua)

Vesey Street (rua)

West Broadway (rua)

World Trade Center (prédio)

Fiterman Hall (prédio)

Fresh Kills Landfill (aterro)

Michael R. Bloomberg (prefeito)

Sept. 11 (data)

9/11 (data)

City Room

April 5, 2010, 6:09 PM

Searching for Sept. 11 Victims, Again

By **ANDY NEWMAN**

The search for the remains of the victims of the [World Trade Center attacks](#) has begun anew.

As the morning sun climbed over what was once the Fresh Kills landfill on Staten Island Monday, workers under the supervision of the medical examiner's office began sifting through 844 cubic yards of dirt and debris excavated from the area around the World Trade Center site since 2007.

Officials are hoping to turn up traces of some of the 1,123 victims of the terrorist attacks whose remains were never found.

"There is soil that in theory could — nobody knows whether it does or not — contain remains from the terrible tragedy that occurred on 9/11," Mayor [Michael R. Bloomberg](#) said.

"We are not going to walk away from our obligation to do anything that is reasonable and practical and affordable to recover and identify any remains that might be there."

The previous sifting effort, which ended in 2007, sorted through than 15,000 yards of debris and turned up 1,772 objects that were deemed potential human remains.

Most of the debris being analyzed in the new search — 630 of the 844 cubic yards — was dug up beneath West Street, between Albany Street and Vesey Street.

Another 105 cubic yards came from the rooftop of [Fiterman Hall](#) at 30 West Broadway, which was damaged on 9/11.

Staten Island (bairro)

West Street (rua)

Albany Street (rua)

Vesey Street (rua)

West Broadway (rua)

World Trade Center (prédio)

Fiterman Hall (prédio)

Fresh Kills Landfill (aterro)

Michael R. Bloomberg (prefeito)

Sept. 11 (data)

9/11 (data)

Terrorist Attacks (tópico)

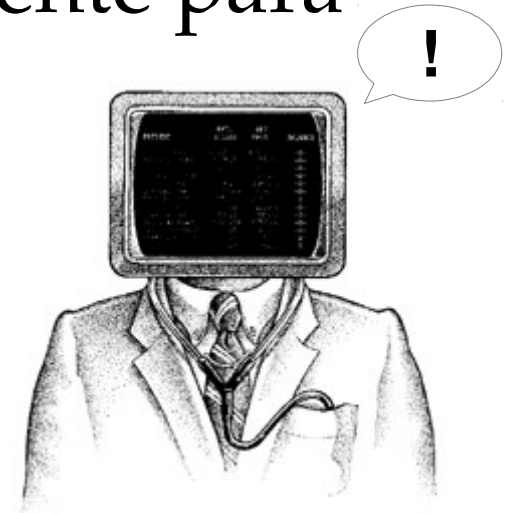
Uma solução é:

Usar uma **base de conhecimento** existente para entender a relação de pessoas, atividades, eventos e tópicos de diversas naturezas com os locais do globo.



Uma solução é:

Usar uma **base de conhecimento** existente para entender a relação de pessoas, atividades, eventos e tópicos de diversas naturezas com os locais do globo.



A **Wikipedia** é uma alternativa!

Propõe-se o uso da técnica de Wikificação automática (Mihalcea & Csomai 2007)

A Universidade Federal de Minas Gerais (UFMG) é uma instituição brasileira de ensino superior de nível federal, sediada na cidade de Belo Horizonte no estado de Minas Gerais. Estando entre as mais prestigiadas universidades do Brasil, é a maior universidade do estado de Minas Gerais, possuindo campi nas cidades de Belo Horizonte (sede), Montes Claros e Tiradentes.

Propõe-se o uso da técnica de Wikificação automática (Mihalcea & Csomai 2007)

A Universidade Federal de Minas Gerais (UFMG) é uma instituição brasileira de ensino superior de nível federal, sediada na cidade de Belo Horizonte no estado de Minas Gerais. Estando entre as mais prestigiadas universidades do Brasil, é a maior universidade do estado de Minas Gerais, possuindo campi nas cidades de Belo Horizonte (sede), Montes Claros e Tiradentes.

W!

A [Universidade Federal de Minas Gerais \(UFMG\)](#) é uma instituição brasileira de [ensino superior](#) de nível federal, sediada na cidade de [Belo Horizonte](#) no estado de [Minas Gerais](#). Estando entre as mais prestigiadas universidades do [Brasil](#), é a maior universidade do estado de Minas Gerais, possuindo [campi](#) nas cidades de Belo Horizonte (sede), [Montes Claros](#) e [Tiradentes](#).

Propõe-se o uso da técnica de Wikificação automática (Mihalcea & Csomai 2007)

A Universidade Federal de Minas Gerais (UFMG) é uma instituição brasileira de ensino superior de nível federal, sediada na cidade de Belo Horizonte no estado de Minas Gerais. Estando entre as mais prestigiadas universidades do Brasil, é a maior universidade do estado de Minas Gerais, possuindo campi nas cidades de Belo Horizonte (sede), Montes Claros e Tiradentes.

W!

A [Universidade Federal de Minas Gerais](#) ([UFMG](#)) é uma instituição brasileira de [ensino superior](#) de nível federal, sediada na cidade de [Belo Horizonte](#) no estado de [Minas Gerais](#). Estando entre as mais prestigiadas universidades do [Brasil](#), é a maior universidade do estado de Minas Gerais, possuindo [campi](#) nas cidades de Belo Horizonte (sede), [Montes Claros](#) e [Tiradentes](#).

Identifica tópicos do texto

Provê conhecimento estruturado

Objetivo:

Identificar o Contexto Geográfico de textos a partir dos tópicos identificados nele com a wikificação



Objetivo:

Identificar o Contexto Geográfico de textos a partir dos tópicos identificados nele com a wikificação



Como?

Uma vez **wikificado**, o documento se conecta ao **grafo** da da Wikipedia, o que torna possível identificar a **correlação** do texto com os **locais** que fazem parte dessa rede.



Ganho esperado:

Além dos nomes de de locais tradicionais, termos e expressões diversas do vocabulário controlado da Wikipedia podem vir a influenciar na identificação do contexto geográfico.



Resultados parciais:

Estudo preliminar para verificar o potencial das evidências da Wikipedia (Alencar et al, GIR 2010).

- Classificação single-label
- Corpo e título de 831 notícias
- 8 estados brasileiros
- Frequencia de expressões do vocabulário da wikipedia como *features*
- Naive Bayes Multinomial



Estágio atual:



- ✓ Reproduziu-se o mecanismo de wikificação proposto em Milne & Witten, 2008
- ✓ Ajustando modelos de aprendizagem para versão em Português
- ✓ Avaliando métricas de correlação tópicos/locais
- ✓ Construindo coleção mais ampla e detalhada de documentos de locais...

Fim Perguntas?

Alencar, R. O., Davis Jr, C. A., and Goncalves, M. A. (2010). *Geographical classification of documents using evidence from wikipedia*. In GIR'10: Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich.

Medelyan, O., Witten, I. H., and Milne, D. (2008). *Topic indexing with wikipedia*. In WIKIAI'08: Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence, Chicago, I.L.

Mihalcea, R. and Csomai, A. (2007). *Wikify!: linking documents to encyclopedic knowledge*. In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 233–242, New York, NY, USA. ACM.

Milne, D. and Witten, I. H. (2008). *Learning to link with wikipedia*. In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, pages 509–518, New York, NY, USA. ACM.

