

# Mineração de eventos: algoritmos e aplicações

Solange O. Rezende  
[solange@icmc.usp.br](mailto:solange@icmc.usp.br)

4<sup>a</sup> ESCOLA AVANÇADA EM BIG DATA ANALYSIS (Outubro de 2020)



# Agenda

- Mineração de Textos e Inteligência Analítica
- Mineração de Eventos
  - Identificação do Problema
  - Pré-Processamento
  - Extração de Padrões
  - Pós-processamento
  - Uso do Conhecimento
- Considerações Finais

# Agenda

- Mineração de Textos e Inteligência Analítica
- Mineração de Eventos
  - Identificação do Problema
  - Pré-Processamento
  - Extração de Padrões
  - Pós-processamento
  - Uso do Conhecimento
- Considerações Finais

# Inteligência Analítica para Textos

## ■ Por qual motivo analisar dados textuais?

- Notícias
- Redes Sociais
- Artigos Científicos
- Boletins Financeiros
- Laudos, Livros, E-mails, Help-desk, etc...

automated data mining survey  
responses computer transcripts  
qualitative root cause  
classification insights  
ad-hoc analysis product  
reviews sentiment voice of the  
customer dashboards consumer  
trends ad-hoc analysis early warning

# Inteligência Analítica para Textos

## ■ Por qual motivo analisar dados textuais?

- Notícias
- Redes Sociais
- Artigos Científicos
- Boletins Financeiros
- Laudos, Livros, E-mails, Help-desk, etc...



Fonte: [Img1]

Textos representam 80% da informação existente nas organizações!

Forma natural do ser humano em transferir conhecimento.

# Inteligência Analítica para Textos

## ■ Por qual motivo analisar dados textuais?

- Notícias
- Redes Sociais
- Artigos Científicos
- Boletins Financeiros
- Laudos, Livros, E-mails, Help-desk, etc...



Fonte: [Img1]

## Inteligência Analítica para Textos

# Inteligência Analítica para Textos

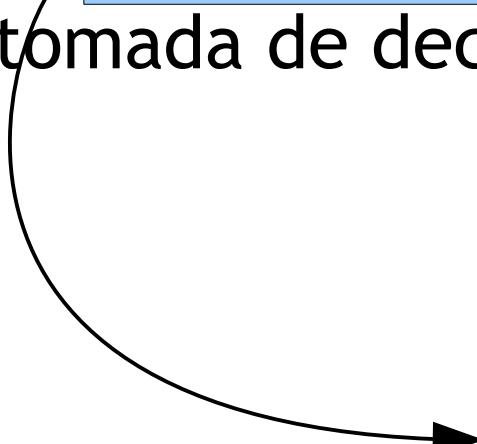
## ■ Definição

É um processo de extração de conhecimento útil a partir de grandes bases de textos usando com apoio de inteligência artificial para apoiar processos de tomada de decisão.

# Inteligência Analítica para Textos

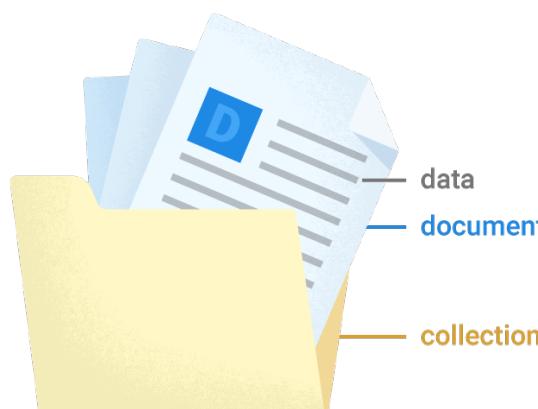
## ■ Definição

É um processo de extração de conhecimento útil a partir de grandes bases de textos usando com apoio de inteligência artificial para apoiar processos de tomada de decisão.

- 
- Aprendizado de Máquina
  - Mineração de Dados e Textos
  - Ciências de Dados e Big Data

# Inteligência Analítica para Textos

- Quais os desafios?
  - Textos são dados NÃO ESTRUTURADOS!



Fonte: [Img2]

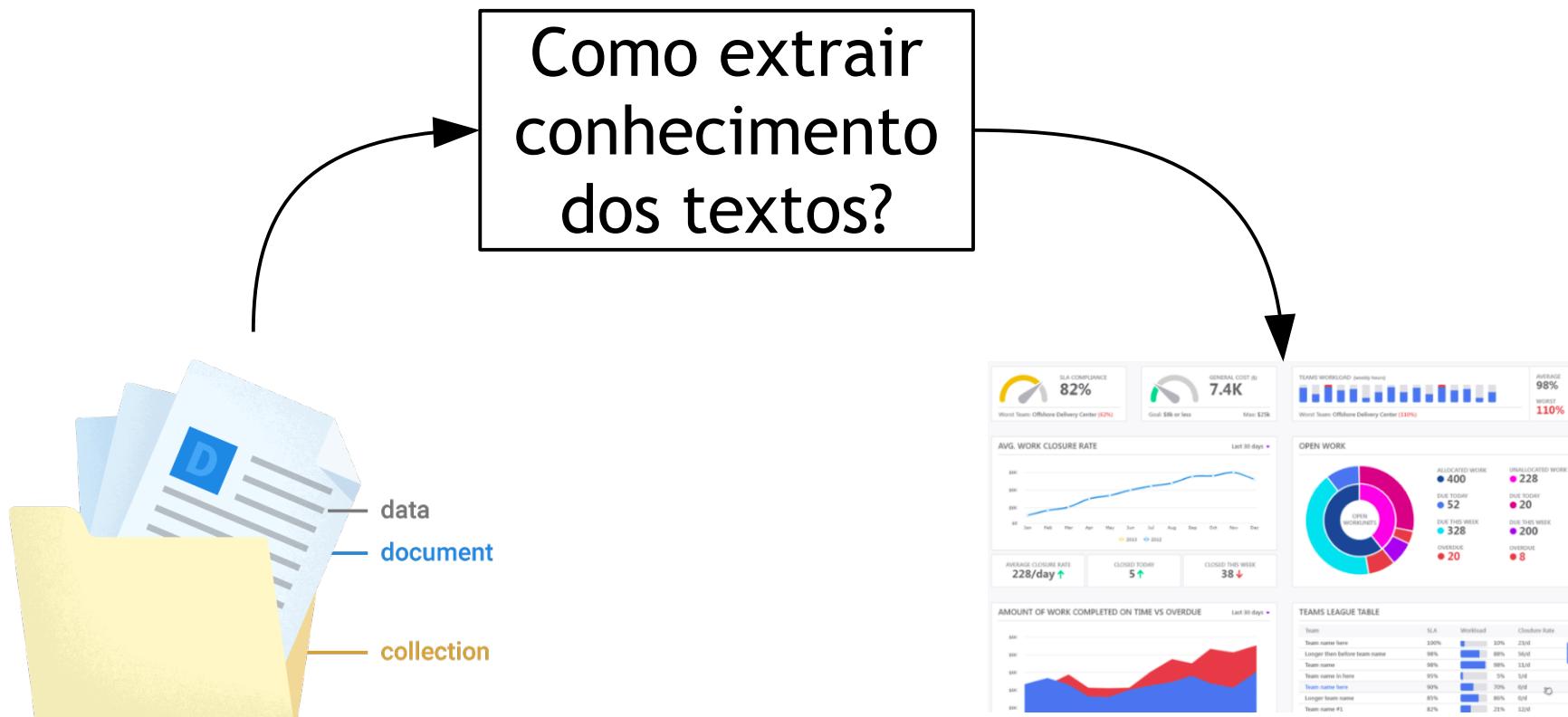
Textos



Tomada de Decisão

# Inteligência Analítica para Textos

- Quais os desafios?
  - Textos são dados NÃO ESTRUTURADOS!



Fonte: [Img2]

Textos

Tomada de Decisão

# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.



# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

### ■ Definir o problema



# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

- Definir o problema
- Quais informações disponíveis?



# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

- Definir o problema
- Quais informações disponíveis?
- Interpretabilidade?



(Rezende et al., 2003)

# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

- Definir o problema
- Quais informações disponíveis?
- Interpretabilidade?
- Privacidade?



(Rezende et al., 2003)

# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

- Definir o problema
- Quais informações disponíveis?
- Interpretabilidade?
- Privacidade?
- Metas e critérios de avaliação



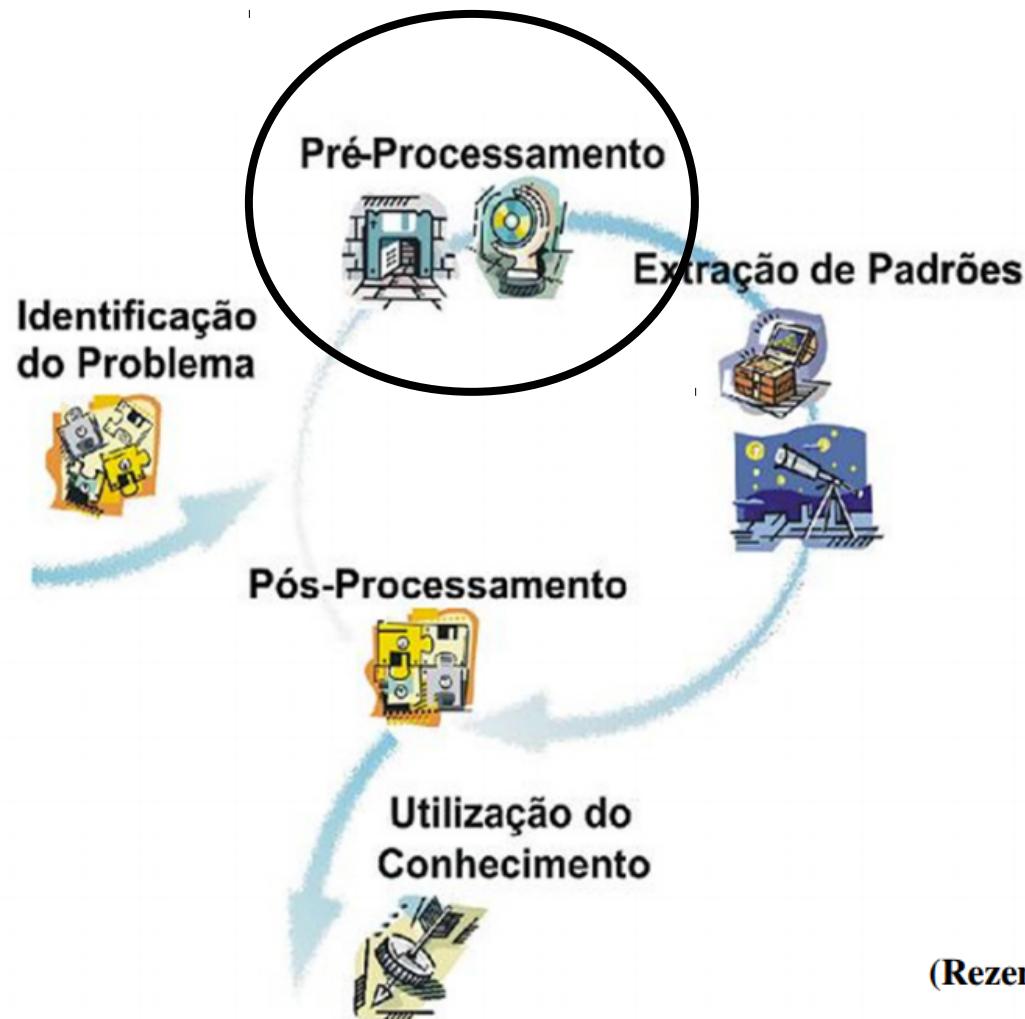
(Rezende et al., 2003)

# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

Obter uma  
representação  
estruturada apropriada  
para a extração de  
padrões.



(Rezende et al., 2003)

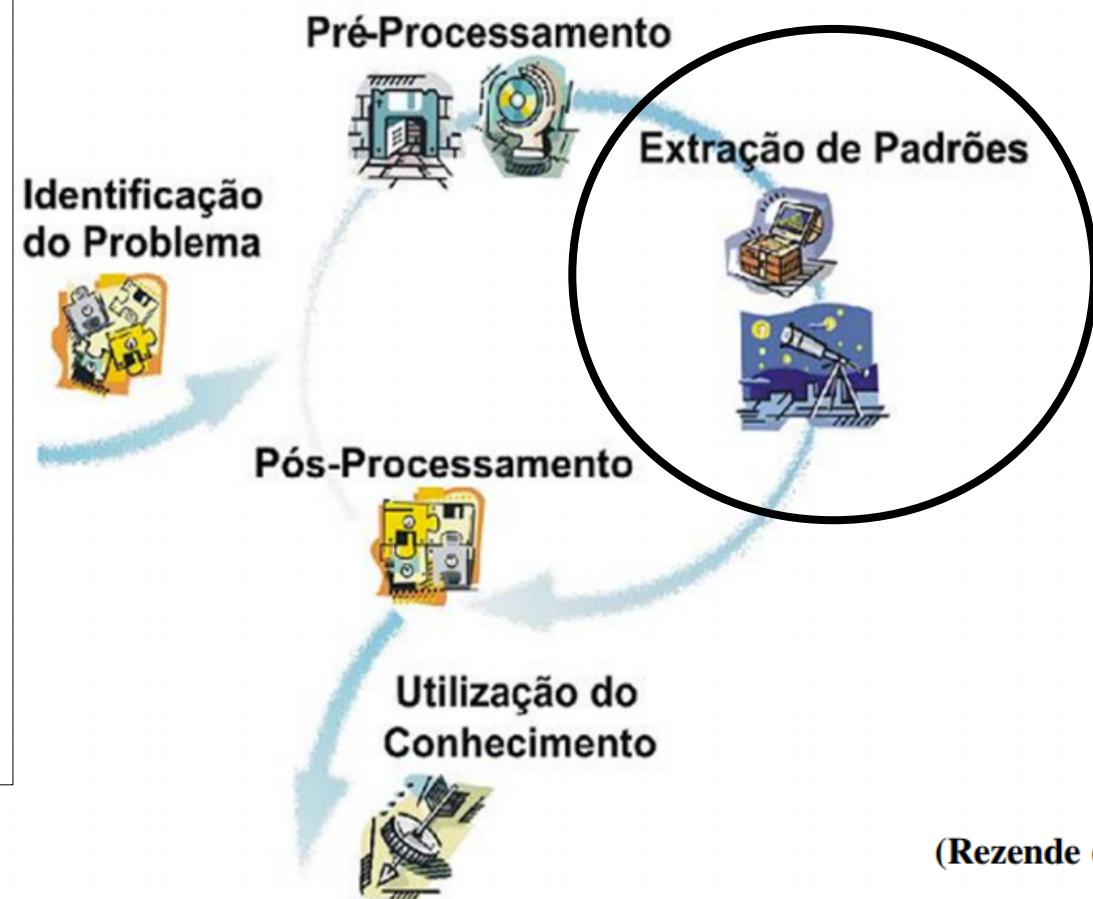
# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

Escolha do algoritmo  
de aprendizado de  
máquina.

Exemplos:  
*Agrupamento*  
*Classificação*



(Rezende et al., 2003)

# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

Avaliação do conhecimento extraído na etapa de extração de padrões



(Rezende et al., 2003)

# Mineração de Textos

## ■ Etapas da Mineração de Textos

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

Apoio à tomada de decisão.

Exemplo:  
Painel de indicadores baseado nos padrões obtidos (*dashboard*).

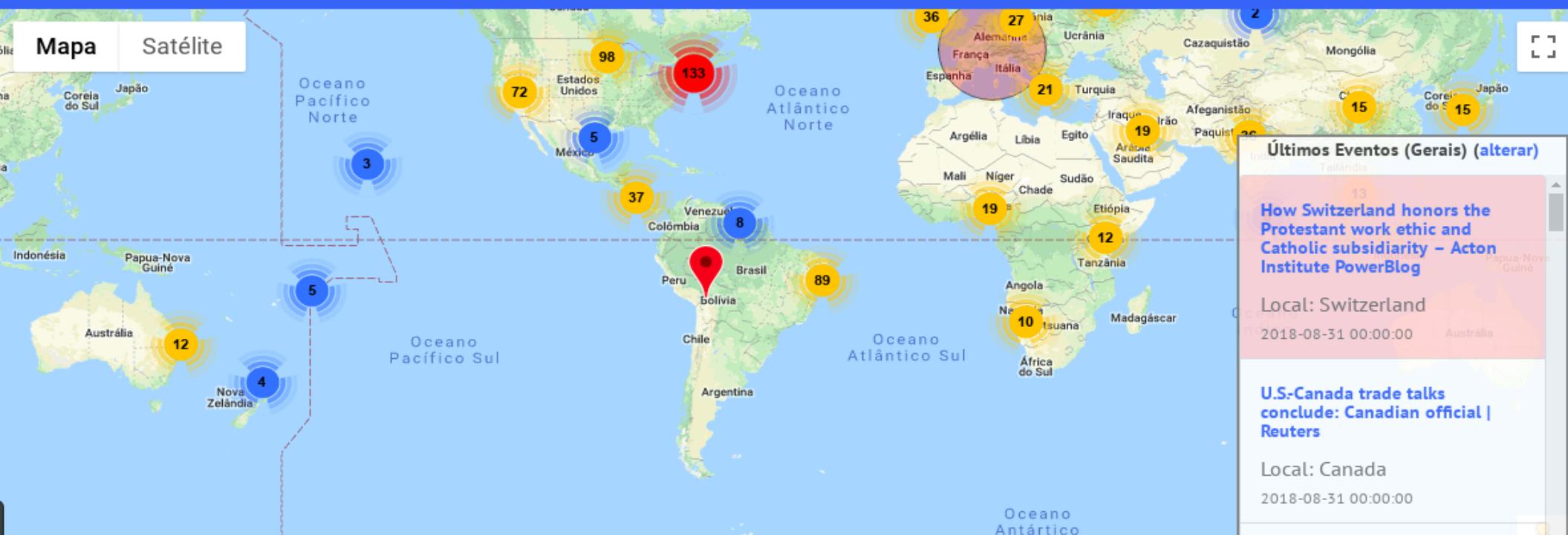


(Rezende et al., 2003)

# Mineração de Eventos

Mapear eventos que ocorrem na web (mundo virtual) para o nosso mundo real. Monitorar fenômenos, tendências e realizar tarefas descritivas e preditivas.

Coleta e monitoramento de eventos em tempo real.



# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Textos possuem várias componentes de informação que juntos representam “Eventos”.

20/04/2017 (AgroNews)

O estado de Mato Grosso do Sul  
é o único que deve contabilizar  
aumento de produção de  
soja para o próximo ano, conforme  
relatório divulgado nesta  
quinta-feira pela Conab (Companhia  
Nacional de Abastecimento)

- O que aconteceu?  
Componente WHAT

# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Textos possuem várias componentes de informação que juntos representam “Eventos”.

20/04/2017 (AgroNews)

O **estado de Mato Grosso do Sul** é o único que deve contabilizar aumento de produção de soja para o próximo ano, conforme relatório divulgado nesta quinta-feira pela Conab (Companhia Nacional de Abastecimento)

- O que aconteceu?  
**Componente WHAT**
- Onde aconteceu?  
**Componente WHERE**

# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Textos possuem várias componentes de informação que juntos representam “Eventos”.

20/04/2017 (AgroNews)

O estado de Mato Grosso do Sul é o único que deve contabilizar aumento de produção de soja para o **próximo ano**, conforme relatório divulgado nesta quinta-feira pela Conab (Companhia Nacional de Abastecimento)

- O que aconteceu?  
**Componente WHAT**
- Onde aconteceu?  
**Componente WHERE**
- Quando aconteceu?  
**Componente WHEN**

# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Textos possuem várias componentes de informação que juntos representam “Eventos”.

20/04/2017 (AgroNews)

O estado de Mato Grosso do Sul é o único que deve contabilizar aumento de produção de soja para o próximo ano, conforme relatório divulgado nesta quinta-feira pela Conab (Companhia Nacional de Abastecimento)

- O que aconteceu?  
Componente WHAT
- Onde aconteceu?  
Componente WHERE
- Quando aconteceu?  
Componente WHEN
- Quem está envolvido?  
Componente WHO

# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Textos possuem várias componentes de informação que juntos representam “Eventos”.

20/04/2017 (AgroNews)

O est...  
é o ú...  
au...  
soja pa...  
rel...  
quinta-feira pela Conab (Companhia

Um evento é algo que ocorre em determinado tempo e local.

Nacional de Abastecimento)

■ O que aconteceu?

Componente WHAT

Componente WHEN

■ Quem está envolvido?

Componente WHO

# Inteligência Analítica para Textos

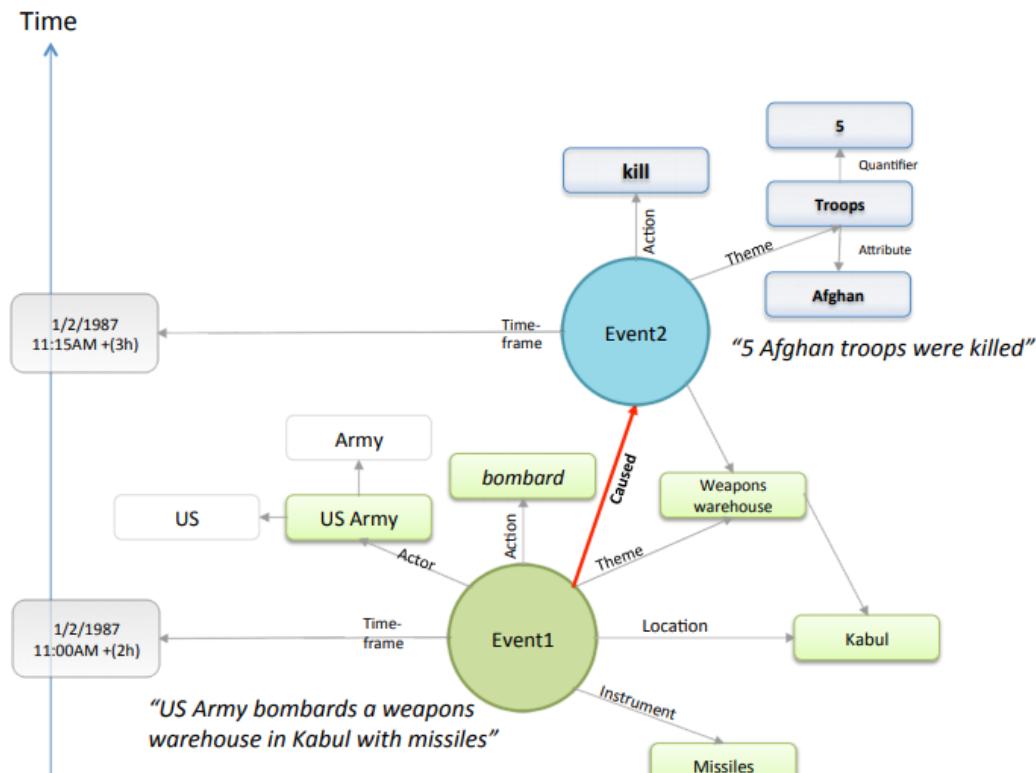
## ■ Quais os desafios?

- Identificar eventos correlacionados e (esperançosamente) a causalidade entre eventos!

# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Identificar eventos correlacionados e (esperançosamente) a causalidade entre eventos!

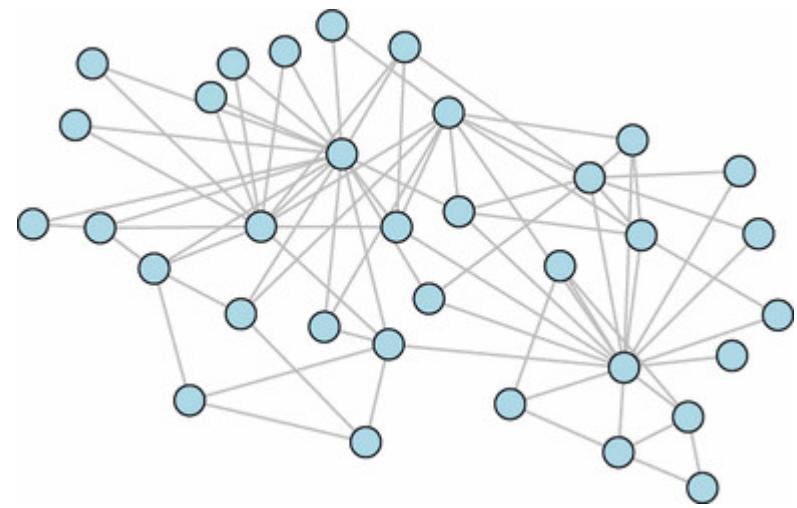
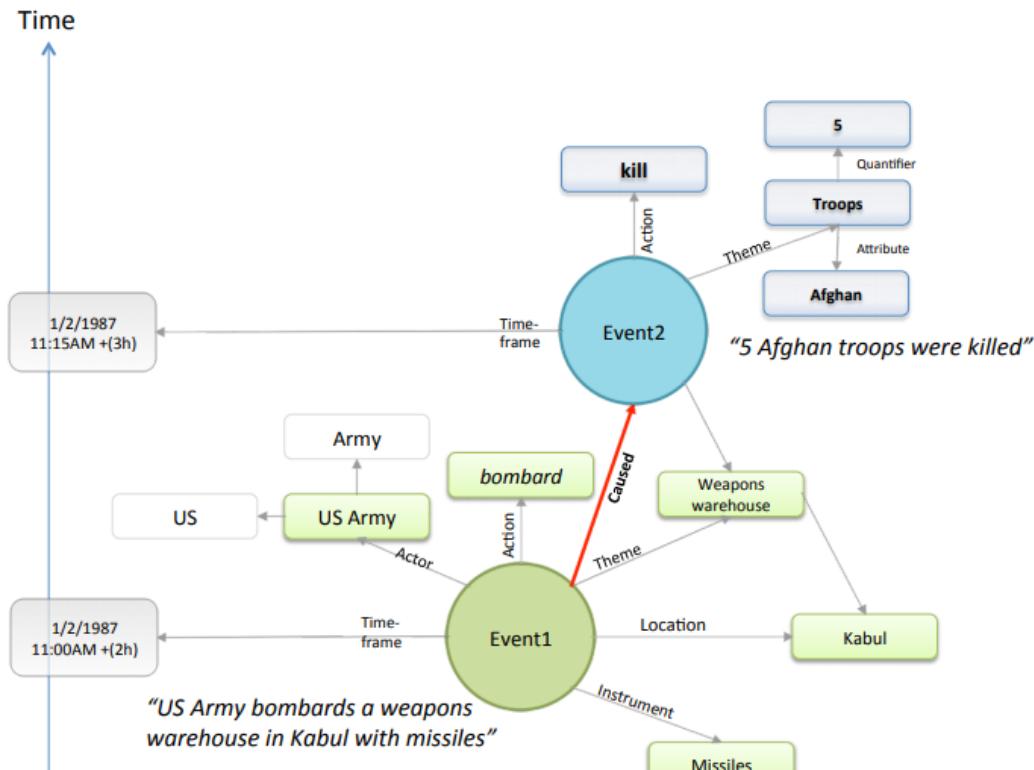


RADINSKY, Kira; HORVITZ, Eric. Mining the web to predict future events. In: Proceedings of the sixth ACM international conference on Web search and data mining. 2013. p. 255-264.

# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Identificar eventos correlacionados e (esperançosamente) a causalidade entre eventos!

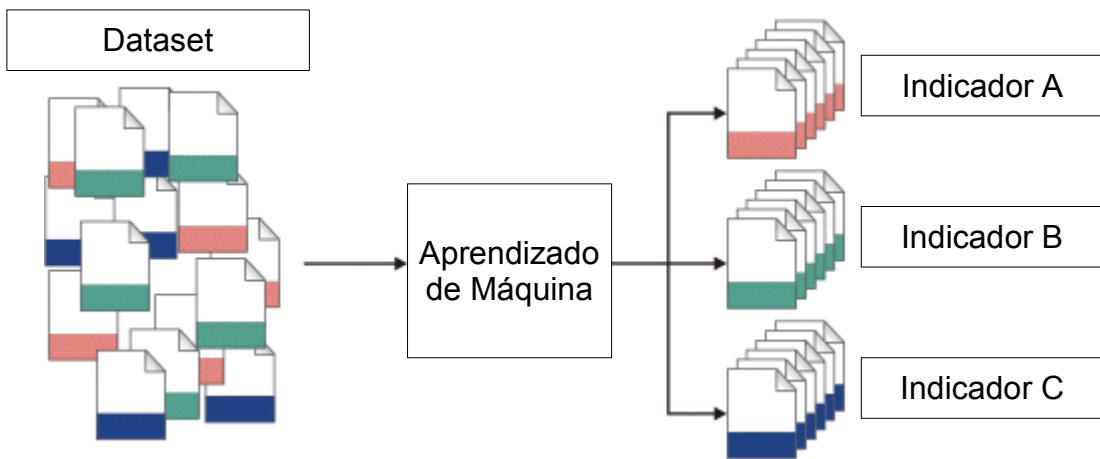


RADINSKY, Kira; HORVITZ, Eric. Mining the web to predict future events. In: Proceedings of the sixth ACM international conference on Web search and data mining. 2013. p. 255-264.

# Inteligência Analítica para Textos

## ■ Quais os desafios?

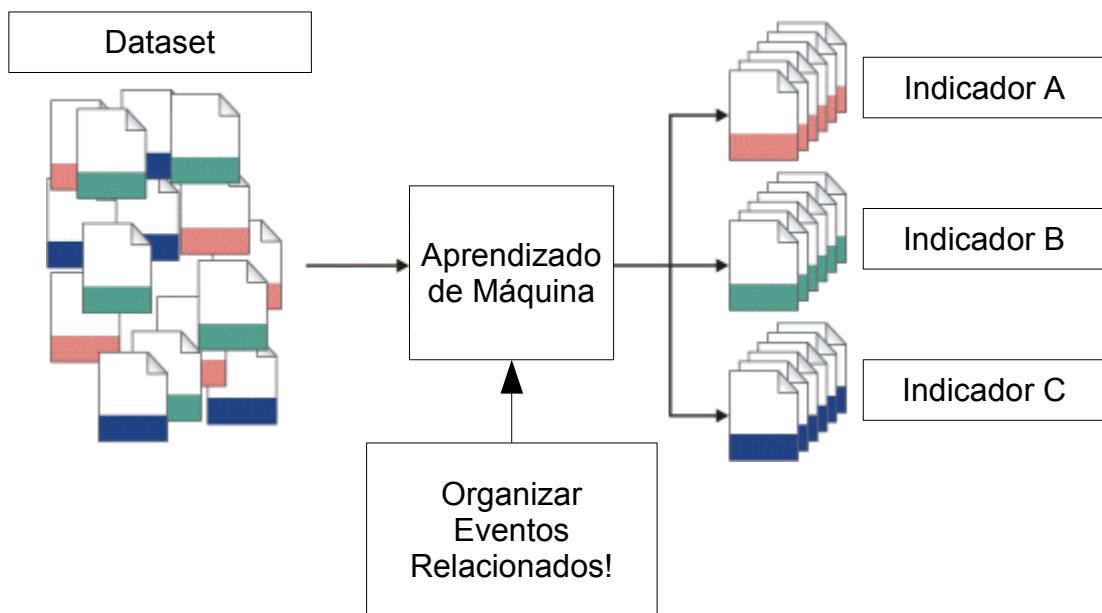
- Aprender indicadores (inteligentes) a partir de eventos extraídos dos textos



# Inteligência Analítica para Textos

## ■ Quais os desafios?

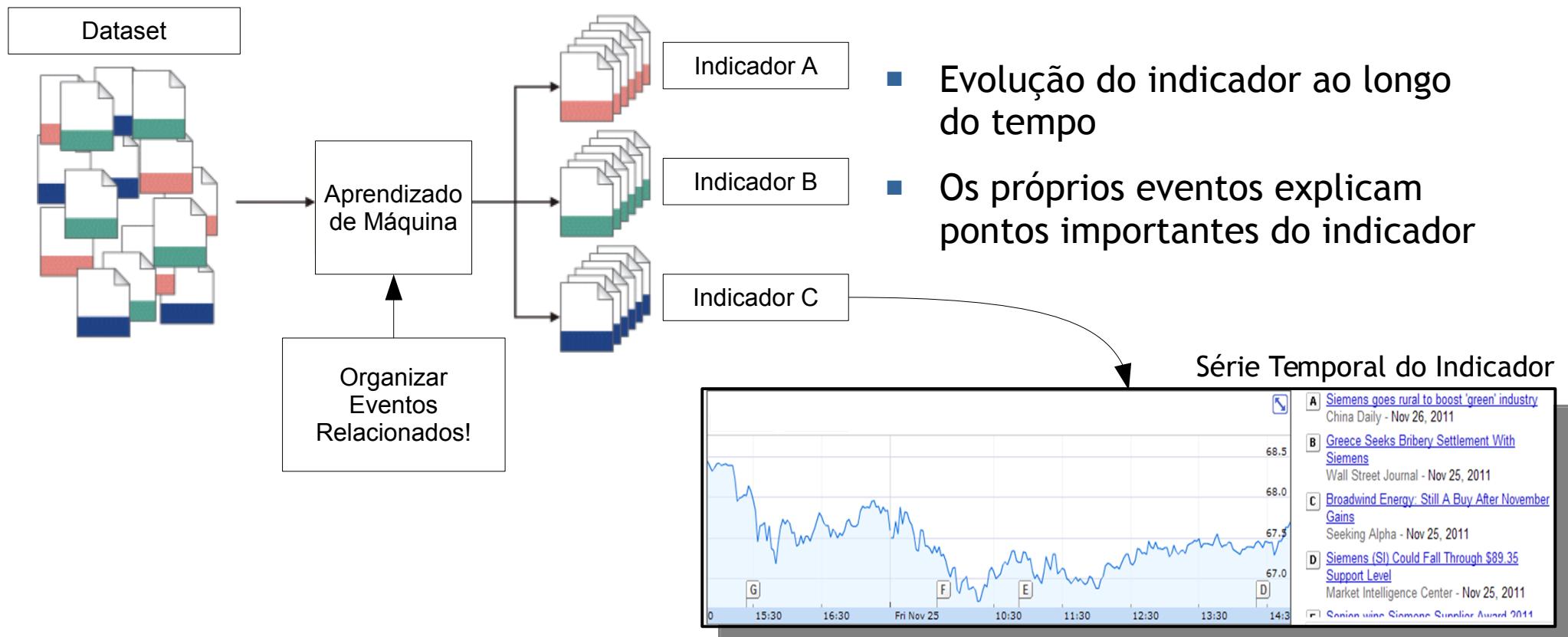
- Aprender indicadores (inteligentes) a partir de eventos extraídos dos textos



# Inteligência Analítica para Textos

## ■ Quais os desafios?

- Aprender indicadores (inteligentes) a partir de eventos extraídos dos textos



# Inteligência Analítica para Textos

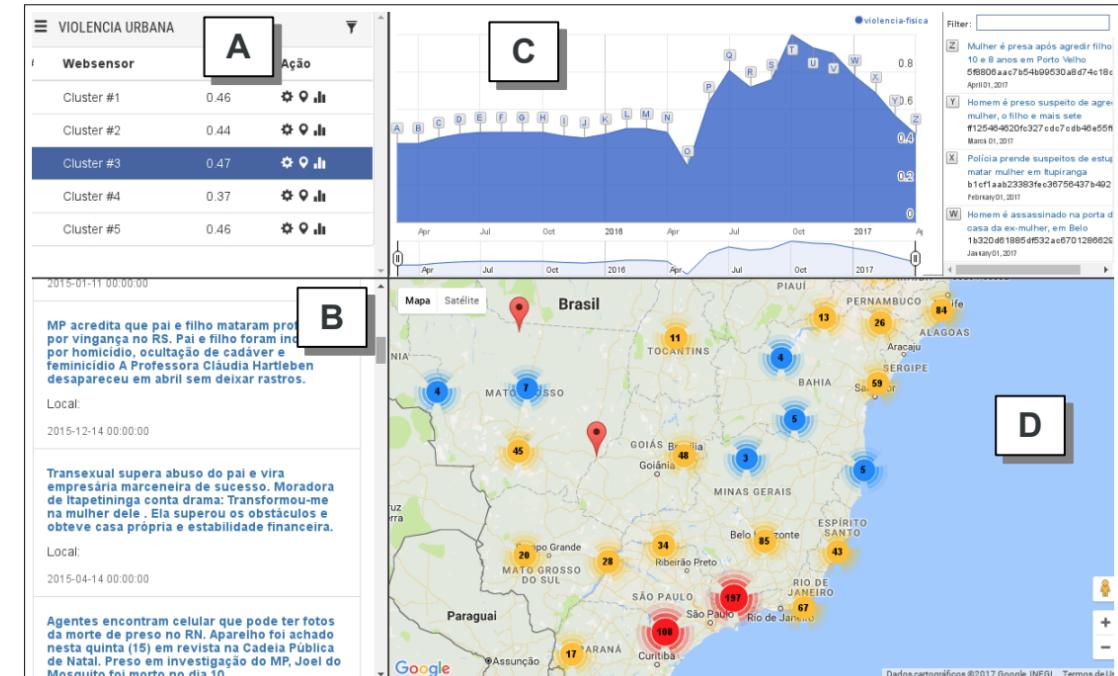
## ■ Aprendizado de Sensores

- A partir de uma amostra de eventos de interesse, aprender um sensor para monitorar (agrupar ou classificar) outros eventos relacionados...

## ■ Diversas aplicações

Exemplo:

Projeto Websensors



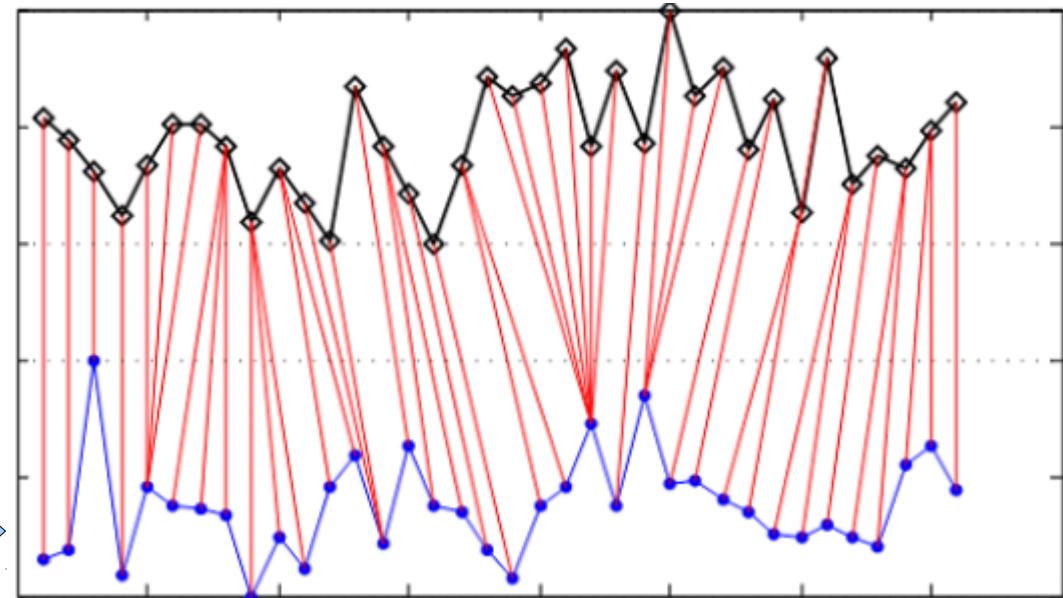
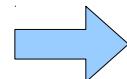
MARCACINI, Ricardo M. et al. Websensors analytics: Learning to sense the real world using web news events. In: Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web. SBC, 2017. p. 169-173.

# Inteligência Analítica para Textos

## ■ Quais os desafios?

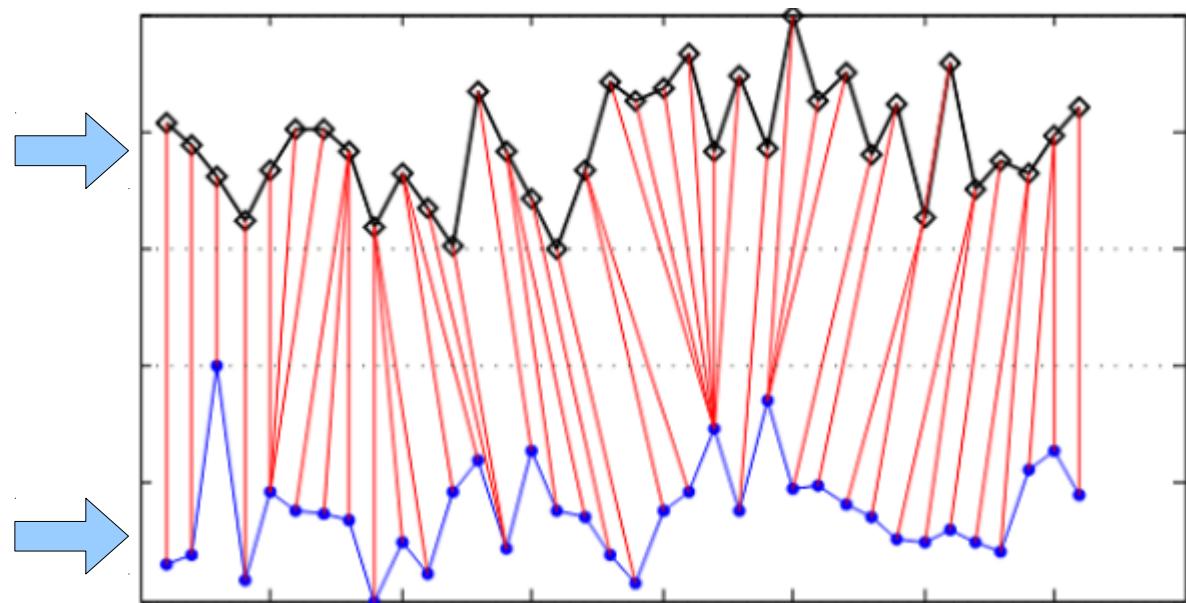
- Correlacionar indicadores obtidos dos textos com “indicadores tradicionais”

- Indicador da empresa:  
Ex: Taxa de Exportação de Produto



# Inteligência Analítica para Textos

## ■ Quais os desafios?

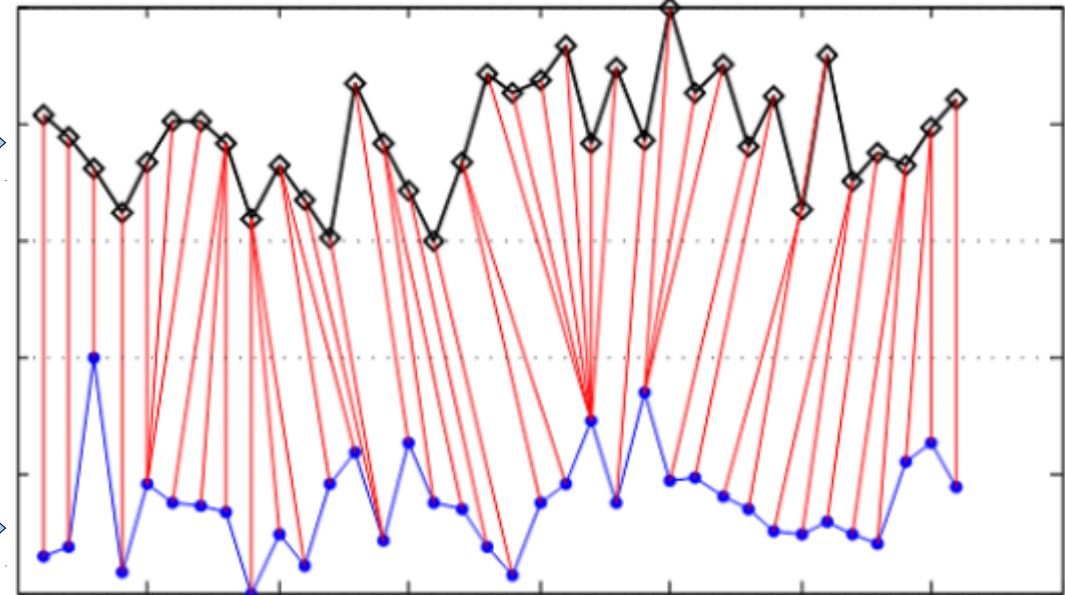
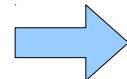
- Correlacionar indicadores obtidos dos textos com “indicadores tradicionais”
  - Indicador “aprendido”:  
Ex: Eventos sobre Investimento em Tecnologia
  - Indicador da empresa:  
Ex: Taxa de Exportação de Produto
- 

# Inteligência Analítica para Textos

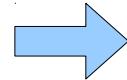
## ■ Quais os desafios?

- Correlacionar indicadores obtidos dos textos com “indicadores tradicionais”

- Indicador “aprendido”:  
Ex: Eventos sobre Investimento em Tecnologia



- Indicador da empresa:  
Ex: Taxa de Exportação de Produto

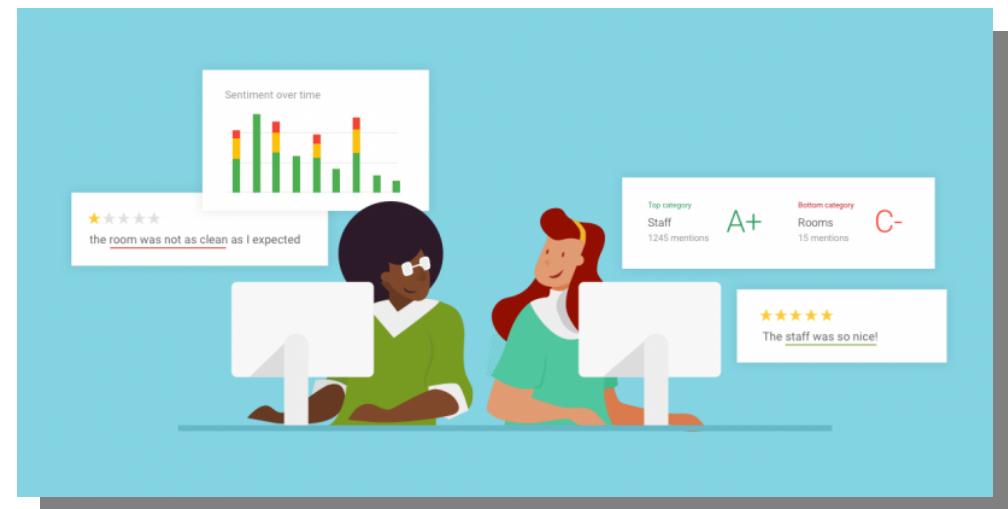


Apoio para tarefas descritivas (entender o comportamento passado) e tarefas preditivas (estimar comportamento futuro)

# Inteligência Analítica para Textos

## ■ Análise de Sentimentos

- Analisar eventos que representam uma opinião.
  - WHO? → Consumidor, Empresa, etc.
  - WHAT? → É a opinião a respeito de algo/algumé.
  - WHEN? → Quando a opinião foi emitida.
  - WHERE? → Local do usuário, empresa, etc.



Birdeye.com, "The benefits of customer sentiment analysis tools." <https://s3.amazonaws.com/blog4.0/wp-content/uploads/2018/04/customer-sentiment-analysis-tools-810x405.png>

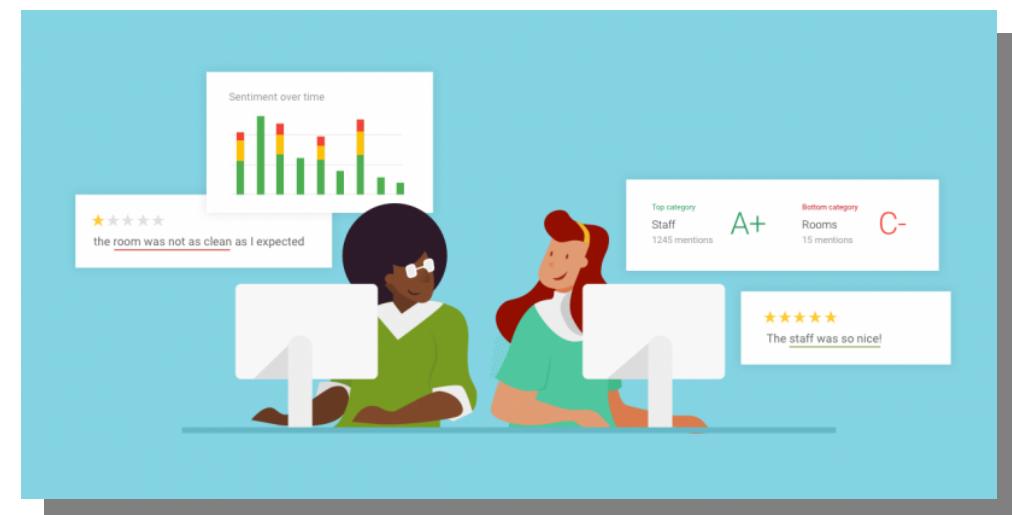
# Inteligência Analítica para Textos

## ■ Análise de Sentimentos

- Analisar eventos que representam uma opinião.
  - WHO? → Consumidor, Empresa, etc.
  - WHAT? → É a opinião a respeito de algo/algum.
  - WHEN? → Quando a opinião foi emitida.
  - WHERE? → Local do usuário, empresa, etc.
- Problema desafiador:
  - Sentimento por Aspecto!

José da Silva (São Paulo, SP)  
Em 12/04/2018:

O atendimento do Hotel ABC é muito bom,  
mas a limpeza do quarto precisa melhorar.

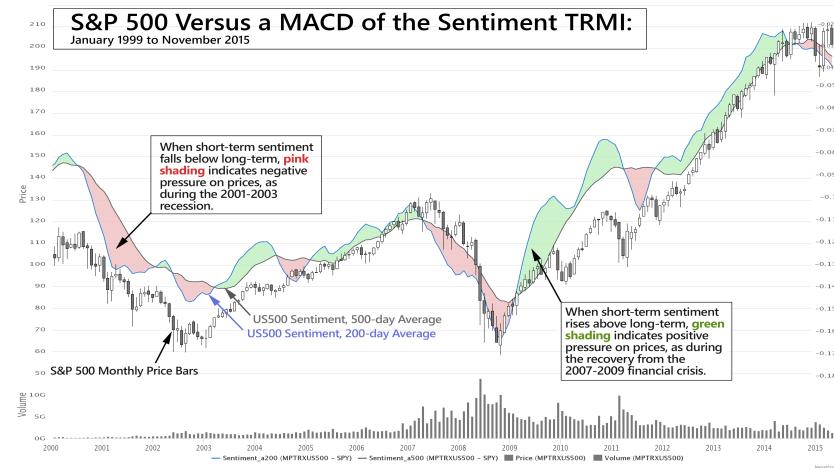


Birdeye.com, "The benefits of customer sentiment analysis tools." <https://s3.amazonaws.com/blog4.0/wp-content/uploads/2018/04/customer-sentiment-analysis-tools-810x405.png>

# Inteligência Analítica para Textos

## ■ Mercado Financeiro

- Recomendar eventos de notícias que podem afetar (positivamente ou negativamente) um ativo financeiro
- Estratégia: selecionar eventos e estimar comportamento futuro do ativo financeiro com base em eventos similares do passado

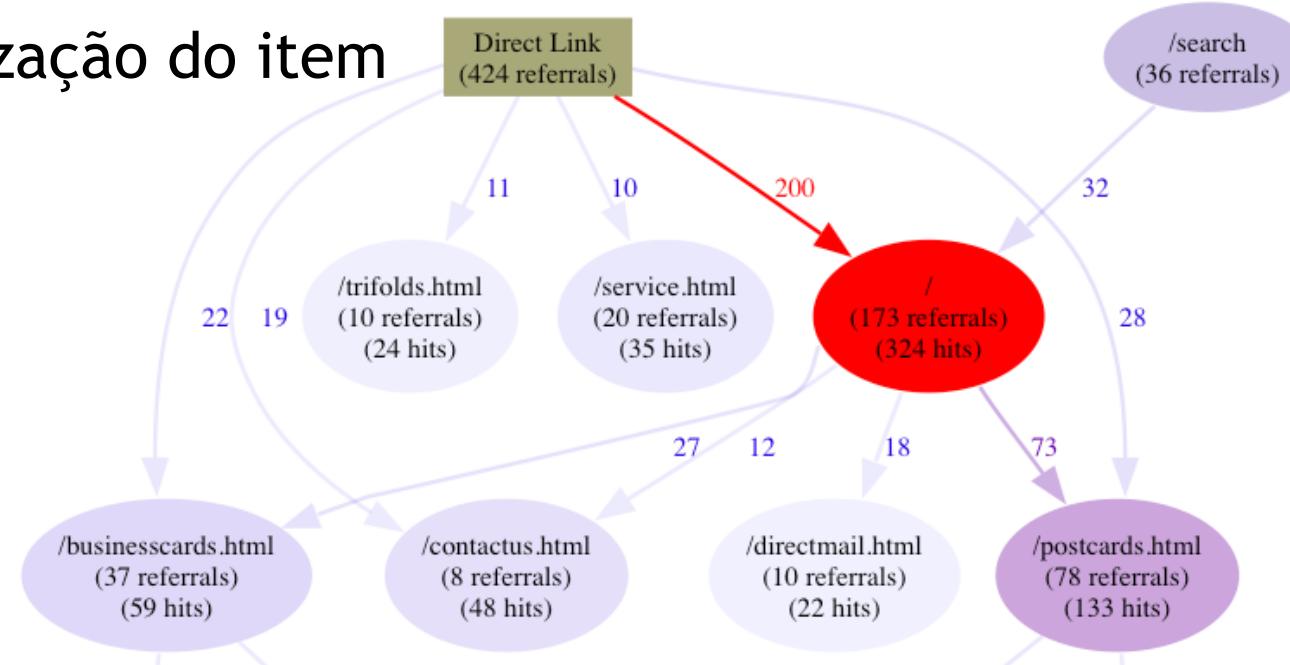


Marketpsych.com, "MarketPsych Report: How Fund Managers Trade onSentiment."  
[http://marketpsych-website.s3.amazonaws.com/images/S%26P500\\_MACD\\_1998-2015.png](http://marketpsych-website.s3.amazonaws.com/images/S%26P500_MACD_1998-2015.png)

# Inteligência Analítica para Textos

## ■ Clickstream

- Analisar eventos com “logs” de navegação/interação
  - WHO? → usuário
  - WHAT? → conteúdo do item
  - WHEN? → data do acesso
  - WHERE? → localização do item

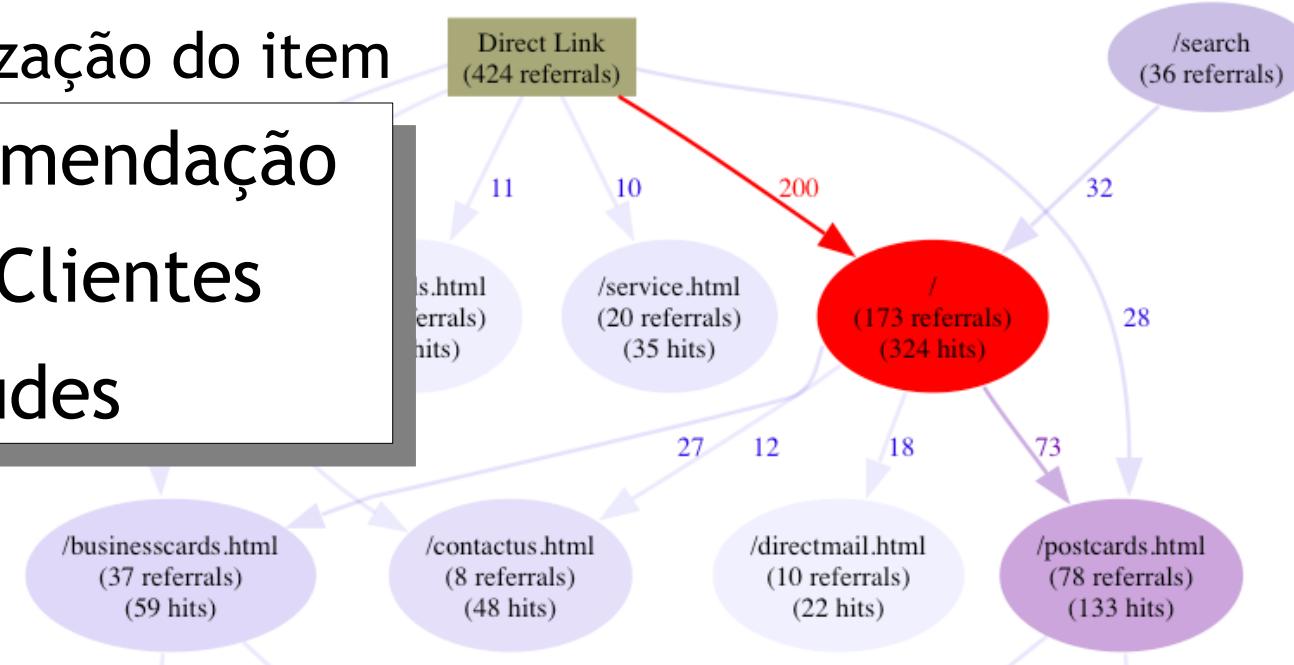


# Mineração de Eventos

## ■ Clickstream

- Analisar eventos com “logs” de navegação/interação
  - WHO? → usuário
  - WHAT? → conteúdo do item
  - WHEN? → data do acesso
  - WHERE? → localização do item

- Sistemas de Recomendação
- Segmentação de Clientes
- Detecção de Fraudes



# Agenda

- Mineração de Textos e Inteligência Analítica
- Mineração de Eventos
  - Identificação do Problema
  - Pré-Processamento
  - Extração de Padrões
  - Pós-processamento
  - Uso do Conhecimento
- Projetos em andamento

# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados se possuem conteúdo similar.

Como extrair e representar a informação textual dos eventos?

### Seeking Life's Bare (Genetic) Necessities

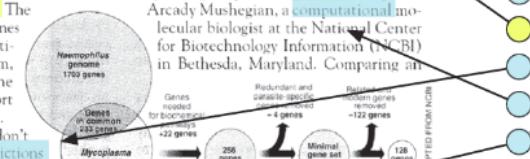
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>10</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

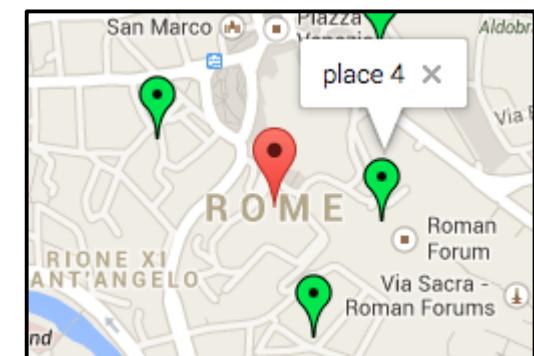
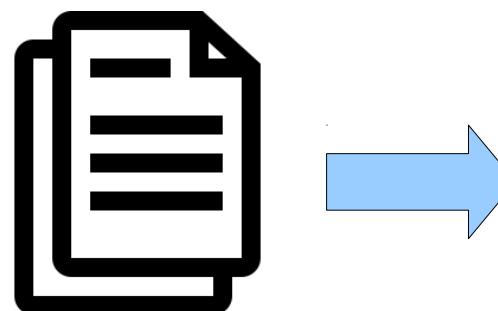
Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados se ocorreram em regiões próximas

Como identificar informação geográfica em eventos e georreferenciar?



# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

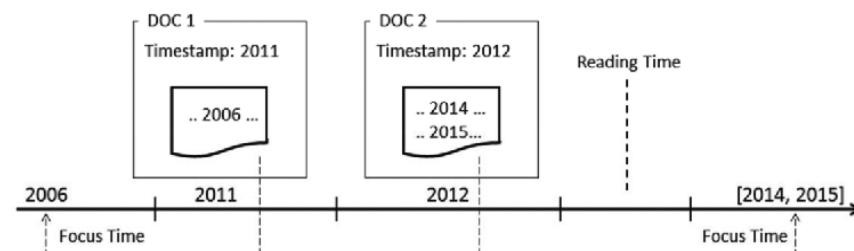
Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados se ocorreram no mesmo período de tempo.

Como extrair informação temporal dos eventos?



# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados conforme informação de domínio.
  - Nomes de pessoas e organizações
  - Entidades do domínio

# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados se possuem conteúdo similar.

Como extrair e representar a informação textual dos eventos?

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>10</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

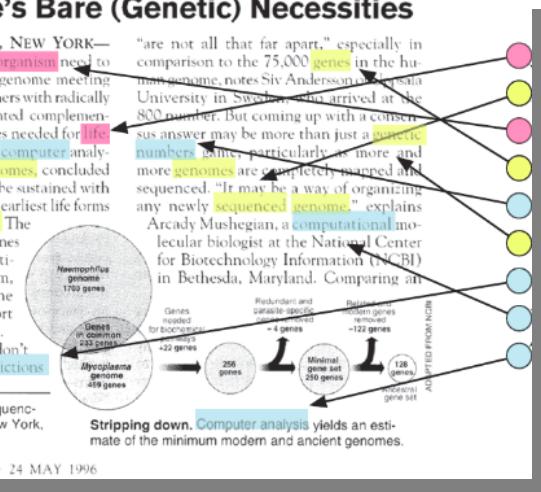
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



# Mineração de Eventos

## Pré-processamento de Textos Modelo Espaço-Vetorial

[11] AGGARWAL, Charu C. Text Preparation and Similarity Computation. In: Machine Learning for Text. Springer, Cham, 2018. p. 17-30.

# Mineração de Eventos

## ■ Pré-processamentos dos textos

### ■ Modelo espaço-vetorial

- Cada objeto (e.g. documentos, eventos, etc.) é representado por um vetor de  $m$  dimensões.
- Cada dimensão é um atributo.
- Cada atributo tem um peso indicando sua relevância para um determinado objeto



# Mineração de Eventos

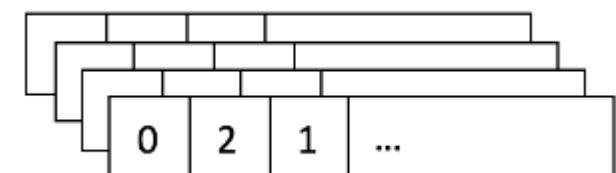
## ■ Pré-processamentos dos textos

### ■ Modelo espaço-vetorial

- Cada objeto (e.g. documentos, eventos, etc.) é representado por um vetor de  $m$  dimensões.
- Cada dimensão é um atributo.
- Cada atributo tem um peso indicando sua relevância para um determinado objeto

Questões do modelo espaço-vetorial:

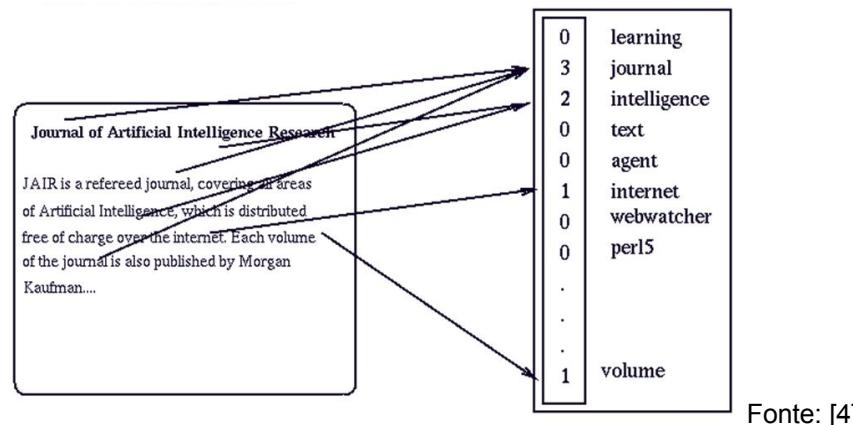
- 1) Quais são os atributos?
- 2) Como definir os pesos dos atributos?



Vetores para 4 objetos

# Mineração de Eventos

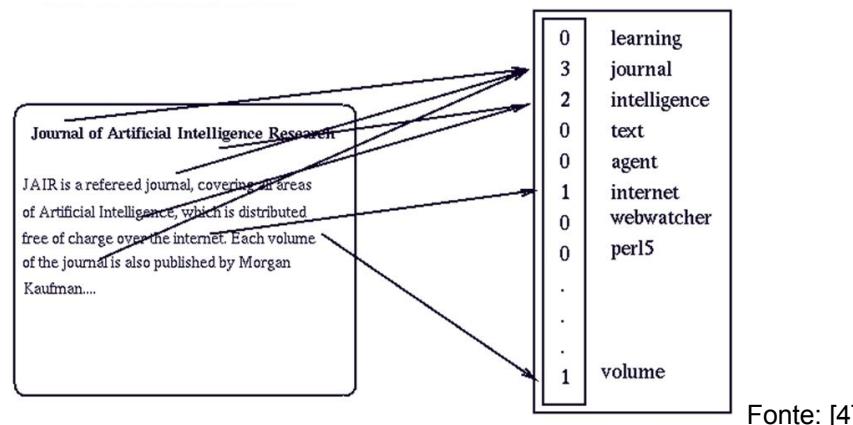
- Pré-processamentos dos textos
  - Modelo espaço-vetorial usando *Bag-of-words*
    - Atributos são extraídas dos textos
    - Peso da palavra é sua frequência objeto
    - A ordem das palavras nos textos não é considerada



Fonte: [4]

# Mineração de Eventos

- Pré-processamentos dos textos
  - Modelo espaço-vetorial usando *Bag-of-words*
    - Atributos são extraídas dos textos
    - Peso da palavra é sua frequência objeto
    - A ordem das palavras nos textos não é considerada

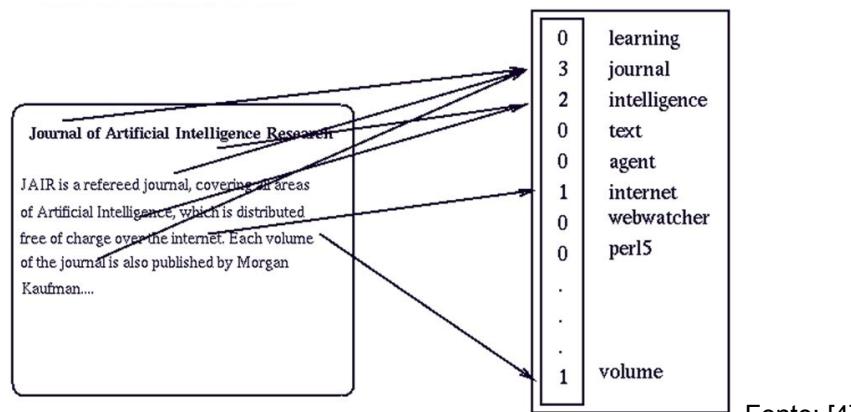


Fonte: [4]

*Bag-of-words* é uma representação que “subestima” o problema.  
Porém, pode ser suficiente para várias aplicações!

# Mineração de Eventos

- Pré-processamentos dos textos
  - Modelo espaço-vetorial usando *Bag-of-words*
    - Atributos são extraídas dos textos
    - Peso da palavra é sua frequência objeto
    - A ordem das palavras nos textos não é considerada



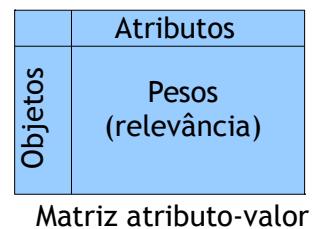
Fonte: [4]

Como tornar a *Bag-of-Words* uma representação mais concisa,  
ou seja, reduzir informação redundante?

# Mineração de Eventos

## ■ Pré-processamento - Informação Textual

- Bag-of-words: representação no modelo espaço-vetorial. Simples (Baseline).



- Pode ser construída com técnicas estatísticas simples
- Permite o uso de diferentes algoritmos de aprendizado de máquina

Exemplo de modelo espaço-vetorial (*bag-of-words*)

Text	This	Is	A	Nice	Hotel	Not	All	at
This is a nice hotel	1	1	1	1	1	0	0	0
Not a nice hotel! not at all	0	0	1	1	1	2	1	1

Fonte: [3]

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Técnicas mais utilizadas:
    - Remoção de pontuações e *stopwords*
    - Radicalização de palavras
    - N-gramas
    - Ponderação por TF-IDF

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Remoção de pontuações e *stopwords*

Dado um texto, remover pontuações, pronomes, preposição e artigos.

Original:

O estudante de Inteligência Artificial foi na livraria comprar livros para estudar.

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Remoção de pontuações e *stopwords*

Dado um texto, remover pontuações, pronomes, preposição e artigos.

Original:

O estudante de Inteligência Artificial foi na livraria comprar livros para estudar.

Identificando pontuação e *stopwords*:

Q estudante de Inteligência Artificial foi na livraria comprar livros para estudar.

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Remoção de pontuações e *stopwords*

Dado um texto, remover pontuações, pronomes, preposição e artigos.

Original:

O estudante de Inteligência Artificial foi na livraria comprar livros para estudar.

Identificando pontuação e *stopwords*:

Q estudante de Inteligência Artificial foi na livraria comprar livros para estudar.

Final:

estudante Inteligência Artificial foi livraria comprar livros estudar

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Técnicas mais utilizadas:
    - Remoção de pontuações e *stopwords*
    - Radicalização de palavras
    - N-gramas
    - Ponderação por TF-IDF

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Radicalização de palavras

Dado um texto converter variações de uma palavra para uma única forma.

Exemplo: {comprar, compras, comprei} → compr

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Radicalização de palavras

Dado um texto converter variações de uma palavra para uma única forma.

Exemplo: {comprar, compras, comprei} → compr

Texto:

estudante Inteligência Artificial foi livraria comprar livros estudar

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Radicalização de palavras

Dado um texto converter variações de uma palavra para uma única forma.

Exemplo: {comprar, compras, comprei} → compr

Texto:

estudante Inteligência Artificial foi livraria comprar livros estudar

Após radicalização:

estud Intelig Artifici fo livr compr livr estud

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Radicalização de palavras

Dado um texto converter variações de uma palavra para uma única forma.

Exemplo: {comprar, compras, comprei} → compr

## Notas importantes:

- Radicalização é dependente da línguagem.
- Alguns estudos reportam que pode prejudicar a extração de conhecimento.
- Erros de radicalização: overstemming e understemming
- Algoritmos de radicalização populares: Porter (várias línguas) e Orengo (português)

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Técnicas mais utilizadas:
    - Remoção de pontuações e *stopwords*
    - Radicalização de palavras
    - N-gramas
    - Ponderação por TF-IDF

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - N-gramas

Consiste em combinar duas ou mais palavras em um termo (composto), com um sentido único.

Exemplo: {Data, Mining} → {Data\_Mining}

Texto:

estud Intelig Artifici fo livr compr livr estud

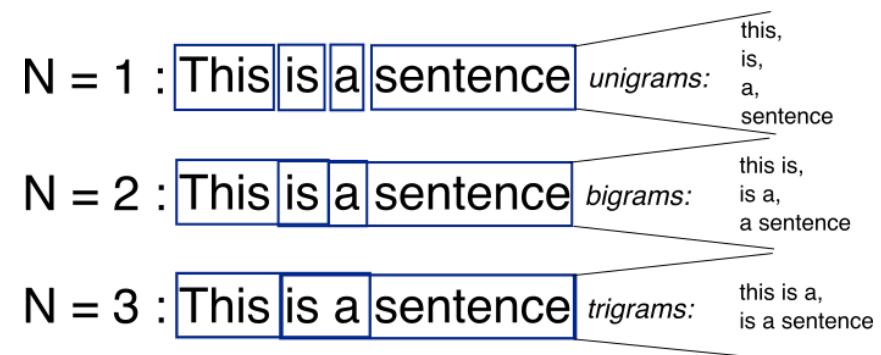
Após identificação de *n-gramas*:

estud Intelig\_Artifici fo livr compr livr estud

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - N-gramas
    - Extração de n-gramas não é um problema trivial.
    - Identificar quando a coocorrência entre duas ou mais palavras é significativa (não ocorre ao acaso).
    - Exemplo:

<https://books.google.com/ngrams>



# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Técnicas mais utilizadas:
    - Remoção de pontuações e *stopwords*
    - Radicalização de palavras
    - N-gramas
    - Ponderação por TF-IDF

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Ponderação por TF-IDF
    - Identificar um *trade-off*:
      - Atributos que são frequentes em um objeto são relevantes.
      - Atributos que ocorrem em muitos objetos não são relevantes.

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Refinando a *Bag-of-words* (representação concisa)
  - Ponderação por TF-IDF

	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

	angeles	los	new	post	times	york
d1	0	0	0.584	0	0.584	0.584
d2	0	0	0.584	1.584	0	0.584
d3	1.584	1.584	0	0	0.584	0

$$w_{x,y} = tf_{x,y} \times \log \left( \frac{N}{df_x} \right)$$

## TF-IDF

Term  $x$  within document  $y$

$tf_{x,y}$  = frequency of  $x$  in  $y$   
 $df_x$  = number of documents containing  $x$   
 $N$  = total number of documents

# Proximidade de Conteúdo

## ■ Pré-processamentos dos textos

### ■ Modelo espaço-vetorial

- Estudamos as técnicas mais básicas da área.
- Representa um (razoável) *baseline* para representação.
- Qualquer nova proposta de representação de textos deve ser melhor do que a representação aqui estudada.

A partir de uma representação estruturada podemos computar a similaridade entre dois eventos textuais!

# Proximidade de Conteúdo

## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Como calcular a proximidade entre conteúdo no modelo espaço-vetorial?

	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3

Vamos considerar quatro eventos.  
Escolhemos (propositadamente) apenas dois atributos.

e1 → O estudo da vida submarina (...)

e2 → A temporada de caça começou (...)

e3 → A caça submarina é ilegal no período (...)

e4 → Multas por caça submarina cresceram (...)

# Proximidade de Conteúdo

## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Como calcular a proximidade entre conteúdo no modelo espaço-vetorial?

	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3

Mais  
relacionados

Vamos considerar quatro eventos.  
Escolhemos (propositadamente) apenas dois atributos.

e1 → O estudo da vida submarina (...)

e2 → A temporada de caça começou (...)

e3 → A caça submarina é ilegal no período (...)

e4 → Multas por caça submarina cresceram (...)

# Proximidade de Conteúdo

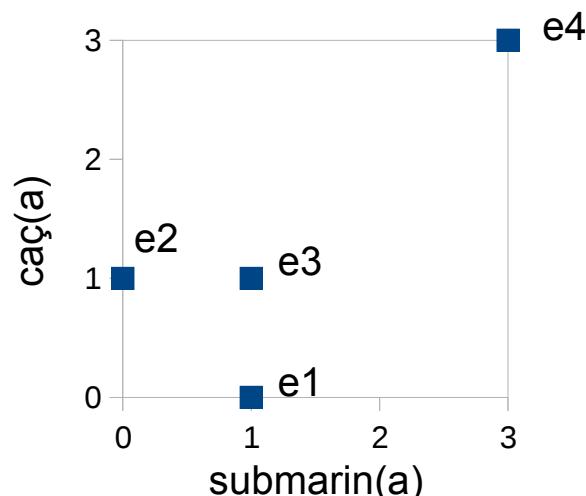
## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Como calcular a proximidade entre conteúdo no modelo espaço-vetorial?

	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3



# Proximidade de Conteúdo

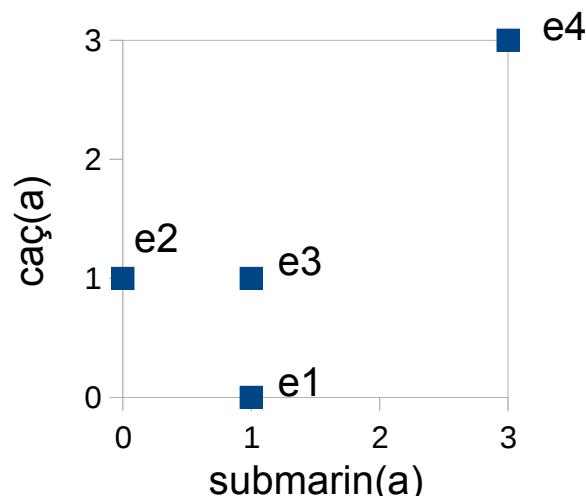
## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ O espaço euclidiano não capturou adequadamente o conceito de proximidade entre os eventos!

	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3



# Proximidade de Conteúdo

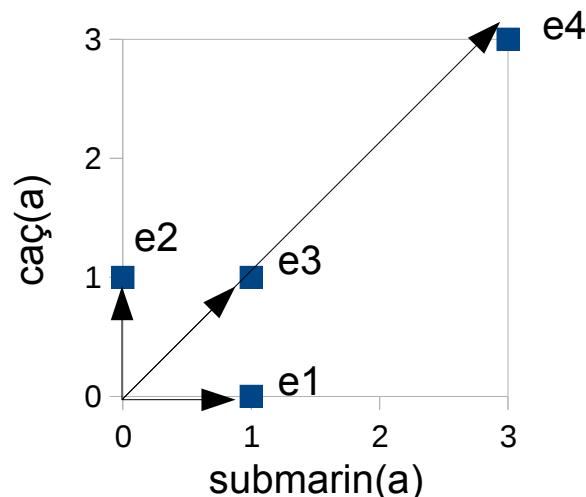
## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Considere utilizar o ângulo entre os vetores!

	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3



# Proximidade de Conteúdo

## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Considere utilizar o ângulo entre os vetores!

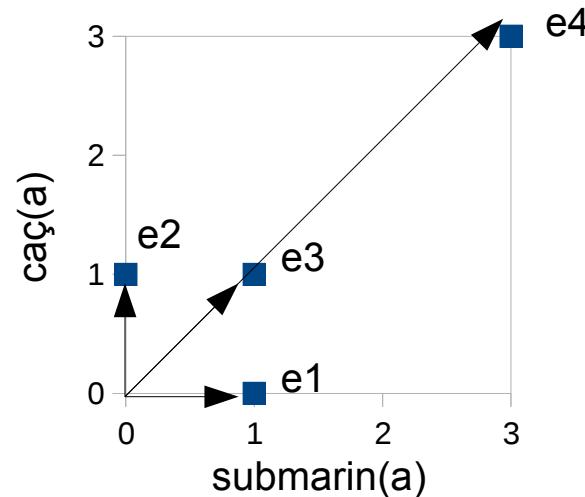
	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3

#### Alguns exemplos:

$$\text{ângulo}(e1, e2) = 90^\circ; \cos(90^\circ) = 0$$

$$\text{ângulo}(e2, e3) = 45^\circ; \cos(45^\circ) = 0.5$$

$$\text{ângulo}(e3, e4) = 0^\circ; \cos(0^\circ) = 1$$



# Proximidade de Conteúdo

## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Sejam os vetores $a_i$ e $a_j$ , com k dimensões:

Proximidade de conteúdo por  
similaridade de cosseno

$$\frac{\sum_k a_{i,k} a_{j,k}}{\sqrt{\sum_k a_{i,k}^2} \sqrt{\sum_k a_{j,k}^2}}$$

Quanto maior, mais  
próximo.

# Proximidade de Conteúdo

## ■ O problema da similaridade

### ■ Premissa No. 1 - Proximidade de Conteúdo.

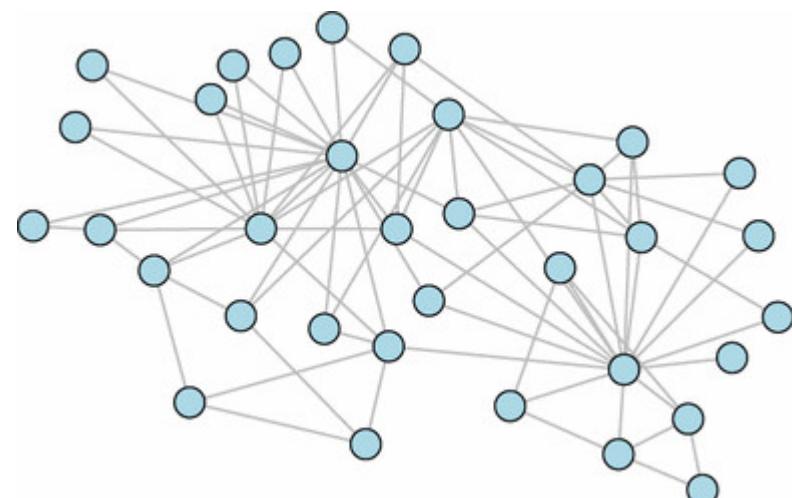
*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Sejam os vetores $a_i$ e $a_j$ , com k dimensões:

Proximidade de conteúdo por  
similaridade de cosseno

$$\frac{\sum_k a_{i,k} a_{j,k}}{\sqrt{\sum_k a_{i,k}^2} \sqrt{\sum_k a_{j,k}^2}}$$

Quanto maior, mais  
próximo.



# Mineração de Eventos

## ■ Prática #1

- Pré-processamento básico de textos
- Representação bag-of-words
- Representação VSM com ponderação TFIDF
- Similaridade Cosseno
- Geração de Rede de Eventos

Link para o código da Prática #1  
está disponível no repositório do curso.

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Considerar apenas a *Bag-of-words* para representar informação textual pode falhar em alguns casos na Mineração de Eventos
- Exemplo:
  - *D1: Obama speaks to the media in Illinois.*
  - *D2: The President greets the press in Chicago.*

*D1 e D2 representam eventos relacionados.  
Não possuem atributos em comum!*

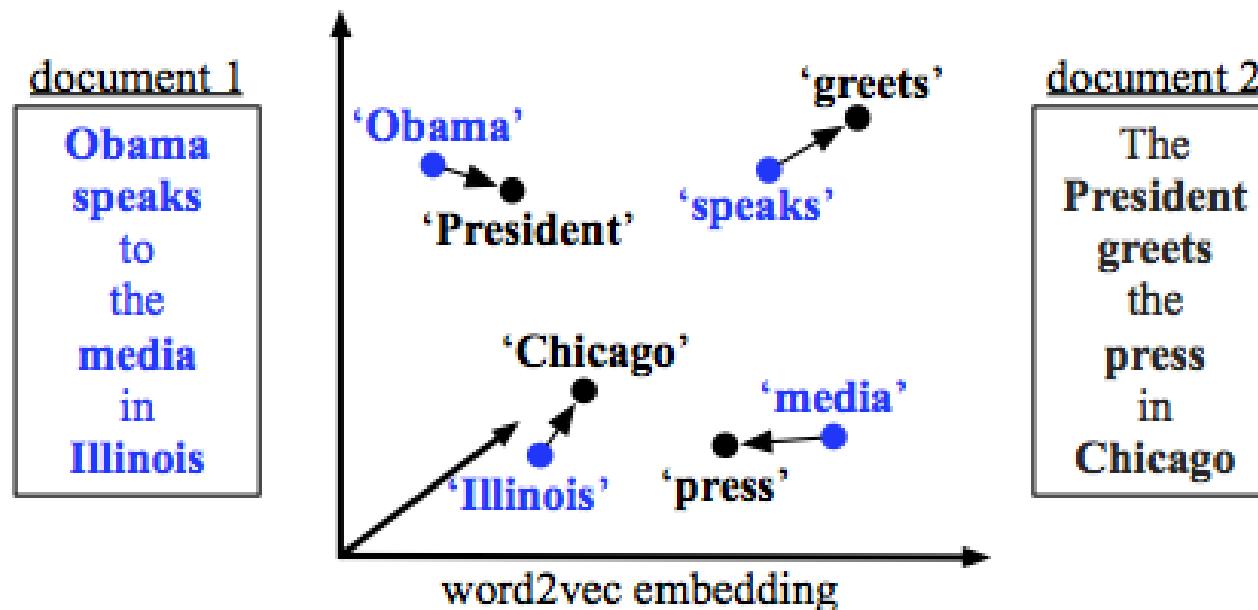
# Mineração de Eventos

- Pré-processamento - Informação Textual
  - Considerar apenas a *Bag-of-words* para representar informação textual pode falhar em alguns casos na Mineração de Eventos
- Exemplo:
  - *D1: Obama speaks to the media in Illinois.*
  - *D2: The President greets the press in Chicago.*

Uma solução para este problema:  
*Word Embedding Models*

# Mineração de Eventos

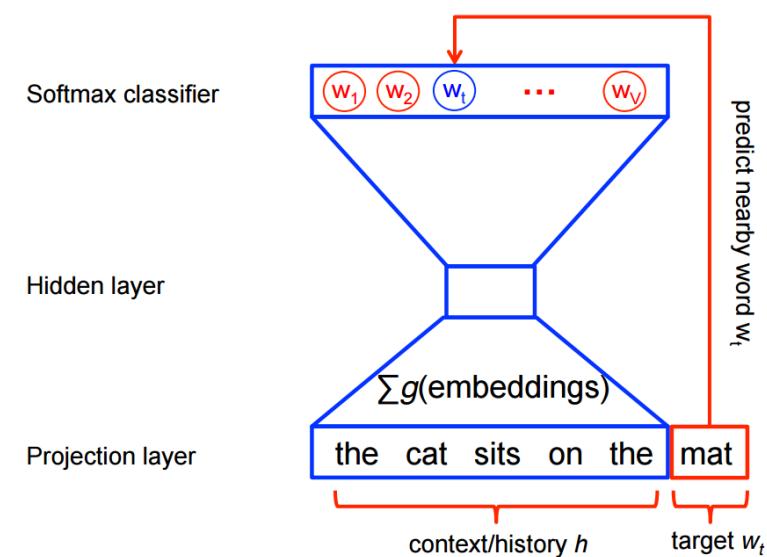
- Pré-processamento - Informação Textual
  - *Word Embedding Models*



# Mineração de Eventos

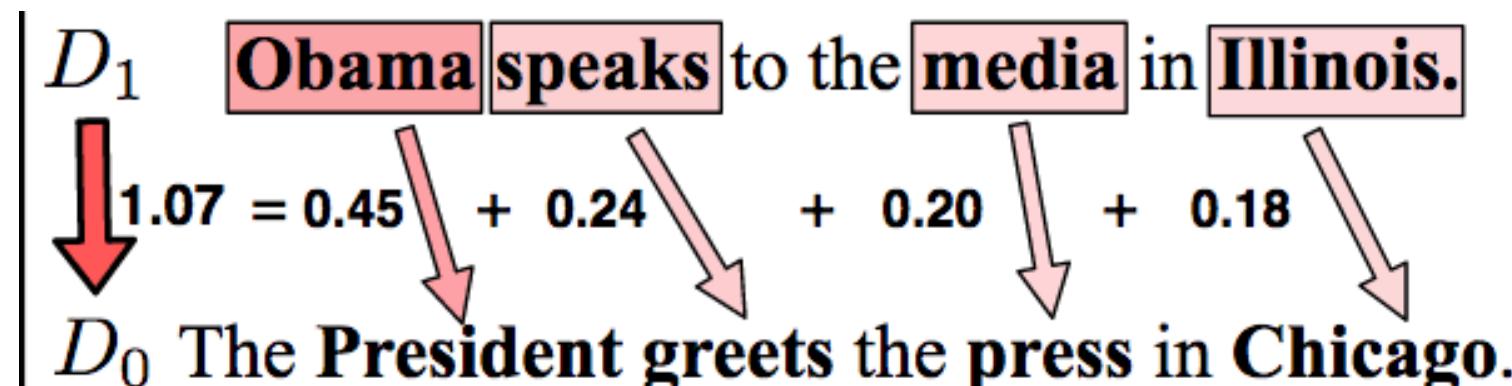
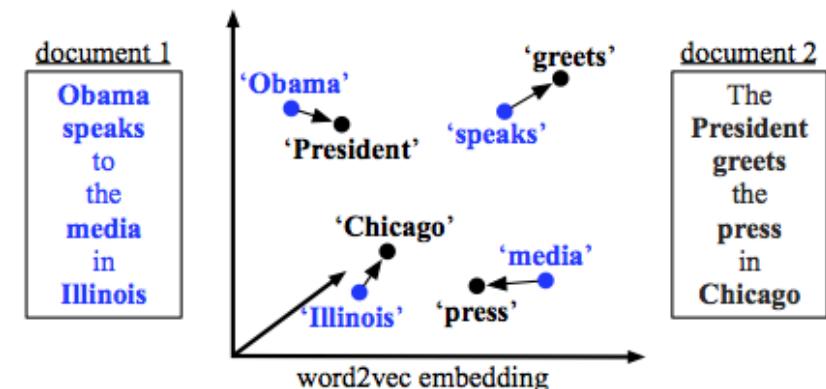
## ■ Pré-processamento - Informação Textual

- *Word Embedding Models*
- Alguns modelos pré-treinados (públicos)
  - Word2Vec (Google)
    - Gogle News dataset (100 bilhões de tokens)
  - Glove (Stanford)
    - Wikipedia (6 bilhões de tokens)
  - Fasttext (Facebook)
    - Wikipedia e Statmt News (16 bilhões de tokens)
    - 157 línguas



# Mineração de Eventos

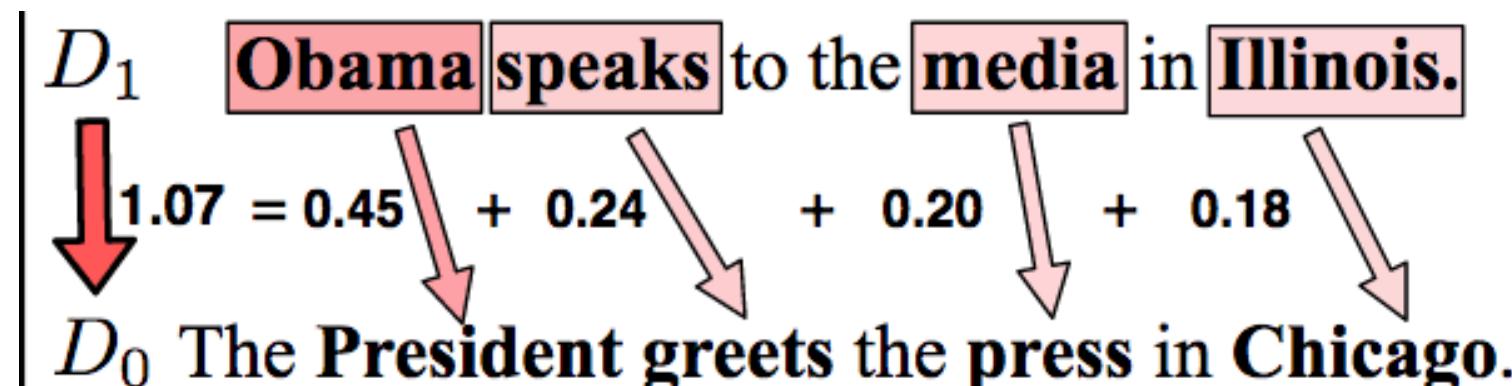
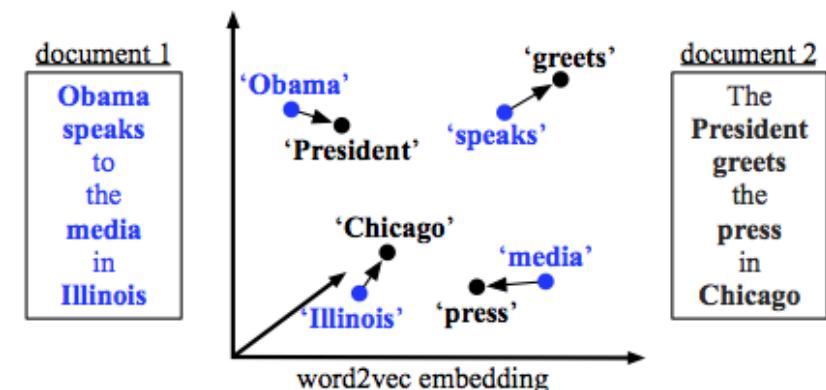
- Pré-processamento - Informação Textual
  - *Word Embedding Models*



KUSNER, Matt et al. From word embeddings to document distances. In: International conference on machine learning. 2015. p. 957-966.

# Mineração de Eventos

- Pré-processamento - Informação Textual
  - *Word Embedding Models*



KUSNER, Matt et al. From word embeddings to document distances. In: International conference on machine learning. 2015. p. 957-966.

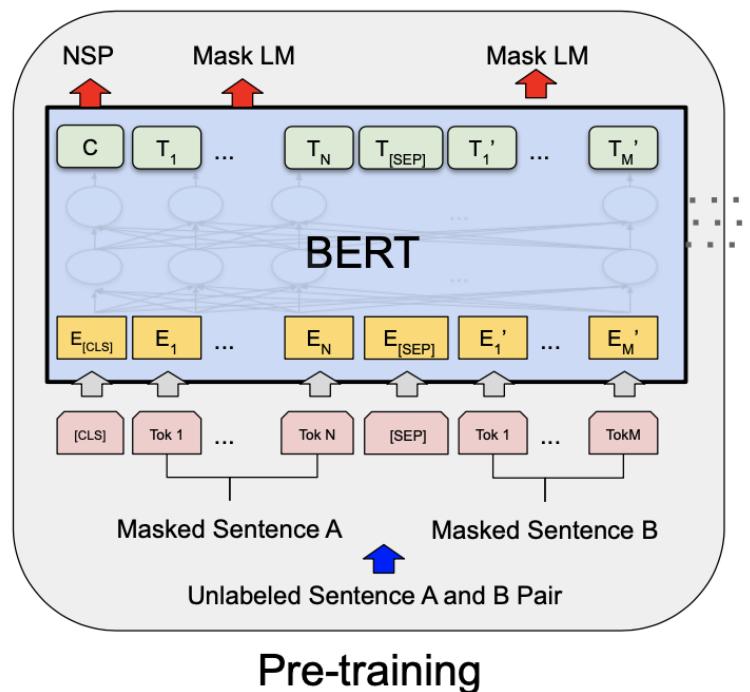
# Mineração de Eventos

- Pré-processamento - Informação Textual
  - *Word Embedding Models Livres de Contexto*
    - Word2Vec, FastText, Glove
    - Após o treinamento, os word vectors são estáticos
    - Exemplo
      - *Eu sentei no banco da praça*
      - *Eu fui no banco conferir o saldo*
  - Mais recentemente
    - *Word Embeddings Contextuais*
    - Exemplo: BERT (Bidirectional Encoder Representations from Transformers)

# Mineração de Eventos

## ■ Pré-processamento - Informação Textual

- *Word Embedding Models Contextuais*



- *Masked Language Model*
  - Durante o treinamento, ~15% das palavras são mascaradas. O modelo tenta prever tais palavras.
- *Next Sentence Prediction*
  - Prever quando uma sentença é sucessora de outra sentença

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

# Mineração de Eventos

## ■ Prática #2

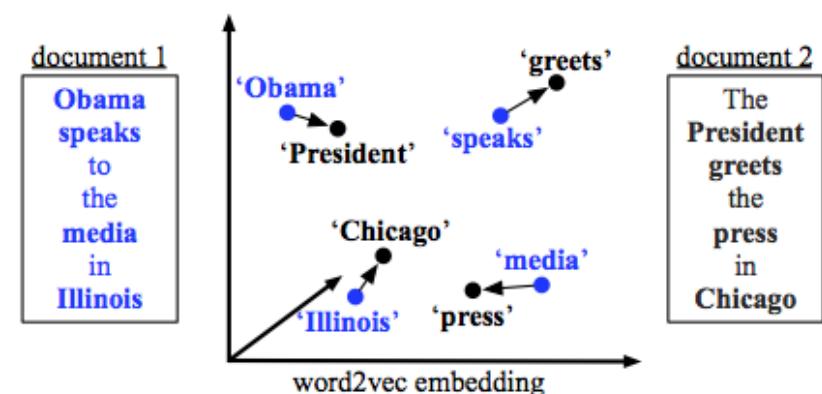
### ■ Proximidade de Conteúdo com Word Embeddings.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar.*

### ■ Experimentar a similaridade cosseno a partir dos word vectors:

Proximidade de conteúdo por similaridade de cosseno

$$\frac{\sum_k a_{i,k} a_{j,k}}{\sqrt{\sum_k a_{i,k}^2} \sqrt{\sum_k a_{j,k}^2}}$$



# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

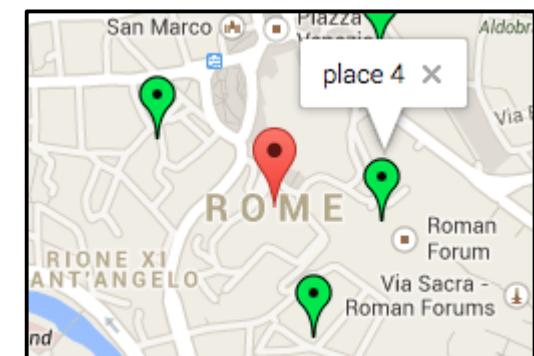
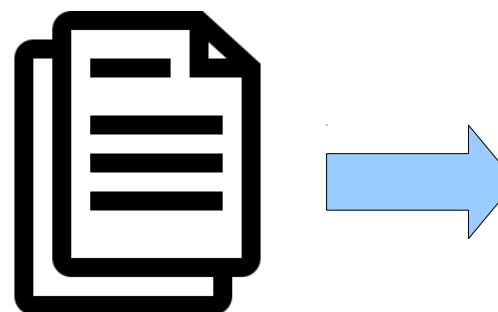
Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados se ocorreram em regiões próximas

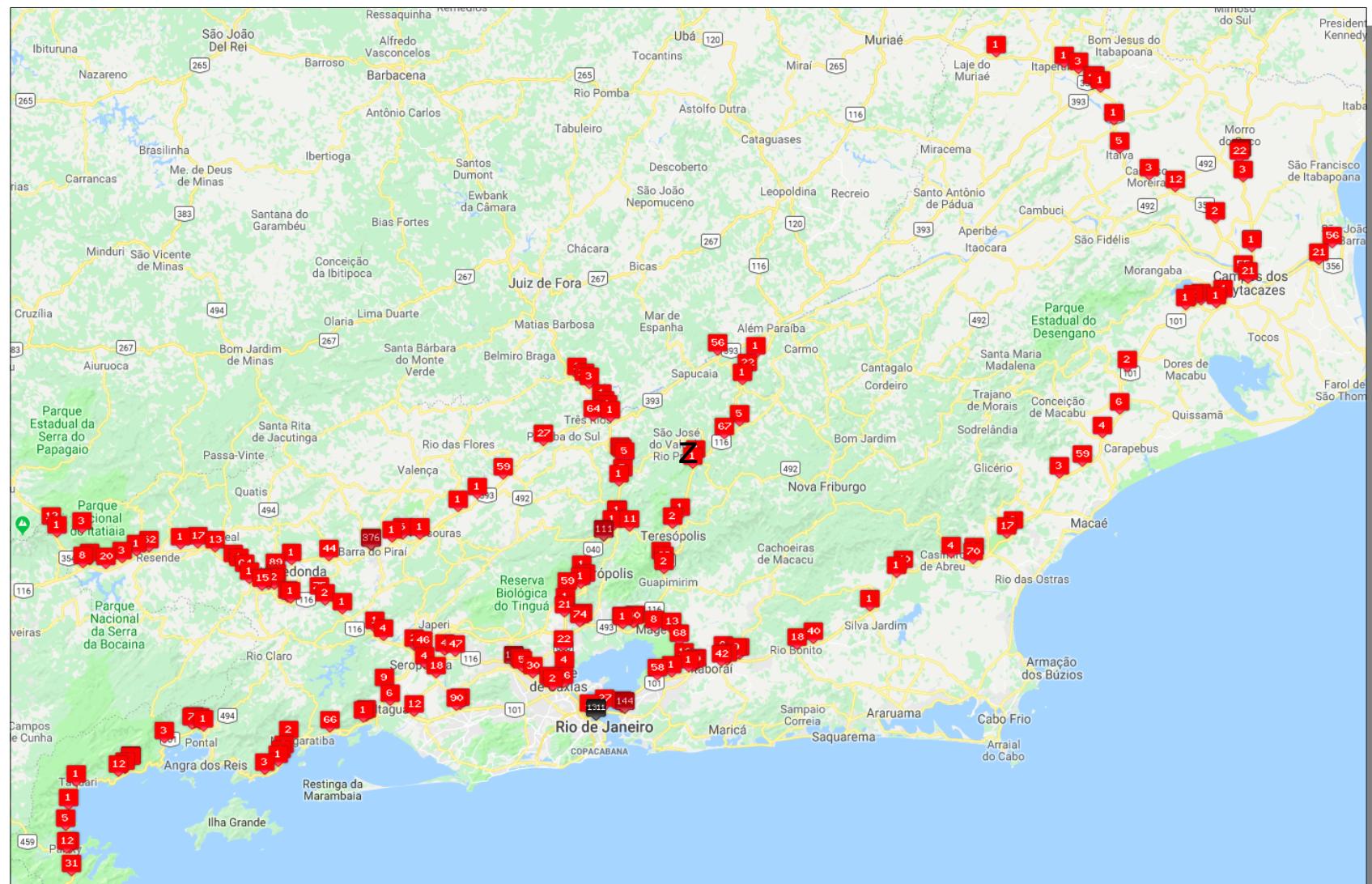
Como identificar informação geográfica em eventos e georreferenciar?



# Mineração de Eventos

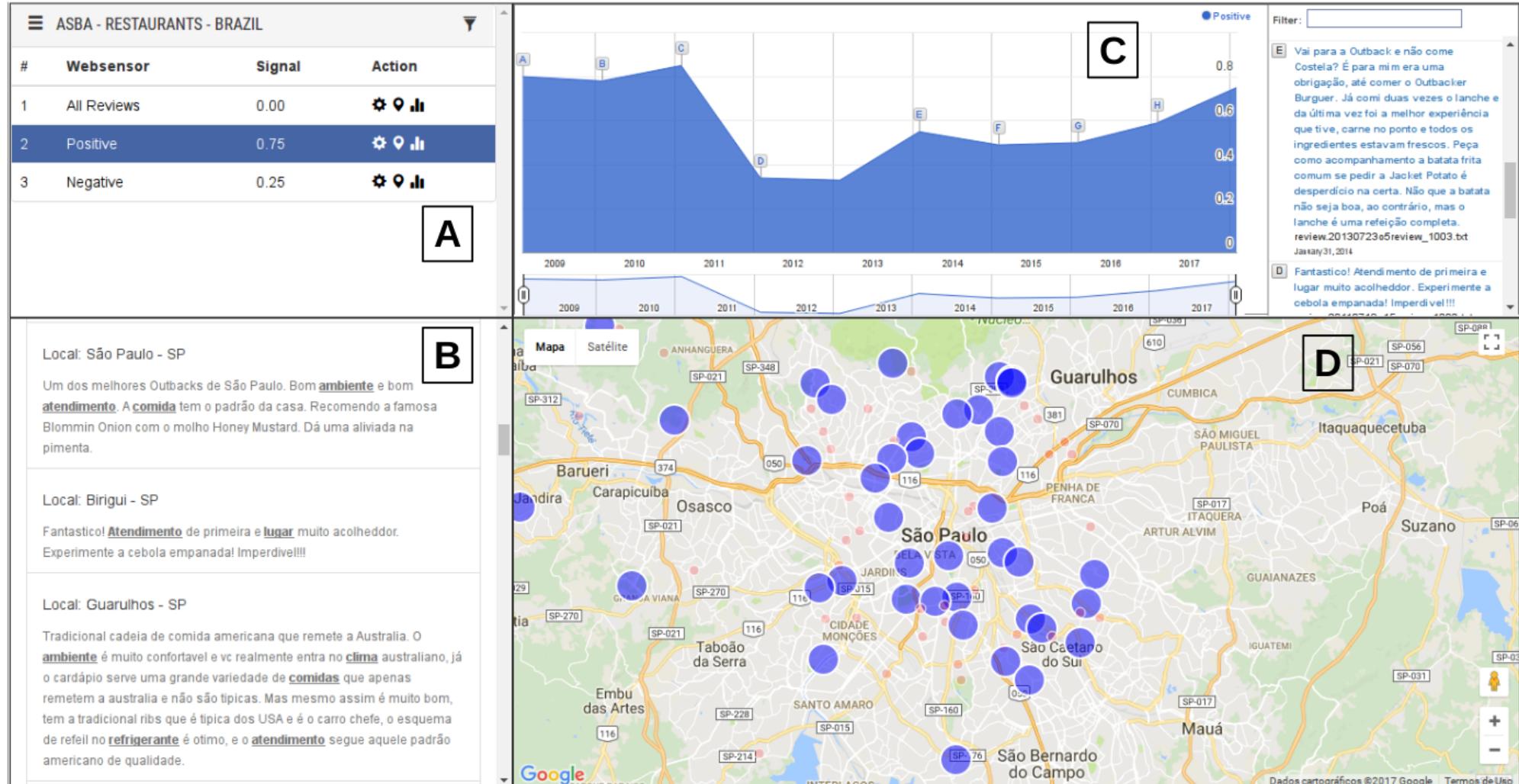
- Pré-processamento - Informação Geográfica
  - A informação geográfica representa um importante conhecimento para a análise de eventos.
  - Muitas aplicações utilizam informação geográfica:
    - Propagação de epidemias e efeitos de desastres naturais.
    - Violência urbana.
    - Acidentes.
    - Protestos.
    - Análise demográfica de consumidores.

# Mineração de Eventos



Acidentes em Rodovias Federais  
Plataforma [Websensors](#)/[Web@Cidadania](#) (2016)

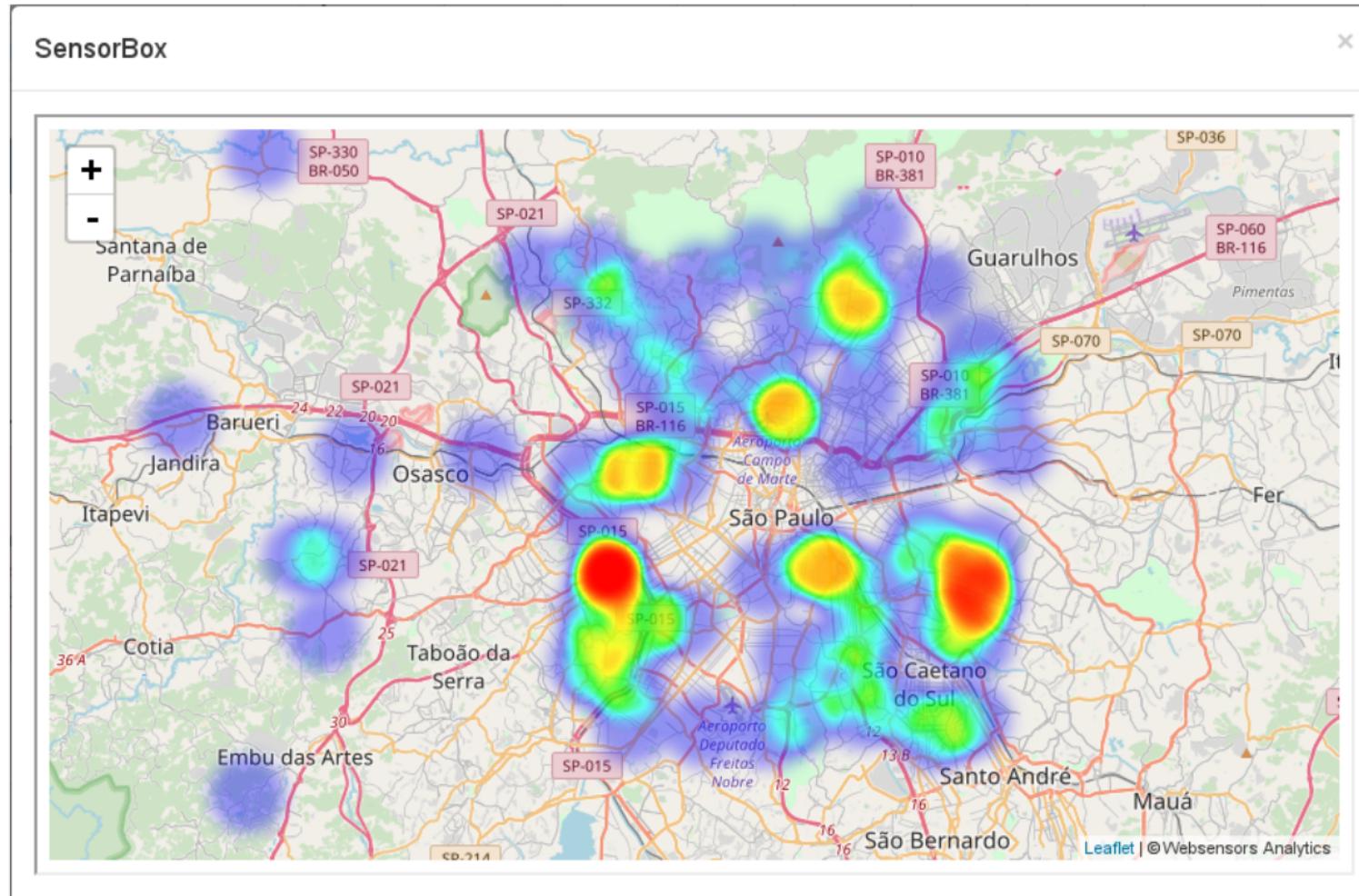
# Mineração de Eventos



Análise de Sentimentos baseadas em Aspectos. Mapeamento dos consumidores.

Marcacini, R. M., Rossi, R. G., Matsuno, I. P., & Rezende, S. O. (2018). Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decision Support Systems*, 114, 70-80.

# Mineração de Eventos



Análise de Sentimentos baseadas em Aspectos. Mapeamento dos consumidores. Mapa de calor com reclamações negativas.

Marcacini, R. M., Rossi, R. G., Matsuno, I. P., & Rezende, S. O. (2018). Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decision Support Systems*, 114, 70-80.

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.

<https://oglobo.globo.com/rio/bairros/biologos-encontram-cavalo-marinho-proximo-as-praias-de-adao-eva-16030979>

- Localidade:
- Relação espacial:

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.

<https://oglobo.globo.com/rio/bairros/biologos-encontram-cavalo-marinho-proximo-as-praias-de-adao-eva-16030979>

- Localidade: praias de Adão e Eva
- Relação espacial: próximo

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.
  - Onde ficam?

<https://oglobo.globo.com/rio/bairros/biologos-encontram-cavalo-marinho-proximo-as-praias-de-adao-eva-16030979>

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.
  - Onde ficam?



Localizadas próximo a Niterói, as Praias de Adão e Eva são praias gêmeas, uma com 250 metros e a outra com 150 metros de extensão, possui águas bastante frias, não muito calmas e de coloração esverdeada, suas areias são brancas e finas.

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.
  - Onde ficam?



Localizadas próximo a Niterói, as Praias de Adão e Eva são praias gêmeas, uma com 250 metros e a outra com 150 metros de extensão, possui águas bastante frias, não muito calmas e de coloração esverdeada, suas areias são brancas e finas.



A Praia de Adão e Eva é uma praia marítima da freguesia de Monte Gordo, concelho de Vila Real de Santo António, no Algarve, Portugal.

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.
  - Onde ficam?



Como desambiguar?

Localizadas próximo a Niterói, as Praias de Adão e Eva são praias gêmeas, uma com 250 metros e a outra com 150 metros de extensão, possui águas bastante frias, não muito calmas e de coloração esverdeada, suas areias são brancas e finas.



A Praia de Adão e Eva é uma praia marítima da freguesia de Monte Gordo, concelho de Vila Real de Santo António, no Algarve, Portugal.

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.
- Desambiguar entidades geográficas:
  - Precisa de informação externa e/ou do contexto.
  - Verificar outras entidades do texto podem ajudar.
  - Verificar informação geográfica da fonte que publicou a notícias/evento.

<https://oglobo.globo.com/rio/bairros/biologos-encontram-cavalo-marinho-proximo-as-praias-de-adao-eva-16030979>

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.

<https://oglobo.globo.com/rio/bairros/biologos-encontram-cavalo-marinho-proximo-as-praias-de-adao-eva-16030979>

- Localidade: praias de Adão e Eva
- Relação espacial: próximo

# Mineração de Eventos

- Extração de Informação Geográfica
  - Identificar informação geográfica a partir de textos apresenta muitos desafios.
  - Por exemplo:

Biólogos encontram cavalo-marinho próximo às praias de Adão e Eva.

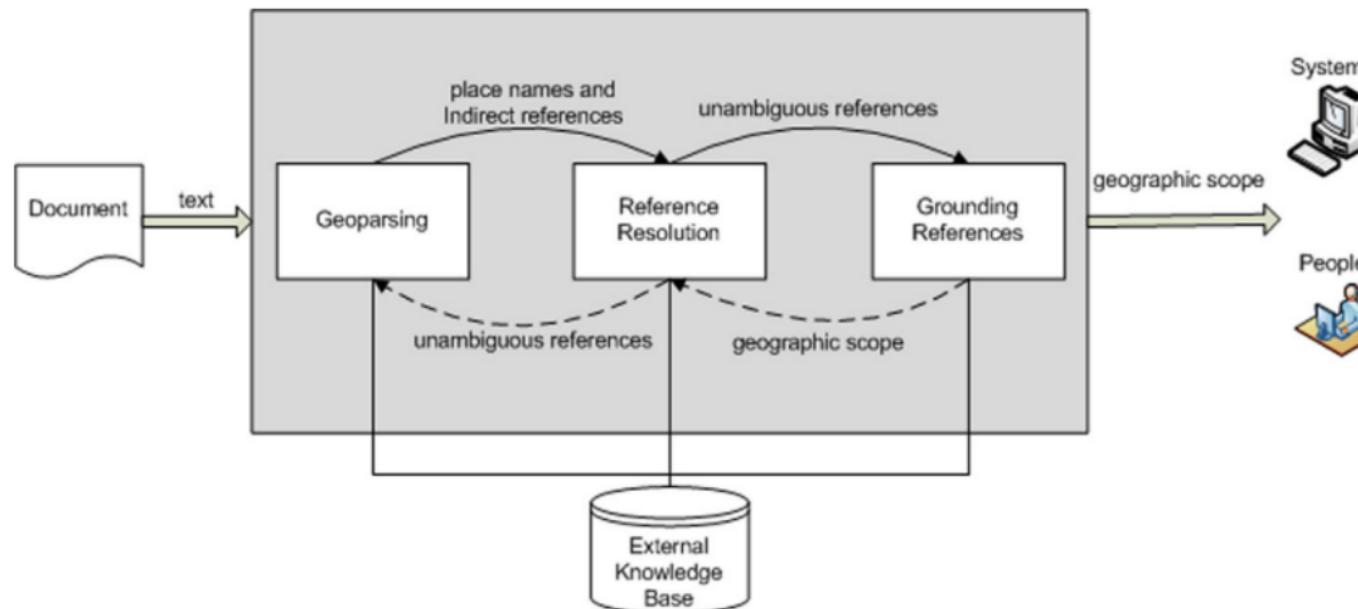
<https://oglobo.globo.com/rio/bairros/biologos-encontram-cavalo-marinho-proximo-as-praias-de-adao-eva-16030979>

- Localidade: praias de Adão e Eva
- Relação espacial: próximo

A palavra “próximo” pode significar metros ou quilômetros!  
Mais ambiguidade...

# Mineração de Eventos

- Extração de Informação Geográfica
  - Muita pesquisa na área, mas há framework geral para esta atividade.



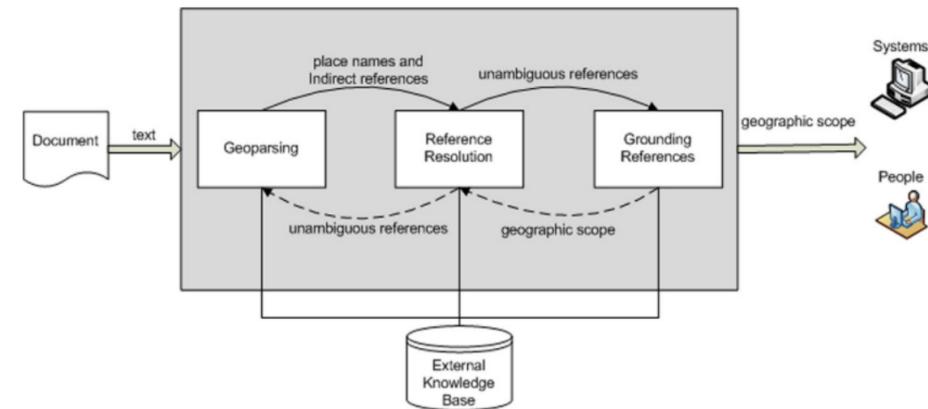
B. R. Monteiro, C. A. Davis Jr, and F. Fonseca, "A survey on the geographic scope of textual documents," *Computers & Geosciences*, vol. 96, pp. 23–34, 2016.

# Mineração de Eventos

## ■ Extração de Informação Geográfica

### ■ Geoparsing:

- Extri palavras do texto que são candidatas a serem entidades geográficas.
- Técnicas *baselines*:
  - Palavras com primeiro caracter capitalizado
  - Identificação de Nomes Próprios e Entidades Nomeadas (Processamento de Ling. Natural)
  - Base de Conhecimento Geográfico



# Mineração de Eventos

## ■ Extração de Informação Geográfica

- Exemplo de PLN para extração de entidades nomeadas [<http://corenlp.run/>]:

— Text to annotate —

Halogen light bulbs to be banned in Europe

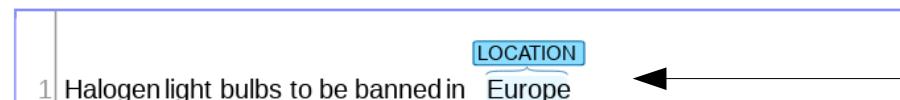
— Annotations —

parts-of-speech  named entities  dependency parse

Part-of-Speech:

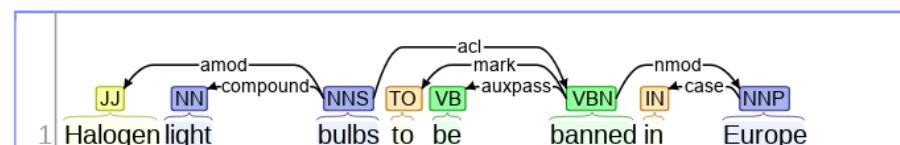


Named Entity Recognition:



Palavra candidata para o Geoparsing.

Basic Dependencies:



# Mineração de Eventos

- Extração de Informação Geográfica
  - Exemplo de uso de uma Base de Conhecimento Geográfico [<http://geonames.org/>]:



The screenshot shows a search interface for 'São Carlos'. The search bar contains 'São Carlos' and the dropdown country selector is set to 'Brazil'. Below the search bar are two buttons: 'search' and 'advanced search'. To the right of the search bar, it says '121 records found for ""São Carlos""'. The main area is a table with columns: 'Name', 'Country', 'Feature class', 'Latitude', and 'Longitude'. The table lists 121 entries, with the first few rows shown below:

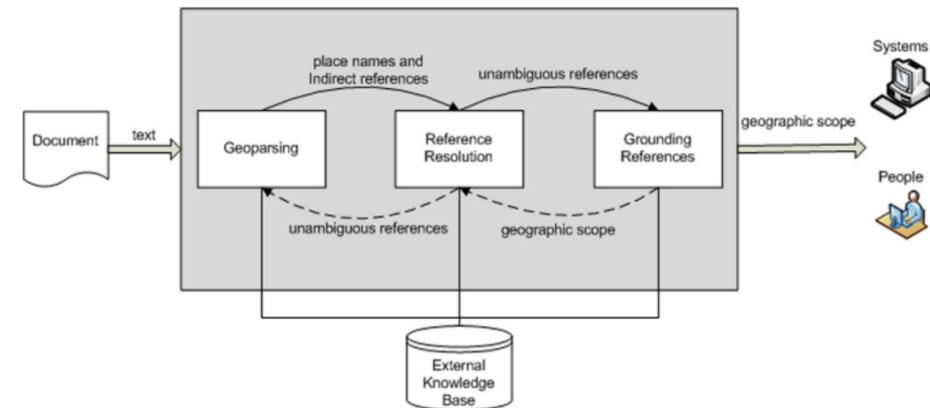
Name	Country	Feature class	Latitude	Longitude
1 ⓘ São Carlos 🌎 QSC, San Karlos, San-Karlos, San-Karlus, Sao Carlos, Sao Carlos do Pinhal, Sao Carlos, São Carlos, São Car...	Brazil, São Paulo São Carlos	populated place population 205,035	S 22° 1' 3"	W 47° 53' 27"
2 ⓘ São Carlos Airport 🌎 Aeroporto Estadual Mario Pereira Lopes, Aeroporto Estadual Mário Pereira Lopes, Aeroporto de São Carlo...	Brazil, São Paulo São Carlos	airport	S 21° 52' 35"	W 47° 54' 11"
3 ⓘ São Carlos	Brazil, São Paulo São Carlos	second-order administrative division population 221,936	S 21° 54' 23"	W 47° 52' 29"
4 ⓘ São Carlos 🌎	Brazil, Santa Catarina São Carlos	second-order administrative division population 10,284	S 27° 2' 40"	W 53° 2' 7"
5 ⓘ São Carlos do Ivaí	Brazil, Paraná São Carlos do Ivaí	second-order administrative division population 6,352	S 23° 21' 17"	W 52° 31' 14"
6 ⓘ São Carlos 🌎	Brazil, Santa Catarina São Carlos	populated place	S 27° 4' 39"	W 53° 0' 14"

# Mineração de Eventos

## ■ Extração de Informação Geográfica

### ■ Geoparsing:

- Extrai palavras do texto que são candidatas a serem entidades geográficas.



### ***Geoparsing (Baseline):***

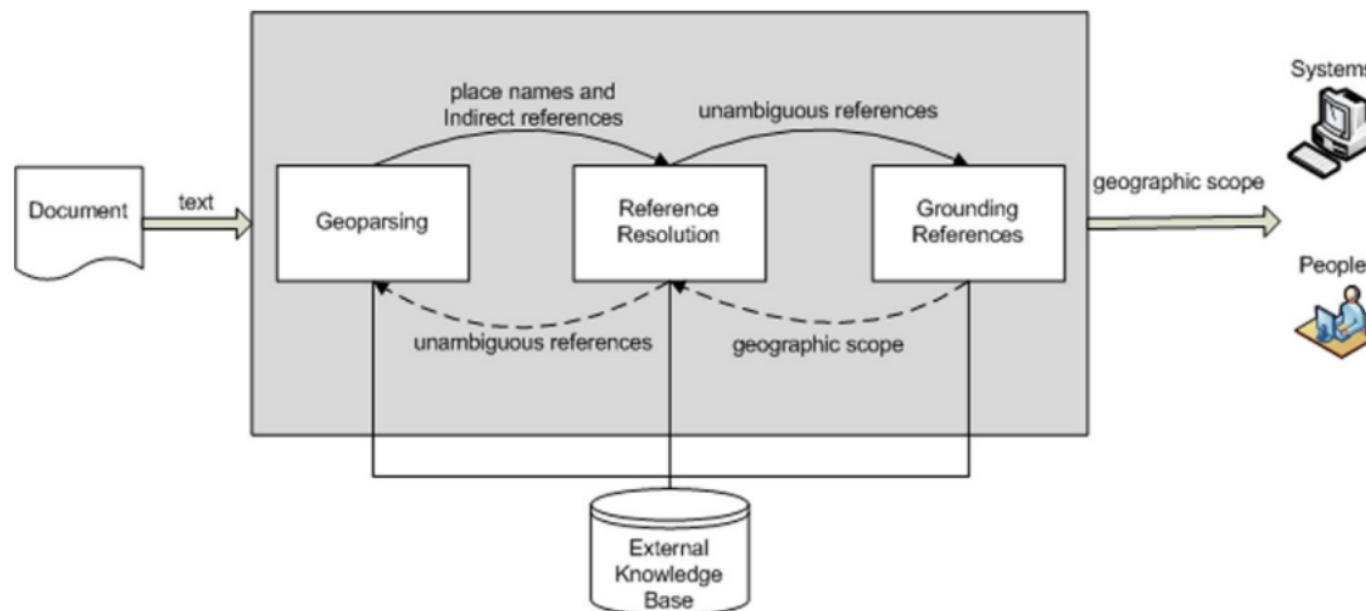
INPUT: texto T e base de conhecimento KDB

- 1) Iniciar conjunto de candidatas a entidades geográficas G
- 2) Extrair tokens de T com a primeira letra capitalizada e inserir em G
- 3) Extrair entidades nomeadas de T e inserir em G
- 4) Remover candidatas em G que não pertencem ao KDB
- 5) Retornar G.

# Mineração de Eventos

## ■ Extração de Informação Geográfica

- Após o Geoparsing, iniciamos a etapa de Reference Resolution.



B. R. Monteiro, C. A. Davis Jr, and F. Fonseca, “A survey on the geographic scope of textual documents,” *Computers & Geosciences*, vol. 96, pp. 23–34, 2016.

# Mineração de Eventos

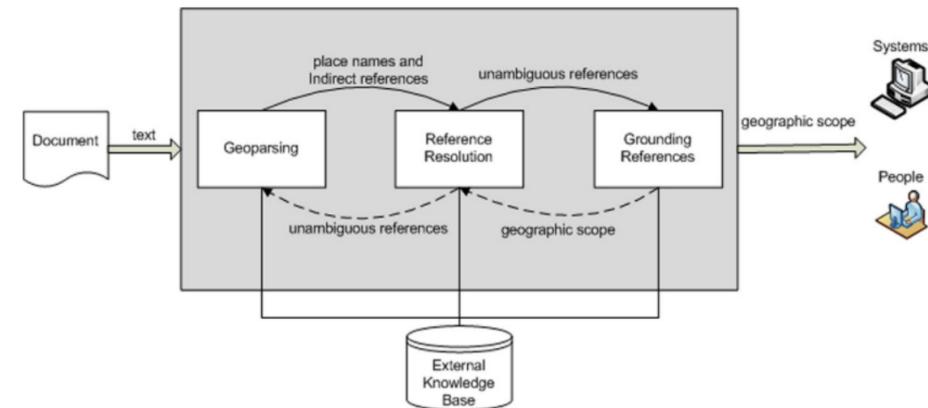
## ■ Extração de Informação Geográfica

### ■ *Reference Resolution:*

- Desambiguar entidades geográficas com base em informação de contexto

- Exemplo:

A entidade “São Paulo” pode estar relacionada ao município, estado da federação, ou até mesmo ao time de futebol. A desambiguação depende do uso de informação de contexto do próprio documento (quando disponível), de outros textos associados, da fonte que publicou o texto, ou outro tipo de conhecimento externo



# Mineração de Eventos

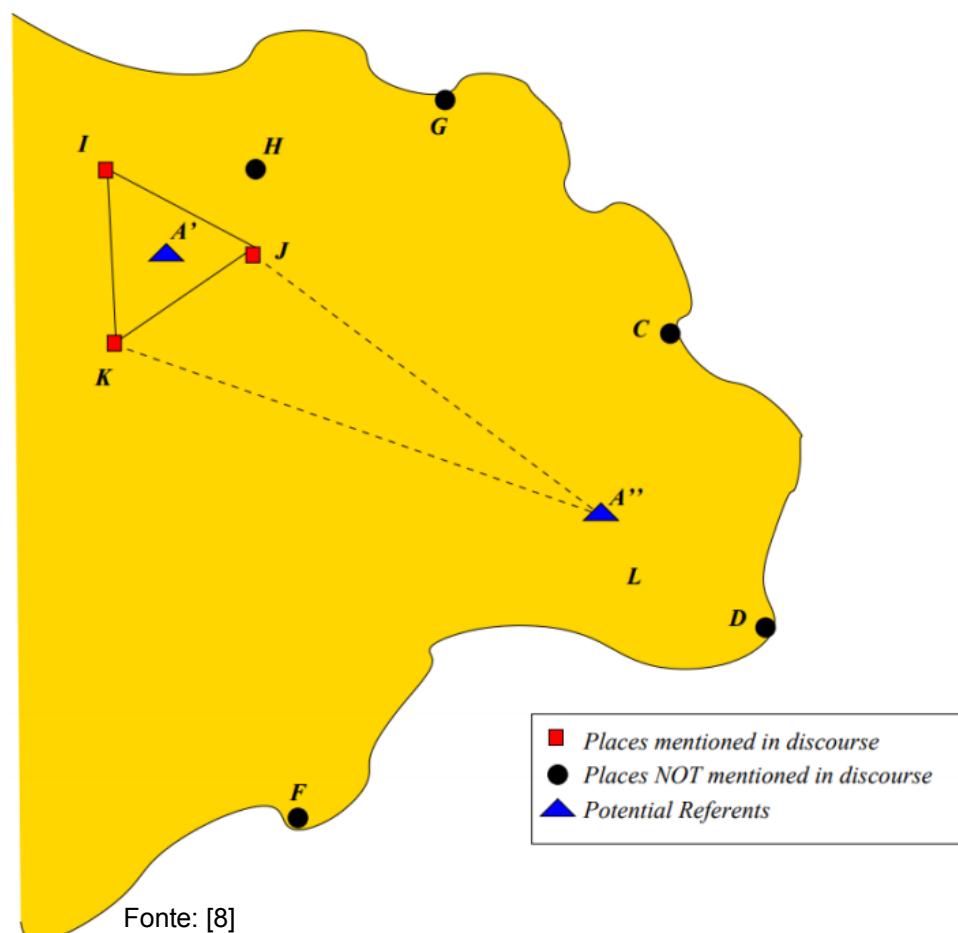
## ■ Extração de Informação Geográfica

### ■ *Reference Resolution:*

- Técnica da Minimalidade Espacial (Leidner, 2008; Leidiner, 2016)
- Considere que foram obtidas quatro entidades geográficas I, J, K e A, em que a entidade A possui duas referências ambíguas  $A = \{A', A''\}$ .
- As entidades não ambíguas são utilizadas para selecionar  $A'$  ou  $A''$
- É necessário calcular a proximidade geográfica entre cada possível entidade geográfica (exige uma Base de Conhecimento Geográfico; e.g. GeoNames)

# Mineração de Eventos

- Extração de Informação Geográfica
  - *Reference Resolution (Minimalidade Espacial)*



A' e A'' são ambíguas.  
Ex. A'=São Carlos-SP  
A''=São Carlos-SC

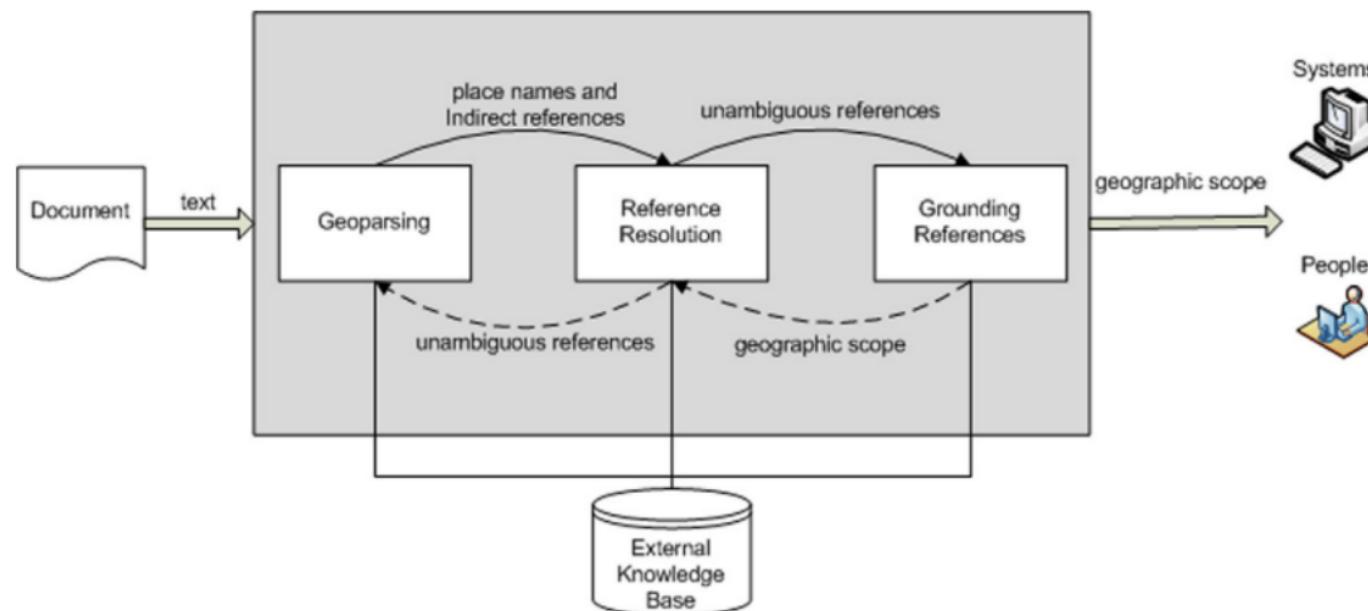
As outras entidades não ambíguas do texto (I, J e K) são utilizadas para desambiguar.

OBS: pode falhar para textos longos mas tem boa acurácia para análise de eventos.

# Mineração de Eventos

## ■ Extração de Informação Geográfica

- Após a *Reference Resolution*, iniciamos a etapa de *Grouding References (Geocoder)*.



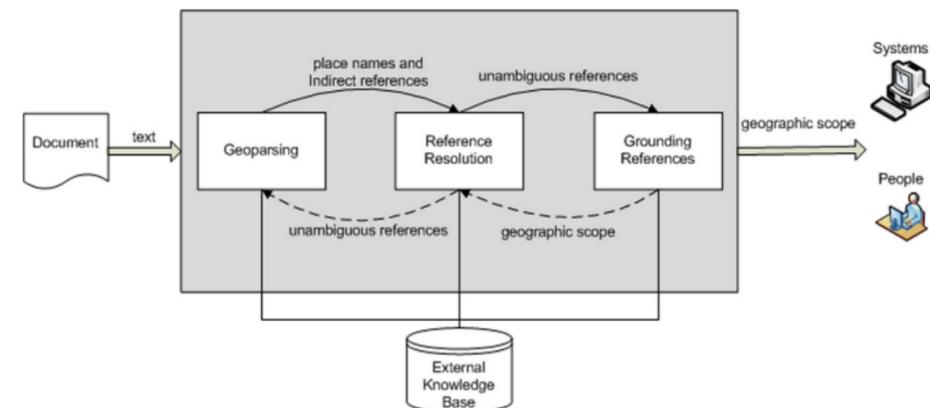
B. R. Monteiro, C. A. Davis Jr, and F. Fonseca, "A survey on the geographic scope of textual documents," *Computers & Geosciences*, vol. 96, pp. 23–34, 2016.

# Mineração de Eventos

## ■ Extração de Informação Geográfica

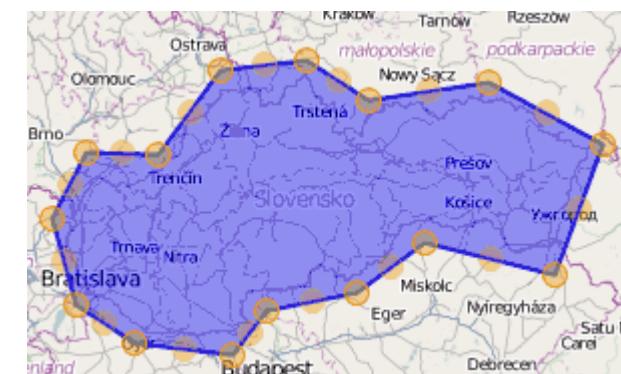
### ■ *Geocoding:*

- Associar uma coordenada (lat,lon) para cada entidade e definir a informação geográfica do documento.



### ■ Estratégias (baselines):

- 1) Utilizar a coordenada (lat,lon) média/moda das várias entidades
- 2) Construir o polígono da região (se três ou mais pontos)
- 3) Coordenada média respeitando o polígono da região.



# Mineração de Eventos

## ■ Prática #3

### ■ Rede de Eventos com Proximidade de Conteúdo+GEO

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar e ocorreram em regiões próximas.*

### ■ Como calcular a proximidade geográfica entre dois eventos?

# Mineração de Eventos

## ■ Prática #3

### ■ Rede de Eventos com Proximidade de Conteúdo+GEO

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar e ocorreram em regiões próximas.*

### ■ Como calcular a proximidade geográfica entre dois eventos?

Uma vez extraídas as  $q$  informações geográficas  $R = \{(lat, lon)_1, \dots, (lat, lon)_q\}$  de um evento, então é possível calcular a distância geográfica entre dois eventos  $i$  e  $j$  com base na Equação 2.5 (OVERELL, 2009), em que  $dist_{gps}$  é alguma função de distância no sistema de coordenadas geográficas.

$$geoDist(R_i, R_j) = \sum_{(lat, lon)_i \in R_i} \sum_{(lat, lon)_j \in R_j} \frac{dist_{gps}((lat, lon)_i, (lat, lon)_j)}{|R_i| \times |R_j|}$$

# Mineração de Eventos

## ■ Prática #3

### ■ Rede de Eventos com Proximidade de Conteúdo+GEO

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar e ocorreram em regiões próximas.*

### ■ Como calcular a proximidade geográfica entre dois eventos?



A distância entre coordenadas (lat,lon) precisa considerar a curvatura da terra!

Haversine Distance ( $r$  = raio da terra):

$$1 + 2r \arcsin \left( \sqrt{\sin^2 \left( \frac{lat_{e_i} - lat_{e_j}}{2} \right) + \cos(lat_{e_i}) \cos(lat_{e_j}) \sin^2 \left( \frac{lon_{e_i} - lon_{e_j}}{2} \right)} \right)$$

# Mineração de Eventos

## ■ Prática #3

### ■ Rede de Eventos com Proximidade de Conteúdo+GEO

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar e ocorreram em regiões próximas.*

### ■ Como calcular a proximidade geográfica entre dois eventos?

$$StJoin(i, j) = \frac{sim_{text}(e_i, e_j)}{1 + geoDist(R_i, R_j)} \quad (2.6)$$

Note que na  $StJoin(i, j)$ , quanto maior o valor da distância  $geoDist(R_i, R_j)$  maior será a penalização e, consequentemente, menor a relação entre os eventos  $i$  e  $j$ . Se os eventos ocorreram exatamente no mesmo local, com  $geoDist(R_i, R_j) = 0$ , então o  $StJoin(i, j)$  será a própria similaridade textual entre os eventos.

# Mineração de Eventos

## ■ Prática #3

### ■ Rede de Eventos com Proximidade de Conteúdo+GEO

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar e ocorreram em regiões próximas.*

### ■ Como calcular a proximidade geográfica entre dois eventos?

$$StJoin(i, j) = \frac{sim_{text}(e_i, e_j)}{1 + geoDist(R_i, R_j)} \quad (2.6)$$

Cosseno / Espaço vetorial

Haversine

Note que na  $StJoin(i, j)$ , quanto maior o valor da distância  $geoDist(R_i, R_j)$  maior será a penalização e, consequentemente, menor a relação entre os eventos  $i$  e  $j$ . Se os eventos ocorreram exatamente no mesmo local, com  $geoDist(R_i, R_j) = 0$ , então o  $StJoin(i, j)$  será a própria similaridade textual entre os eventos.

# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

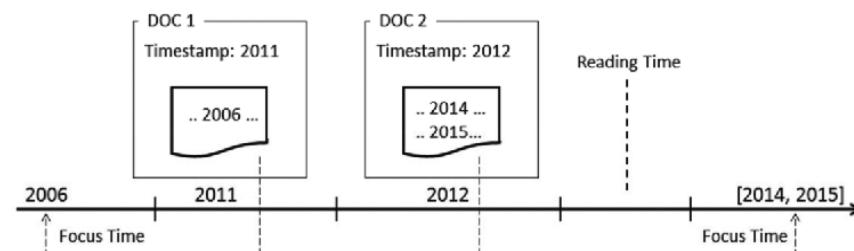
Informação  
Geográfica

Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados se ocorreram no mesmo período de tempo.

Como extrair informação temporal dos eventos?



# Mineração de Eventos

- Análise de eventos é dependente da informação temporal
  - Ordenar as informações no tempo
  - Identificar sazonalidades
  - Identificar tendências
  - Construir sensores!
- Desafio:
  - Extrair informação temporal de texto puro para um formato no modelo de calendário (gregoriano) YYYYMMDD

# Mineração de Eventos

- Pré-processamento - Informação Temporal
  - Informação Temporal na Forma Explícita
    - É a forma mais simples e direta.
    - Ex.: “A tempestade ocorreu em 3 de Outubro de 2017”.
    - Expressões regulares resolvem o problema.

# Mineração de Eventos

- Pré-processamento - Informação Temporal
  - Informação Temporal na Forma Explícita
    - É a forma mais simples e direta.
    - Ex.: “*A tempestade ocorreu em 3 de Outubro de 2017.*”.
    - Expressões regulares resolvem o problema.
  - Informação Temporal na Forma Implícita
    - Necessidade de algum recurso externo.
    - Ex.: “*Muito congestionamento no dia de natal*”

# Mineração de Eventos

- Pré-processamento - Informação Temporal
  - Informação Temporal na Forma Explícita
    - É a forma mais simples e direta.
    - Ex.: “*A tempestade ocorreu em 3 de Outubro de 2017.*
    - Expressões regulares resolvem o problema.
  - Informação Temporal na Forma Implícita
    - Necessidade de algum recurso externo.
    - Ex.: “*Muito congestionamento no dia de natal*”
  - Informação Temporal na Forma Relativa
    - Necessidade de padronizar as datas.
    - Ex.: “*O sindicato irá protestar na próxima semana*”

# Mineração de Eventos

Enter text in Portuguese:

Os computadores serão comprados na próxima semana.

Click for special characters:

á â ã à ç é ê í ó ô õ ú

Caps Lock

Specify the time when this text was created:

- Use the current time (2018-08-31, 12:20:12)

## Output

Creation time: 2018-08-31, 12h 19m 55s

Text	
Os computadores serão <b>comprados</b> na <b>próxima semana</b> .	
Date	
Year:	2018
Week:	36
comprados OVERLAPS próxima semana próxima semana IS AFTER 2018-08-31, 12:19:55 (the document creation time)	

<http://lxcenter.di.fc.ul.pt/services/en/LXServicesTimeAnalyzer.html>

# Mineração de Eventos

Enter text in Portuguese:

Os computadores serão comprados na próxima semana.

Click for special characters:

á â ã à ç é ê í ó ô õ ú

Caps Lock

Specify the time when this text was created:

- Use the current time (2018-08-31, 12:20:12)

## Output

Creation time: 2018-08-31, 12h 19m 55s

Text

Os computadores serão comprados na próxima semana.

Conjunto de regras de conversão  
de expressão temporal para datas.

Date

Year:

2018

Week:

36

comprados OVERLAPS próxima semana

próxima semana IS AFTER 2018-08-31, 12:19:55 (the document creation time)

<http://lxcenter.di.fc.ul.pt/services/en/LXServicesTimeAnalyzer.html>

# Proximidade Temporal

## ■ Prática #4

### ■ Proximidade Temporal.

*Dois eventos A e B podem estar relacionados se A e B apresentam conteúdo similar e ocorreram em datas próximas.*

### ■ Como incorporar a informação temporal na similaridade entre eventos?

# Proximidade Temporal

Nesta equação, dois eventos  $e_i$  e  $e_j$  possuem similaridade textual  $sim_{text}(e_i, e_j)$  (e.g. similaridade cosseno). As informações de datas entre os eventos são representadas por  $d_i$  e  $d_j$ , respectivamente. O decaimento exponencial (e) pode ser visto como uma gaussiana, na qual o parâmetro  $\mu$  indica a abertura da gaussiana e, consequentemente, a força desse decaimento. Quanto maior o valor de  $\mu$ , menor a penalização da similaridade ao considerar datas muito distantes.

$$TAS(i, j) = sim_{text}(e_i, e_j) e^{\frac{-|d_i - d_j|^2}{2\mu^2}}$$

Note que se dois eventos tiverem as mesmas informações de datas, então a similaridade entre eles será a própria similaridade textual.

# Proximidade Temporal

Cosseno / Espaço vetorial

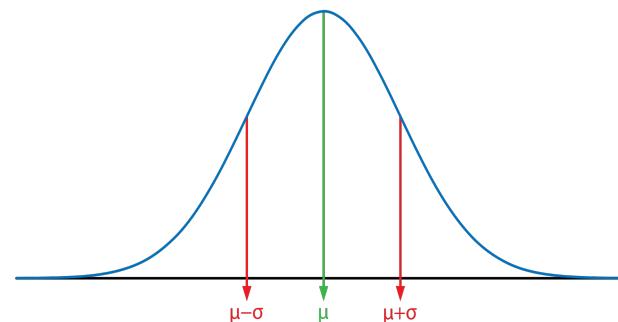
$$TAS(i, j) = sim_{text}(e_i, e_j) e^{\frac{-|d_i - d_j|^2}{2\mu^2}}$$

Note que se dois eventos tiverem as mesmas informações de datas, então a similaridade entre eles será a própria similaridade textual.

# Proximidade Temporal

Cosseno / Espaço vetorial

$$TAS(i, j) = sim_{text}(e_i, e_j) e^{-\frac{|d_i - d_j|^2}{2\mu^2}}$$



**Decaimento temporal**  
(penaliza a similaridade entre eventos que ocorreram em datas distantes)

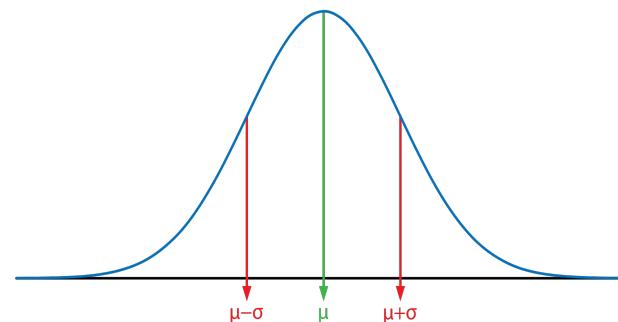
Note que se dois eventos tiverem as mesmas informações de datas, então a similaridade entre eles será a própria similaridade textual.

# Proximidade Temporal

## Prática #4

Cosseno / Espaço vetorial

$$TAS(i, j) = sim_{text}(e_i, e_j) e^{-\frac{|d_i - d_j|^2}{2\mu^2}}$$



**Decaimento temporal**  
(penaliza a similaridade entre eventos que ocorreram em datas distantes)

Note que se dois eventos tiverem as mesmas informações de datas, então a similaridade entre eles será a própria similaridade textual.

# Mineração de Eventos

## ■ Pré-processamento

Informação  
Textual

Informação  
Geográfica

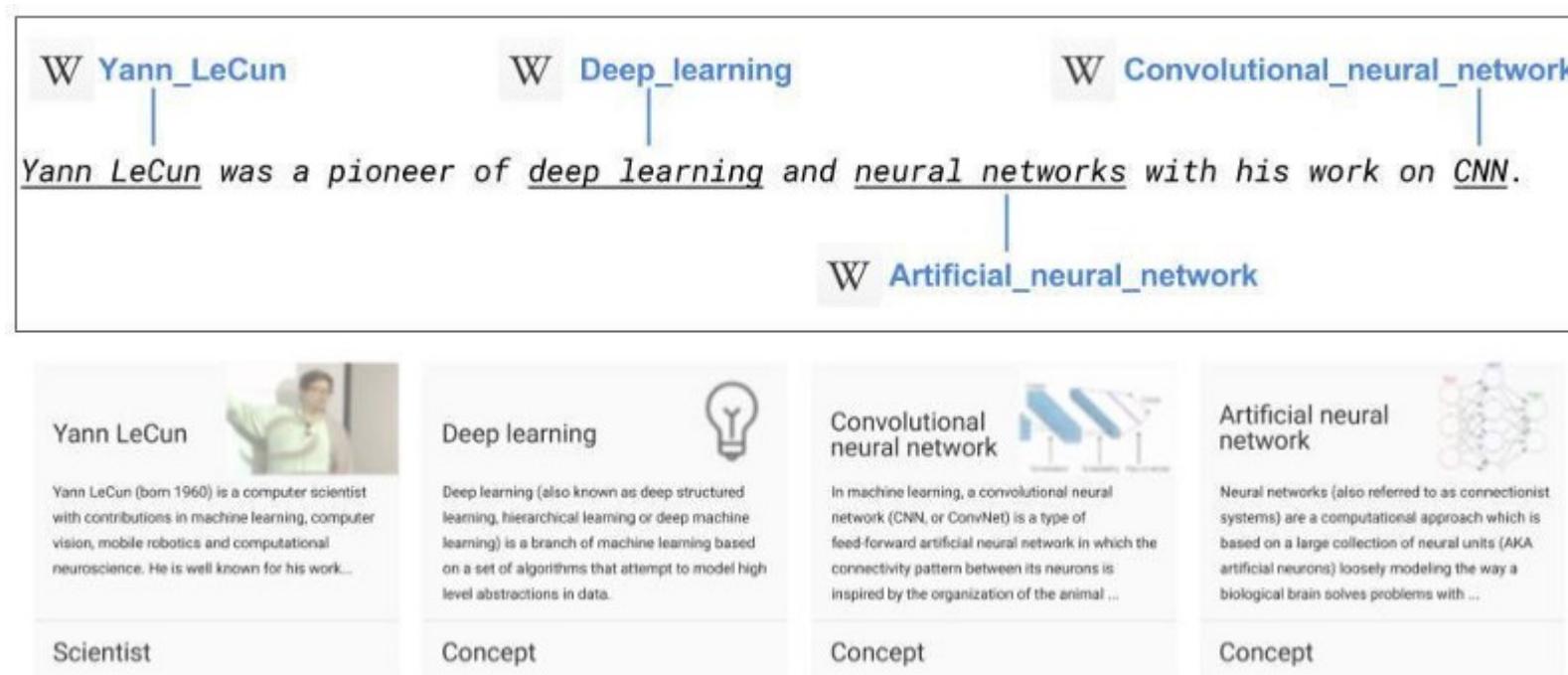
Informação  
Temporal

Informação de  
Domínio

- Eventos podem estar relacionados conforme informação de domínio.
  - Nomes de pessoas e organizações
  - Entidades do domínio

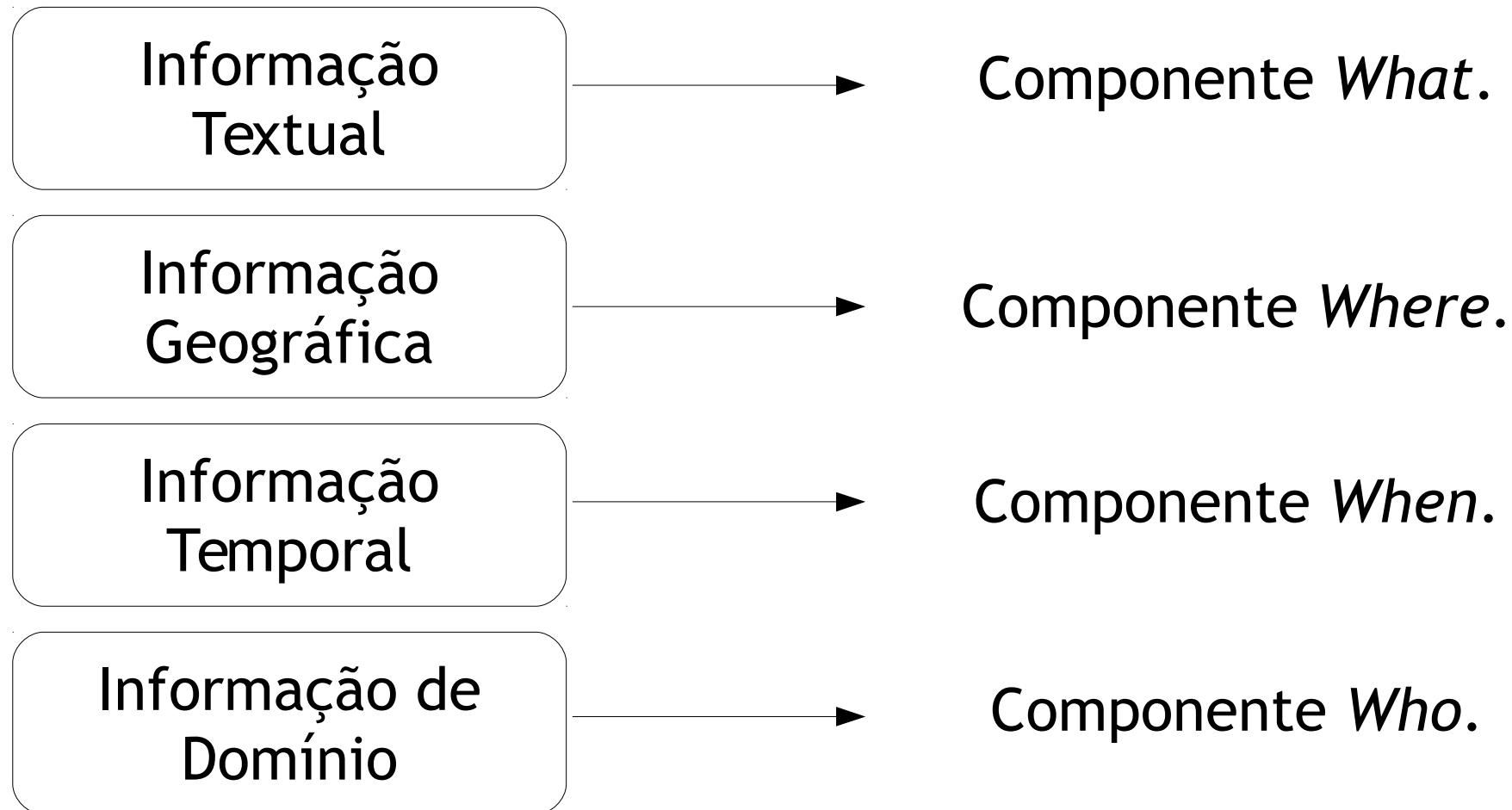
# Mineração de Eventos

- Pré-processamento - Informação de Domínio
  - Usa uma base de conhecimento externo
  - Identifica e cria link com entidades



# Mineração de Eventos

## ■ Pré-processamento



# Mineração de Eventos

## ■ Pré-processamento

- Representação estruturada = Rede de Eventos
- Dois eventos  $e_i$  e  $e_j$  estão relacionados se sua distância for menor do que um determinado limiar.

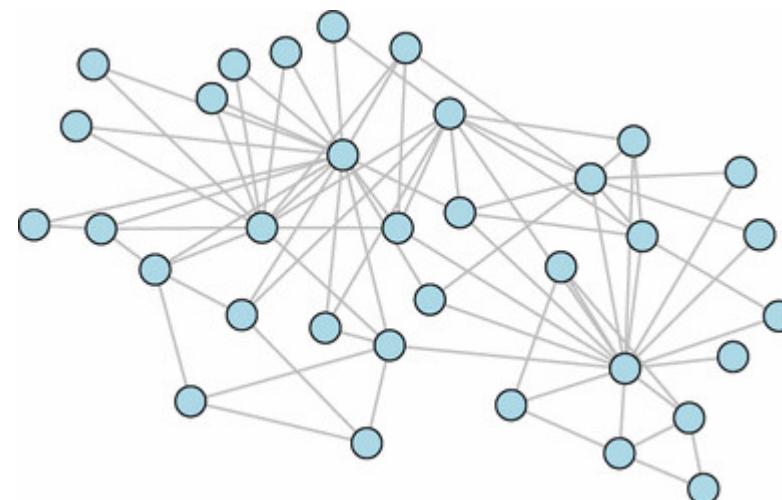
$$d(\mathbf{e}_i, \mathbf{e}_j) = \alpha d_{what}(\mathbf{e}_i, \mathbf{e}_j) + \beta d_{where}(\mathbf{e}_i, \mathbf{e}_j) + \gamma d_{when}(\mathbf{e}_i, \mathbf{e}_j) + \omega d_{who}(\mathbf{e}_i, \mathbf{e}_j)$$

# Mineração de Eventos

## ■ Pré-processamento

- Representação estruturada = Rede de Eventos
- Dois eventos  $e_i$  e  $e_j$  estão relacionados se sua distância for menor do que um determinado limiar.

$$d(\mathbf{e}_i, \mathbf{e}_j) = \alpha d_{what}(\mathbf{e}_i, \mathbf{e}_j) + \beta d_{where}(\mathbf{e}_i, \mathbf{e}_j) + \gamma d_{when}(\mathbf{e}_i, \mathbf{e}_j) + \omega d_{who}(\mathbf{e}_i, \mathbf{e}_j)$$



# Mineração de Eventos

## ■ Extração de Padrões

- Vamos ver exemplos de dois algoritmos:

- Label Propagation

Permite inserir informação de domínio (rótulos) para identificar eventos de interesse.

# Mineração de Eventos

## ■ Extração de Padrões

- Vamos ver exemplos de dois algoritmos:

- Label Propagation

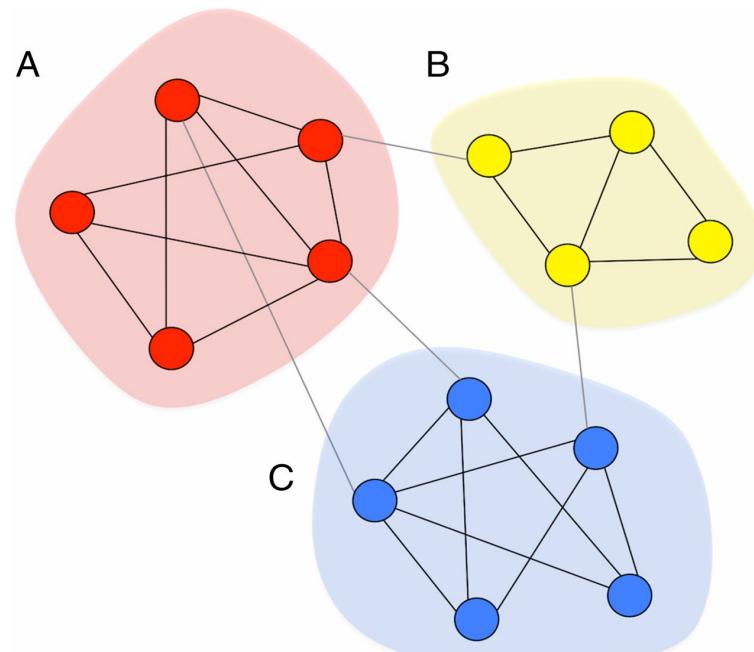
Permite inserir informação de domínio (rótulos) para identificar eventos de interesse.

- DeepWalk / Node2Vec

Permite aprender uma representação a partir da rede de eventos. A representação pode ser empregada em outros métodos de aprendizado de máquina.

# Mineração de Eventos

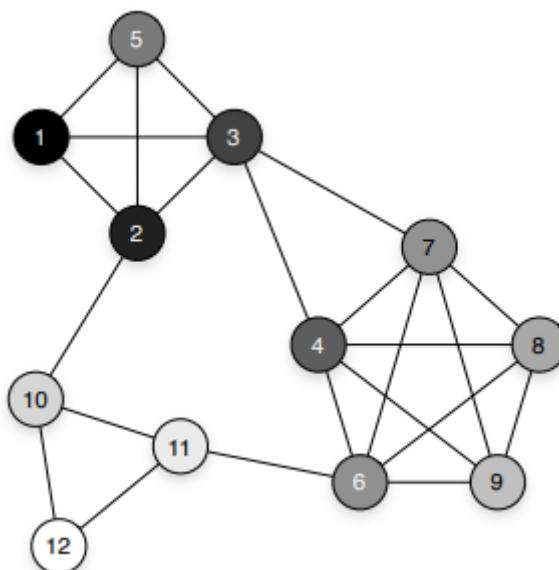
- Extração de Padrões - Label Propagation
  - Como definir eventos de interesse/relacionados?



# Mineração de Eventos

## ■ Extração de Padrões - Label Propagation

Cenário para aprendizado não supervisionado.



Fonte: [10]

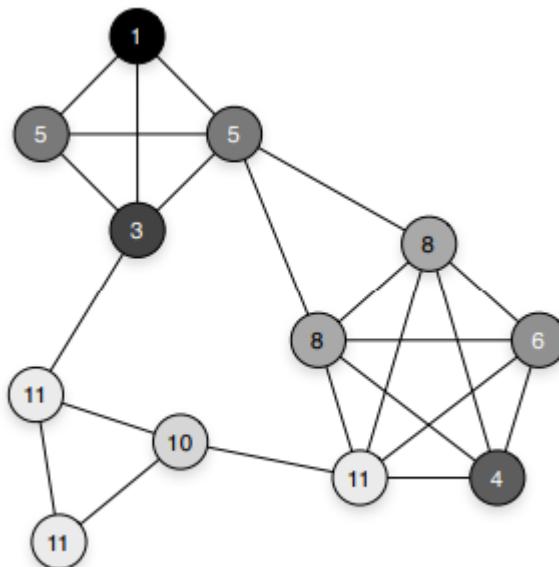
Inicialização

1. Inicializar cada nó com rótulos únicos (sem repetição).
2. Selecionar um nó A da rede (aleatoriamente)
3. [Propagação] Definir o rótulo de A com base nos rótulos dos vizinhos (e.g. maioria)
  - 3.1. Em caso de empate, selecionar rótulo aleatoriamente
  - 3.2. Considerar peso das arestas para alterar a estratégia de propagação
4. Repetir 2 e 3 até convergência (estabilidade dos rótulos)

# Mineração de Eventos

## ■ Extração de Padrões - Label Propagation

Cenário para aprendizado não supervisionado.



Fonte: [10]

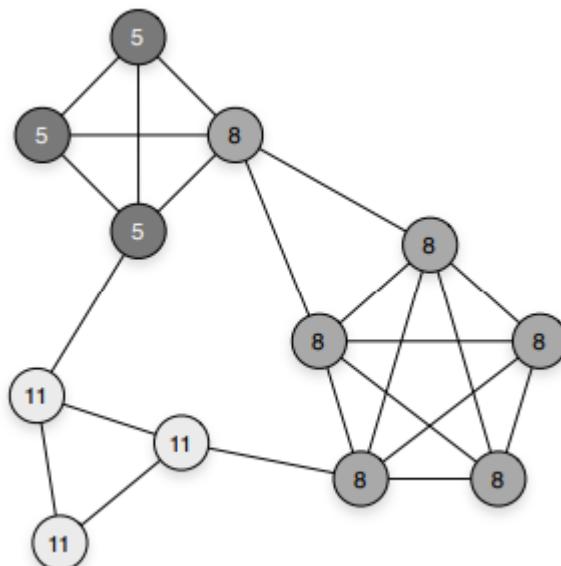
Iteração #1

1. Inicializar cada nó com rótulos únicos (sem repetição).
2. Selecionar um nó A da rede (aleatoriamente)
3. [Propagação] Definir o rótulo de A com base nos rótulos dos vizinhos (e.g. maioria)
  - 3.1. Em caso de empate, selecionar rótulo aleatoriamente
  - 3.2. Considerar peso das arestas para alterar a estratégia de propagação
4. Repetir 2 e 3 até convergência (estabilidade dos rótulos)

# Mineração de Eventos

## ■ Extração de Padrões - Label Propagation

Cenário para aprendizado não supervisionado.



Fonte: [10]

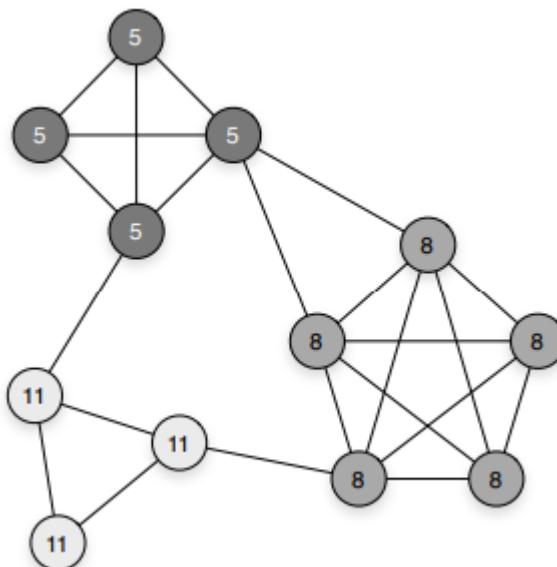
Iteração #2

1. Inicializar cada nó com rótulos únicos (sem repetição).
2. Selecionar um nó A da rede (aleatoriamente)
3. [Propagação] Definir o rótulo de A com base nos rótulos dos vizinhos (e.g. maioria)
  - 3.1. Em caso de empate, selecionar rótulo aleatoriamente
  - 3.2. Considerar peso das arestas para alterar a estratégia de propagação
4. Repetir 2 e 3 até convergência (estabilidade dos rótulos)

# Mineração de Eventos

## ■ Extração de Padrões - Label Propagation

Cenário para aprendizado não supervisionado.



Fonte: [10]

Iteração #3  
(Convergência!)

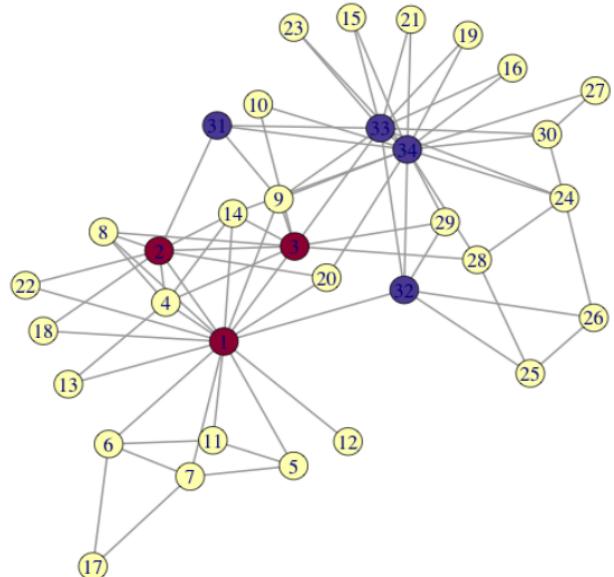
1. Inicializar cada nó com rótulos únicos (sem repetição).
2. Selecionar um nó A da rede (aleatoriamente)
3. [Propagação] Definir o rótulo de A com base nos rótulos dos vizinhos (e.g. maioria)
  - 3.1. Em caso de empate, selecionar rótulo aleatoriamente
  - 3.2. Considerar peso das arestas para alterar a estratégia de propagação
4. Repetir 2 e 3 até convergência (estabilidade dos rótulos)

# Proximidade de Rótulos

## ■ Propagação de Rótulos

Cenário para aprendizado semissupervisionado.

A estratégia mais simples é que usuários definam o rótulo de alguns nós. Esses nós (previamente rotulados) não devem ter os rótulos alterados durante a propagação.

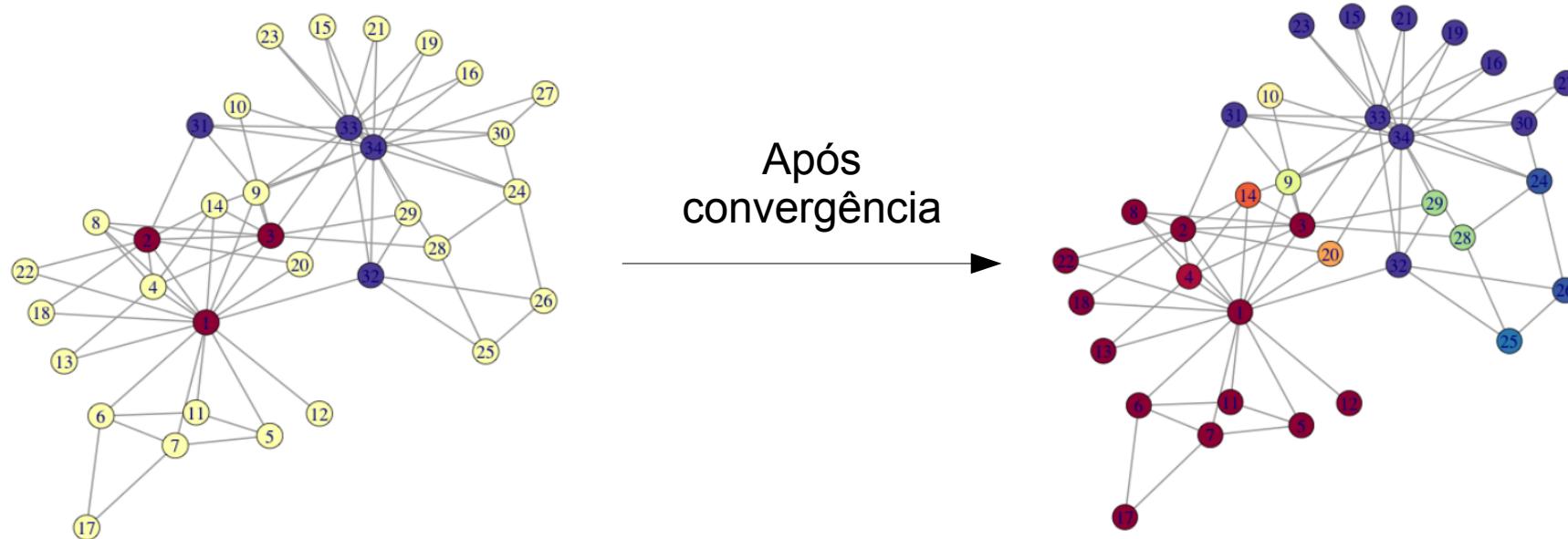


# Proximidade de Rótulos

## ■ Propagação de Rótulos

Cenário para aprendizado semissupervisionado.

A estratégia mais simples é que usuários definam o rótulo de alguns nós. Esses nós (previamente rotulados) não devem ter os rótulos alterados durante a propagação.



# Proximidade de Rótulos

## ■ Análise da Complexidade

- Complexidade:  $\mathcal{O}(cn+cm)$

c = #iterações; n = #nós; m = #arestas

- Caso médio:  $\mathcal{O}(m)$

- Poucas iterações para convergência
- Para grandes redes, número de arestas  $m \gg n$

# Proximidade de Rótulos

## ■ Análise da Complexidade

- Complexidade:  $\mathcal{O}(cn+cm)$

c = #iterações; n = #nós; m = #arestas

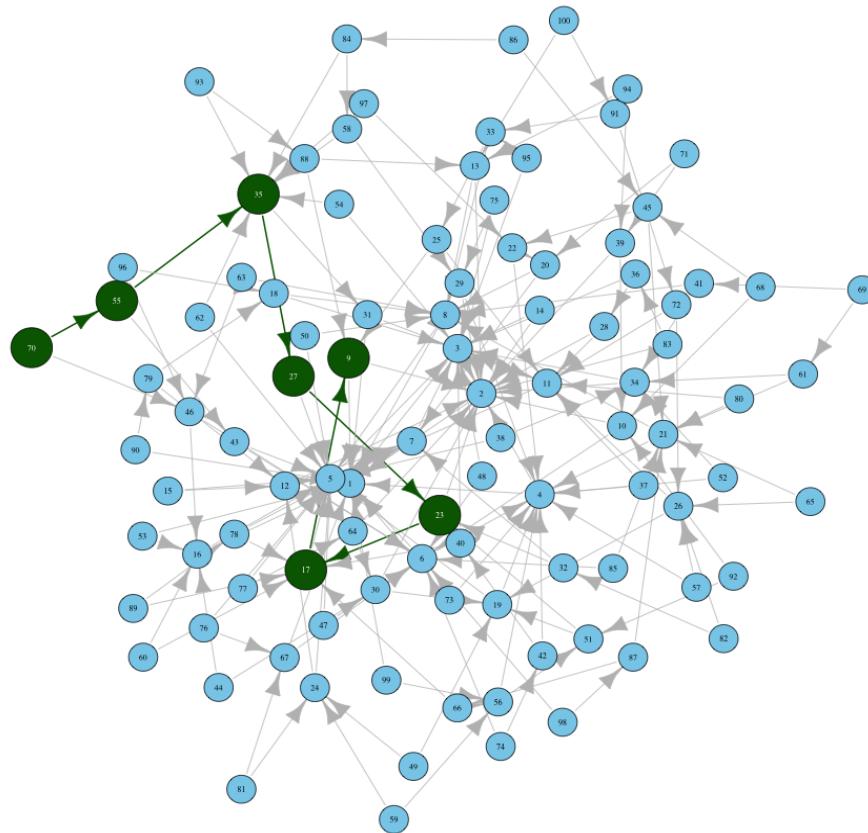
- Caso médio:  $\mathcal{O}(m)$

- Poucas iterações para convergência
- Para grandes redes, número de arestas  $m \gg n$

**Apropriado para Big Data Analytics!**

# Proximidade de Rótulos

Identificar eventos de interesse a partir de nós rotulados, considerando informação textual, temporal e geográfica!



Visualizar resultados conforme a aplicação



# Proximidade de Rótulos

## Exemplo Prático #5 Propagação de Rótulos

# Mineração de Eventos

## ■ Extração de Padrões

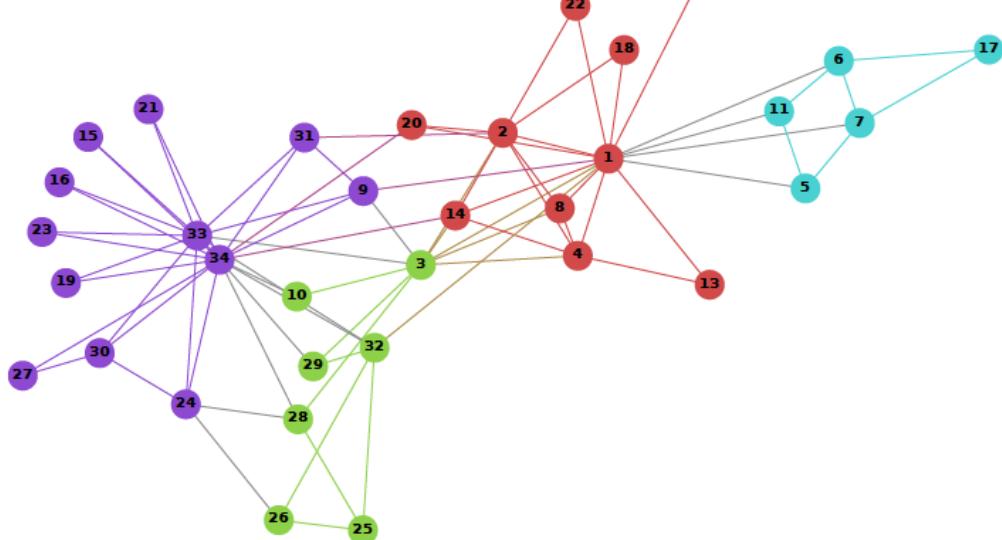
- DeepWalk/Node2Vec (aprendizado de representações)

Habilita o uso de algoritmos tradicionais:

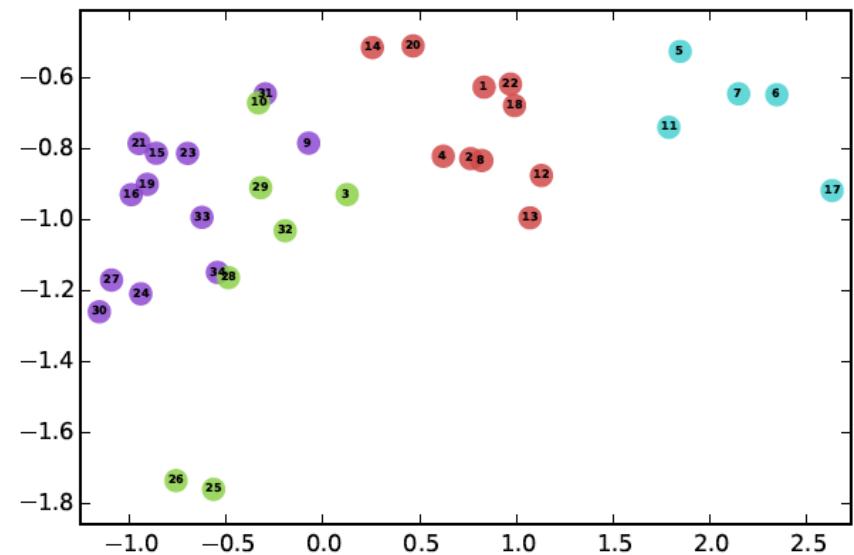
- Visualização
- Agrupamento
- Classificação

# Mineração de Eventos

- Extração de Padrões - DeepWalk/Node2Vec
  - Como extrair representações vetoriais a partir da rede de eventos?



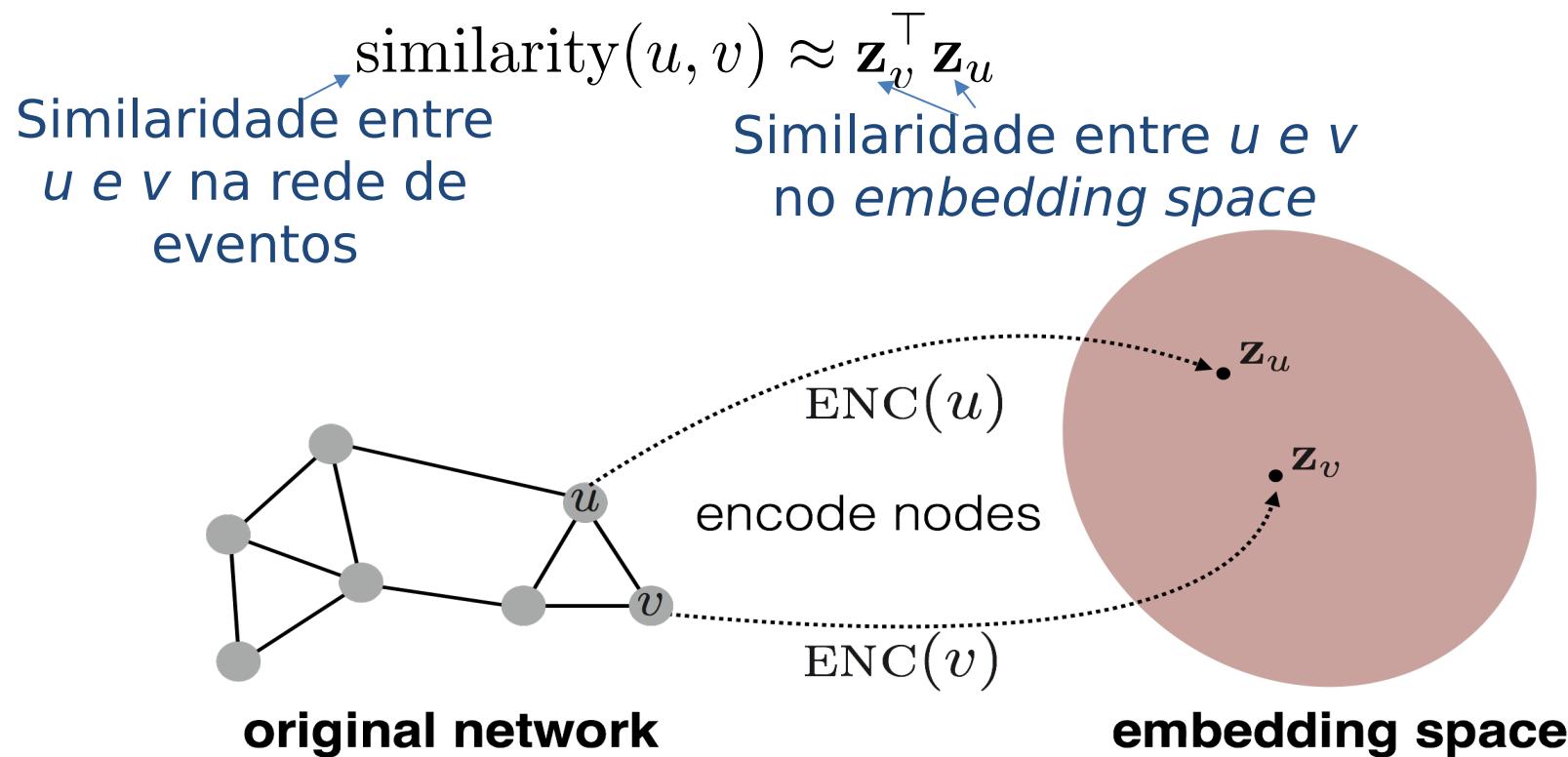
Entrada: Rede



Saída: Representação em baixa dimensionalidade (espaço-vetorial)

# Mineração de Eventos

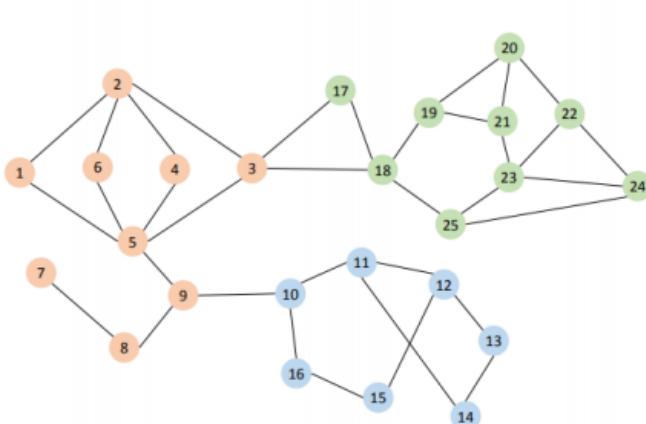
- Extração de Padrões - DeepWalk/Node2Vec
  - Temos que realizar um *encoding* que transforma a similaridade na rede na similaridade no novo espaço



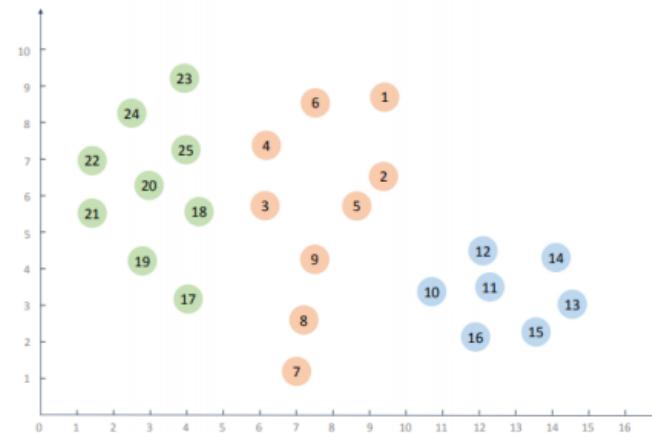
# Mineração de Eventos

## ■ DeepWalk / Node2Vec

- Rede de eventos:  $\mathcal{N} = \langle \mathcal{O}, \mathcal{R}, \mathcal{W} \rangle$
- Aprender uma função de mapeamento:  $\mathbf{F} : \mathcal{N} \rightarrow \mathbb{R}^m$ 
  - Modelo espaço-vetorial  $m$ -dimensional
  - Cada objeto na rede possui um vetor de características



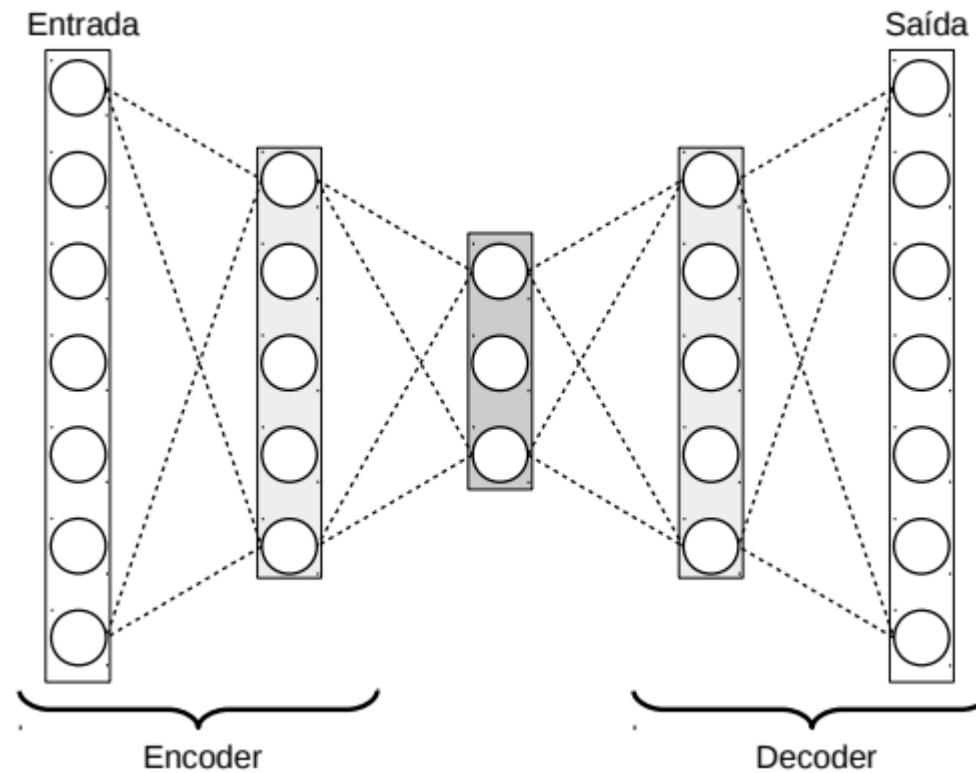
(a) Entrada: Rede de Informação



(b) Saída: Nós representados em um espaço vetorial (e.g. euclidiano).

# Mineração de Eventos

## ■ Autoencoders



# Proximidade de Rótulos

Exemplo Prático #6  
DeepWalk/Node2Vec

# Mineração de Eventos

## ■ Pós-Processamento

### Critérios de Validação Interna

- **Coerência Semântica:** eventos relacionados possuem informação textual similar?
- **Coesão Geográfica:** eventos relacionados ocorreram em regiões próximas?
- **Conectividade Temporal:** eventos relacionados possuem datas de ocorrência próximas?

# Mineração de Eventos

## ■ Pós-Processamento

### Critérios de Validação Interna

- **Coerência Semântica:** eventos relacionados possuem informação textual similar?
- **Coesão Geográfica:** eventos relacionados ocorreram em regiões próximas?
- **Conectividade Temporal:** eventos relacionados possuem datas de ocorrência próximas?

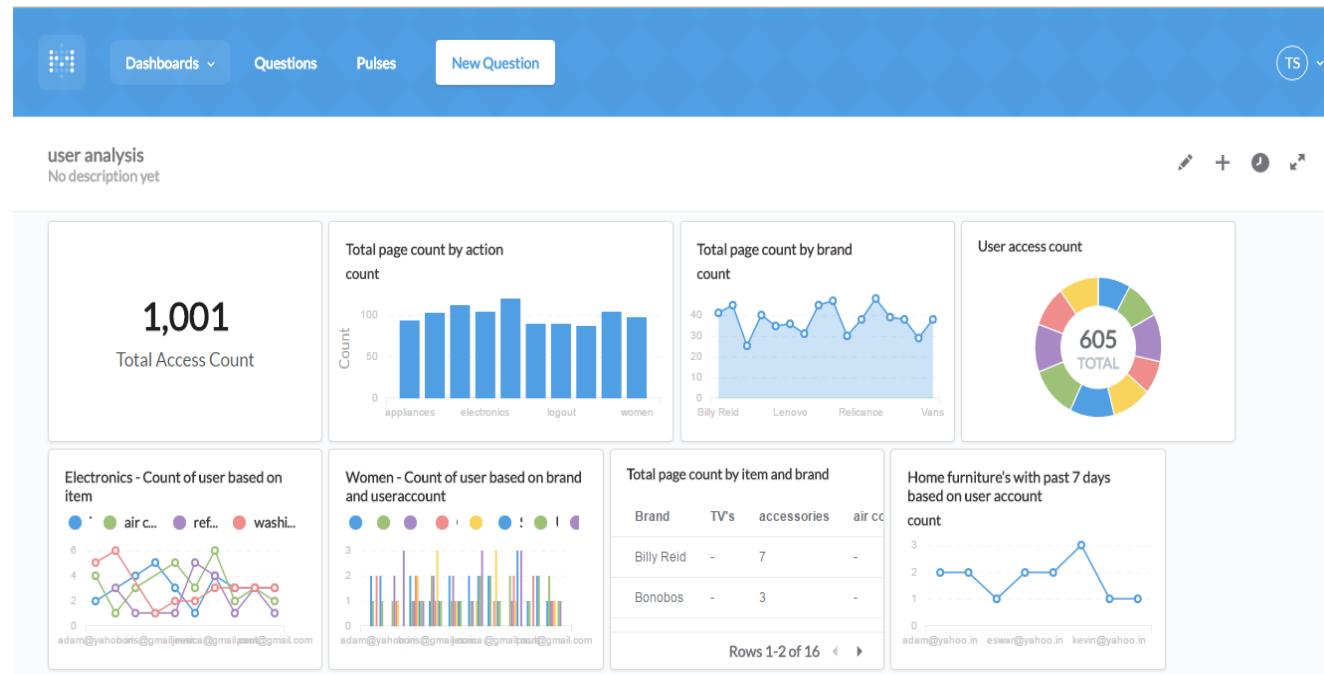
### Critérios de Validação Externa

- Precisão, Revocação, F1-Score, Acurácia
- Dependem de um conjunto verdade

# Mineração de Eventos

## ■ Uso do Conhecimento

- Sugestão de plataformas (open source) de visualização de dados para inteligência analítica
  - Metabase, Apache Superset, Redash, Google Datastudio



# Agenda

- Mineração de Textos e Inteligência Analítica
- Mineração de Eventos
  - Identificação do Problema
  - Pré-Processamento
  - Extração de Padrões
  - Pós-processamento
  - Uso do Conhecimento
- Considerações Finais

# Projetos em Andamento

## ■ Websensors

### Metodologia para Mineração de Eventos

■ <https://websensors.net.br/>

The screenshot displays the Websensors website's "Nosso Framework" section and a "Coleta e monitoramento de eventos em tempo real" (Real-time event collection and monitoring) map.

**Nosso Framework**

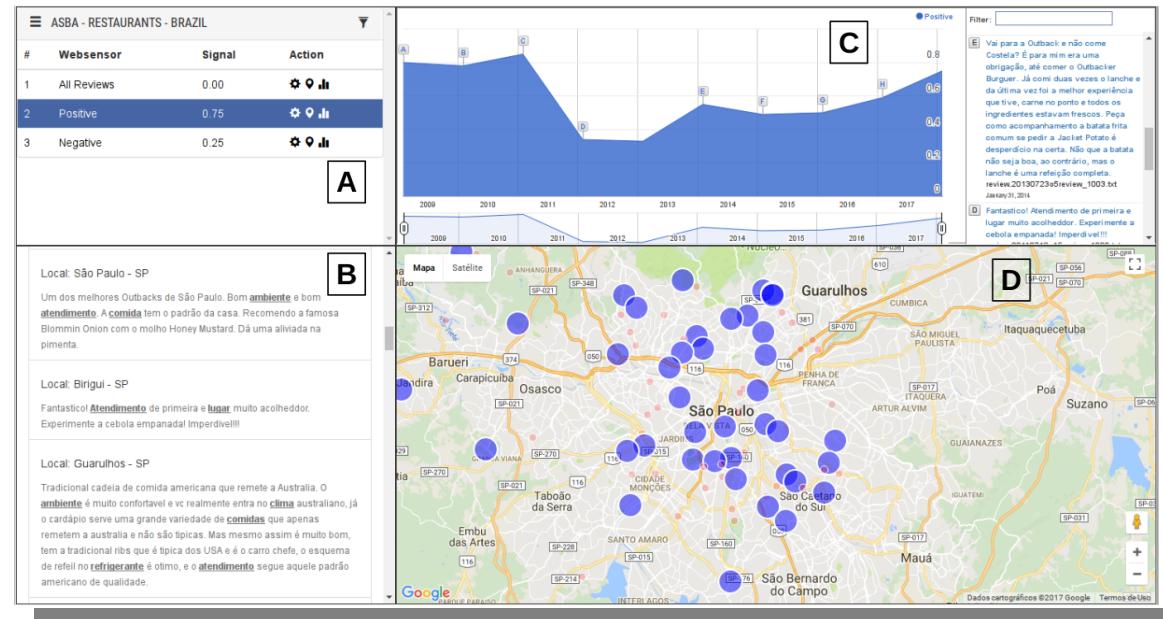
- Informações locais sobre seu negócio**: Shows a central building icon surrounded by various data visualization icons like pie charts, bar graphs, and line charts.
- Nosso conhecimento global sobre Eventos relacionados**: A world map with numerous red location pins and a magnifying glass icon. Below it, three bullet points explain their data processing:
  - Processamos diariamente milhares de eventos da web
  - Nossos algoritmos de Aprendizado de Máquina identificam eventos relacionados
  - Eventos com Representação 4W: What? Where? When? Why?
- Websensors**: A box containing:
  - Inteligência Analítica Preditiva
  - Websensors permitem otimizar modelos preditivos existentes
  - Websensors são indicadores inteligentes atualizados em tempo real

**Coleta e monitoramento de eventos em tempo real.**

A world map showing real-time event monitoring. Events are represented by colored circles (blue, orange, yellow) with numbers indicating their count. A callout box shows a specific event: "Putin and Aliyev to discuss Karabakh | ARMENPRESS Armenian News Agency". The map includes labels for continents, oceans, and countries.

# Projetos em Andamento

- Mineração de Eventos para Comércio Eletrônico
  - Foco em eventos positivos / negativos



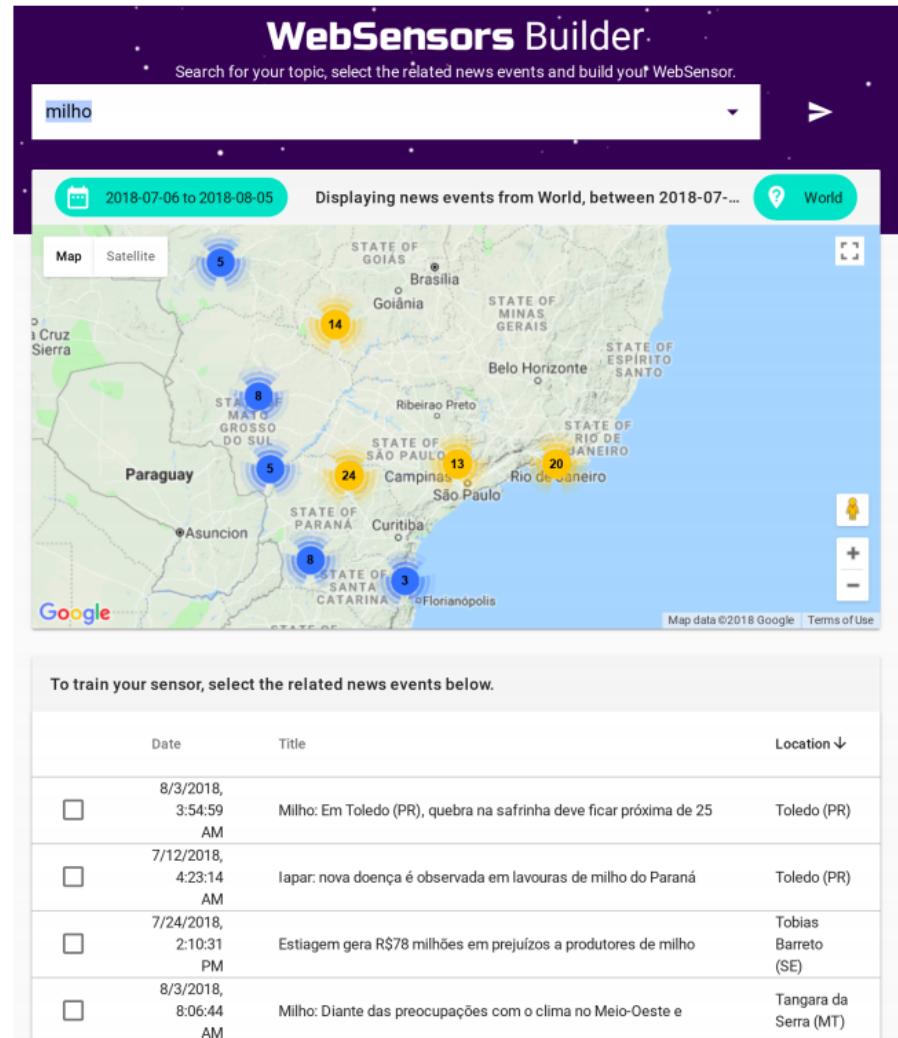
# Projetos em Andamento

## ■ Mineração de Eventos para Agronegócios

- Monitorar o impacto de eventos (leis, notícias, boletins) para esse setor
- Otimizar predição em *commodities* agrícolas

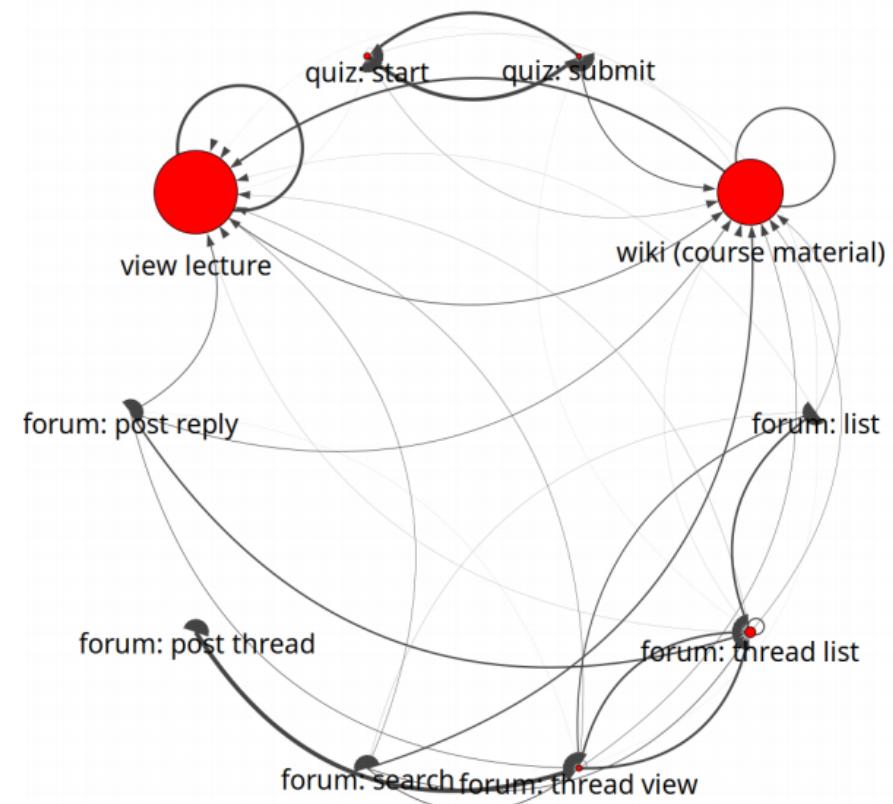
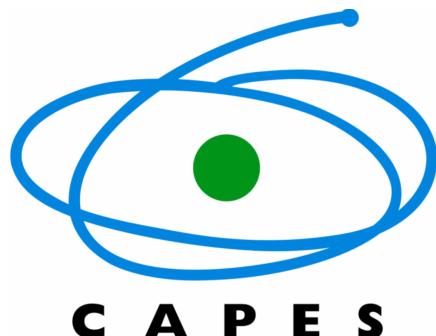


Conselho Nacional de Desenvolvimento  
Científico e Tecnológico



# Projetos em Andamento

- Mineração de Eventos para Predição de Evasão
  - Ambientes EAD
  - Analisa redes clickstream



<http://porvir.org/conheca-4-projetos-inovadores-para-cursos-a-distancia/>

# Mineração de eventos: algoritmos e aplicações

Solange O. Rezende  
[solange@icmc.usp.br](mailto:solange@icmc.usp.br)

4<sup>a</sup> ESCOLA AVANÇADA EM BIG DATA ANALYSIS (Outubro de 2020)

Agradecimentos:

- Brucce Neves dos Santos (Monitoria e Organização) - [brucce.neves@gmail.com](mailto:brucce.neves@gmail.com)
- Ricardo Marcacini (Exemplos e Práticas) – [ricardo.marcacini@gmail.com](mailto:ricardo.marcacini@gmail.com)

