Agrupamento de Textos e suas Aplicações em Inteligência Analítica

Pré-processamento de Textos: *Bag-of-Words* e Similaridade Cosseno

Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

Cursos de Extensão – Difusão de Conhecimento – Dezembro de 2021

ICMC SÃO CARLOS USP

LABIC Laboratório de Inteligência Computacional ICMC.USP SÃO CARLOS

WebSensors

Agrupamento de Textos

- Pré-processamento

Informação Textual

Informação Geográfica

Informação Temporal

Agrupamento de Textos

- Pré-processamento

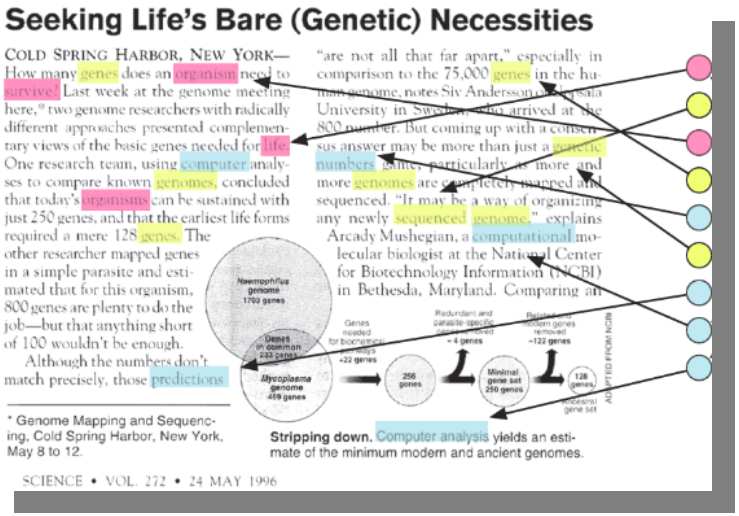
Informação Textual

Informação Geográfica

Informação Temporal

- Documentos textuais 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,* 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

Agrupamento de Textos

- Pré-processamento

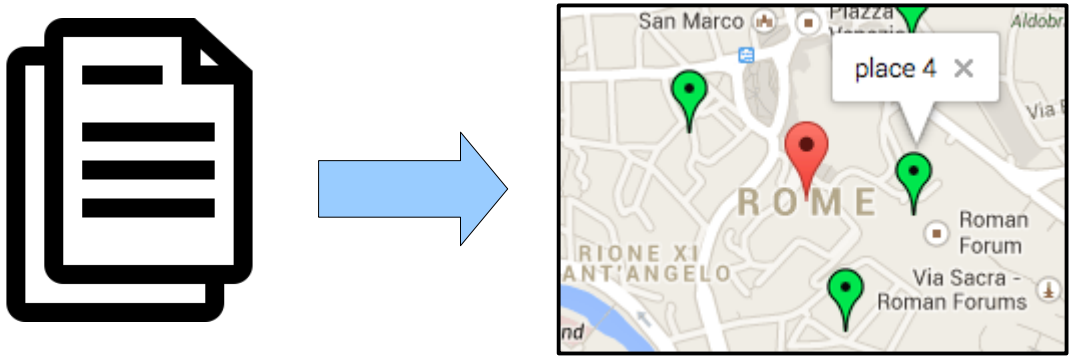
Informação Textual

Informação Geográfica

Informação Temporal

- Textos podem estar relacionados de acordo com entidades geográficas

  Como identificar informação geográfica em dados textuais?



Agrupamento de Textos

- Pré-processamento

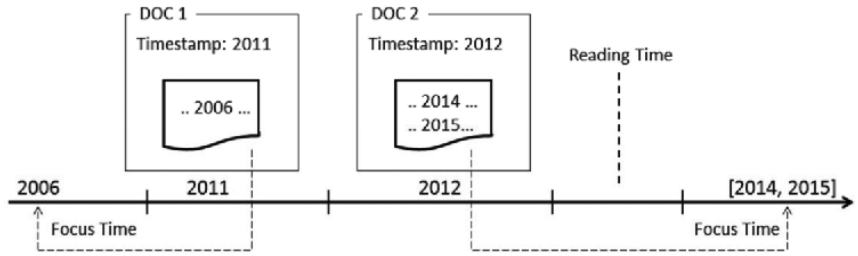
Informação Textual

Informação Geográfica

Informação Temporal

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

  Como extrair informação temporal dos textos?



Agrupamento de Textos

- Pré-processamento

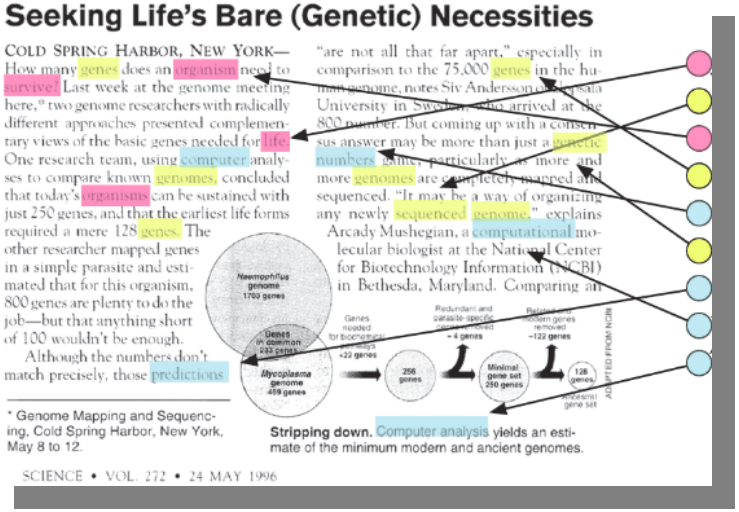
Informação Textual

Informação Geográfica

Informação Temporal

- Documentos textuais 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,* 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

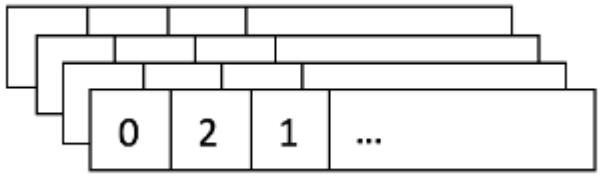
Pré-processamento de Textos

Modelo Espaço-Vetorial

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

Agrupamento de Textos

- 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



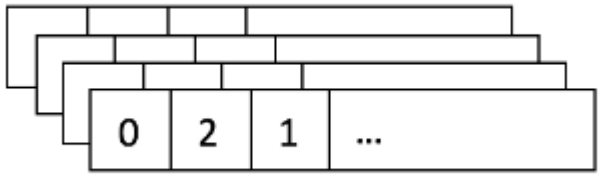
Vetores para 4 objetos

Agrupamento de Textos

- 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:

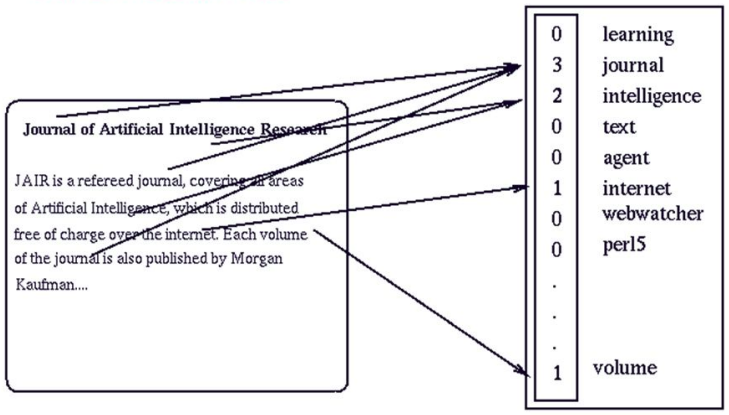
1) Quais são os atributos?
2) Como definir os pesos dos atributos?



Vetores para 4 objetos

Agrupamento de Textos

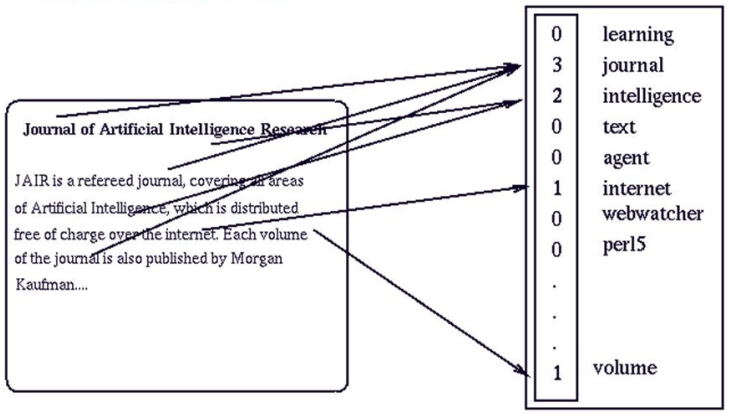
- Pré-processamentos dos textos
  - Modelo espaço-vetorial usando *Bag-of-words*
    - Atributos são extraídas dos textos
    - Peso da palavra é sua frequência objeto
    - A ordem das palavras nos textos não é considerada



Fonte: [4]

Agrupamento de Textos

- 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

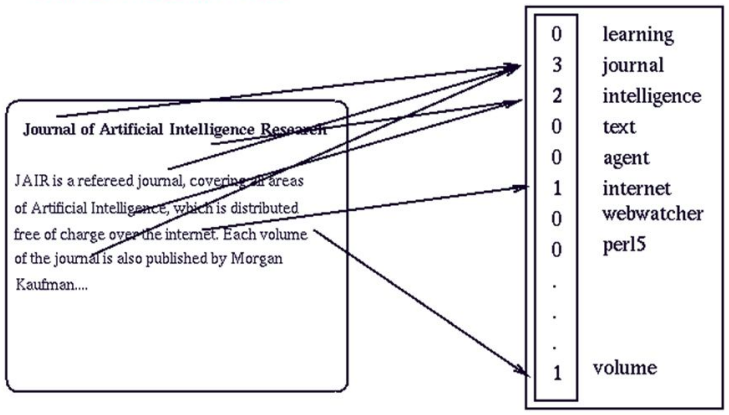


Fonte: [4]

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

Agrupamento de Textos

- 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

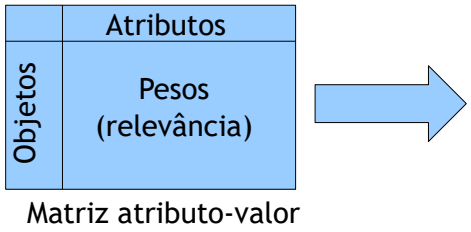


Fonte: [4]

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

Agrupamento de Textos

- Pré-processamento – Informação Textual
  - Bag-of-words: representação no modelo espaço-vetorial. Simples (Baseline).



Matriz atributo-valor

- 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]

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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.

Agrupamento de Textos

- 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*:

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

Agrupamento de Textos

- 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*:

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

Final:

estudante Inteligência Artificial foi livraria comprar livros estudar

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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)

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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

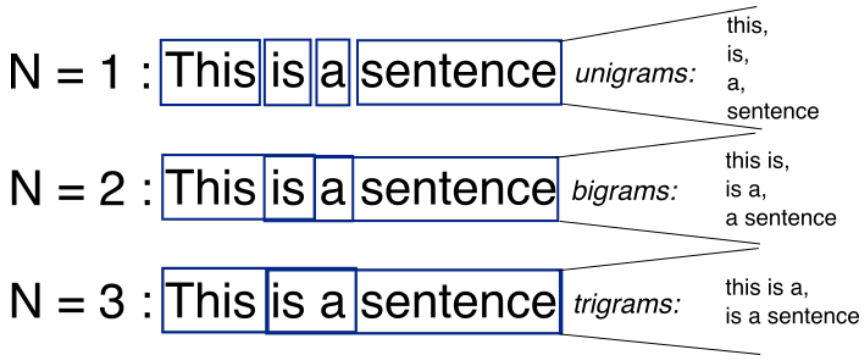
Após identificação de *n-gramas*:

estud Intelig_Artifici fo livr compr livr estud

Agrupamento de Textos

- 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



Agrupamento de Textos

- 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

Agrupamento de Textos

- 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.

Agrupamento de Textos

- 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

Agrupamento de Textos

- 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 documentos textuais!

Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - 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 (propositalmente) 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 (...)

Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - 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

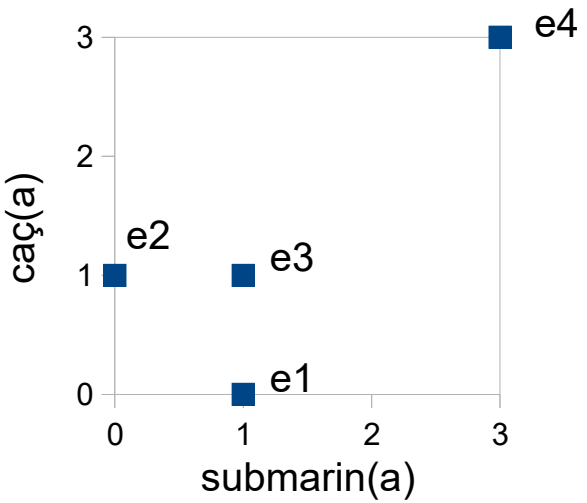
Vamos considerar quatro eventos.
Escolhemos (propositalmente) apenas dois atributos.

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 (...)

Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - Como calcular a proximidade entre conteúdo no modelo espaço-vetorial?

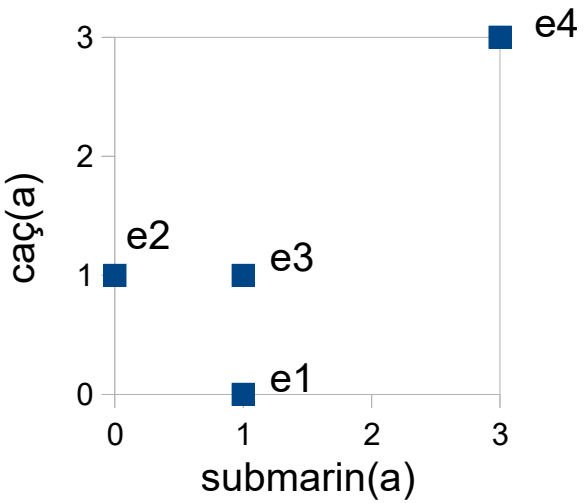
	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3



Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - O espaço euclidiano não capturou adequadamente o conceito de proximidade entre os eventos!

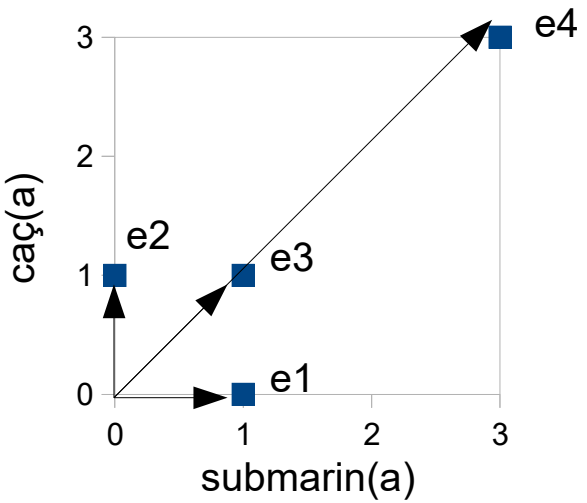
	caç(a)	submarin(a)
e1	0	1
e2	1	0
e3	1	1
e4	3	3



Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - Considere utilizar o ângulo entre os vetores!

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

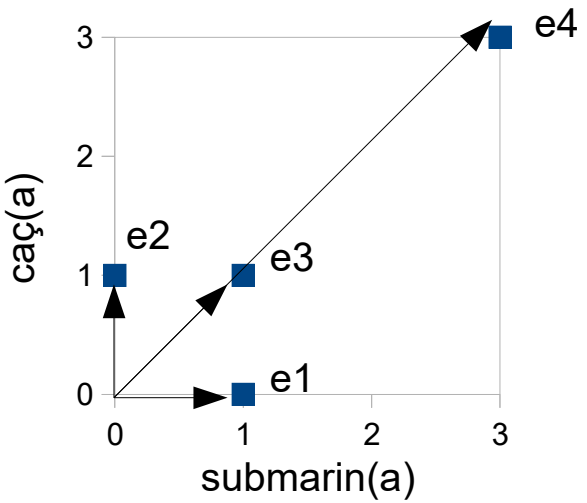


Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - 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:
ângulo(e1,e2) = 90º; cos(90º) = 0
ângulo(e2,e3) = 45º; cos(45º) = 0.5
ângulo(e3,e4) = 0º; cos(0º) = 1



Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - 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.

Agrupamento de Textos

- O problema da similaridade
  - Proximidade de Conteúdo.
  - 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.