

## POR QUE ESTUDAR PROBABILIDADE?

- Probabilidade e língua
  - Probabilidade dos fenômenos lingüísticos
    - **Descrição, caracterização**
      - Caracterização de discursos políticos, detecção de mudanças históricas, contraste de discurso oral vs. textual, estudo de fenômenos sintáticos, probabilidades das colocações, etc.
    - **Previsão**
      - Que traduções são possíveis, qual a palavra correta mais provável dada uma palavra com ortografia errada, qual a chance de uma sentença ser importante no texto, etc.

3

## POR QUE ESTUDAR PROBABILIDADE?

- Probabilidade e língua
  - Probabilidade dos fenômenos lingüísticos
    - Às vezes, esses “números mágicos” são intuitivos
      - Calculados naturalmente por nós
    - Às vezes, exigem raciocínio mais sofisticado

4

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

5

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

6

Qual a probabilidade da palavra “que” ocorrer?  
“chance”

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio **que** atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que** administra os aeroportos no país) informou **que** a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

Qual a probabilidade da palavra “que” ocorrer?  $3/125 = 0.024 = 2.4\%$   
“chance”

7

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto **de** Congonhas, na zona sul **de** São Paulo. Apesar disso, equipes do Corpo **de** Bombeiros permaneciam no local às 10h para o trabalho **de** rescaldo. Não há informação **de** feridos. **De** acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João **de** Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato **de** propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto **de** Congonhas.

E “de”?

8

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto **de** Congonhas, na zona sul **de** São Paulo. Apesar disso, equipes do Corpo **de** Bombeiros permaneciam no local às 10h para o trabalho **de** rescaldo. Não há informação **de** feridos. **De** acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João **de** Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato **de** propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto **de** Congonhas.

E “de”?  $9/125 = 0.072 = 7.2\%$

9

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E “incêndio”?

10

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E “incêndio”?  $4/125 = 0.032 = 3.2\%$

11

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E qualquer palavra do texto?

12

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E qualquer palavra do texto?  $125/125 = 1 = 100\%$

13

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O **fogo** atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E “incêndio” ou “fogo”?

14

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o **incêndio** que atingiu uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o **incêndio** teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O **fogo** atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do **incêndio** ainda serão investigadas. Apesar do **incêndio**, a Infraero (estatal que administra os aeroportos no país) informou que a fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E “incêndio” ou “fogo”?  $4/125 + 1/125 = 5/125 = 0.04 = 4\%$

15

## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio **que atingiu** uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que administra** os aeroportos no país) informou **que a** fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E de um verbo seguir “que”?

16



## EXEMPLO

### o Exemplo (125 palavras)

Foi controlado o incêndio **que atingiu** uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros permaneciam no local às 10h para o trabalho de rescaldo. Não há informação de feridos. De acordo com a corporação, o incêndio teve início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros foram encaminhados para o local. O fogo atingiu vários barracos, mas as equipes ainda não tinham o número exato de propriedades atingidas. As causas do incêndio ainda serão investigadas. Apesar do incêndio, a Infraero (estatal **que administra** os aeroportos no país) informou **que a** fumaça não comprometeu os pousos e decolagens no aeroporto de Congonhas.

E de um verbo seguir “que”?  $2/3 = 0.666 = 66.6\%$

17

## EXEMPLO

### o Exemplo (125 palavras)

Foi **controlado** o incêndio **que atingiu** uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros **permaneciam** no local às 10h para o trabalho de rescaldo. Não **há** informação de feridos. De acordo com a corporação, o incêndio **teve** início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros **foram encaminhados** para o local. O fogo **atingiu** vários barracos, mas as equipes ainda não **tinham** o número exato de propriedades **atingidas**. As causas do incêndio ainda **serão investigadas**. Apesar do incêndio, a Infraero (estatal **que administra** os aeroportos no país) **informou** que a fumaça não **comprometeu** os pousos e decolagens no aeroporto de Congonhas.

E de “que” preceder um verbo?

18

## EXEMPLO

### o Exemplo (125 palavras)

Foi **controlado** o incêndio **que atingiu** uma favela na região do aeroporto de Congonhas, na zona sul de São Paulo. Apesar disso, equipes do Corpo de Bombeiros **permaneciam** no local às 10h para o trabalho de rescaldo. Não **há** informação de feridos.

De acordo com a corporação, o incêndio **teve** início por volta das 8h20 na rua João de Lery, no bairro Parque Jabaquara. Ao todo, 17 carros dos bombeiros **foram encaminhados** para o local. O fogo **atingiu** vários barracos, mas as equipes ainda não **tinham** o número exato de propriedades **atingidas**.

As causas do incêndio ainda **serão investigadas**. Apesar do incêndio, a Infraero (estatal **que administra** os aeroportos no país) **informou** que a fumaça não **comprometeu** os pousos e decolagens no aeroporto de Congonhas.

E de “que” preceder um verbo?  $2/15 = 0.133 = 13.3\%$

19

## PROBABILIDADES

- o Probabilidade: resultado entre 0 e 1, ou 0 e 100%
- o  $P(\text{evento impossível}) = 0$
- o  $P(\text{qualquer coisa}) = 1$  (ou 100%)
- o  $P(A) \text{ ou } P(B) = P(A) + P(B)$ 
  - $P(\text{qualquer coisa}) = P(\text{uma coisa}) + P(\text{segunda coisa}) + \dots + P(\text{enésima coisa})$
- o Probabilidade condicional  $P(A | B) = P(A \cap B) / P(B)$
- o  $P(A \cap B) = P(B) * P(A | B) = P(A) * P(B | A)$ 
  - $P(A \cap B) = P(A) * P(B)$ , se eventos independentes
  - $P(A_1 \cap \dots \cap A_n) = P(A_1) * P(A_2 | A_1) * P(A_3 | A_1 \cap A_2) \dots$
- o A e B são condicionalmente independentes se  $P(A \cap B | C) = P(A | C) * P(B | C)$

20

## BAYES

### ○ Teorema de Bayes

- $P(A | B) = P(B | A) * P(A) / P(B)$ 
  - Pode-se inverter: usar  $P(B | A)$  em vez de  $P(A | B)$ 
    - Por que isso é interessante?

21

## BAYES

### ○ Teorema de Bayes

- $P(A | B) = P(B | A) * P(A) / P(B)$ 
  - Útil quando não se tem, é difícil ou ilógico calcular  $P(A | B)$   
→ pode-se usar o inverso
    - Por exemplo, o que é melhor?
      - $P(\text{doença} | \text{sintoma})$
      - $P(\text{sintoma} | \text{doença})$

22

## BAYES

### ○ Teorema de Bayes

- $P(A | B) = P(B | A) * P(A) / P(B)$ 
  - Útil quando não se tem, é difícil ou ilógico calcular  $P(A | B)$   
→ pode-se usar o inverso

$$P(\text{doença} | \text{sintoma}) = P(\text{sintoma} | \text{doença}) * P(\text{doença}) / P(\text{sintoma})$$

- o sintoma é o que se observa, e a doença é o que se quer descobrir
  - $P(\text{doença} | \text{sintoma})$
- ... mas quem causa o sintoma é a doença, e não o inverso
  - $P(\text{sintoma} | \text{doença})$
- $P(\text{doença} | \text{sintoma})$  pode ser “tendencioso” e “temporal”

23

## BAYES

### ○ Teorema de Bayes

- Exemplo: “chegou com dor de cabeça no médico”
  - $P(\text{sarampo} | \text{dor de cabeça}) = P(\text{dor de cabeça} | \text{sarampo}) * P(\text{sarampo}) / P(\text{dor de cabeça})$
  - $P(\text{malária} | \text{dor de cabeça}) = P(\text{dor de cabeça} | \text{malária}) * P(\text{malária}) / P(\text{dor de cabeça})$
  - A maior probabilidade ganha e indica o diagnóstico final!
    - Atenção:  $P(\text{dor de cabeça})$  é constante. Faz diferença no resultado?

24

## BAYES

### ○ Teorema de Bayes

- Exemplo: “chegou com dor de cabeça no médico”
  - $P(\text{sarampo} | \text{dor de cabeça}) = \frac{P(\text{dor de cabeça} | \text{sarampo}) * P(\text{sarampo})}{P(\text{dor de cabeça})}$
  - $P(\text{malária} | \text{dor de cabeça}) = \frac{P(\text{dor de cabeça} | \text{malária}) * P(\text{malária})}{P(\text{dor de cabeça})}$
  - A maior probabilidade ganha e indica o diagnóstico final!
    - Atenção:  $P(\text{dor de cabeça})$  é constante. Faz diferença no resultado?
      - Ao comparar hipóteses, pode-se usar  $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

25

## EXERCÍCIO

- Sabe-se que catafóras são raras: de todas as sentenças de um corpus, sabe-se que somente uma fração de 0.008 delas contêm catafóras
- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos
- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

26

## EXERCÍCIO

- Sabe-se que catafóras são raras: de todas as sentenças de um corpus, sabe-se que somente uma fração de 0.008 delas contém catafóras
- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos
- Um sentenças foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução:  $P(\text{catafóra}) = 0.008$                        $P(\text{sem catafóra}) = 0.992$   
 $P(\text{sim} | \text{catafóra}) = 0.98$                        $P(\text{não} | \text{catafóra}) = 0.02$   
 $P(\text{sim} | \text{sem catafóra}) = 0.03$                        $P(\text{não} | \text{sem catafóra}) = 0.97$

$P(\text{catafóra} | \text{sim}) = P(\text{sim} | \text{catafóra}) * P(\text{catafóra}) = 0.98 * 0.008 = 0.0078$   
 $P(\text{sem catafóra} | \text{sim}) = P(\text{sim} | \text{sem catafóra}) * P(\text{sem catafóra}) = 0.03 * 0.992 = \mathbf{0.0298}$

27

## EXERCÍCIO

O que aconteceu?

- Sabe-se que catafóras são raras: de todas as sentenças de um corpus, sabe-se que somente uma fração de 0.008 delas contém catafóras
- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos
- Um sentenças foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução:  $P(\text{catafóra}) = 0.008$                        $P(\text{sem catafóra}) = 0.992$   
 $P(\text{sim} | \text{catafóra}) = 0.98$                        $P(\text{não} | \text{catafóra}) = 0.02$   
 $P(\text{sim} | \text{sem catafóra}) = 0.03$                        $P(\text{não} | \text{sem catafóra}) = 0.97$

$P(\text{catafóra} | \text{sim}) = P(\text{sim} | \text{catafóra}) * P(\text{catafóra}) = 0.98 * 0.008 = 0.0078$   
 $P(\text{sem catafóra} | \text{sim}) = P(\text{sim} | \text{sem catafóra}) * P(\text{sem catafóra}) = 0.03 * 0.992 = \mathbf{0.0298}$

28

## EXERCÍCIO

E se a probabilidade de ocorrência de catáforas fosse uniforme no cópuz?

- Sabe-se que catáforas são raras: de todas as sentenças de um cópuz, sabe-se que somente uma fração de 0.008 delas contém catáforas
- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos
- Um sentenças foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução:  $P(\text{catáfora}) = 0.008$                        $P(\text{sem catáfora}) = 0.992$   
 $P(\text{sim} | \text{catáfora}) = 0.98$                        $P(\text{não} | \text{catáfora}) = 0.02$   
 $P(\text{sim} | \text{sem catáfora}) = 0.03$                        $P(\text{não} | \text{sem catáfora}) = 0.97$

$P(\text{catáfora} | \text{sim}) = P(\text{sim} | \text{catáfora}) * P(\text{catáfora}) = 0.98 * 0.008 = 0.0078$   
 $P(\text{sem catáfora} | \text{sim}) = P(\text{sim} | \text{sem catáfora}) * P(\text{sem catáfora}) = 0.03 * 0.992 = \mathbf{0.0298}$

29

## EXERCÍCIO

E a probabilidade da sentença ser catafórica? Já conseguimos?

- Sabe-se que catáforas são raras: de todas as sentenças de um cópuz, sabe-se que somente uma fração de 0.008 delas contém catáforas
- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos
- Um sentenças foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Resolução:  $P(\text{catáfora}) = 0.008$                        $P(\text{sem catáfora}) = 0.992$   
 $P(\text{sim} | \text{catáfora}) = 0.98$                        $P(\text{não} | \text{catáfora}) = 0.02$   
 $P(\text{sim} | \text{sem catáfora}) = 0.03$                        $P(\text{não} | \text{sem catáfora}) = 0.97$

$P(\text{catáfora} | \text{sim}) = P(\text{sim} | \text{catáfora}) * P(\text{catáfora}) = 0.98 * 0.008 = 0.0078$   
 $P(\text{sem catáfora} | \text{sim}) = P(\text{sim} | \text{sem catáfora}) * P(\text{sem catáfora}) = 0.03 * 0.992 = \mathbf{0.0298}$

30

## EXERCÍCIO

- Sabe-se que catafóras são raras: de todas as sentenças de um cópuz, sabe-se que somente uma fração de 0.008 delas contém catafóras
- Existe um sistema de PLN que diz se sentenças são ou não catafóricas
  - O sistema retorna sim – um verdadeiro positivo (as sentenças são catafóricas e o sistema diz que são) – em 98% dos casos
  - O sistema retorna não – um verdadeiro negativo (as sentenças não são catafóricas e o sistema diz que não são) – em 97% dos casos
- Um sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

### Normalizando...

$$P(\text{catafóra} \mid \text{sim}) = 0.0078 \quad \rightarrow 0.0078 / (0.0078 + 0.0298) = 0.21 = 21\%$$

$$P(\text{sem catafóra} \mid \text{sim}) = 0.0298 \quad \rightarrow 0.0298 / (0.0078 + 0.0298) = 0.79 = 79\%$$

31

## DISTRIBUIÇÕES

- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Exemplo?

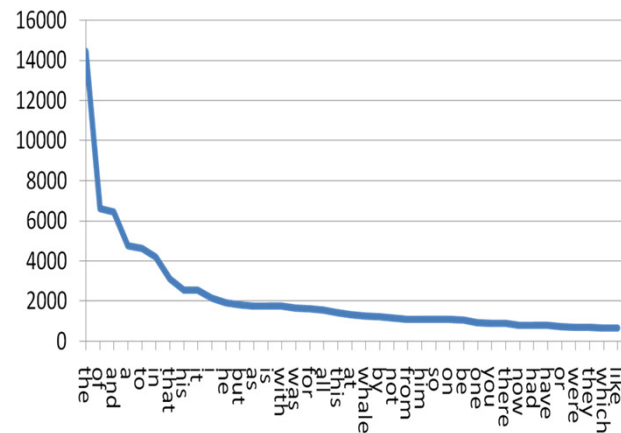
32



## DISTRIBUIÇÕES

- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Por que conhecer esses comportamentos é importante?

Lei de Zipf



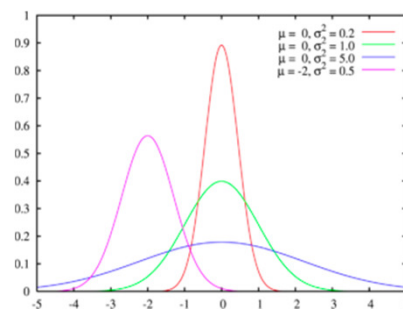
## DISTRIBUIÇÕES

- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Com alguns parâmetros, podemos **descrever ou prever número médio, variações e onde encontrar** os **fenômenos modelados**
      - Em geral, parâmetros são **média** ( $\mu$ ), **freqüência, desvio padrão** ( $\sigma$ ) ou **variância** ( $\sigma^2$ )
      - Distribuições podem ser discretas ou contínuas

## DISTRIBUIÇÕES

### Exemplos

- Poisson
- Geométrica
- Uniforme
- Normal (ou gaussiana)
  - Contínua
- Weibull
- Pareto
- Etc.



35

## EXEMPLO DE USO DE ESTATÍSTICA

### Análise sintática automática

- Gramáticas probabilísticas

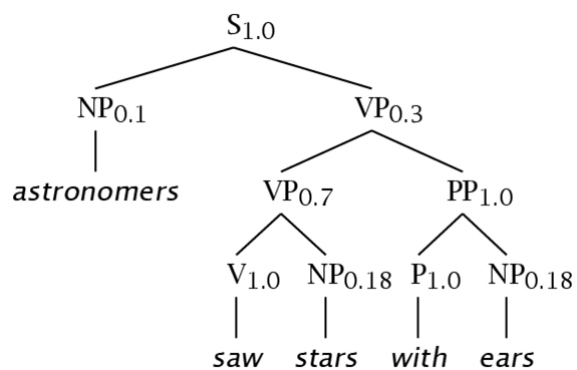
$S \rightarrow NP VP$	1.0	$V \rightarrow saw$	1.0
$PP \rightarrow P NP$	1.0	$NP \rightarrow astronomers$	0.1
$VP \rightarrow V NP$	0.7	$NP \rightarrow ears$	0.18
$VP \rightarrow VP PP$	0.3	$NP \rightarrow saw$	0.04
$NP \rightarrow NP PP$	0.4	$NP \rightarrow stars$	0.18
$P \rightarrow with$	1.0	$NP \rightarrow telescopes$	0.1

- De onde se conseguem as probabilidades de cada regra?

36

## EXEMPLO DE USO DE ESTATÍSTICA

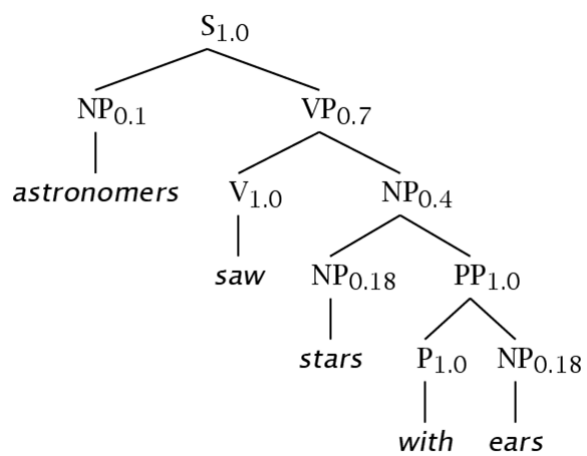
### ○ Possibilidade 1



$$P(1) = 1.0 * 0.1 * 0.3 * 0.7 * 1.0 * 0.18 * 1.0 * 1.0 * 0.18 = 0.0006804$$

## EXEMPLO DE USO DE ESTATÍSTICA

### ○ Possibilidade 2



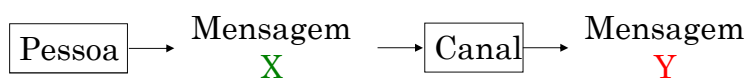
$$P(2) = 1.0 * 0.1 * 0.7 * 1.0 * 0.4 * 0.18 * 1.0 * 1.0 * 0.18 = 0.0009072 \text{ (>anterior)}$$

## MODELO *NOISY-CHANNEL*

- Shannon, 1948
  - Teoria da Informação
- Modelo probabilístico
  - No coração do renascimento da estatística no PLN na década de 70
- Transmissão de mensagens pela linha telefônica
  - Capacidade de transmissão por um canal (*channel*)
  - Ocorrência de ruídos (*noise*)
  - Quantidade de informação necessária para recuperação da mensagem original

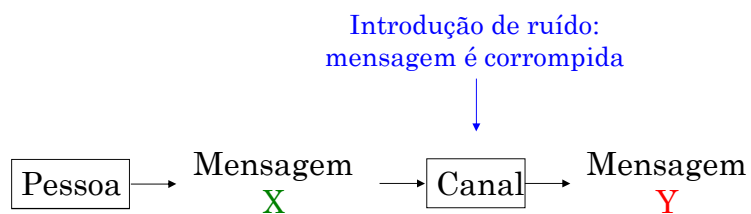
39

## MODELO *NOISY-CHANNEL*



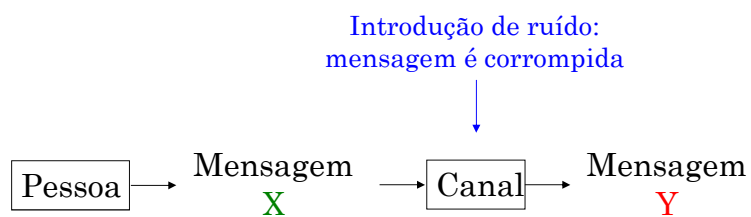
40

## MODELO *NOISY-CHANNEL*



41

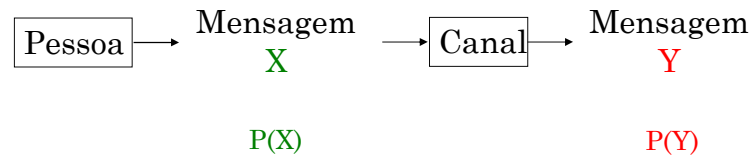
## MODELO *NOISY-CHANNEL*



Quanta informação pode ser transmitida?  
 Quanta informação pode ser transmitida para  
 minimizar a ocorrência de ruídos?  
 Como e que tipos de ruído ocorrem?  
 Como recuperar a mensagem original X a partir de Y?

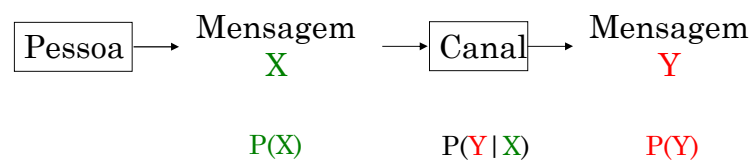
42

### MODELO *NOISY-CHANNEL*



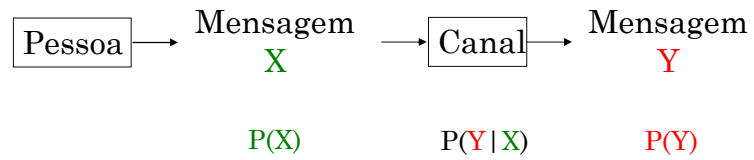
43

### MODELO *NOISY-CHANNEL*



44

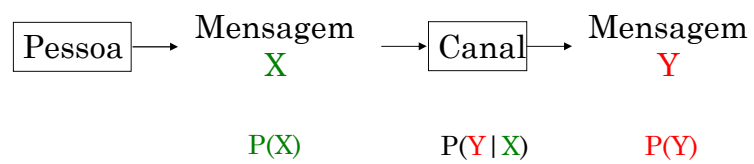
## MODELO *NOISY-CHANNEL*



Determinar  $X$  a partir de  $Y$ :  $P(X|Y)$

45

## MODELO *NOISY-CHANNEL*



Determinar  $X$  a partir de  $Y$ :  $P(X|Y)$

Teorema  
de Bayes

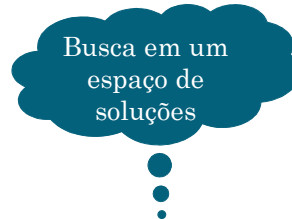
$$P(X|Y) = P(Y|X) \times P(X) / P(Y)$$

46

## TEOREMA DE BAYES

$$P(\mathbf{X} | \mathbf{Y}) = P(\mathbf{Y} | \mathbf{X}) \times P(\mathbf{X}) / P(\mathbf{Y})$$

- $\mathbf{Y}$  é observado
- Deve-se escolher  $\mathbf{X}$  que maximize  $P(\mathbf{X} | \mathbf{Y})$ : **decodificação**



$$P(\mathbf{X} | \mathbf{Y}) = P(\mathbf{Y} | \mathbf{X}) \times \underbrace{P(\mathbf{X}) / P(\mathbf{Y})}_{\text{constante}}$$

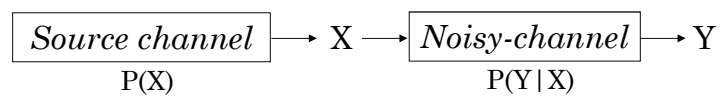
↓  
constante

$$P(\mathbf{X} | \mathbf{Y}) = P(\mathbf{Y} | \mathbf{X}) \times P(\mathbf{X})$$

47

## MODELO *NOISY-CHANNEL*

- Generalizando o modelo



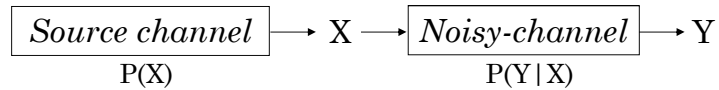
- Conjuntos  $P(\mathbf{X})$  e  $P(\mathbf{Y} | \mathbf{X})$  são os parâmetros do modelo
- $P(\mathbf{Y} | \mathbf{X})$ 
  - **História gerativa**
    - Como  $\mathbf{X}$  se transforma em  $\mathbf{Y}$
  - **Principal parte** do modelo, responsável por seu sucesso ou fracasso

48



## MODELO *NOISY-CHANNEL*

- Generalizando o modelo



### Transmissão de bits:

$P(X) \sim \text{uniforme}$  → pode ser ignorado, portanto  
 $P(0)=P(1)=0.5$

### $P(Y|X)$

$$P(0 \rightarrow 0) = P(0|0) = 0.6$$

$$P(0 \rightarrow 1) = P(1|0) = 0.4$$

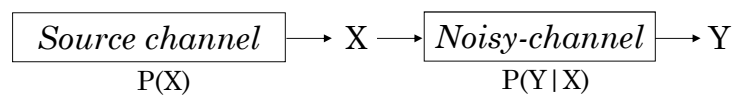
$$P(1 \rightarrow 1) = P(1|1) = 0.3$$

$$P(1 \rightarrow 0) = P(0|1) = 0.7$$

49

## MODELO *NOISY-CHANNEL*

- Generalizando o modelo

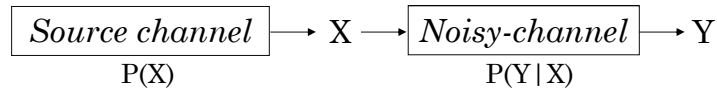


- O processo pode ser **tão complexo quanto se queira**
  - Dependente do **problema** modelado
  - Em vez de 1 bit, podem-se ter **bytes, sinais sonoros, palavras, sentenças, textos**, etc.
  - Em geral,  $P(X)$  **não segue distribuição uniforme**

50

## MODELO *NOISY-CHANNEL*

- Generalizando o modelo

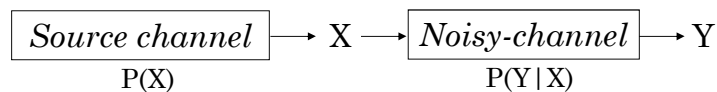


- O processo pode ser tão complexo quanto se queira
  - $P(Y|X)$  pode ser uma **composição de probabilidades** condicionais
    - No exemplo anterior: em vez de  $P(\text{bit } Y | \text{bit } X)$  ser simplesmente a probabilidade de um bit virar outro, poderia ser isso CONJUGADO à probabilidade de o receptor ter problemas técnicos/operacionais
      - $P(\text{bit } Y | \text{bit } X) = ?$

51

## MODELO *NOISY-CHANNEL*

- Generalizando o modelo

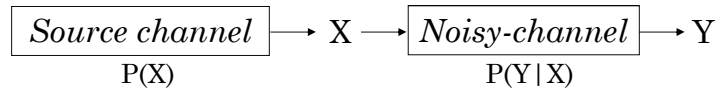


- O processo pode ser tão complexo quanto se queira
  - $P(Y|X)$  pode ser uma **composição de probabilidades** condicionais
    - No exemplo anterior: em vez de  $P(\text{bit } Y | \text{bit } X)$  ser simplesmente a probabilidade de um bit virar outro, poderia ser isso CONJUGADO à probabilidade de o receptor ter problemas técnicos/operacionais
      - $P(\text{bit } Y | \text{bit } X) = p_{\text{conversão\_bit}}(Y|X) * p_{\text{problema\_recepção}}(X)$

52

## MODELO *NOISY-CHANNEL*

- Generalizando o modelo

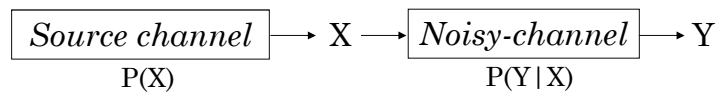


- O processo pode ser tão complexo quanto se queira
  - $P(Y|X)$  pode ser uma **composição de probabilidades** condicionais
    - No exemplo anterior: em vez de  $P(\text{bit } Y | \text{bit } X)$  ser simplesmente a probabilidade de um bit virar outro, poderia ser isso **CONJUGADO** à probabilidade de o receptor ter problemas técnicos/operacionais
      - $P(\text{bit } Y | \text{bit } X) = c(Y|X) * r(X)$

53

## MODELO *NOISY-CHANNEL*

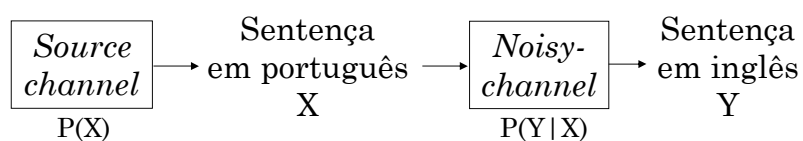
- Generalizando o modelo



Aplicação	Entrada (X)	Saída (Y)	P(X)	P(Y X)
Tradução Automática	Seqüência de palavras	Seqüência de palavras	Modelo de língua	Modelo de tradução
<i>Optical Character Recognition (OCR)</i>	Texto	Texto com erros	Prob. do texto	Modelo de erros de OCR
Reconhecimento de Fala	Seqüência de palavras	Sinal acústico	Prob. de seqüência de palavras	Modelo acústico

## TRADUÇÃO AUTOMÁTICA

- Tradução de uma sentença em inglês para português

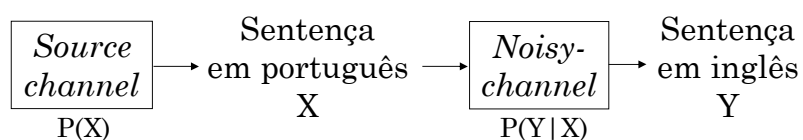


- Do que precisamos?

55

## TRADUÇÃO AUTOMÁTICA

- Tradução de uma sentença em inglês para português



- Do que precisamos?
  - Saber como calcular  $P(X)$  e  $P(Y|X)$

56

## P(Y | X)

### o História gerativa → modelo de tradução

- Como uma sentença se traduz na outra
  - o Por exemplo, palavras são traduzidas e depois reordenadas
    - o 2 parâmetros: tradução (t) e reordenação (r)

O cão preto morreu.

The black dog died.

$P(\text{tradução}) = t(\text{the} | \text{o}) \times t(\text{dog} | \text{cão}) \times t(\text{black} | \text{preto}) \times t(\text{died} | \text{morreu}) \times r(1 | 1) \times r(3 | 2) \times r(2 | 3) \times r(4 | 4)$

57

## P(X)

### o Modelo de língua

- Como prever a probabilidade de uma sentença traduzida a partir de “*The boy fell.*”?

P(O menino caiu.)?

P(O menino diminuiu.)?

P(O garoto caiu.)?

58

## P(X)

### o Modelo de língua baseado em n-gramas

- A probabilidade de uma sentença é a multiplicação da probabilidade de seus n-gramas (calculados a partir do conjunto de dados) ponderados

$$\begin{aligned}
 P(\text{O menino caiu.}) = & \\
 & \text{peso}_1 \times P(\text{O}) \times P(\text{menino}) \times P(\text{caiu}) \times P(.) + \\
 & \text{peso}_2 \times P(\text{O, menino}) \times P(\text{menino, caiu}) \times P(\text{caiu, .}) + \\
 & \text{peso}_3 \times P(\text{O, menino, caiu}) \times P(\text{menino, caiu, .}) + \\
 & \text{peso}_4 \times P(\text{O, menino, caiu, .})
 \end{aligned}$$

### o Distribuição uniforme

- Toda sentença é igualmente provável

59

## ENTROPIA

### o Preocupação de Shannon com a informação sendo veiculada em um canal

- Mais dados
  - o Mais longas são as mensagens
  - o Maior a probabilidade de erros

### o Questões

- Como medir a quantidade de informação?
- Como otimizar seu envio?

- o Entropia

60

## ENTROPIA

- **Entropia**: grau de desordem/surpresa de um conjunto de dados
  - Quanto menor a entropia, mais previsível e organizado é o conjunto de dados
    - Melhor para transmissão!

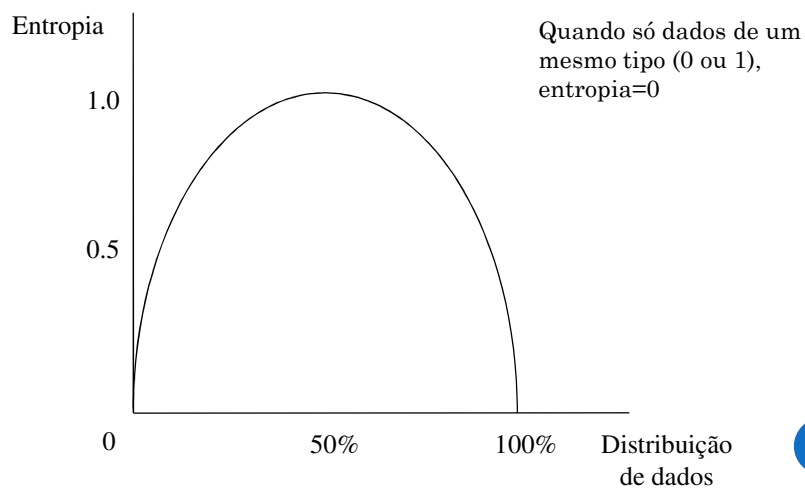
61

## ENTROPIA

- Originalmente, para calcular o **número de bits necessários** para a codificação de uma mensagem
  - Quanto menor a entropia, menos bits são necessários para codificar a mensagem
    - 1 bit: 0 ou 1 → 2 possibilidades
    - 2 bits: 00, 01, 10 ou 11 → 4 possibilidades
    - 3 bits: 000, 001, 010, 011, 100, 101, 110 ou 111 → 8 possibilidades
    - Etc.

62

## ENTROPIA



63

## ENTROPIA

- A entropia é 0 se todos os exemplos são do mesmo tipo
  - Uma seqüência de letras iguais tem entropia igual a 0  
→ não há surpresa, sabe-se o que esperar
- A entropia é 1 quando a coleção contém número igual de exemplos de cada tipo
  - Maior desordem possível
- Se a coleção contém número diferente de exemplos de cada tipo, a entropia varia entre 0 e 1
- Em geral, quanto menor a entropia de um fenômeno em PLN, teoricamente é “mais fácil automatizá-lo”

64



## ENTROPIA

- Genericamente, para qualquer número de tipos de exemplos de um conjunto de dados  $S$ , a **entropia** de  $S$  é dada pela fórmula

$$Entropia(S) = \sum_{i=1}^T -p_i * \log_2(p_i)$$

em que  $p_i$  é a **proporção de exemplos** de  $S$  pertencendo ao tipo  $i$  e  $T$  é o número total de tipos

- Por que esse “menos”? Por que  $\log_2$ ?



## ENTROPIA

- Exemplo: **língua polinésia simplificada**

- Letras dessa língua e suas frequências

<b>p</b>	<b>t</b>	<b>k</b>	<b>a</b>	<b>i</b>	<b>u</b>
1/8	1/4	1/8	1/4	1/8	1/8

- Entropia da língua

$$Entropia(S) = -1/8 * \log_2(1/8) - 1/4 * \log_2(1/4) - 1/8 * \log_2(1/8) - 1/4 * \log_2(1/4) - 1/8 * \log_2(1/8) - 1/8 * \log_2(1/8)$$

$$Entropia(S) = \mathbf{2,5 \text{ bits}}$$

<b>p</b>	<b>t</b>	<b>k</b>	<b>a</b>	<b>i</b>	<b>u</b>
100	00	101	01	110	111

## ENTROPIA

### o Exemplo: língua polinésia simplificada

- Letras dessa língua e suas frequências

p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8

- Entropia da língua

$$\text{Entropia}(S) = -1/8 \cdot \log_2(1/8) - 1/4 \cdot \log_2(1/4) - 1/8 \cdot \log_2(1/8) - 1/4 \cdot \log_2(1/4) - 1/8 \cdot \log_2(1/8)$$

$$\text{Entropia}(S) = \mathbf{2,5 \text{ bits}}$$

Menores códigos para letras mais frequentes

p	t	k	a	i	u
100	00	101	01	110	111

67

## ENTROPIA

### o Há diferentes formas de se calcular

- Por exemplo, para línguas, pode-se considerar a formação silábica em vez das letras

68