

POR QUE ESTUDAR PROBABILIDADE

- Probabilidade e língua
 - Probabilidade dos fenômenos lingüísticos
 - **Descrição**
 - 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.

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

EXEMPLO

○ 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.

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?
"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"

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"?

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"? $9/125 = 0.072 = 7.2\%$

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"?

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"? $4/125 = 0.032 = 3.2\%$

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?

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? $125/125 = 1 = 100\%$

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"?

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"? $4/125 + 1/125 = 5/125 = 0.04 = 4\%$

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 "que" seguida por um verbo?

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 "que" seguida por um verbo? $2/3 = 0.666 = 66.6\%$

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 um verbo precedido por “que”?

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 um verbo precedido por “que”? $2/15 = 0.133 = 13.3\%$

PROBABILIDADES

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

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?

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?
 - $P(\text{doença}|\text{sintoma})$
 - $P(\text{sintoma}|\text{doença})$



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”

BAYES

○ Teorema de Bayes

• Exemplo

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



BAYES

○ Teorema de Bayes

• Exemplo

- $P(\text{sarampo}|\text{dor de cabeça}) = P(\text{dor de cabeça}|\text{sarampo}) * \frac{P(\text{sarampo})}{\cancel{P(\text{dor de cabeça})}}$
- $P(\text{malária}|\text{dor de cabeça}) = P(\text{dor de cabeça}|\text{malária}) * \frac{P(\text{malária})}{\cancel{P(\text{dor de cabeça})}}$
- 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 $P(A|B) = P(B|A) * P(A)$
 - E se a probabilidade é importante?



EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que cataforas são raras: de todas as sentenças de um corpus, sabe-se que somente uma fração de 0.008 delas contém cataforas
- 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?

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que cataforas são raras: de todas as sentenças de um corpus, sabe-se que somente uma fração de 0.008 delas contém cataforas
- 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:

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

$$P(\text{catafóra}|\text{sim}) = P(\text{sim}|\text{catafóra}) \cdot P(\text{catafóra}) = 0.98 \cdot 0.008 = 0.0078$$

$$P(\text{sem catafóra}|\text{sim}) = P(\text{sim}|\text{sem catafóra}) \cdot P(\text{sem catafóra}) = 0.03 \cdot 0.992 = \mathbf{0.0298}$$

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um corpus, 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: $P(\text{catáfora}) = 0.008$ $P(\text{sem catáfora}) = 0.992$
 $P(\text{sim}|\text{catáfora}) = 0.98$ $P(\text{não}|\text{catáfora}) = 0.02$
 $P(\text{sim}|\text{sem catáfora}) = 0.03$ $P(\text{não}|\text{sem catáfora}) = 0.97$

$$P(\text{catáfora}|\text{sim}) = P(\text{sim}|\text{catáfora}) * P(\text{catáfora}) = 0.98 * 0.008 = 0.0078$$

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

O que aconteceu?

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um corpus, 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: $P(\text{catáfora}) = 0.008$ $P(\text{sem catáfora}) = 0.992$
 $P(\text{sim}|\text{catáfora}) = 0.98$ $P(\text{não}|\text{catáfora}) = 0.02$
 $P(\text{sim}|\text{sem catáfora}) = 0.03$ $P(\text{não}|\text{sem catáfora}) = 0.97$

$$P(\text{catáfora}|\text{sim}) = P(\text{sim}|\text{catáfora}) * P(\text{catáfora}) = 0.98 * 0.008 = 0.0078$$

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

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

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

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

$$P(\text{catafóra}|\text{sim}) = P(\text{sim}|\text{catafóra}) \cdot P(\text{catafóra}) = 0.98 \cdot 0.008 = 0.0078$$

E a probabilidade da sentença ser catafórica? Já conseguimos? $P(\text{sem catafóra}|\text{sim}) = 0.03 \cdot 0.992 = 0.0298$

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- 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
- Um sentenças foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

Normalizando...

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

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

DISTRIBUIÇÕES

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

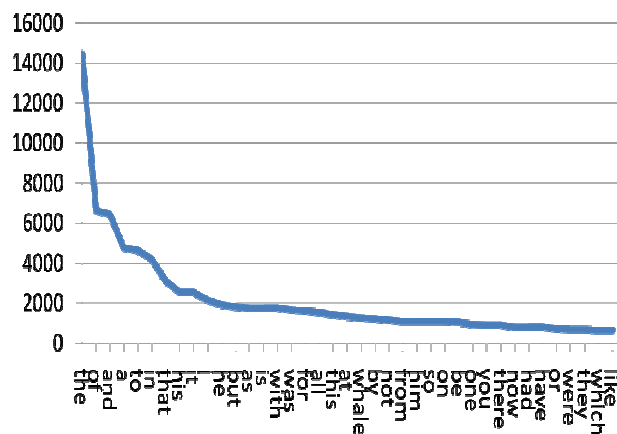


DISTRIBUIÇÕES

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

Lei de Zipf

→ distribuição zeta
(versão da distribuição
de Pareto)



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** (μ), **frequência**, **desvio padrão** (σ) ou **variância** (σ^2)

DISTRIBUIÇÃO BINOMIAL

- Discreta
- Eventos com duas possíveis saídas
 - Sim ou não, 0 ou 1, ocorre ou não ocorre, etc.
- Eventos independentes
- Muito apreciada em textos
 - Por exemplo
 - Frequência de uma palavra em um corpus
 - Porcentagem de sentenças em um corpus que têm um artigo definido ou qualquer outro fenômeno em particular
 - Quão comum um verbo é utilizado transitivamente

DISTRIBUIÇÃO BINOMIAL

- Número S de sucessos de N tentativas, com probabilidade P de sucesso em cada tentativa
 - $B(S; N, P)$

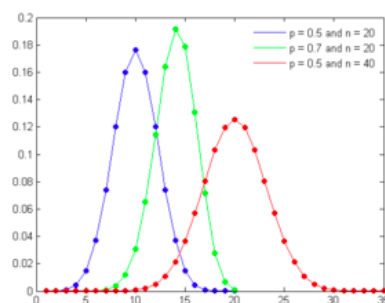
$$B(S; N, P) = \binom{N}{S} * P^S * (1 - P)^{N-S}$$

$$\binom{N}{S} = \frac{N!}{(N - R)! * R!}$$

- Média esperada = $N * P$
- Variância = $N * P * (1 - P)$
 - Desvio padrão = $\sqrt{\text{variância}}$

DISTRIBUIÇÃO BINOMIAL

- Exemplo
 - Jogando uma moeda “honesta”
 - $P=0.50$, pois mesma chance para cara ou coroa
 - Com $N=20$, espera-se que 10 sejam de um mesmo lado
 - Desbalanceamento é raro



DISTRIBUIÇÃO BINOMIAL

o Distribuição da palavra “Kennedy” no Brown Corpus

- N = número de palavras no cópús (1.000.000)
- P = chance de escolher uma palavra aleatoriamente e ela ser “Kennedy”
- Média esperada de ocorrências da palavra
 - o $N \cdot P$ = número de ocorrências de “Kennedy”
 - o $1.000.000 \cdot P = 140$
 - o $P = 140/1.000.000$

DISTRIBUIÇÃO BINOMIAL

o Distribuição da palavra “Kennedy” no Brown Corpus

- N = número de palavras no cópús (1.000.000)
- P = chance de escolher uma palavra aleatoriamente e ela ser “Kennedy”
- Média esperada de ocorrências da palavra
 - o $N \cdot P$ = número de ocorrências de “Kennedy”
 - o $1.000.000 \cdot P = 140$
 - o $P = 140/1.000.000$

o Evidência no cópús

- Divisão em 10 segmentos (cada um com N=100.000 palavras)
- Ocorrência de “Kennedy” em cada segmento: 58, 57, 2, 12, 6, 1, 4, 0, 0, 0
 - o Variância de 539 para esses 10 números
- Segundo a distribuição binomial
 - o Variância = $N \cdot P \cdot (1-P) = 100.000 \cdot 140/1.000.000 \cdot (1-140/1.000.000) \approx 14$
 - o $14 \neq 539!!!$
 - o O cópús não segue a distribuição binomial!
 - Por que?

DISTRIBUIÇÃO BINOMIAL

○ Distribuição da palavra “Kennedy” no Brown Corpus

- N = número de palavras no corpus (1.000.000)
- P = chance de escolher uma palavra aleatoriamente e ela ser “Kennedy”
- Média esperada de ocorrências da palavra
 - $N \cdot P$ = número de ocorrências de “Kennedy”
 - $1.000.000 \cdot P = 140$
 - $P = 140/1.000.000$

○ Evidência no corpus

- Divisão em 10 segmentos (cada um com N=100.000 palavras)
- Ocorrência de “Kennedy” em cada segmento: 58, 57, 2, 12, 6, 1, 4, 0, 0, 0
 - Variância de 539 para esses 10 números
- Segundo a distribuição binomial
 - Variância = $N \cdot P \cdot (1-P) = 100.000 \cdot 140/1.000.000 \cdot (1-140/1.000.000) \approx 14$

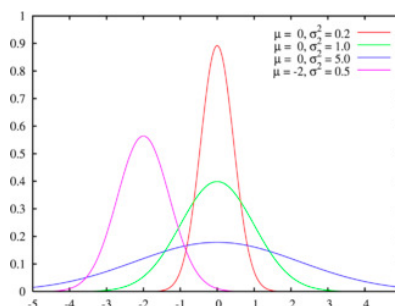
1. Ocorrência de palavras não é um evento independente
2. Uma das leis de Zipf: palavras de conteúdo tendem a se agrupar

DISTRIBUIÇÕES

○ Muitas outras (busquem algumas e suas aplicações)

- Discretas e contínuas

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



EXEMPLO DE USO DE ESTATÍSTICA

○ Análise sintática automática

- Gramáticas probabilísticas

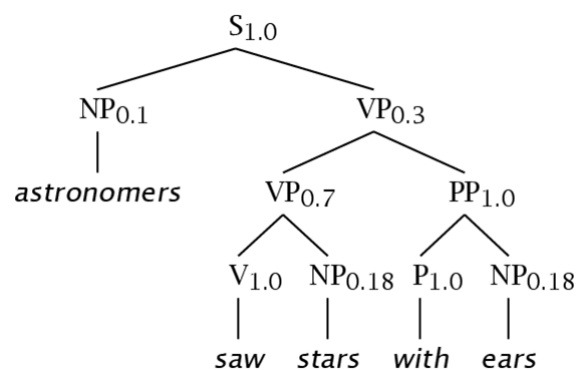
$S \rightarrow NP VP$	1.0	$V \rightarrow \textit{saw}$	1.0
$PP \rightarrow P NP$	1.0	$NP \rightarrow \textit{astronomers}$	0.1
$VP \rightarrow V NP$	0.7	$NP \rightarrow \textit{ears}$	0.18
$VP \rightarrow VP PP$	0.3	$NP \rightarrow \textit{saw}$	0.04
$NP \rightarrow NP PP$	0.4	$NP \rightarrow \textit{stars}$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

- De onde se conseguem as probabilidades de cada regra?



EXEMPLO DE USO DE ESTATÍSTICA

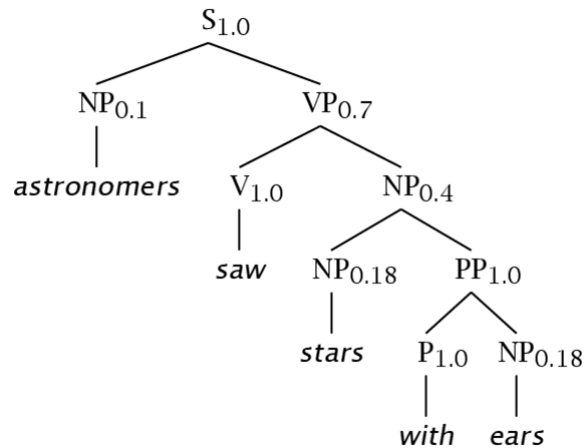
○ Possibilidade 1



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

EXEMPLO DE USO DE ESTATÍSTICA

○ Possibilidade 2

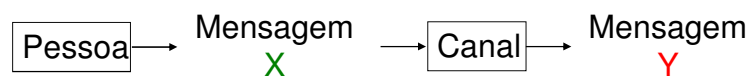


$$P(2) = 1.0 * 0.1 * 0.7 * 1.0 * 0.4 * 0.18 * 1.0 * 1.0 * 0.18 = \mathbf{0.0009072} (>\text{anterior})$$

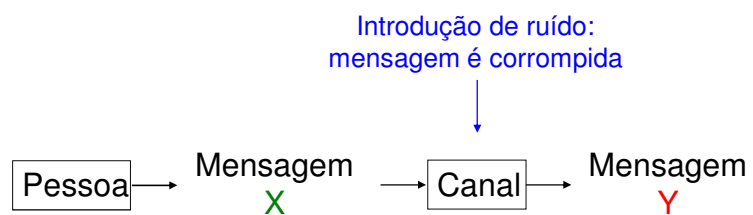
MODELO NOISY-CHANNEL

- 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

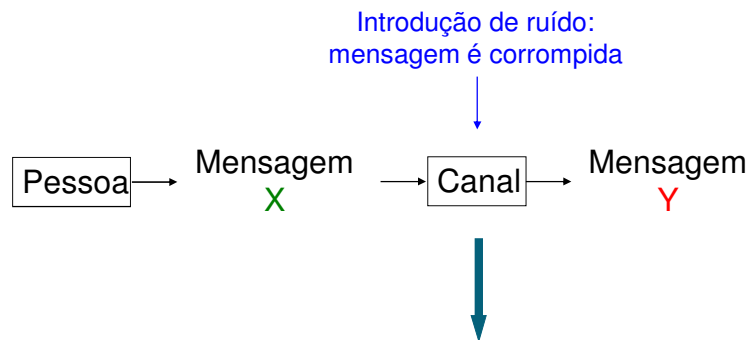
MODELO *NOISY-CHANNEL*



MODELO *NOISY-CHANNEL*

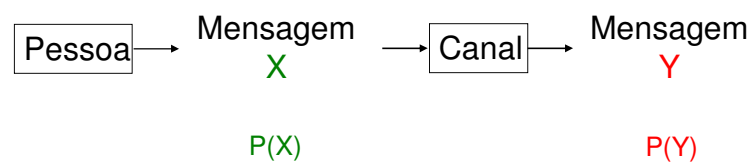


MODELO *NOISY-CHANNEL*

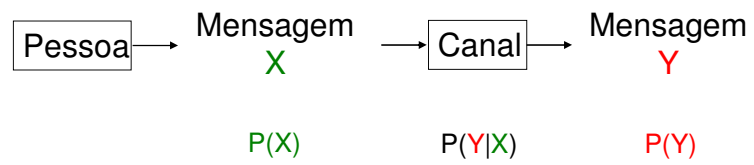


Quanta informação pode ser transmitida?
 Quanta informação pode ser transmitida para minimizar
 a ocorrência de ruídos?
 Como e que tipos de ruído ocorrem?
 Como recuperar a mensagem original X a partir de Y ?

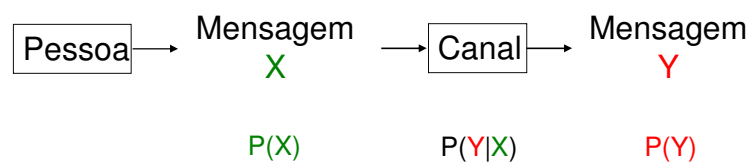
MODELO *NOISY-CHANNEL*



MODELO *NOISY-CHANNEL*

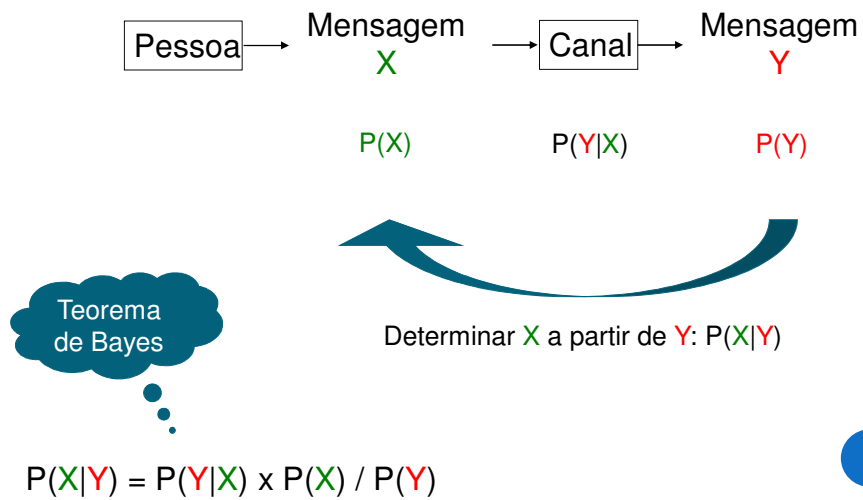


MODELO *NOISY-CHANNEL*



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

MODELO NOISY-CHANNEL



TEOREMA DE BAYES

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

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

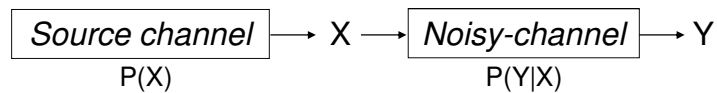
$$P(X|Y) = P(Y|X) \times \frac{P(X)}{\text{constante}}$$

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

Busca em um espaço de soluções

MODELO *NOISY-CHANNEL*

- Generalizando o modelo



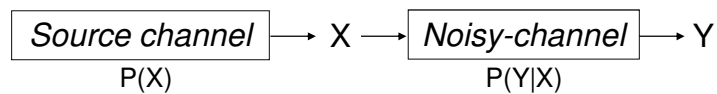
- Conjuntos $P(X)$ e $P(Y|X)$ são os parâmetros do modelo

- $P(Y|X)$

- História gerativa
 - Como X se transforma em Y
- Principal parte do modelo, responsável por seu sucesso ou fracasso

MODELO *NOISY-CHANNEL*

- Generalizando o modelo



Transmissão de bits:

$P(X)$ ~uniforme

$P(0)=P(1)=0.5$

$P(Y|X)$

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

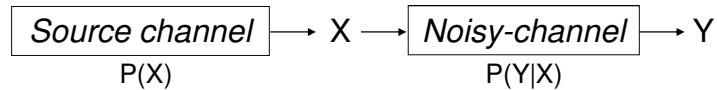
$P(0 \rightarrow 1) = P(1|0) = 0.4$

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

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

MODELO *NOISY-CHANNEL*

Generalizando o modelo



Transmissão de bits:

$P(X) \sim \text{uniforme}$
 $P(0)=P(1)=0.5$

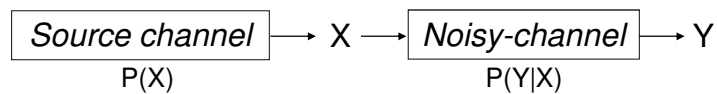
$P(Y|X)$

$P(0 \rightarrow 0) = P(0|0) = 0.6$
 $P(0 \rightarrow 1) = P(1|0) = 0.4$
 $P(1 \rightarrow 1) = P(1|1) = 0.3$
 $P(1 \rightarrow 0) = P(0|1) = 0.7$

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

MODELO *NOISY-CHANNEL*

Generalizando o modelo



Transmissão de bits:

$P(X) \sim \text{uniforme}$
 $P(0)=P(1)=0.5$

$P(Y|X)$

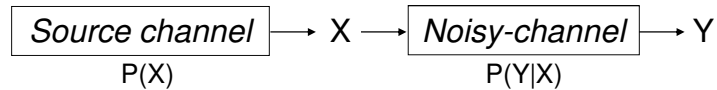
$P(0 \rightarrow 0) = P(0|0) = 0.6$
 $P(0 \rightarrow 1) = P(1|0) = 0.4$
 $P(1 \rightarrow 1) = P(1|1) = 0.3$
 $P(1 \rightarrow 0) = P(0|1) = 0.7$

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

$$Y=0 \rightarrow X=$$

MODELO *NOISY-CHANNEL*

Generalizando o modelo



Transmissão de bits:

$P(X) \sim \text{uniforme}$
 $P(0)=P(1)=0.5$

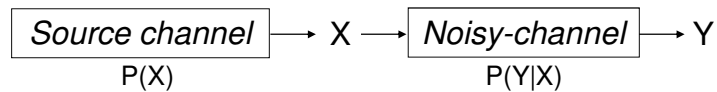
$P(Y|X)$
 $P(0 \rightarrow 0) = P(0|0) = 0.6$
 $P(0 \rightarrow 1) = P(1|0) = 0.4$
 $P(1 \rightarrow 1) = P(1|1) = 0.3$
 $P(1 \rightarrow 0) = P(0|1) = 0.7$

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

$$Y=0 \rightarrow X=0$$

MODELO *NOISY-CHANNEL*

Generalizando o modelo



Transmissão de bits:

$P(X) \sim \text{uniforme}$
 $P(0)=P(1)=0.5$

$P(Y|X)$
 $P(0 \rightarrow 0) = P(0|0) = 0.6$
 $P(0 \rightarrow 1) = P(1|0) = 0.4$
 $P(1 \rightarrow 1) = P(1|1) = 0.3$
 $P(1 \rightarrow 0) = P(0|1) = 0.7$

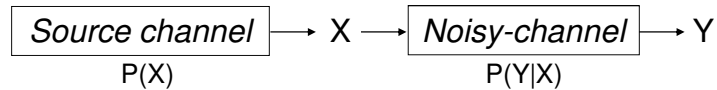
$$P(X|Y) = P(Y|X) \times P(X) = P(Y|X)$$

$$Y=0 \rightarrow X=0$$

$$Y=1 \rightarrow X=$$

MODELO *NOISY-CHANNEL*

Generalizando o modelo



Transmissão de bits:

$P(X) \sim \text{uniforme}$

$P(0)=P(1)=0.5$

$P(Y|X)$

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

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

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

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

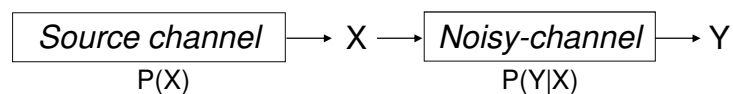
$$P(X|Y) = P(Y|X) \times P(X) = P(Y|X)$$

$Y=0 \rightarrow X=0$

$Y=1 \rightarrow X=0$

MODELO *NOISY-CHANNEL*

Generalizando o modelo

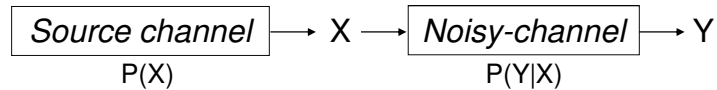


O processo pode ser tão complexo quanto se queira

- 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**

MODELO *NOISY-CHANNEL*

○ Generalizando o modelo

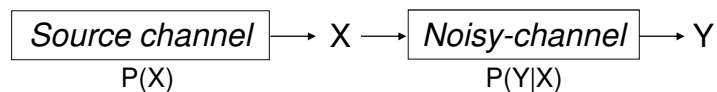


○ O processo pode ser tão complexo quanto se queira

- $P(Y|X)$ 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
 - $P(\text{bit } Y|\text{bit } X) = ?$

MODELO *NOISY-CHANNEL*

○ Generalizando o modelo

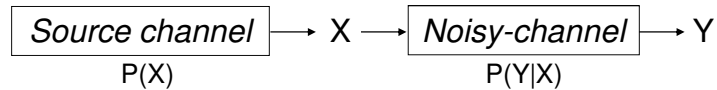


○ O processo pode ser tão complexo quanto se queira

- $P(Y|X)$ 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
 - $P(\text{bit } Y|\text{bit } X) = p_{\text{conversão_bit}}(Y|X) * p_{\text{problema_recepção}}(X)$

MODELO *NOISY-CHANNEL*

Generalizando o modelo

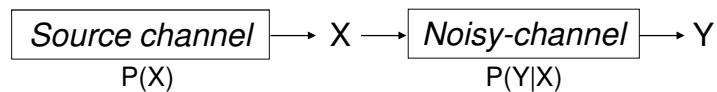


O processo pode ser tão complexo quanto se queira

- $P(Y|X)$ 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
 - $P(\text{bit } Y | \text{bit } X) = c(Y|X) * r(X)$

MODELO *NOISY-CHANNEL*

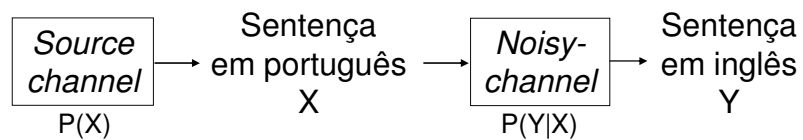
Generalizando o modelo



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

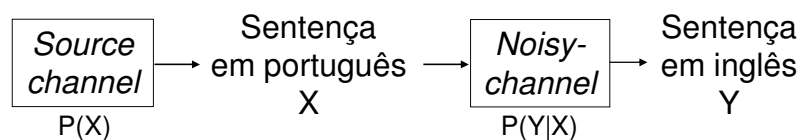
- Tradução de uma sentença em inglês para português



- Do que precisamos?

TRADUÇÃO AUTOMÁTICA

- Tradução de uma sentença em inglês para português



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

$P(Y|X)$

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.

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

$P(X)$

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 de dados)

$$\begin{aligned}
 P(\text{O menino caiu.}) = & \\
 & w_1 \times P(\text{O}) \times P(\text{menino}) \times P(\text{caiu}) \times P(.) + \\
 & w_2 \times P(\text{O, menino}) \times P(\text{menino, caiu}) \times P(\text{caiu, .}) + \\
 & w_3 \times P(\text{O, menino, caiu}) \times P(\text{menino, caiu, .}) + \\
 & w_4 \times P(\text{O, menino, caiu, .})
 \end{aligned}$$

o Distribuição uniforme

- Toda sentença é igualmente provável

ENTROPIA

- Preocupação de Shannon com a informação sendo veiculada em um canal
 - Mais dados
 - Mais longas são as mensagens
 - Maior a probabilidade de erros
 - Questões
 - Como medir a quantidade de informação?
 - Como otimizar seu envio?
- Entropia

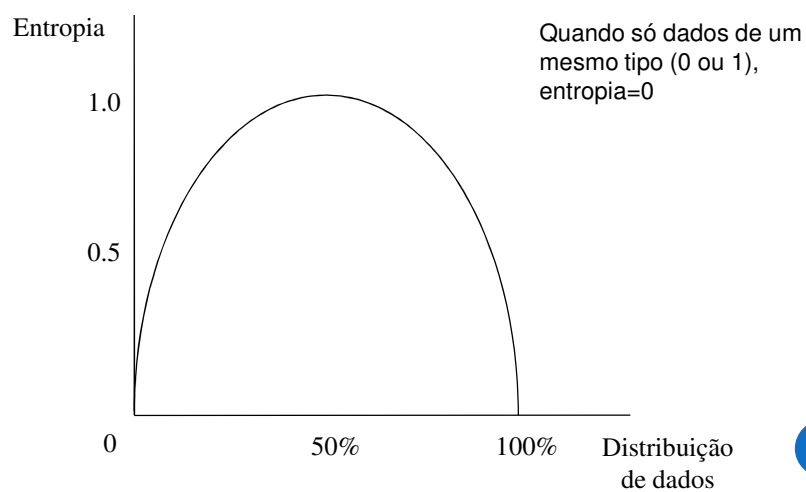
ENTROPIA

- 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!

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.

ENTROPIA



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”

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

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

- Entropia da língua

$$\begin{aligned} \text{Entropia}(S) &= -1/8 \log_2(1/8) - 1/4 \log_2(1/4) - 1/8 \log_2(1/8) \\ &\quad - 1/4 \log_2(1/4) - 1/8 \log_2(1/8) - 1/8 \log_2(1/8) \end{aligned}$$

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

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

ENTROPIA

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

$$\begin{aligned} \text{Entropia}(S) &= -1/8 \log_2(1/8) - 1/4 \log_2(1/4) - 1/8 \log_2(1/8) \\ &\quad - 1/4 \log_2(1/4) - 1/8 \log_2(1/8) - 1/8 \log_2(1/8) \end{aligned}$$

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

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

Menores códigos para letras mais frequentes

ENTROPIA

- Há diferentes formas de se calcular
 - Por exemplo, para línguas, pode-se considerar a formação silábica em vez das letras

