FUNDAMENTOS MATEMÁTICOS

1

SCC5908 Introdução ao Processamento de Língua Natural

Thiago A. S. Pardo

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.

2

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

3

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.

4

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.

5

Qual a probabilidade da palavra "que" ocorrer?
"chance"

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.

6

Qual a probabilidade da palavra "que" ocorrer? 3/125 = 0.024 = 2.4%
"chance"

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.

7

E "de"?

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.

8

E "de"? 9/125 = 0.072 = 7.2%

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.

9

E "incêndio"?

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.

10

E "incêndio"? 4/125 = 0.032 = 3.2%

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.

11

E qualquer palavra do texto?

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.

12

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

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.

13

E "incêndio" ou "fogo"?

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.

14

E "incêndio" ou "fogo"? 4/125 + 1/125 = 5/125 = 0.04 = 4%

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.

15

E "que" seguida por um verbo?

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.

16

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

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.

17

E um verbo precedido por "que"?

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.

18

E um verbo precedido por "que"? 2/15 = 0.133 = 13.3%

PROBABILIDADES

- Probabilidade: resultado entre 0 e 1, ou 0 e 100%
- P(evento impossível) = 0
- P(qualquer coisa) = 1 (ou 100%)
- P(A) ou P(B) = P(A) + P(B)
  - P(qualquer coisa) = P(uma coisa) + P(segunda coisa) + ... + P(enésima coisa)
- Probabilidade condicional $P(A \mid B) = P(A \cap B) / P(B)$
- $P(A \cap B) = P(B) * P(A \mid B) = P(A) * P(B \mid A)$
  - $P(A \cap B) = P(A) * P(B)$, se eventos independentes
  - $P(A_1 \cap ... \cap A_n) = P(A_1) * P(A_2 \mid A_1) * P(A_3 \mid A_1 \cap A_2) ...$

19

BAYES

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

20

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(doença | sintoma)
        - P(sintoma | doença)

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

P(doença | sintoma) = P(sintoma | doença) * P(doença) / P(sintoma)

- o sintoma é o que se observa, e a doença é o que se quer descobrir
  - P(doença | sintoma)
- ... mas quem causa o sintoma é a doença, e não o inverso
  - P(sintoma | doença)
- P(doença | sintoma) pode ser "tendencioso" e "temporal"

22

BAYES

- Teorema de Bayes
  - Exemplo
    - P(sarampo | dor de cabeça) = P(dor de cabeça | sarampo) * P(sarampo) / P(dor de cabeça)
    - P(malária | dor de cabeça) = P(dor de cabeça | malária) * P(malária) / P(dor de cabeça)
      - A maior probabilidade ganha e indica o diagnóstico final!
        - Atenção: P(dor de cabeça) é constante. Faz diferença no resultado?

23

BAYES

- Teorema de Bayes
  - Exemplo
    - P(sarampo | dor de cabeça) = P(dor de cabeça | sarampo) * P(sarampo) ~~/ P(dor de cabeça)~~
    - P(malária | dor de cabeça) = P(dor de cabeça | malária) * P(malária) ~~/ P(dor de cabeça)~~
      - A maior probabilidade ganha e indica o diagnóstico final!
        - Atenção: P(dor de cabeça) é constante. Faz diferença no resultado?
          - Ao comparar hipóteses, pode-se usar P(A | B) = P(B | A) * P(A)

24

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm 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
- Uma sentença foi rotulada como catafórica pelo sistema. É possível afirmar que ela é catafórica? Qual a probabilidade de ela ser catafórica de fato?

25

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm 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(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim | catáfora) = 0.98 P(não | catáfora) = 0.02

P(sim | sem catáfora) = 0.03 P(não | sem catáfora) = 0.97

P(catáfora | sim) = P(sim | catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora | sim) = P(sim | sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

26

O que aconteceu?

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm 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(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim | catáfora) = 0.98 P(não | catáfora) = 0.02

P(sim | sem catáfora) = 0.03 P(não | sem catáfora) = 0.97

P(catáfora | sim) = P(sim | catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora | sim) = P(sim | sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

27

E se a probabilidade de ocorrência de catáforas fosse **uniforme** no córpus?

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm 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(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim | catáfora) = 0.98 P(não | catáfora) = 0.02

P(sim | sem catáfora) = 0.03 P(não | sem catáfora) = 0.97

P(catáfora | sim) = P(sim | catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora | sim) = P(sim | sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

28

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

E a probabilidade da sentença ser catafórica? Já conseguimos?

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm 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(catáfora) = 0.008 P(sem catáfora) = 0.992

P(sim | catáfora) = 0.98 P(não | catáfora) = 0.02

P(sim | sem catáfora) = 0.03 P(não | sem catáfora) = 0.97

P(catáfora | sim) = P(sim | catáfora)*P(catáfora) = 0.98*0.008 = 0.0078

P(sem catáfora | sim) = P(sim | sem catáfora)*P(sem catáfora) = 0.03*0.992 = **0.0298**

29

EXERCÍCIO (EM GRUPOS DE 2 ALUNOS)

- Sabe-se que catáforas são raras: de todas as sentenças de um córpus, sabe-se que somente uma fração de 0.008 delas contêm 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?

Normalizando...

P(catáfora | sim) = 0.0078 → 0.0078 / (0.0078+0.0298) = **0.21 = 21%**

P(sem catáfora | sim) = 0.0298 → 0.0298 / (0.0078+0.0298) = **0.79 = 79%**

30

DISTRIBUIÇÕES

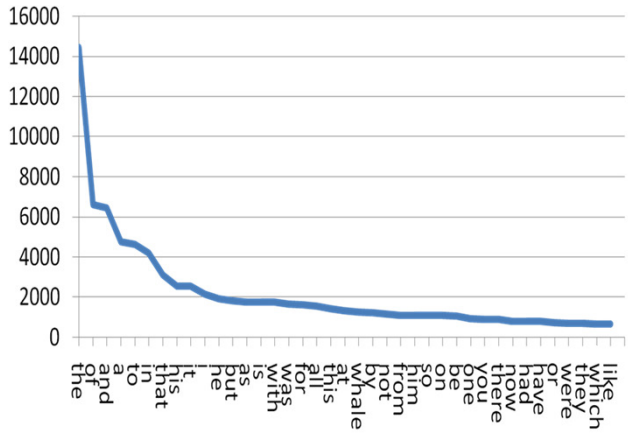
- Os dados, em geral, seguem determinados padrões
  - Comportamentos
    - Exemplo?

31

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

33

DISTRIBUIÇÃO BINOMIAL

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

34

DISTRIBUIÇÃO BINOMIAL

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

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

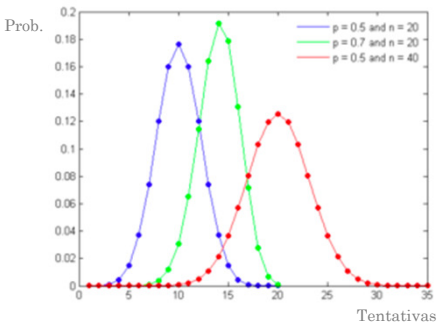
$$\binom{N}{S} = \frac{N!}{(N-S)! * S!}$$

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

35

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



36

DISTRIBUIÇÃO BINOMIAL

- Distribuição da palavra "Kennedy" no Brown Corpus
  - N = número de palavras no córpus (1.000.000)
  - P = chance de escolher uma palavra aleatoriamente e ela ser "Kennedy"
  - Média esperada de ocorrências da palavra

    N*P = número de ocorrências de "Kennedy"

    1.000.000*P = 140

    P = 140/1.000.000

37

DISTRIBUIÇÃO BINOMIAL

- Distribuição da palavra "Kennedy" no Brown Corpus
  - N = número de palavras no córpus (1.000.000)
  - P = chance de escolher uma palavra aleatoriamente e ela ser "Kennedy"
  - Média esperada de ocorrências da palavra

    N*P = número de ocorrências de "Kennedy"

    1.000.000*P = 140

    P = 140/1.000.000
- Evidência no córpus
  - 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
    - Desvio padrão de 23 e variância de 539 para esses 10 números
  - Segundo a distribuição binomial
    - Variância = N*P*(1-P) = 100.000*140/1.000.000*(1-140/1.000.000) ≈ 14
    - $14 \neq 539$!!!
    - O córpus não segue a distribuição binomial!
      - Por que?

38

DISTRIBUIÇÃO BINOMIAL

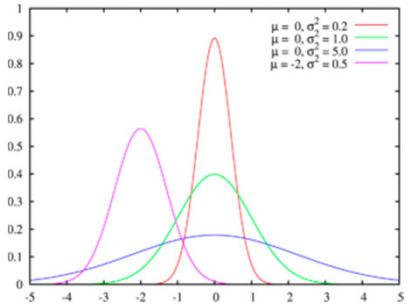
- Distribuição da palavra "Kennedy" no Brown Corpus
  - N = número de palavras no córpus (1.000.000)
  - P = chance de escolher uma palavra aleatoriamente e ela ser "Kennedy"
  - Média esperada de ocorrências da palavra
    N*P = número de ocorrências de "Kennedy"
    1.000.000*P = 140
    P = 140/1.000.000
- Evidência no córpus
  - 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
    - Desvio padrão de 23 e variância de 539 para esses 10 números
  - Segundo a distribuição binomial
    - Variância = N*P*(1-P) = 100.000*140/1.000.000*(1-140/1.000.000) ≈ 14

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

39

DISTRIBUIÇÕES

- Muitas outras (busquem algumas e suas aplicações)
  - Discretas e contínuas
    - Poisson
    - Geométrica
    - Uniforme
    - **Normal (ou gaussiana)**
      - Contínua
    - Weibull
    - Pareto
    - Etc.



40

EXEMPLO DE USO DE ESTATÍSTICA

- Análise sintática automática
  - Gramáticas probabilísticas

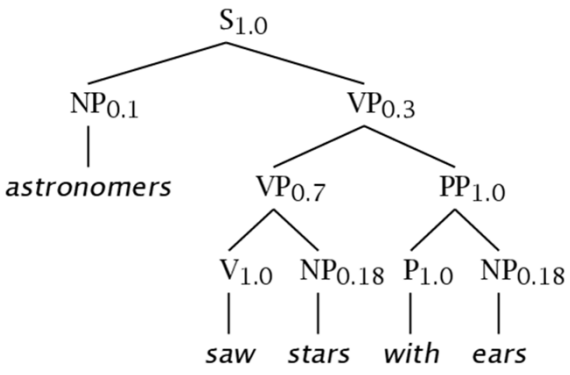
S → NP VP	1.0	V → *saw*	1.0
PP → P NP	1.0	NP → *astronomers*	0.1
VP → V NP	0.7	NP → *ears*	0.18
VP → VP PP	0.3	NP → *saw*	0.04
NP → NP PP	0.4	NP → *stars*	0.18
P → *with*	1.0	NP → *telescopes*	0.1

  - De onde se conseguem as probabilidades de cada regra?

41

EXEMPLO DE USO DE ESTATÍSTICA

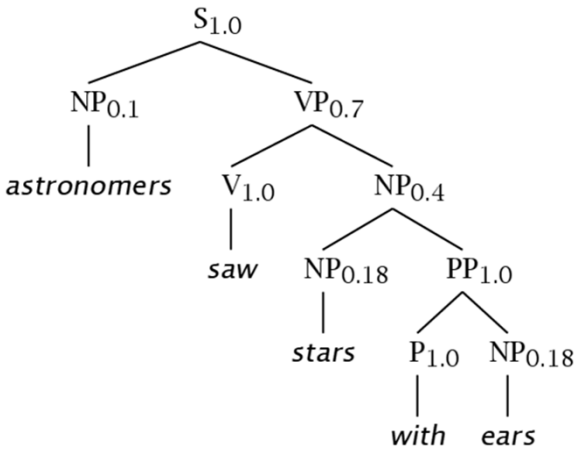
- Possibilidade 1



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



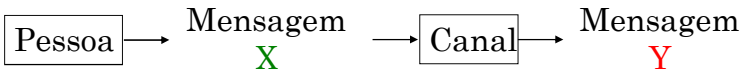
P(2) = 1.0 * 0.1 * 0.7 * 1.0 * 0.4 * 0.18 * 1.0 * 1.0 * 0.18 = **0.0009072** (>anterior)

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

44

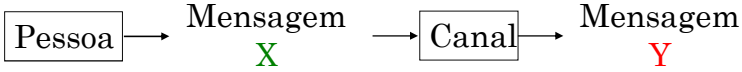
MODELO *NOISY-CHANNEL*



45

MODELO *NOISY-CHANNEL*

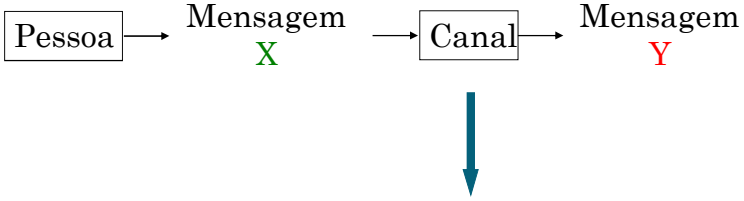
Introdução de ruído:
mensagem é corrompida



46

MODELO *NOISY-CHANNEL*

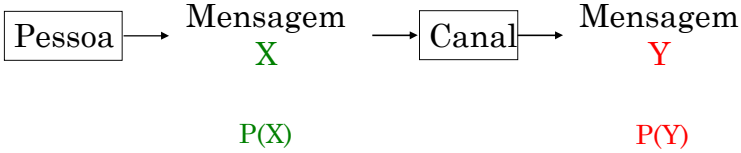
Introdução de ruído:
mensagem é corrompida



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?

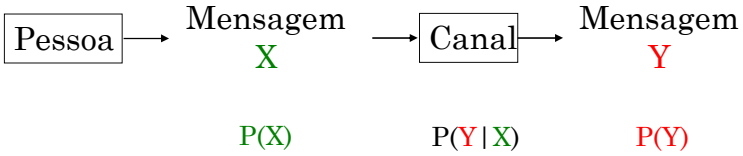
47

MODELO *NOISY-CHANNEL*



48

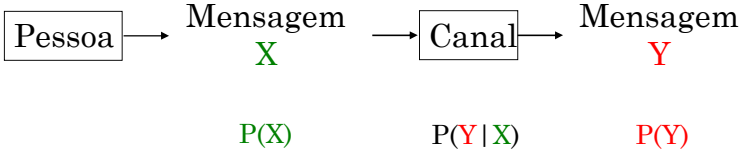
MODELO *NOISY-CHANNEL*



P(X) P(Y | X) P(Y)

49

MODELO *NOISY-CHANNEL*

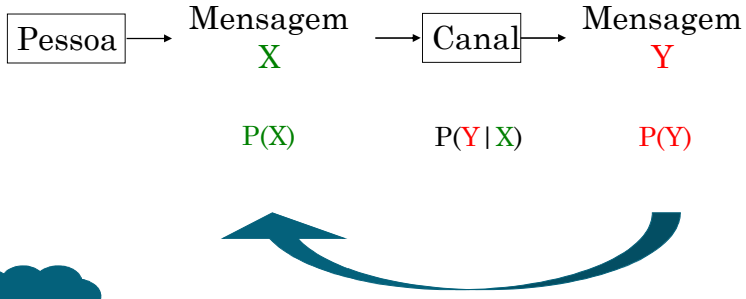


P(X) P(Y | X) P(Y)

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

50

MODELO *NOISY-CHANNEL*



Teorema de Bayes

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

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

51

TEOREMA DE BAYES

Busca em um espaço de soluções

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

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

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

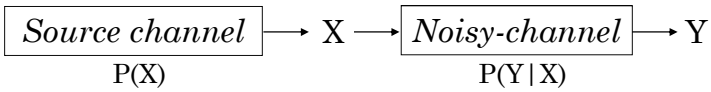
constante

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

52

MODELO *NOISY-CHANNEL*

- Generalizando o modelo

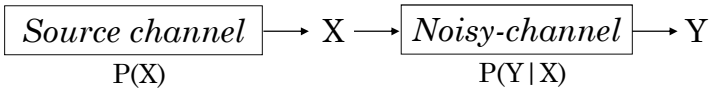


- Conjuntos P(X) e P(Y | X) são os parâmetros do modelo
- P(Y | X)
  - História gerativa
    - Como X se transforma em Y
  - Principal parte do modelo, responsável por seu sucesso ou fracasso

53

MODELO *NOISY-CHANNEL*

- Generalizando o modelo



Transmissão de bits:

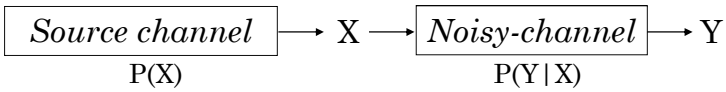
P(X) ~uniforme → pode ser ignorado, portanto
P(0)=P(1)=0.5

P(Y | X)
P(0 → 0) = P(0 | 0) = 0.6
P(0 → 1) = P(1 | 0) = 0.4
P(1 → 1) = P(1 | 1) = 0.3
P(1 → 0) = P(0 | 1) = 0.7

54

MODELO *NOISY-CHANNEL*

- Generalizando o modelo

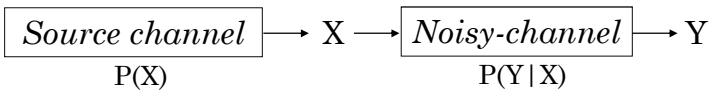


- O processo pode ser tão complexo quanto se queira
  - Dependente do problema modelado
  - Em vez de 1 bit, podem-se ter bytes, sinais sonoros, palavras, sentenças, textos, etc.
  - Em geral, P(X) não segue distribuição uniforme

55

MODELO *NOISY-CHANNEL*

- Generalizando o modelo

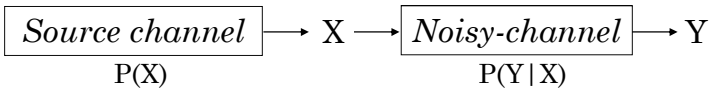


- O processo pode ser tão complexo quanto se queira
  - P(Y | X) pode ser uma composição de probabilidades condicionais
    - No exemplo anterior: em vez de P(bit Y | bit X) ser simplesmente a probabilidade de um bit virar outro, poderia ser isso CONJUGADO à probabilidade de o receptor ter problemas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

56

MODELO *NOISY-CHANNEL*

- Generalizando o modelo

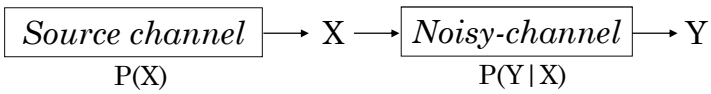


- O processo pode ser tão complexo quanto se queira
  - P(Y | X) pode ser uma composição de probabilidades condicionais
    - No exemplo anterior: em vez de P(bit Y | bit X) ser simplesmente a probabilidade de um bit virar outro, poderia ser isso CONJUGADO à probabilidade de o receptor ter problemas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

57

MODELO *NOISY-CHANNEL*

- Generalizando o modelo

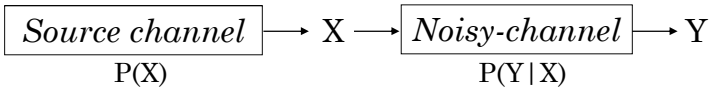


- O processo pode ser tão complexo quanto se queira
  - P(Y | X) pode ser uma composição de probabilidades condicionais
    - No exemplo anterior: em vez de P(bit Y | bit X) ser simplesmente a probabilidade de um bit virar outro, poderia ser isso CONJUGADO à probabilidade de o receptor ter problemas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

58

MODELO *NOISY-CHANNEL*

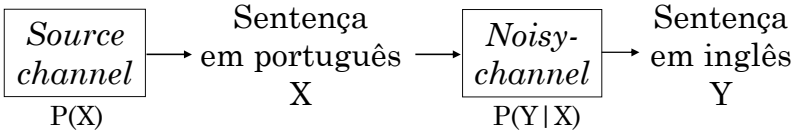
- Generalizando o modelo



Aplicação	Entrada (X)	Saída (Y)	P(X)	P(Y|X)
Tradução Automática	Seqüência de palavras	Seqüência de palavras	Modelo de língua	Modelo de tradução
Optical Character Recognition (OCR)	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

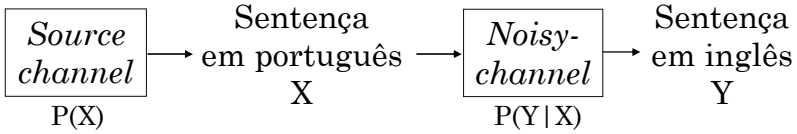


- Do que precisamos?

60

TRADUÇÃO AUTOMÁTICA

- Tradução de uma sentença em inglês para português



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

61

P(Y | X)

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

O cão preto morreu.

The black dog died.

P(tradução) = t(the | o) x t(dog | cão) x t(black | preto) x t(died | morreu) x r(1 | 1) x r(2 | 3) x r(3 | 2) x r(4 | 4)

62

P(X)

- 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

P(O menino caiu.) =
$peso_1$ x P(O) x P(menino) x P(caiu) x P(.) +
$peso_2$ x P(O,menino) x P(menino,caiu) x P(caiu,.) +
$peso_3$ x P(O,menino,caiu) x P(menino,caiu,.) +
$peso_4$ x P(O,menino,caiu,.)

- Distribuição uniforme
  - Toda sentença é igualmente provável

63

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

64

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!

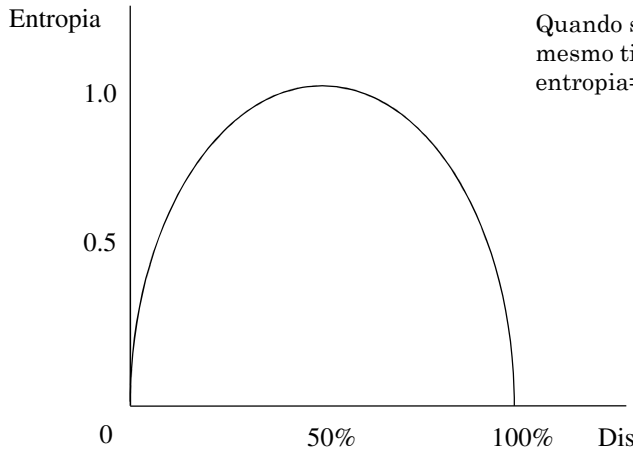
65

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.

66

ENTROPIA



67

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"

68

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 freqüências

p	t	k	a	i	u
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) = **2,5 bits**

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

70

ENTROPIA

- Exemplo: língua polinésia simplificada
  - Letras dessa língua e suas freqüências

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

  - Entropia da língua

Entropia(S) = $-1/8*\log_2(1/8)$ $-1/4*\log_2(1/4)$ $-1/8*\log_2(1/8)$
$-1/4*\log_2(1/4)$ $-1/8*\log_2(1/8)$
Entropia(S) = **2,5 bits**

Menores códigos para letras mais freqüentes

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

71

ENTROPIA

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

72