PROCESSAMENTO DE LINGUAGEM NATURAL

Modelos n-gramas

TÓPICOS

1. Introdução
2. Modelos de linguagem
3. N-gramas

LINGUAGENS

- O que é a linguagem?

“Sistema de símbolos de um vocabulário que, quando colocados numa determinada ordem e expressos num determinado contexto, emitem um significado.”

ORDEM DAS PALAVRAS

- O que define a ordem das palavras?
  - Cada macaco no seu _____
    - galho
  - Macaco cada no galho seu X Cada macaco no seu galho

- Como ensinar o computador a definir qual é a próxima palavra ou a ordem correta delas?

ORDEM DAS PALAVRAS

- Geralmente definida com base na probabilidade de ocorrência
  - Cada macaco no seu galho
    - P(galho | cada macaco no seu)
  - Macaco cada no galho seu X Cada macaco no seu galho
    - P(macaco cada no galho seu) < P(cada macaco no seu galho)

ORDEM DAS PALAVRAS

- **Aplicações de geração de texto**
  - **Sequência já gerada: Cada macaco no seu _____**

Próxima palavra	Probabilidade
galinheiro	0,0003
anzol	0,000002
galho	0,004
toca	0,00005

ORDEM DAS PALAVRAS

- **Aplicações de tradução**
  - **Probabilidade de uma sentença**

Possíveis sentenças	Probabilidade
Macaco cada no galho seu	0,00003
Cada macaco no seu galho	0,0005
Galho no seu cada macaco	0,000004

ORDEM DAS PALAVRAS

- **Aplicações de correção**
  - **Probabilidade de uma sentença corrigida**
    - **Sentença original: saudade corta como aço de navaia**
    - **Opções de correção: navalha ou navais**

Possíveis sentenças	Probabilidade
saudade corta como aço de navalha	0,005
saudade corta como aço de navais	0,00002

TÓPICOS

1. Introdução
2. Modelos de linguagem
3. N-gramas

MODELO DE LINGUAGEM

- Modelo computacional que infere a probabilidade de uma sequência de palavras
- Usado para prever
  - A próxima palavra dada uma sequência
  - P(w5 | w1, w2, w3, w4)
  - A ocorrência de uma sequência de palavras
  - P(w1, w2, w3, w4, w5)

MODELO DE LINGUAGEM

- **Regra da cadeia**
  - **A probabilidade de uma sentença pode ser estimada pela multiplicação das probabilidades de cada palavra, estimada com base nas palavras anteriores**

P(w1, w2, w3, w4, w5) = P(w1) x P(w2|w1) x P(w3|w1 , w2) x P(w4|w1, w2, w3) x P(w5|w1, w2, w3, w4)

P(Cada macaco no seu galho) = P(Cada) x P(macaco|Cada) x P(no|Cada macaco) x P(seu|Cada macaco no) x P(galho|Cada macaco no seu)

MODELO DE LINGUAGEM

- **Estimando probabilidades**
  - **A partir de um grande *corpus***
  - **Com base nas contagens das sequências**

$$P(\text{galho}|\text{Cada macaco no seu}) = \frac{\text{freq}(\text{Cada macaco no seu galho})}{\text{freq}(\text{Cada macaco no seu})}$$

  - **Mas não é necessário "olhar" para toda a sequência**

P(<s> Cada macaco no seu galho </s>) = P(Cada|<s>) x P(macaco|Cada) x P(no|~~Cada~~ macaco) x P(seu|~~Cada macaco~~ no) x P(galho|~~Cada macaco no~~ seu) x P(</s>|~~Cada macaco no seu~~ galho)

TÓPICOS

1. Introdução
2. Modelos de linguagem
3. N-gramas

N-GRAMA

- Sequência de *n* palavras (*tokens*)
  - unigrama = 1 palavra (*token*)
    - <S>, Cada, macaco, no, seu, galho, </S>
  - bigrama = 2 palavras (*tokens*)
    - <S> Cada, Cada macaco, macaco no, no seu, seu galho, galho </S>
  - trigrama = 3 palavras (*tokens*)
    - <S> Cada macaco, Cada macaco no, macaco no seu, no seu galho, seu galho </S>

N-GRAMA

- **Treinamento de um modelo de linguagem**

Colab - NLTK

$$P(\text{galho}|\ \text{seu}) = \frac{\text{freq(seu galho)}}{\text{freq(seu)}}$$

Modelo de bigrama

$$P(\text{galho}|\text{no seu}) = \frac{\text{freq(no seu galho)}}{\text{freq(no seu)}}$$

Modelo de trigrama

OUTROS “N-GRAMAS”

- Expressões multipalavras
  - Combinações de palavras que
    - juntas representam algo mais do que a simples composição das suas ideias
    - não podem ter partes substituídas por sinônimos
  - Dica: http://mwetoolkit.sourceforge.net/

- Entidades nomeadas
  - Uma ou mais palavras que
    - têm um papel no mundo
  - Dica: BERT para NER

O QUE VIMOS?

- Introdução
- Modelos de linguagem
- N-gramas

PRÓXIMA VIDEOAULA

- Representação vetorial de textos

REFERÊNCIAS

- Curso de Processamento de Linguagem Natural
  - Profa. Helena Caseli (UFSCar)
- Curso de Processamento de Linguagem Natural
  - Prof. Thiago Pardo (ICMC-USP)
- Curso de Linguística Computacional
  - Prof. Thiago Castro Ferreira (UFMG)