Processamento de Linguagem Natural

Modelos de Linguagem

Yuri Malheiros (yuri@ci.ufpb.br)

Introdução

- É possível prever quais as próximas palavras que alguém vai falar?
  - "Ao sair de casa desligue a _____”

Introdução

- É possível prever quais as próximas palavras que alguém vai falar?
  - "Ao sair de casa desligue a _____”

- É provável que a palavra seja “luz" ou “televisão"
- É bem menos provável que a palavra seja “geladeira" ou “bola"

Introdução

- Qual a frase é mais provável?
  - “Eu acendi o fósforo”
  - “Eu ascendi o fósforo”

Modelos de Linguagem

- Modelos que atribuem probabilidade a palavras ou sequências de palavras são chamadas de modelos de linguagem
- Probabilidade de uma palavra: $P(w_1)$
- Probabilidade de uma sequência de palavras: $P(w_1, w_2, w_3, ..., w_n)$
- Probabilidade da próxima palavra: $P(w_n|w_1, w_2, ..., w_{n-1})$

Modelos de Linguagem

- As probabilidades são calculadas a partir do processamento de uma grande massa de dados textuais

- As probabilidades carregam muito conhecimento

Modelos de Linguagem

- Qual frase está correta: “ele é maior" ou “ele é mais grande" ?
- Podemos descobrir calculando a probabilidade de cada uma dessas frases
- P(ele é maior) > P(ele é mais grande)

Modelos de Linguagem

- Complete a frase: “A Terra é _____”

Modelos de Linguagem

- Complete a frase: “A Terra é _____”

- Probabilidades altas:
  - P(azul|A Terra é)
  - P(redonda|A Terra é)

Modelos de Linguagem

- Complete a frase: “A Terra é _____”

- Probabilidades altas:
  - P(azul|A Terra é)
  - P(redonda|A Terra é)

- Probabilidades baixas:
  - P(amarela|A Terra é)
  - P(banana|A Terra é)

Modelos de Linguagem

- Como essas probabilidades são calculadas?
- Vamos aprender como funciona um modelo que atribui probabilidades para sequências de palavras
- As sequências são chamadas de **n-gramas**
  - **n** é o tamanho da sequência
  - 2-gramas (bigramas) é uma sequência de duas palavras
  - 3-gramas (trigramas) é uma sequência de três palavras

N-gramas

- Como calcular a probabilidade de uma palavra dado um contexto (palavras anteriores)?
  - $P(w \mid h)$

N-gramas

- P(transparente|a água é tão)

- Precisamos de um corpus grande
- Contamos o número de vezes que a sequência “a água é tão” aparece
- Contamos o número de vezes que a sequência “a água é tão transparente” aparece
- Dividimos a segunda pela primeira

N-gramas

- $P(\text{transparente}|\text{a água é tão}) = \frac{\text{Cont(a água é tão transparente)}}{\text{Cont(a água é tão)}}$

N-gramas

- Esse método pode funcionar em alguns casos
- Mas existem tantas possibilidades de construção de texto, que mesmo um corpus muito grande pode não ser suficiente
  - Frases novas e criativas podem ser escritas
  - Pode ser que uma sequência de palavras não exista no corpus
    - Cont(a água do sítio da minha avó Severina no sertão da Paraíba)

N-gramas

- Nós podemos computar a probabilidade de uma sequência de palavras assim:
  - $P(w_1, w_2, w_3, w_4) = P(w_1) * P(w_2|w_1) * P(w_3|w_1,w_2) * P(w_4|w_1,w_2,w_3)$
- Por exemplo:
  - P("penso logo existo") = P(penso) * P(logo|penso) * P(existo|penso logo)

N-gramas

- P("a necessidade é a mãe da invenção") =
  P(a) * P(necessidade|a) * P(é|a necessidade) * P(a|a necessidade é) *
  P(mãe|a necessidade é a) * P(da|a necessidade é a mãe) *
  P(invenção|a necessidade é a mãe da)

- Muito complexo para sequências grandes
- Muitas possibilidades para conseguir calcular as probabilidades

N-gramas

- Usando um modelo n-gramas vamos simplificar o cálculo das probabilidades
- Ao invés de condicionar o cálculo da probabilidade a todas as palavras anteriores, vamos fazer uma aproximação
- Usaremos apenas um número restrito de palavras anteriores

N-gramas

- Usando um modelo bigrama, aproximaremos a probabilidade de uma palavra dado o seu histórico, usando apenas uma palavra anterior
- Ao invés de:
  - P(invenção|a necessidade é a mãe da)
- Usaremos
  - P(invenção|da)
- Ou seja
  - P(invenção|a necessidade é a mãe da) $\approx$ P(invenção|da)

N-gramas

- A suposição de que a probabilidade de uma palavra depende apenas de palavras anteriores próximas é chamada de suposição de Markov

- Utilizando o modelo bigrama calculamos P("a necessidade é a mãe da invenção”) assim:
  - P(a) * P(necessidade|a) * P(é|necessidade) * P(a|é) * P(mãe|a) * P(da|mãe) * P(invenção|da)

N-gramas

- Para calcular P(w_2|w_1), temos:

$$P(w_2 \mid w_1) = \frac{cont(w_1, w_2)}{cont(w_1)}$$

- $cont(w_1, w_2)$ é a quantidade de vezes que o par w_1 w_2 aparece
- $cont(w_1)$ é a quantidade de vezes que w_1 aparece

N-gramas

- Para um modelo trigrama, temos:
  - P(invenção|a necessidade é a mãe da) ≈ P(invenção|mãe da)

- P("a necessidade é a mãe da invenção") =
  P(a) * P(necessidade|a) * P(é|a necessidade) * P(a|necessidade é) *
  P(mãe|é a) * P(da|a mãe) * P(invenção|mãe da)

N-gramas

- Para calcular P(w_3|w_1,w_2), temos:

$$P(w_3 \mid w_1, w_2) = \frac{cont(w_1, w_2, w_3)}{cont(w_1, w_2)}$$

- $cont(w_1, w_2, w_3)$ é a quantidade de vezes que o trio w_1 w_2 w_3 aparece
- $cont(w_1, w_2)$ é a quantidade de vezes que o par w_1 w_2 aparece

Exemplo

- Vamos implementar uma função para calcular a probabilidade de bigramas no corpus com as obras de Machado de Assis

```
import re
from collections import Counter

regex = "[a-zA-ZçÇãÃõÕáÁéÉíÍóÓúÚâÂêÊîÎôÔûÛàÀ]+"

data = open("machado.txt").read()

tokens = re.findall(regex, data)
tokens_count = Counter(tokens)

def p_bigram(w1, w2):
    count_w1 = tokens_count[w1]
    count_w1w2 = 0

    for i in range(len(tokens)-1):
        if tokens[i] == w1 and tokens[i+1] == w2:
            count_w1w2 += 1

    return count_w1w2/count_w1


print(p_bigram("o", "dia"))
print(p_bigram("o", "homem"))
print(p_bigram("a", "sol"))
```

Método de Visualização de Shannon

- Iniciamos com um bigrama (essa escolha pode ser aleatória de acordo com sua probabilidade)
- Escolha um bigrama <w2, w3> de forma aleatória acordo com sua probabilidade
- Continua o passo anterior até um critério de parada

w_1 w_2

w_2 w_3

w_3 w_4

w_4 w_5

w_5 <parada>

Método de Visualização de Shannon

- Iniciamos com um bigrama (essa escolha pode ser aleatória de acordo com sua probabilidade)
- Escolha um bigrama <w2, w3> de forma aleatória acordo com sua probabilidade
- Continua o passo anterior até um critério de parada

w_1 w_2

w_2 w_3

w_3 w_4

w_4 w_5

w_5 <parada>

frase gerada: w_1 w_2 w_3 w_4 w_5

Método de Visualização de Shannon

- Podemos usar o método para qualquer n-grama
- Vamos gerar frases baseadas nos livros de Machado de Assis usando diferentes n-gramas

Método de Visualização de Shannon

- Unigrama:
  - por dela luís mesa é onde e fossem dias logo os. me ele gente de logo o vez e amor sua tudo andava a andar, seus trazer repetiu cumprimentar
  - ia-se quisessem replicou. quem daqui ficava de convencional natural a se, a que pode da que as sem e, moscas com aqui, a de que febre não “seu” jorge muito gente

Método de Visualização de Shannon

- Bigrama:
  - Ao ver; a olhar para o Padre Bernardes. O interior, era vê-lo. Ele sorriu. Estela sem filhos, não ter confessado que a esqueceu todos com indiferença.
  - Agora só explica o último, se daqui a um instante em você se chegarem a dor; mas era, sou o erro.
  - Novidade não serem felizes. Nem eu só da senhora deixou de tal aspecto. Era assim como na queda do dedo de leite da igreja, com o pai.

Método de Visualização de Shannon

- Trigrama:
  - Teria percorrido meia página, mas não tratava de concluir, por exemplo. — E depois a hora presente. E repetiu, até chegar à porta, e que o ministro da justiça, pensava na presidência do Rio de Janeiro, por mais ridícula que pareça algo excessivo.
  - Camargo era pouco mais velho do que perturbar a placidez do médico se refugiara na Tijuca. Uma vez que Lalau ia obedecer constrangida; e pagou.
  - Mas era tarde, havendo algumas pessoas da casa e o pai de Eugênia; e ele gostam muito um do outro lado, um rumor próximo; era só a voz arrastada.

Método de Visualização de Shannon

- 4-grama:
  - Então, afastando-me, respondi: — Você é rico, continuou ele depois de alguns passos, com as palavras idôneas e castas que a situação exigia
  - A dona da casa, e especialmente no espírito do moço, que passeava ao longo do terraço, ouvindo as saudações e os cochichos.
  - D. Cláudia colheu as rosas do último baile do ano. Estácio não se animou a dizer nada, observou ela

Aplicações

- Garantir que uma sequência de palavras tem sentido (alta probabilidade)
  - Com um corpus especializado, podemos avaliar se uma sequência de palavras é rara ou não
- Correção ortográfica/gramatical
- Classificação
- Responder perguntas
- Geração de texto / previsão de próxima palavra