Processamento de Linguagem Natural

Modelagem de Linguagens

Prof.: Hansenclever Bassani (Hans) hfb@cin.ufpe.br
Site da disciplina: www.cin.ufpe.br/~hfb/pln/

Baseado nos slides do curso de Stanford no Coursera
por Daniel Jurafsky e Christopher Manning.

Tradução: Ygor Sousa
Revisão: Hansenclever Bassani

Modelagem de Linguagens

Introdução a N-grams

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Modelos de Linguagem Probabilísticos

- Objetivo de Hoje: atribuir uma probabilidade a uma sentença
  - Tradução Automática:
    - » P(“**high** winds tonight”) > P(“**large** winds tonight”)
  - Correção Ortográfica
    - » The office is about fifteen **minuets** from my house
      - P(“about fifteen **minutes** from”) > P(“about fifteen **minuets** from”)
  - Reconhecimento de Discurso
    - » P(I saw a van) >> P(eyes awe of an)
  - + Sumarização, pergunta-resposta, etc., etc.!!

Por quê?

Modelagem de Linguagens Probabilísticas

- Objetivo: computar a probabilidade de uma sentença ou sequência da palavras:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5,\ldots w_n)$

- Tarefa Relacionada: probabilidade de uma próxima palavra:

  $P(w_5|w_1,w_2,w_3,w_4)$

- Um modelo que calcula qualquer uma destas:

  $P(W)$ or $P(w_n|w_1,w_2,\ldots w_{n-1})$ é chamado de **Modelo de Linguagem**.

- Melhor: **Gramática** Mas **modelo de linguagem** ou **LM** é o padrão

Como calcular P(W)

- Como calcular esta probabilidade conjunta:
  - P(its, water, is, so, transparent, that)
- Intuição: vamos contar com a Regra da Cadeia de Probabilidade

Lembrete: A Regra da Cadeia

- Probabilidade condicional de B dado A:
  - $P(A|B) = P(A,B)/P(A) \rightarrow P(A,B) = P(B|A)P(A)$
- Mais variáveis:

  $P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$
- Regra da cadeia em geral

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

A Regra da Cadeia aplicada para calcular Probabilidade Conjunta de palavras em sentença

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P("its water is so transparent") =

P(its) × P(water|its) × P(is|its water) × P(so|its water is) × P(transparent|its water is so)

Como estimar estas probabilidades

- Poderíamos apenas contar e dividir?

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- Não! As frases possíveis são muitas!
- Nós nunca vamos ver dados suficiente para os estimar

Suposição de Markov

Andrei Markov

- Suposição simplificada:

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$$

- Ou talvez

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$$

Suposição de Markov

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

- Em outras palavras, nós aproximamos cada componente no produto

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-k} \ldots w_{i-1})$$

Caso mais simples: Modelo Unigram

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Algumas sentenças automaticamente geradas de um modelo unigram

- `fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass`
- `thrift, did, eighty, said, hard, 'm, july, bullish`
- `that, or, limited, the`

Modelo Bigram

- Condição da palavra anterior:

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

- texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen
- outside, new, car, parking, lot, of, the, agreement, reached
- this, would, be, a, record, november

Modelos N-gram

- Podemos extender para trigrams, 4-grams, 5-grams
- Em geral este é um modelo de linguagem insuficiente
  - Porque linguagem tem **dependências de longa distância**:

  "The **computer** which I had just put into the machine room on the fifth floor **crashed**."

- Mas nós podemos sempre ir longe com modelos N-gram

Modelagem de Linguagens

Introdução a N-grams

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Modelagem de Linguagens

Estimar Probabilidades

N-gram

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Estimar Probabilidades Bigram

- Estimativa de Máxima Verossimilhança

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

Um Exemplo em Bigram

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(\texttt{I} \mid \texttt{<s>}) = \tfrac{2}{3} = .67 \qquad P(\texttt{Sam} \mid \texttt{<s>}) = \tfrac{1}{3} = .33 \qquad P(\texttt{am} \mid \texttt{I}) = \tfrac{2}{3} = .67$$

$$P(\texttt{</s>} \mid \texttt{Sam}) = \tfrac{1}{2} = 0.5 \qquad P(\texttt{Sam} \mid \texttt{am}) = \tfrac{1}{2} = .5 \qquad P(\texttt{do} \mid \texttt{I}) = \tfrac{1}{3} = .33$$

Mais exemplos: Sentenças do Projeto de Restaurante de Berkeley

- can you tell me about any good cantonese restaurants close by
- mid priced thai food is what i'm looking for
- tell me about chez panisse
- can you give me a listing of the kinds of food that are available
- i'm looking for a good place to eat breakfast
- when is caffe venezia open during the day

Contagem de Bigram Bruto

- Saída de 9222 sentenças

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Probabilidades de Bigram Bruto

$$P(A,B) = P(B|A)P(A)$$

- Normalizado por unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Resultado:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Estimativas Bigram de Probabilidades de Sentença

P(<s> I want english food </s>) =

P(I|<s>)

× P(want|I)

× P(english|want)

× P(food|english)

× P(</s>|food)

= 0.25 x 0.33 x 0.011 x 0.5 x 0.68 = 0.000031

Que Tipos de Conhecimento?

- P(english|want) = .0011
- P(chinese|want) = .0065
- P(to|want) = .66
- P(eat | to) = .28
- P(food | to) = 0
- P(want | spend) = 0
- P (i | <s>) = .25

Pontos Práticos

- Nós fazemos tudo em espaço de log
  - Evitar underflow
  - (além do que adição é mais rápida que multiplicação)

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Toolkit de Modelagem de Linguagens

- SRILM
  - http://www.speech.sri.com/projects/srilm/
  - Em C++

Google N-Gram Release, Agosto de 2006

AUG
3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

...

That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

https://books.google.com/ngrams

Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Modelagem de Linguagens

Estimar Probabilidades

N-gram

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Modelagem de Linguagens

Avaliação e Perplexidade

Avaliação: Quão bom é o nosso modelo?

- Nosso modelo de linguagem prefere boas sentenças às ruins?
  - Atribui probabilidades mais altas à sentenças “reais” ou “frequentemente observadas”
    - Do que sentenças “não gramaticalmente bem formadas” ou “raramente observadas”?
- Treinamos parâmetros do nosso modelo por um **conjunto de treinamento**.
- Testamos a performance do modelo com dados não vistos anteriormente.
  - Um **conjunto de teste** é um conjunto de dados que é totalmente diferente do de treinamento, totalmente não utilizado.
  - Uma **métrica de avaliação** nos mostra o quão bom nosso modelo foi com o conjunto de testes.

Avaliação Extrínseca de Modelos N-gram

- Melhor avaliação para comparar modelos A e B
  - Coloque cada modelo em uma tarefa
    - Corretor Ortográfico, Reconhecedor de Discurso, Sistema MT
  - Executar a tarefa, obter uma precisão para A e B
    - Quantas palavras com erros ortográficos corrigidas corretamente
    - Quantas palavras traduzidas corretamente
  - Compare precisão de A e B

Dificuldade de Avaliação Extrínseca (in-vivo) de Modelos N-gram

- Avaliação Extrínseca
  - Demorado; pode levar dias ou semanas
- Então
  - As vezes utilizar avaliação **intrínseca**: **perplexidade**
  - Má aproximação
    - A menos que os dados de teste se pareçam com os dados de treinamento
    - Então, **geralmente só é útil em experimentos piloto**
  - Mas é útil para pensar.

Intuição de Perplexidade

- O Jogo de Shannon:
  - Quão bem podemos prever a próxima palavra?

    I always order pizza with cheese and ____

    The 33rd President of the US was ____

    I saw a ____

  - Unigrams são terríveis neste jogo. (Por quê?)

mushrooms 0.1

pepperoni 0.1

anchovies 0.01

….

fried rice 0.0001

….

and 1e-100

- Um modelo melhor de um texto
  - é aquele que atribui uma probabilidade mais elevada para a palavra que ocorre efetivamente

Perplexidade

O melhor modelo de linguagem é aquele que melhor prediz uma base de teste nunca vista

- Apresenta a mais alta P(sentença)

Perplexidade é a probabilidade inversa do conjunto de teste, normalizado pelo número de palavras:

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

Regra de Cadeia:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

Para bigrams:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

Minimizar perplexidade é o mesmo que maximizar probabilidade

O Jogo de Intuição de Shannon para Perplexidade

- De Josh Goodman
- Quão difícil é a tarefa de reconhecer dígitos '0,1,2,3,4,5,6,7,8,9'
  - Perplexidade = 10
- Quão difícil é reconhecer (30,000) nomes na Microsoft.
  - Perplexidade = 30,000
- Se um sistema tem que reconhecer
  - Operador (1 em 4)
  - Vendas (1 em 4)
  - Suporte Técnico (1 em 4)
  - 30,000 nomes (1 em 120,000 cada)
  - Perplexidade é 53 = $[(1/4)x (1/4) x (1/4) x (1/120,000)]^{(1/4)}$
- Perplexidade é um fator de ramificação equivalente ponderado

Perplexidade como Fator de Ramificação

- Vamos supor uma sentença que consiste de dígitos aleatórios
- Qual é a perplexidade de uma sentença de acordo com um modelo que atribui P=1/10 para cada dígito?

$$\begin{aligned} \mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}} \\ &= \frac{1}{10}^{-1} \\ &= 10 \end{aligned}$$

Menor perplexidade = melhor modelo

- Treinamento 38 milhões de palavras e teste 1.5 milhões, WSJ

Ordenação N-gram	Unigram	Bigram	Trigram
Perplexity	962	170	109

Modelagem de Linguagens

Avaliação e Perplexidade

Modelagem de Linguagens

Generalização e Zeros

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

O Método de Visualização de Shannon

- Escolha um bigram aleatório (<s>, w) de acordo com sua probabilidade
- Agora escolha um bigram aleatório (w, x) de acordo com sua probabilidade
- E assim por diante até que nós escolhemos </s>
- E por fim junte as palavras

```
<s> I
    I want
      want to
           to eat
              eat Chinese
                  Chinese food
                          food </s>
I want to eat Chinese food
```

Aproximação de Shakespeare

Unigram

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
Every enter now severally so, let
Hill he late speaks; or! a more to leg less first you enter
Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.
Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.
This shall forbid it should be branded, if renown made it empty.
Indeed the duke; and had a very good friend.
Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;
Will you not tell me who I am?
It cannot be but so.
Indeed the short and the long. Marry, 'tis a noble Lepidus.

Shakespeare como Corpus

- N=884,647 tokens, V=29,066
- Shakespeare produziu 300,000 tipos de bigram de um total de V^2= 844 milhões possíveis bigrams.
  - Assim, 99,96% dos possíveis bigrams nunca foram vistos (tem entradas zero na tabela)
- Quadrigram pior: O que está sendo apresentado parece Shakespeare porque ***é*** Shakespeare

O Jornal de Wall Street não é Shakespeare (sem ofensa)

Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Os Perigos do Sobretreinamento (*Overfitting*)

- N-grams só funciona bem para predição de palavras se o corpo de teste se parece com o corpo de treino
  - Na vida real, isso geralmente não acontece
  - Nós precisamos treinar modelos robustos que generalizam!
  - Um tipo de generalização: Zeros!
    - Coisas que nunca ocorrem no conjunto de treino
      - Mas ocorrem no conjunto de teste

Zeros

- Conjunto de Treinamento:
  … denied the allegations
  … denied the reports
  … denied the claims
  … denied the request

  P(“offer” | denied the) = 0

- Conjunto de Teste:
  … denied the offer
  … denied the loan

Bigrams de Probabilidade Zero

- Bigrams com zero probabilidade
  - significa que nós vamos atribuir probabilidade 0 ao conjunto de teste!
- E portanto nós não podemos calcular perplexidade (não podemos dividir por 0)!

Modelagem de Linguagens

Generalização e Zeros

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO