UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Informática
Departamento de Informática Aplicada

Aula 2: Conceitos Básicos de Textos e Modelos de Linguagem Probabilísticos

Prof. Dennis Giovani Balreira

INF01221 - Tópicos Especiais em Computação XXXVI:
Processamento de Linguagem Natural

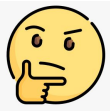
Conteúdo

- Conceitos básicos de textos:
  - Dados e informação
  - Tipos de dados
  - Níveis de conhecimento da linguagem
  - Terminologia (palavra, sentença, texto, corpus, dataset)
  - Propriedades estatísticas dos textos

- Modelos de linguagem probabilísticos:
  - Introdução e definição formal
  - Modelo N-grama

Conceitos básicos de textos

Dado e informação

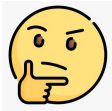
- Dado vs. informação? 

Dado e informação

- Dado:
  - Fato bruto e não processado
  - Pode ser número, texto, observação, medição ou descrição que, isoladamente, não têm um significado claro ou contexto
  - Exemplo: Números como "10", "15", "20" ou palavras soltas como "azul", "verde", "amarelo"

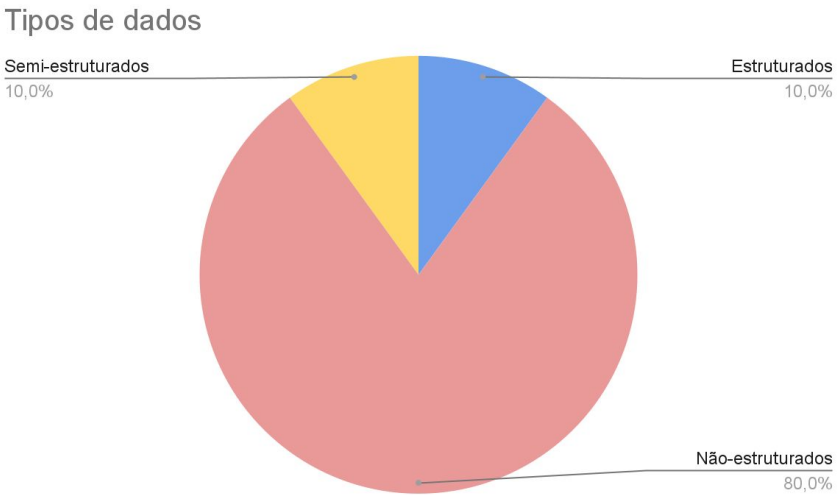
- Informação:
  - Resultado do processamento, organização e estruturação dos dados
  - Fornece contexto e significado, permitindo que os dados sejam compreendidos e usados para tomar decisões
  - Exemplo: "A temperatura média desta semana foi de 20°C" ou "As vendas aumentaram 15% no último trimestre"

Tipos de dados

- Estruturados vs. Semi-estruturados vs. Não-estruturados 

Tipos de dados

- Estruturados: podem ser armazenados em bancos de dados relacionais em formato de tabelas SQL
- Semi-estruturados: apesar de não estarem em um bancos de dados relacionais, apresentam algumas propriedades organizacionais
- Não-estruturados: podem ter uma estrutura interna, mas não se enquadram perfeitamente em uma tabela



Dados Estruturados

- São dados organizados que podem ser facilmente armazenados e consultados
  - Podem ser armazenados em bancos de dados relacionais em formato de tabelas SQL

- Necessitam de um esquema bem definido (nomes de tabelas, atributos e tipos)

- Ex.: Dados de clientes no banco de dados da empresa

Students Table

Student	ID*
John Smith	084
Jane Bloggs	100
John Smith	182
Mark Antony	219

Participants Table

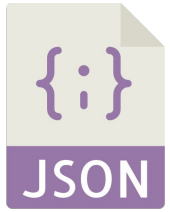
ID*	Activity*
084	Tennis
084	Swimming
100	Squash
100	Swimming
182	Tennis
219	Golf
219	Swimming
219	Squash

Activities Table

Activity*	Cost
Golf	$47
Sailing	$50
Squash	$40
Swimming	$15
Tennis	$36

Dados Semi-Estruturados

- Não estão organizados em um esquema rígido como os dados estruturados, mas possuem uma organização que facilita a análise e o processamento
- Contêm elementos estruturais que ajudam a identificar dados individuais e suas relações
  - Tags, marcadores, etc.
- Ex: JSON, XML, e-mail, CSV, etc.




```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```

Dados Não-Estruturados

- Podem ter uma estrutura interna, mas não se enquadram perfeitamente em uma tabela
- São os dados mais abundantes existentes
  - Estimativas dizem que correspondem a pelo menos 80% dos dados de interesse para as organizações
- Ex: Texto livre (páginas web, mensagens de texto, notícias, etc.), áudio, vídeo

WIKIPEDIA
The Free Encyclopedia

Tipos de dados

- Estruturados: podem ser armazenados em bancos de dados relacionais em formato de tabelas SQL
- Não-estruturados: podem ter uma estrutura interna, mas não se enquadram perfeitamente em uma tabela
- Semi-estruturados: apesar de não estarem em um bancos de dados relacionais, apresentam algumas propriedades organizacionais

Qual o foco da disciplina?

Tipos de dados

- Estruturados: podem ser armazenados em bancos de dados relacionais em formato de tabelas SQL
- Não-estruturados: podem ter uma estrutura interna, mas não se enquadram perfeitamente em uma tabela
- Semi-estruturados: apesar de não estarem em um bancos de dados relacionais, apresentam algumas propriedades organizacionais

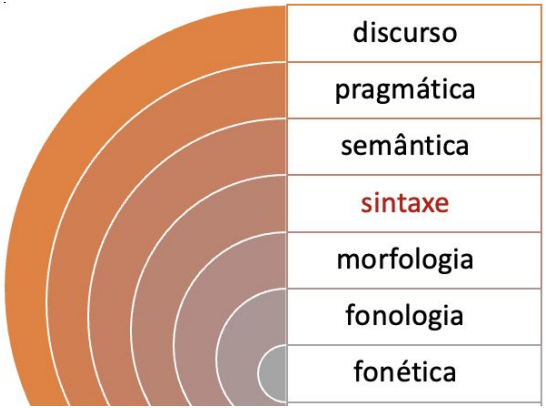
Grande parte dos dados de valor para as organizações é não-estruturada e encontra-se representada sob a forma de texto

Foco principal da disciplina!

Níveis de conhecimento da linguagem

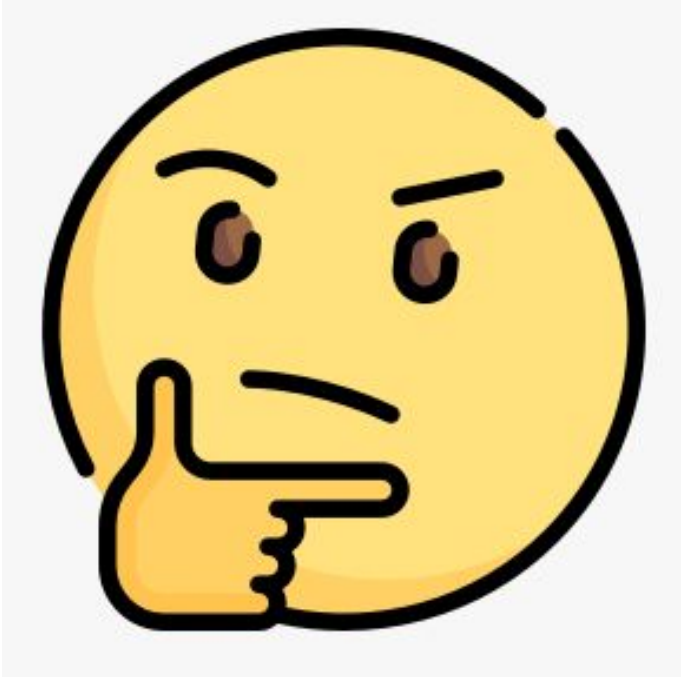
- Referem-se às diferentes camadas ou aspectos que podem ser examinados para entender como as línguas funcionam

- Principais níveis de análise linguística:
  - Fonética e fonologia: conhecimento sobre sons linguísticos
  - Morfologia: conhecimento dos componentes significativos das palavras
  - Sintaxe: conhecimento das relações estruturais entre as palavras
  - Semântica: conhecimento do significado
  - Pragmática: conhecimento da relação de significado com os objetivos e intenções de quem fala
  - Discurso: conhecimento sobre unidades linguísticas maiores do que uma única sentença



Terminologia da linguagem

- Palavra:
- Sentença (frase):
- Texto:
- Corpus (plural Corpora):
- Dataset:



Terminologia da linguagem

- Palavra: unidade mínima de significado completo na linguagem

cachorro

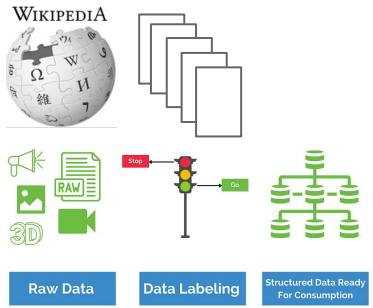
- Sentença (frase): unidade gramatical completa composta por uma ou mais palavras

O cachorro correu pelo parque

- Texto: sequência de palavras organizadas de maneira coesa e coerente

O cachorro correu pelo parque. Ele estava feliz e cheio de energia

- Corpus (plural Corpora): conjunto de textos puros
  - Ex: artigos da Wikipedia

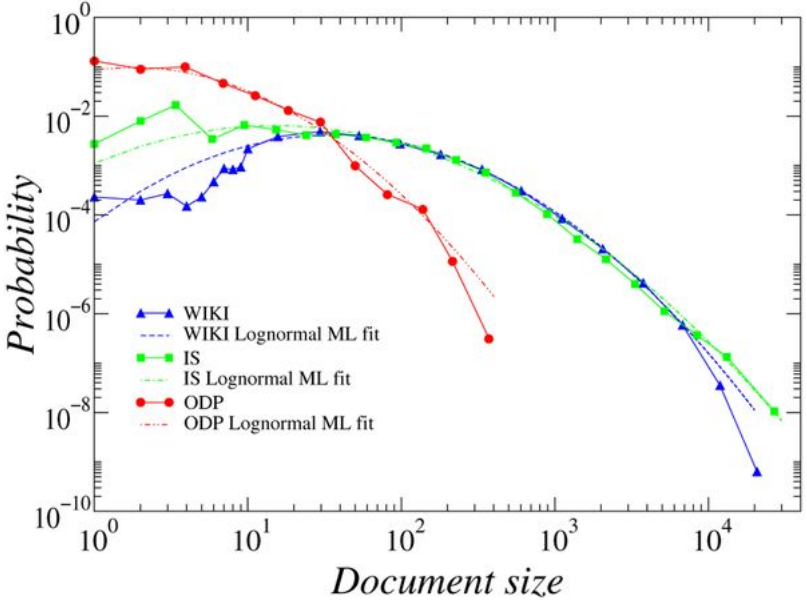


- Dataset: conjunto de dados anotados
  - Ex: dataset para a detecção de discurso de ódio (dados + classe esperada)

Propriedades estatísticas dos textos

- Referem-se às características numéricas e padrões observados em conjuntos de textos que podem ser analisados usando métodos estatísticos
  - Fornecem *insights* sobre a estrutura, o estilo e o conteúdo dos textos

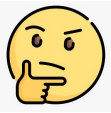
- Exemplos:
  - Distribuição das palavras
  - Frequência de termos
  - Tamanho médio de sentenças
  - Complexidade lexical
  - Distribuição de frequência de sentenças
  - Similaridade textual



Serrano, M. Ángeles, Alessandro Flammini, and Filippo Menczer. **"Modeling statistical properties of written text."** *PloS one* 4.4 (2009): e5372.

Propriedades estatísticas dos textos

O cachorro correu pelo parque. O cachorro brincou com a bola e correu novamente. A bola rolou até o lago, e o cachorro nadou para buscá-la. O sol brilhava no céu, e o cachorro feliz balançava o rabo enquanto corria pelo gramado. No final do dia, o cachorro deitou na grama, cansado, mas contente.

Quais as palavras mais frequentes? 

Propriedades estatísticas dos textos

O cachorro correu pelo parque. O cachorro brincou com a bola e correu novamente. A bola rolou até o lago, e o cachorro nadou para buscá-la. O sol brilhava no céu, e o cachorro feliz balançava o rabo enquanto corria pelo gramado. No final do dia, o cachorro deitou na grama, cansado, mas contente.

Quais as palavras mais frequentes?

Todas outras palavras aparecem apenas uma vez!

o	8
cachorro	5
e	3
a	2
bola	2
correu	2
no	2
pelo	2

Propriedades estatísticas dos textos

- Exemplo IMDB (internet movie database) corpus:
  - Cada documento possui filme, sinopse, artistas e seus papéis com 230.721 documentos

- Há algum padrão “em geral”?
  - Diferentes idiomas?
  - Diferentes assuntos?

- Duas propriedades fundamentais a serem vistas:
  - Lei de Zipf
  - Lei de Heaps

Frequência dos termos encontrados, do 1º ao 20º no IMDB.

rank	term	frequency	rank	term	frequency
1	the	1586358	11	year	250151
2	a	854437	12	he	242508
3	and	822091	13	movie	241551
4	to	804137	14	her	240448
5	of	657059	15	artist	236286
6	in	472059	16	character	234754
7	is	395968	17	cast	234202
8	i	390282	18	plot	234189
9	his	328877	19	for	207319
10	with	253153	20	that	197723

Lei de Zipf

- Criada por George Kingsley Zipf (anos 40)
- Analisou o texto Ulisses (James Joyce), contando a frequência de cada palavra
  - A palavra mais comum apareceu 8000 vezes
  - A décima mais comum, 800 vezes
  - A centésima mais comum, 80 vezes
  - A milésima, 8 vezes
- Constatou que a frequência dos termos diminui rapidamente à medida que o rank aumenta
- Palavras mais comuns em inglês:
  - https://www.youtube.com/watch?v=fCn8zs912OE


George Kingsley Zipf (1902-1950)

rank	term	frequency	rank	term	frequency
1	the	1586358	11	year	250151
2	a	854437	12	he	242508
3	and	822091	13	movie	241551
4	to	804137	14	her	240448
5	of	657059	15	artist	236286
6	in	472059	16	character	234754
7	is	395968	17	cast	234202
8	i	390282	18	plot	234189
9	his	328877	19	for	207319
10	with	253153	20	that	197723

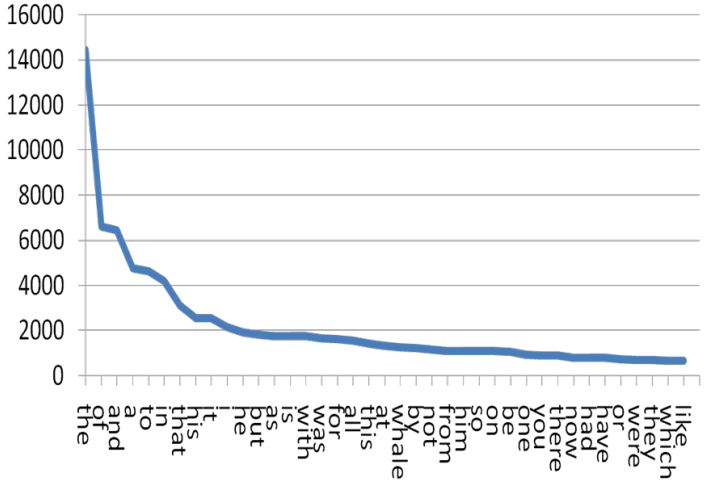
Lei de Zipf

- O termo mais frequente é responsável por 10% do texto
  - O segundo representa 5%
  - O terceiro cerca de 3%
  - Os 10 mais frequentes juntos cerca de 30%
  - Os 20 mais frequentes juntos cerca de 36%
  - Os 50 mais frequentes juntos representam cerca de 45%

- A lei de Zipf diz que, em um corpus, a frequência de ocorrência $f(n)$ de um termo está ligada a sua ordem n

$$f(n) = \frac{K}{n}$$, onde K é uma constante

Para inglês K = 0.1

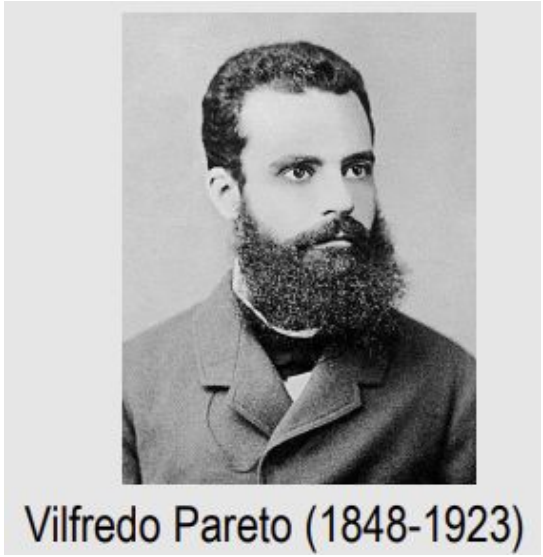


Lei de Zipf

- A lei está relacionada ao princípio de Pareto
  - Há uma distribuição desigual entre causas e efeitos
  - Identificar os "20% críticos" é uma estratégia eficaz

- Regra 80/20
  - 20% das pessoas detém 80% da riqueza
  - 20% do código contém 80% dos erros
  - 20% dos jogadores de League of Legends são responsáveis por 80% dos comportamentos tóxicos

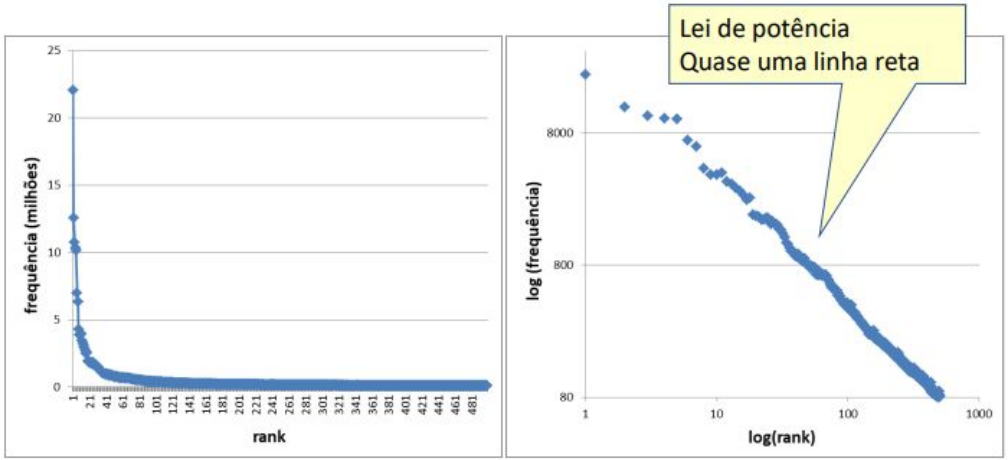
- Sobre os termos:
  - Metade dos termos ocorrem apenas umas vez
  - 75% dos termos ocorrem 3 vezes ou menos
  - 83% dos termos ocorrem 5 vezes ou menos
  - 90% dos termos ocorrem 10 vezes ou menos


Vilfredo Pareto (1848-1923)

Lei de Zipf

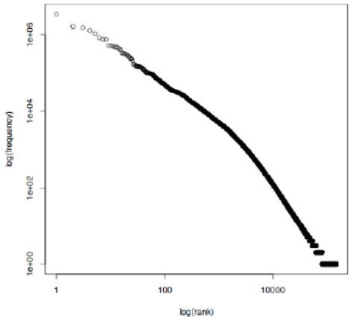
- Pode ser melhor observada plotando-se um gráfico em escala logarítmica, com os eixos log (rank) e log (frequencia)

- Exemplo:
  - 500 palavras mais frequentes em inglês
  - Calculados sobre um corpus de 450 milhões de palavras (https://www.wordfrequency.info/free.asp)

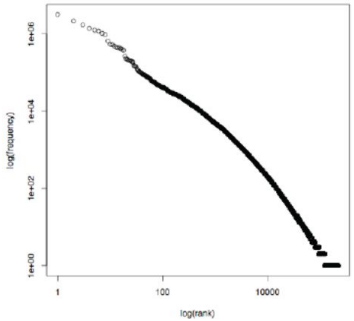


Lei de Zipf

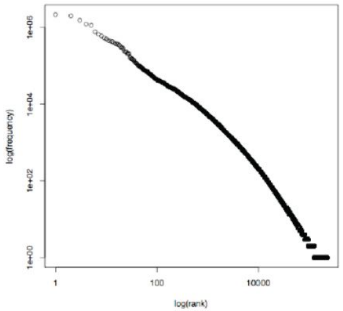
- Generaliza para múltiplos idiomas



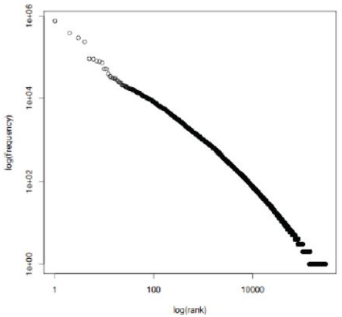
Inglês



Espanhol



Português

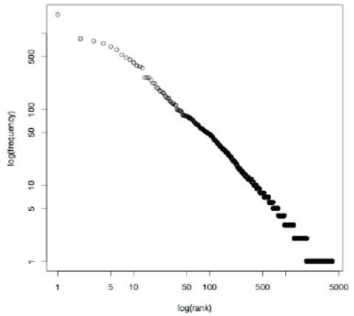


Húngaro

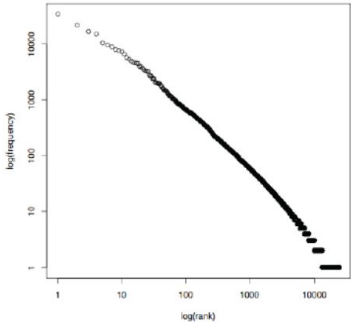
https://ils.unc.edu/courses/2020_fall/inls509_001/lectures/04-StatisticalPropertiesOfText.pdf

Lei de Zipf

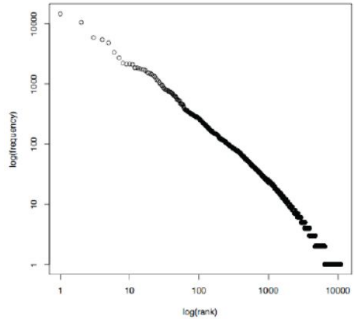
- Generaliza para diversos textos



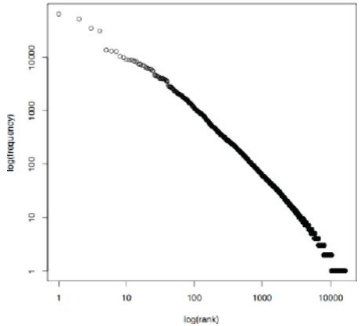
Alice no país
das maravilhas



Guerra e paz



A origem das
espécies



A Bíblia

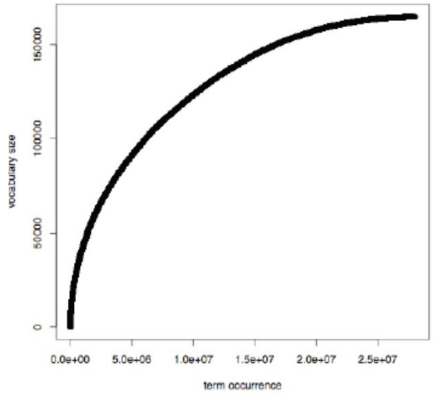
https://ils.unc.edu/courses/2020_fall/inls509_001/lectures/04-StatisticalPropertiesOfText.pdf

Lei de Heaps

- Atribuída a Harold Heaps (1978)
  - Mas originalmente descoberta por Gustav Herdan (1960)
- É uma lei empírica que descreve o número de termos distintos em função do tamanho da coleção
- O número de termos novos aumenta dramaticamente no início, mas depois a taxa de crescimento é menor
- Novos termos sempre irão aparecer (erros de ortografia, nomes próprios, palavras inventadas), mas em menor quantidade

$$V(N) = k \cdot N^{\beta}$$

V = tamanho do vocabulário (número de termos distintos)
N = tamanho do corpus (número de termos)
k = constante ($10 \leq k \leq 100$)
β= constante ($\beta \approx 0.50$)



Material complementar

- Dados e Informação, Tipos de Dados, Níveis de Conhecimento da Linguagem
  - Tutorial da Khan Academy: Explica os conceitos fundamentais de dados e tipos de dados. https://pt.khanacademy.org/computing/computer-science
  - Introdução à Linguística Computacional: Cobre os níveis de conhecimento da linguagem (fonologia, morfologia, sintaxe, semântica, etc.). http://www.nltk.org/book/

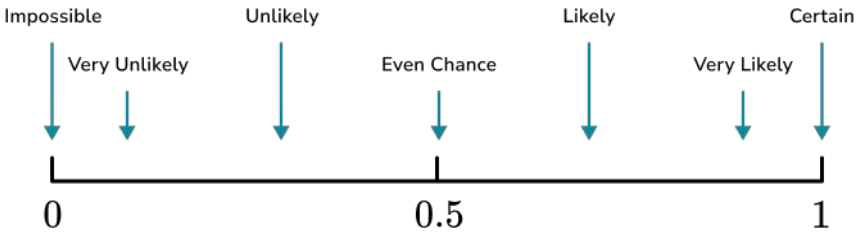
- Terminologia (Palavra, Sentença, Texto, Corpus, Dataset)
  - NLTK Documentation: A documentação da biblioteca NLTK cobre conceitos básicos, como a definição de corpus, tokens, tipos de palavras, etc. https://www.nltk.org/
  - Pandas: Biblioteca para manipulação e análise de dados que pode ser usada para trabalhar com dados textuais. https://pandas.pydata.org/

Modelos de Linguagem Probabilísticos

Modelos de Linguagem Probabilísticos

- Modelos de linguagem probabilísticos são técnicas que utilizam a probabilidade para prever a próxima palavra ou sequência de palavras em uma sentença
  - Baseiam-se em textos previamente observados
  - São a base de técnicas estado da arte (GPT, BERT, etc.)

- Estimam a probabilidade de uma palavra ou sequência de palavras aparecer em um determinado contexto

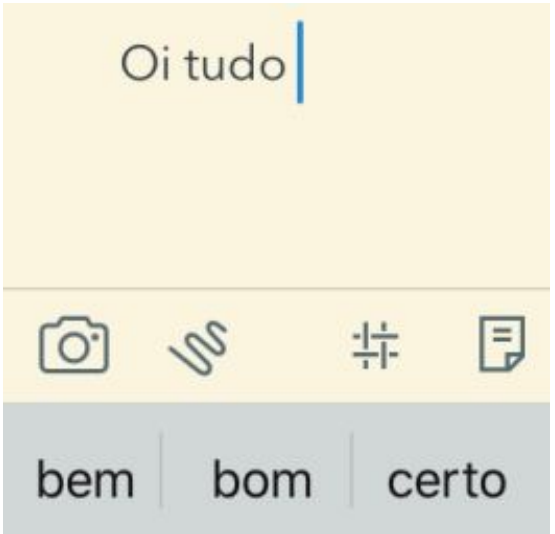


Modelos de Linguagem Probabilísticos

- Qual frase está correta?
  - (A) Aquela personagem é menos importante?
  - (B) Aquela personagem é menas importante?
    - Se P(A) > P(B), então “provavelmente” (A) está correta

- Textos possuem uma certa “previsibilidade” de palavras
  - Dada uma grande quantidade de texto, é possível “prever” com funções de probabilidade quais podem ser os próximos termos
    - Muito do conhecimento já foi “decorado” de tanto ver termos nos textos

- Completar frases é a ideia das palavras sugeridas em smartphones
  - Probabilidade de ocorrer a palavra w após “Oi tudo”



Modelos de Linguagem Probabilísticos

- Como completar as frases abaixo?

  “Pessoal, lembrem de entregar o tema de casa até _____”

  a próxima aula.
  hoje à noite.
  amanhã.
  às 23h55.
  …

  “Não aguento mais jogar _____.

  LOLzin.
  Valorant.
  sozinho.
  com meus amigos.
  …

  “Ao sair de casa desligue a _____.

  luz.
  bola.
  …

Modelos de Linguagem Probabilísticos

- Como completar as frases abaixo?

  “Pessoal, lembrem de entregar o tema de casa até _____”

  a próxima aula. [45%]
  hoje à noite. [20%]
  amanhã. [10%]
  às 23h55. [5%]
  …

  “Não aguento mais jogar _____.

  LOLzin. [50%]
  Valorant. [30%]
  sozinho. [5%]
  com meus amigos. [3%]
  …

  “Ao sair de casa desligue a _____.

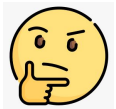
  luz [70%]
  bola [2%]
  …

Modelos de Linguagem Probabilísticos

- A modelagem probabilística de linguagem é a tarefa que atribui uma probabilidade a uma sequência de palavras

- Mas como adaptar para palavras e textos?
  - Considerando o número de ocorrências com frases a partir de um corpus
  - Com esta informação, é necessário contar as frequências das frases no texto

- Precisamos do que?



Modelos de Linguagem Probabilísticos

- A modelagem probabilística de linguagem é a tarefa que atribui uma probabilidade a uma sequência de palavras

- Mas como adaptar para palavras e textos?
  - Considerando o número de ocorrências com frases a partir de um corpus
  - Com esta informação, é necessário contar as frequências das frases no texto

- Precisamos de um corpus com as frequências das palavras e frases já contadas!

Probabilidade condicional

- Notação: $P(B|A)$
  - Probabilidade de que o evento B ocorra, dado que A já ocorreu (A é verdade)
    Ex: A = “está chovendo”
    B = “alguém está usando guarda-chuva”

$P(B|A)$ = “probabilidade de alguém estar usando guarda-chuva dado que já sabemos que está chovendo”

Probabilidade condicional

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Probabilidade dos eventos ocorrerem juntos

Probabilidade de A ocorrer

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

Probabilidade conjunta

- Notação: $P(B|A)$
  - Probabilidade de que o evento B ocorra, dado que A já ocorreu (A é verdade)
    Ex: A = “está chovendo”
    B = “alguém está usando guarda-chuva”

$P(B|A)$ = “probabilidade de alguém estar usando guarda-chuva dado que já sabemos que está chovendo”

Probabilidade condicional

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Probabilidade dos eventos ocorrerem juntos

Probabilidade de A ocorrer

Reescrevendo:

Probabilidade conjunta

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

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

Regra da cadeia (probabilidade)

- Notação: $P(B|A)$
  - Probabilidade de que o evento B ocorra, dado que A já ocorreu (A é verdade)
    Ex: A = “está chovendo”
    B = “alguém está usando guarda-chuva”

$P(B|A)$ = “probabilidade de alguém estar usando guarda-chuva dado que já sabemos que está chovendo”

Probabilidade condicional

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Probabilidade dos eventos ocorrerem juntos

Probabilidade de A ocorrer

Reescrevendo: $P(A,B) = P(B|A) \cdot P(A)$

Probabilidade conjunta

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

Regra da cadeia!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Regra da cadeia (probabilidade)

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

- Permite expressar a probabilidade conjunta de vários eventos em termos de probabilidades condicionais
  Ex:

$$P(A, B, C, D) = P(A) \cdot P(B|A) \cdot P(C|A, B) \cdot P(D|A, B, C)$$

A = “está nublado”
B = “está chovendo”
C = “alguém usa guarda-chuva”
D = “rua está molhada”

$P(D|A, B, C)$ = probabilidade da rua estar molhada, dado que está nublado, chovendo e alguém está usando um guarda-chuva

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

Para palavras

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

C(W): contagem (frequência) das palavras (em ordem) W no texto

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

Calcular P(“The water of Walden Pond”) =

P(“The”) x P("water" | "The") x P("of" | "The water”) x
P("Walden" | "The water of") x P("Pond" | "The water of Walden”)

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

Calcular P("The water of Walden Pond") =

P("The") x P("water" | "The") x P("of" | "The water") x
P("Walden" | "The water of") x P("Pond" | "The water of Walden")

$$P(\text{"The"}) = \frac{\text{Contagem de "The"}}{\text{Número total de palavras}} = \frac{2}{10} = 0.2$$

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

Calcular P("The water of Walden Pond") =

0.2 x P("water" | "The") x P("of" | "The water") x
P("Walden" | "The water of") x P("Pond" | "The water of Walden")

$$P(\text{"water"}|\text{"The"}) = \frac{\text{Contagem de "The water"}}{\text{Contagem de "The"}} = \frac{2}{2} = 1.0$$

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

Calcular P("The water of Walden Pond") =

0.2 x 1.0 x P("of" | "The water") x
P("Walden" | "The water of") x P("Pond" | "The water of Walden")

$$P(\text{"of"}|\text{"The water"}) = \frac{\text{Contagem de "The water of"}}{\text{Contagem de "The water"}} = \frac{1}{2} = 0.5$$

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

Calcular P("The water of Walden Pond") =

0.2 x 1.0 x 0.5 x

P("Walden" | "The water of") x P("Pond" | "The water of Walden")

$$P(\text{"Walden"}|\text{"The water of"}) = \frac{\text{Contagem de "The water of Walden"}}{\text{Contagem de "The water of"}} = \frac{1}{1} = 1.0$$

Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = \frac{C(w_1, w_2, \ldots, w_n)}{C(w_1, w_2, \ldots, w_{n-1})}$$

Calcular P("The water of Walden Pond") =

0.2 x 1.0 x 0.5 x

1.0 x P("Pond" | "The water of Walden")

$$P(\text{"Pond"}|\text{"The water of Walden"}) = \frac{\text{Contagem de "The water of Walden Pond"}}{\text{Contagem de "The water of Walden"}} = \frac{1}{1} = 1.0$$

Modelos de Linguagem Probabilísticos

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

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

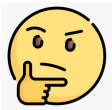
- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

Calcular P("The water of Walden Pond") =
0.2 x 1.0 x 0.5 x
1.0 x 1.0 = **0.1 (ou 10%)**

Problemas desta abordagem probabilística "genérica"?



Modelos de Linguagem Probabilísticos

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

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

- Utilização da Regra da Cadeia para palavras!

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$$

Exemplo: "The water of Walden Pond is clear. The water stinks."

Calcular P("The water of Walden Pond") =
0.2 x 1.0 x 0.5 x
1.0 x 1.0 = **0.1 (ou 10%)**

Problemas desta abordagem probabilística "genérica"?

- Precisamos de muito texto para conseguir estimar "exatamente" estas sequências
- Muito custoso computacionalmente!

Modelos de Linguagem Probabilísticos

Considerando o seguinte corpus (que é um conjunto fictício de poemas sobre Pets):

O gato dorme na cama
O gato dorme na sala
O cão dorme na cama
O cão dorme no sofá

- Outro exemplo:
  - Calcular a probabilidade da frase “o gato dorme no sofá”:

    P(o,gato,dorme,no,sofá) =

    P(o) * P(gato|o) * P(dorme|o,gato) * P(no|o,gato,dorme) * P(sofá|o,gato,dorme,no)

Modelos de Linguagem Probabilísticos

Considerando o seguinte corpus (que é um conjunto fictício de poemas sobre Pets):

O gato dorme na cama
O gato dorme na sala
O cão dorme na cama
O cão dorme no sofá

- Outro exemplo:
  - Calcular a probabilidade da frase “o gato dorme no sofá”:

    P(o,gato,dorme,no,sofá) =

    P(o) * P(gato|o) * P(dorme|o,gato) * P(no|o,gato,dorme) * P(sofá|o,gato,dorme,no)

Frequência de palavras:

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

TOTAL: 20 ÚNICAS: 9
o: 4x

o,gato: 2x

o,gato,dorme: 2x

o,gato,dorme,no: 0x

o,gato,dorme,no,sofá: 0x

Modelos de Linguagem Probabilísticos

Considerando o seguinte corpus (que é um conjunto fictício de poemas sobre Pets):

O gato dorme na cama
O gato dorme na sala
O cão dorme na cama
O cão dorme no sofá

- Outro exemplo:
  - Calcular a probabilidade da frase “o gato dorme no sofá”:

    P(o,gato,dorme,no,sofá) =

    P(o) * P(gato|o) * P(dorme|o,gato) * P(no|o,gato,dorme) * P(sofá|o,gato,dorme,no)

Frequência de palavras:

TOTAL: 20 ÚNICAS: 9
o: 4x
o,gato: 2x
o,gato,dorme: 2x
o,gato,dorme,no: 0x
o,gato,dorme,no,sofá: 0x

Probabilidades:

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

P(o) = 4 / 20 = 0,2
P(gato|o) = 2 / 4 = 0,5
P(dorme|o,gato) = 2 / 2 = 1
P(no|o,gato,dorme) = 0 / 2 = 0
P(sofá|o,gato,dorme,no) = 0 / 0 = erro

Como arrumar probabilidades inexistentes ou incorretas?

Modelos de Linguagem Probabilísticos

Considerando o seguinte corpus (que é um conjunto fictício de poemas sobre Pets):

> O gato dorme na cama
> O gato dorme na sala
> O cão dorme na cama
> O cão dorme no sofá

- Outro exemplo:
  - Calcular a probabilidade da frase “o gato dorme no sofá”:

    P(o,gato,dorme,no,sofá) =

    P(o) * P(gato|o) * P(dorme|o,gato) * P(no|o,gato,dorme) * P(sofá|o,gato,dorme,no)

Frequência de palavras:

TOTAL: 20 ÚNICAS: 9
o: 4x
o,gato: 2x
o,gato,dorme: 2x
o,gato,dorme,no: 0x
o,gato,dorme,no,sofá: 0x

Probabilidades:

$$P(A_n|A_1, A_2, \ldots, A_{n-1}) = \frac{P(A_1, A_2, \ldots, A_n)}{P(A_1, A_2, \ldots, A_{n-1})}$$

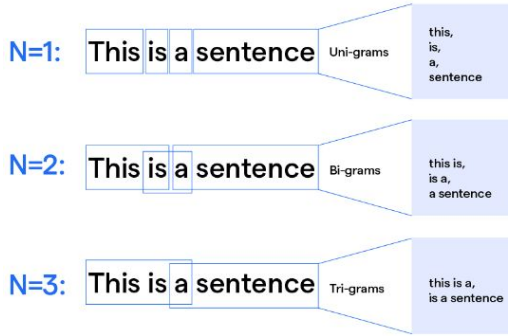
P(o) = 4 / 20 = 0,2
P(gato|o) = 2 / 4 = 0,5
P(dorme|o,gato) = 2 / 2 = 1
P(no|o,gato,dorme) = (0 + 1) / (2 + 9) = 0,09
P(sofá|o,gato,dorme,no) = (0 + 1) / (0 + 9) = 0,11
P(o,gato,dorme,no,sofá) = 0,2 * 0,5 * 1 * 0,09 * 0,11 = 0,00099

É possível evitar probabilidade zero, com a Suavização de Laplace:
- Adiciona 1 no numerador
- Adiciona |V| (número de palavras distintas) no denominador

Modelos probabilísticos

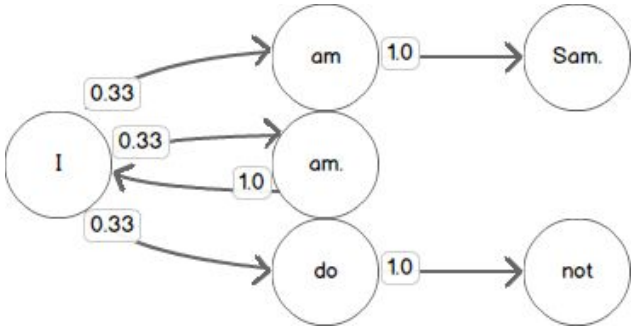
- N-gramas: sequência de N elementos adjacentes em uma cadeia. Por exemplo, em uma frase, um bigrama (2-grama) considera pares de palavras consecutivas
  Exemplo: Em um modelo de bigramas, a probabilidade de uma palavra é condicionada pela palavra anterior
  - Mais detalhes na sequência

- Cadeias de Markov
- Gramáticas Probabilísticas
- Redes Bayesianas
- Baseados em Redes Neurais



Modelos probabilísticos

- N-gramas
- Cadeias de Markov: modelos matemáticos que descrevem sistemas que transitam de um estado para outro em uma série de etapas, onde a probabilidade de transição para o próximo estado depende apenas do estado atual (propriedade de Markov)
  Exemplo: Em um jogo de tabuleiro onde cada posição representa um estado, a probabilidade de mover para a próxima posição depende apenas da posição atual, não das posições anteriores
- Gramáticas Probabilísticas
- Redes Bayesianas
- Baseados em Redes Neurais



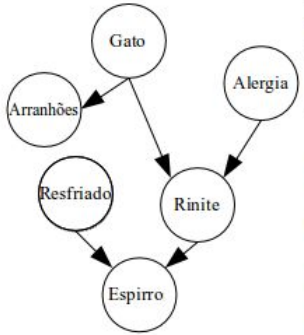
Modelos probabilísticos

$S \rightarrow NP\ VP$	.8
$S \rightarrow VP$	.2
$NP \rightarrow D\ N$	.4
$NP \rightarrow NP\ PP$	.4
$NP \rightarrow PN$	.2
$VP \rightarrow V\ NP$	.7
$VP \rightarrow VP\ PP$	.3
$PP \rightarrow P\ NP$	1

$D \rightarrow the$	.8
$D \rightarrow a$	.2
$N \rightarrow flight$	1
$PN \rightarrow john$	.9
$PN \rightarrow schiphol$	.1
$V \rightarrow booked$	1
$P \rightarrow from$	1

- N-gramas
- Cadeias de Markov
- Gramáticas Probabilísticas: extensão das gramáticas formais, onde cada regra gramatical é associada a uma probabilidade. São usadas para modelar a estrutura de linguagens com incerteza
  Exemplo: Em uma gramática probabilística livre de contexto para a língua inglesa, a regra que produz uma sentença (S) a partir de um sujeito (NP) e um predicado (VP) poderia ter uma probabilidade associada
- Redes Bayesianas
- Baseados em Redes Neurais

Modelos probabilísticos



	V	F
P(Alergia)	45%	55%

	V	F
P(Gato)	20%	80%

	V	F
P(Arranhões|Gato)	95%	5%
P(Arranhões|!Gato)	15%	85%

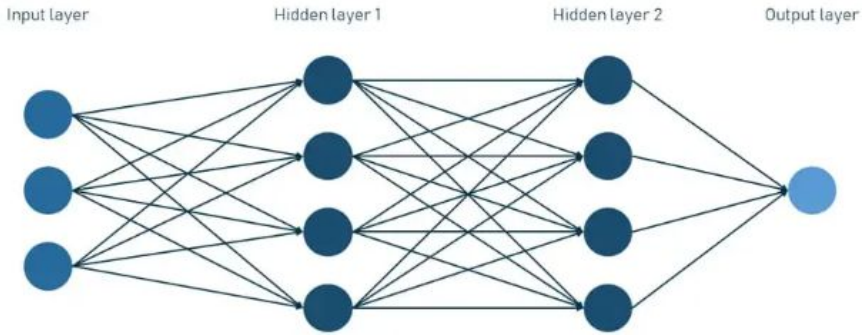
	V	F
P(Rinite|Alergia, Gato)	99%	80%
P(Rinite|Alergia, !Gato)	10%	90%
P(Rinite|!Alergia, Gato)	15%	85%
P(Rinite|!Alergia, !Gato)	0,50%	99,50%

	V	F
P(Espirro|Resfriado,Rinite)	99%	80%
P(Espirro|!Resfriado,Rinite)	90%	10%
P(Espirro|Resfriado,!Rinite)	85%	15%
P(Espirro|!Resfriado,!Rinite)	1%	99%

- N-gramas
- Cadeias de Markov
- Gramáticas Probabilísticas
- Redes Bayesianas: modelos gráficos probabilísticos que representam um conjunto de variáveis e suas dependências condicionais por meio de um grafo direcionado acíclico
  Exemplo: Em um sistema de diagnóstico médico, uma rede bayesiana pode modelar a relação entre sintomas e doenças, onde os nós representam sintomas e doenças, e as arestas representam dependências condicionais entre eles
- Baseados em Redes Neurais

Modelos probabilísticos

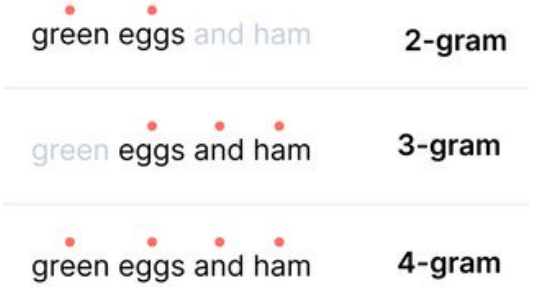
- N-gramas
- Cadeias de Markov
- Gramáticas Probabilísticas
- Redes Bayesianas



- Baseados em Redes Neurais: modelos que utilizam redes neurais artificiais para modelar relações complexas entre variáveis de entrada e saída, sendo capazes de aprender padrões e fazer previsões com base em grandes volumes de dados
  Exemplo: Grandes modelos de linguagem como BERT, GPT e variações, baseados em arquitetura Transformer (estado-da-arte). Serão estudados mais para frente da disciplina

N-gramas

- Um N-grama é uma sequência contígua de N elementos (palavras ou caracteres) de um dado texto ou fala
  - O modelo calcula a probabilidade de cada palavra em uma sequência com base nas N-1 palavras anteriores
- Técnica fundamental em PLN, utilizado para modelar a probabilidade de sequência de palavras em um texto
  - Simples, eficazes e muito utilizados em diversas aplicações
- Podem ser vistos como um caso particular da definição geral de modelos probabilísticos (contexto limitado)
  - Em vez de calcular a probabilidade de toda a sequência anterior, foca apenas nas palavras mais próximas



Suposição de Markov*

- Permite simplificar a regra da cadeia assumindo que apenas os últimos “eventos” realmente importam

Suposição de primeira ordem: $P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdot \ldots \cdot P(A_n|A_1, A_2, \ldots, A_{n-1})$

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2) \cdot \ldots \cdot P(A_n|A_{n-1})$$

Para palavras wk:

$$P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot \ldots \cdot P(w_n|w_1, w_2, \ldots, w_{n-1})$$

*Formalmente assume que o futuro estado de um sistema depende apenas do seu estado atual, e não de toda a sequência de estados anteriores.

N-gramas

- Unigrama (1-grama): $P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2) \cdot P(w_3) \cdot \ldots \cdot P(w_n)$
- Bigrama (2-grama): $P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_2) \cdot \ldots \cdot P(w_n|w_{n-1})$
- Trigrama (3-grama): $P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot P(w_4|w_2, w_3) \cdot \ldots \cdot P(w_n|w_{n-2}, w_{n-1})$
- N-grama: $P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot \ldots \cdot P(w_n|w_{n-(n-1)}, \ldots, w_{n-1})$

N-gramas

$$P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot \ldots \cdot P(w_n|w_{n-(n-1)}, \ldots, w_{n-1})$$

- Exemplo: Dado o texto W = "The cat sat on the mat.", calcule P(W) considerando:
  - Unigramas:

$$P(\text{"The cat sat on the mat"}) = P(\text{"The"}) \cdot P(\text{"cat"}) \cdot P(\text{"sat"}) \cdot P(\text{"on"}) \cdot P(\text{"the"}) \cdot P(\text{"mat"})$$

  - Bigramas:

$$P(\text{"The cat sat on the mat"}) = P(\text{"The"}) \cdot P(\text{"cat"}|\text{"The"}) \cdot P(\text{"sat"}|\text{"cat"}) \cdot P(\text{"on"}|\text{"sat"}) \cdot P(\text{"the"}|\text{"on"}) \cdot P(\text{"mat"}|\text{"the"})$$

  - Trigramas:

$$P(\text{"The cat sat on the mat"}) = P(\text{"The"}) \cdot P(\text{"cat"}|\text{"The"}) \cdot P(\text{"sat"}|\text{"The cat"}) \cdot P(\text{"on"}|\text{"cat sat"}) \cdot P(\text{"the"}|\text{"sat on"}) \cdot P(\text{"mat"}|\text{"on the"})$$

N-gramas

$$P(w_1, w_2, \ldots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot \ldots \cdot P(w_n|w_{n-(n-1)}, \ldots, w_{n-1})$$

- Exemplo: Dado o texto W = "The cat sat on the mat.", calcule P(W) considerando:
  - Unigramas:

$$P(\text{"The cat sat on the mat"}) = P(\text{"The"}) \cdot P(\text{"cat"}) \cdot P(\text{"sat"}) \cdot P(\text{"on"}) \cdot P(\text{"the"}) \cdot P(\text{"mat"})$$

$$P(\text{"The cat sat on the mat"}) = 0.33 \times 0.17 \times 0.17 \times 0.17 \times 0.33 \times 0.17 \approx 0.000049$$

  - Bigramas:

$$P(\text{"The cat sat on the mat"}) = P(\text{"The"}) \cdot P(\text{"cat"}|\text{"The"}) \cdot P(\text{"sat"}|\text{"cat"}) \cdot P(\text{"on"}|\text{"sat"}) \cdot P(\text{"the"}|\text{"on"}) \cdot P(\text{"mat"}|\text{"the"})$$

$$P(\text{"The cat sat on the mat"}) = 0.33 \times 0.5 \times 1.0 \times 1.0 \times 1.0 \times 0.5 = 0.0825$$

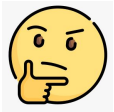
  - Trigramas:

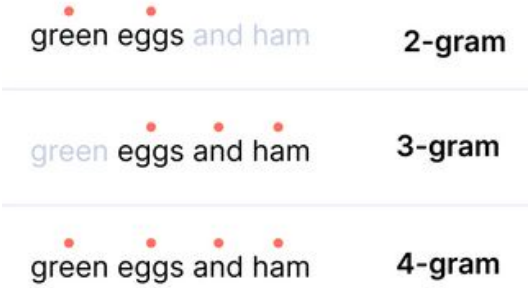
$$P(\text{"The cat sat on the mat"}) = P(\text{"The"}) \cdot P(\text{"cat"}|\text{"The"}) \cdot P(\text{"sat"}|\text{"The cat"}) \cdot P(\text{"on"}|\text{"cat sat"}) \cdot P(\text{"the"}|\text{"sat on"}) \cdot P(\text{"mat"}|\text{"on the"})$$

$$P(\text{"The cat sat on the mat"}) = 0.33 \cdot 1.0 \cdot 1.0 \cdot 1.0 \cdot 1.0 \cdot 1.0$$

N-gramas

- São adequados para tarefas onde a relação entre palavras é de curto alcance e o contexto limitado é suficiente
  - Simples de implementar e interpretar
  - Útil para tarefas com vocabulário moderado

- Limitações? 



N-gramas

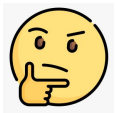
- São adequados para tarefas onde a relação entre palavras é de curto alcance e o contexto limitado é suficiente
  - Simples de implementar e interpretar
  - Útil para tarefas com vocabulário moderado

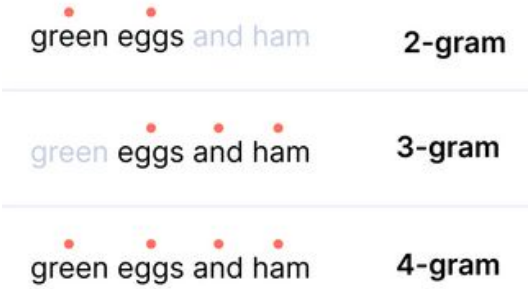
- Limitações:
  - Não capturam dependências de longo prazo, (consideram contexto limitado de N-1 palavras)

"**The soups** that I made from that new cookbook I bought yesterday **were** amazingly delicious."

  - Difícil de gerar novas sequências com significados semelhantes

- Como podemos resolver estes problemas?





N-gramas

- São s… de a relaç… entre pal… é de curt… al… é suficient… …rpr… …om vo…

…ram depe… …o limitado de N…

"T… that I made … new … amazingly delicious."

  - Difícil de gerar novas seq…ências c… …gnifica… …emelhant…

- Como podemos resolver estes problemas

LARGE LANGUAGE MODELS!!!

Google BERT

…eggs … …an… …-gram

gre… …iam 4-gram

T5 Text-to-Text Transfer Transformer

Considerações gerais

- As probabilidades devem ser calculadas a partir do processamento de uma grande quantidade de texto para que possam “carregar” conhecimento (contexto!)
  - Quanto mais texto, melhor o resultado
  - Mas (muito) maior a complexidade

- Computar as probabilidades é um desafio:
  - Existem muitas combinações possíveis de palavras
  - Os modelos ficam muito pesados

- Exemplos de uso:
  - Geração de texto
  - Tradução automática
  - Análise de sentimento

Material complementar

- Modelo N-grama:
  - Artigo sobre N-gramas em NLP: Tutorial que cobre o básico dos modelos N-grama em processamento de linguagem natural. https://towardsdatascience.com/introduction-to-n-grams-572577ce3da0
  - Scikit-learn: Biblioteca Python para modelos probabilísticos e aprendizado de máquina. https://scikit-learn.org/stable/
  - Natural Language Toolkit (NLTK): Inclui ferramentas específicas para a construção de modelos N-grama. http://www.nltk.org/howto/ngram.html

Próxima aula

- Expressões regulares

- Pré-processamento de texto:
  - Segmentação de sentenças
  - Normalização:
    - Correção ortográfica
    - Case folding (conversão para minúsculas)
    - Remoção de acentos, caracteres especiais e pontuação
    - Expansão de contrações
    - Remoção de stop words
    - Lematização
    - Stemming
  - Tokenização

- Distância mínima de edição

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Informática
Departamento de Informática Aplicada

Obrigado pela atenção!
Dúvidas?

Prof. Dennis Giovani Balreira
(Material adaptado da Profa. Viviane Moreira e do Prof. Dan Jurafski)

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INF01221 - Tópicos Especiais em Computação XXXVI:
Processamento de Linguagem Natural

.inf
INSTITUTO DE INFORMÁTICA
UFRGS