

SCC0633/5908 Processamento de Linguagem Natural



Que lugar é este?

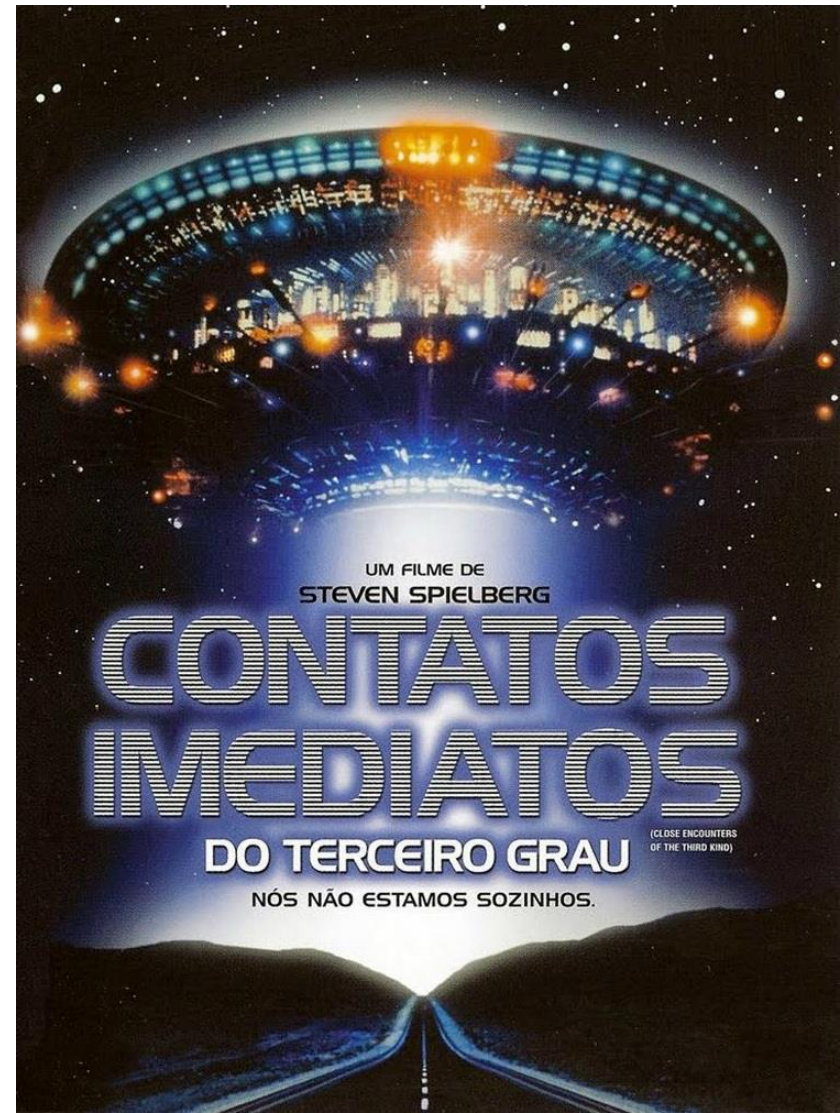


[E a comunicação acontece...]



Contatos Imediatos do Terceiro Grau

- 1977
 - Não apenas Steven Spielberg, mas também o compositor John Williams (Star Wars, Harry Potter, Jurassic Park, Indiana Jones, Superman, etc.)



[Tipos de linguagens]

- Nem toda linguagem é necessariamente verbal (escrita ou oral)
- Linguagens não verbais utilizam outros meios comunicativos, como gestos, sons, cores, imagens, expressões faciais e corporais e símbolos
 - Fotografias, placas, acenos de mão, etc.
 - Libras - Língua Brasileira de Sinais (segunda língua oficial do Brasil desde 2002)
 - Conta com alfabeto, estrutura linguística e gramatical próprios!
- Também há linguagens híbridas/mistas, como ocorre com histórias em quadrinhos, charges e outdoors
 - Mensagens de Instagram e outras redes sociais podem se enquadrar aqui!



MODELAGEM ESTATÍSTICA & DISTRIBUCIONAL

SCC5908 Introdução ao Processamento de Língua Natural
SCC0633 Processamento de Linguagem Natural

Prof. Thiago A. S. Pardo

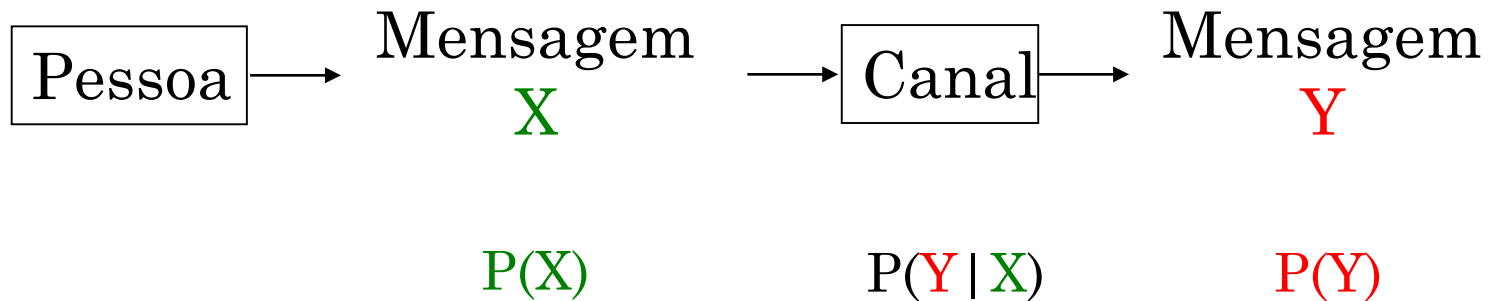
RELEMBRANDO

NA AULA PASSADA

- Córpus e sua importância para PLN
- Modelos matemáticos e estatísticos
 - Leis de Zipf
 - Probabilidades
 - Teorema de Bayes
 - Modelo *Noisy-Channel*

MODELO *NOISY-CHANNEL*

RELEMBRANDO: O QUE ERA ISSO E QUAL SUA IMPORTÂNCIA?



Teorema de Bayes



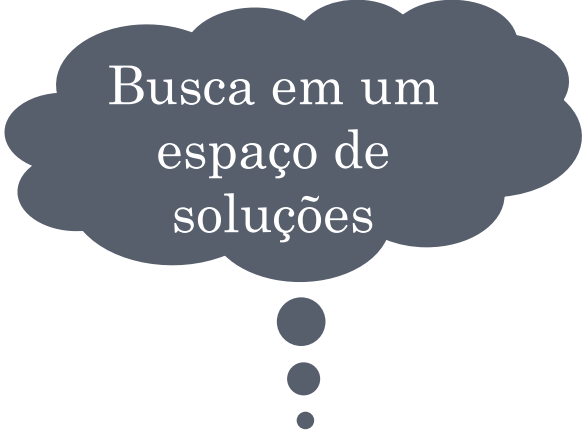
Determinar X a partir de Y: $P(X | Y)$

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

TEOREMA DE BAYES

$$P(\mathbf{X} | \mathbf{Y}) = P(\mathbf{Y} | \mathbf{X}) \times P(\mathbf{X}) / P(\mathbf{Y})$$

- \mathbf{Y} é observado
- Deve-se escolher \mathbf{X} que maximize $P(\mathbf{X} | \mathbf{Y})$: decodificação



Busca em um
espaço de
soluções

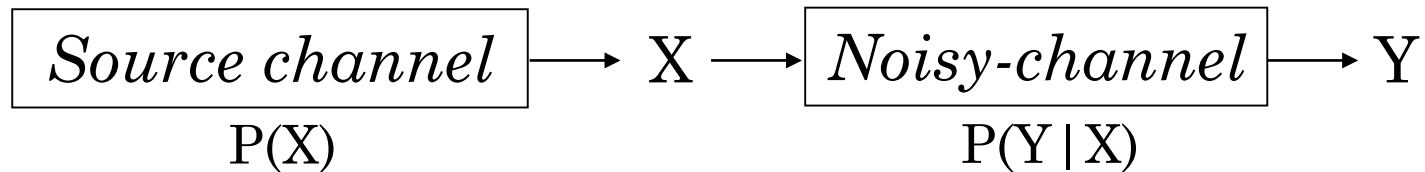
$$P(\mathbf{X} | \mathbf{Y}) = P(\mathbf{Y} | \mathbf{X}) \times P(\mathbf{X}) / P(\mathbf{Y})$$

↓
constante


$$P(\mathbf{X} | \mathbf{Y}) = P(\mathbf{Y} | \mathbf{X}) \times P(\mathbf{X})$$

MODELO *NOISY-CHANNEL*

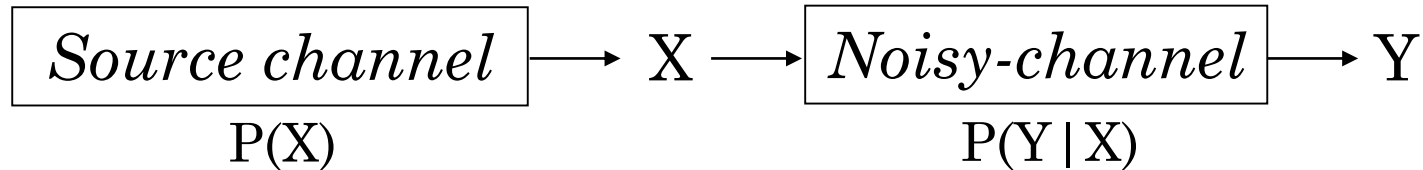
- Generalizando o modelo



- 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

MODELO *NOISY-CHANNEL*

- Generalizando o modelo



Transmissão de bits:

$P(X) \sim$ uniforme \rightarrow pode ser ignorado, portanto

$$P(0)=P(1)=0.5$$

$P(Y|X)$

$$P(0 \rightarrow 0) = P(0|0) = 0.6$$

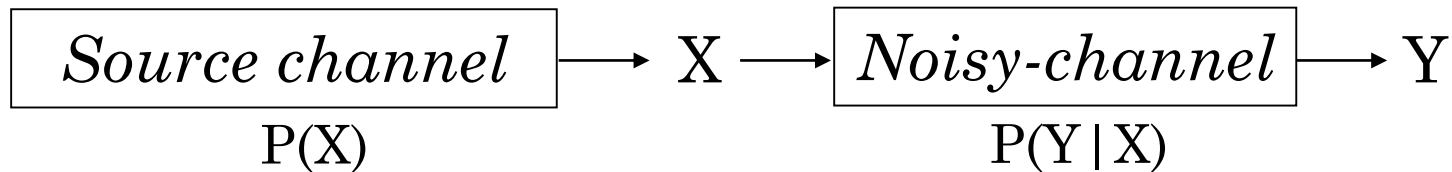
$$P(0 \rightarrow 1) = P(1|0) = 0.4$$

$$P(1 \rightarrow 1) = P(1|1) = 0.3$$

$$P(1 \rightarrow 0) = P(0|1) = 0.7$$

MODELO *NOISY-CHANNEL*

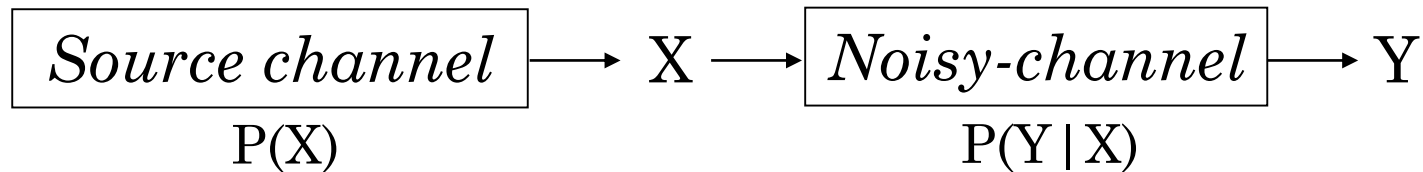
- Generalizando o modelo



- 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
 - **Modelagem ao contrário!**

MODELO *NOISY-CHANNEL*

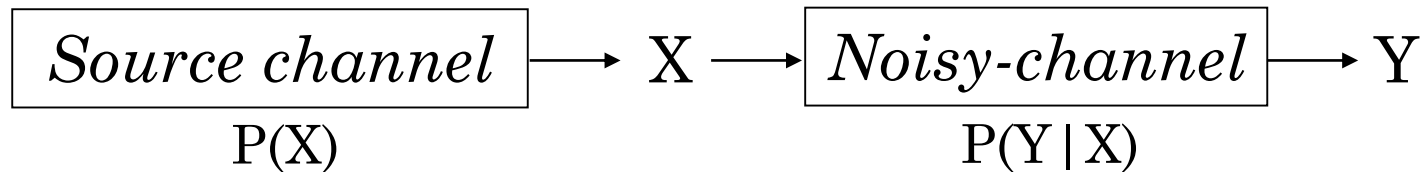
- Generalizando o modelo



- 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 $P(\text{bit } Y | \text{bit } X)$ ser simplesmente a probabilidade de um bit virar outro, poderia ser isso **CONJUGADO** à probabilidade de o receptor ter problemas técnicos/operacionais
 - $P(\text{bit } Y | \text{bit } X) = ?$

MODELO *NOISY-CHANNEL*

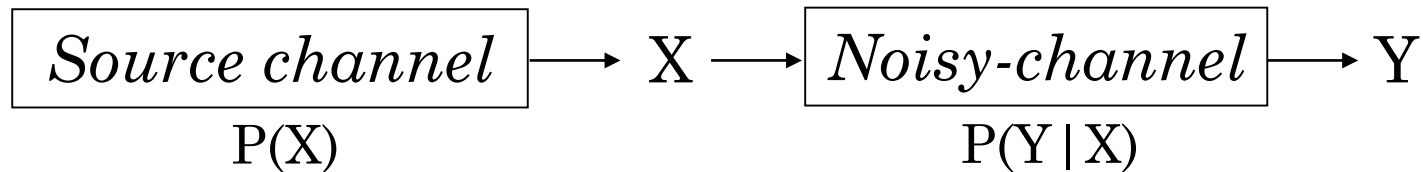
- Generalizando o modelo



- 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 $P(\text{bit } Y | \text{bit } X)$ ser simplesmente a probabilidade de um bit virar outro, poderia ser isso **CONJUGADO** à probabilidade de o receptor ter problemas técnicos/operacionais
 - $P(\text{bit } Y | \text{bit } X) = p_{\text{conversão_bit}}(Y | X) * p_{\text{problema_recepção}}(X)$

MODELO *NOISY-CHANNEL*

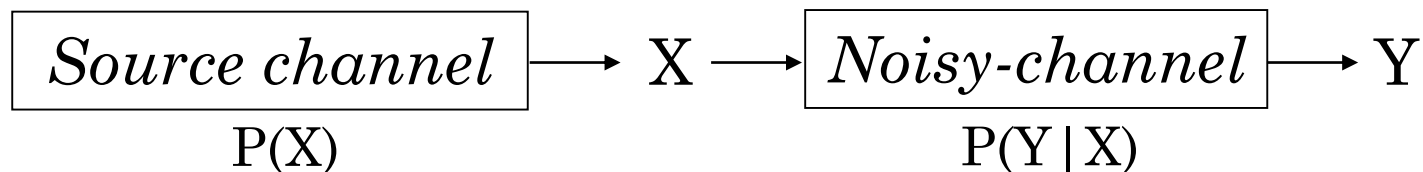
- Generalizando o modelo



- 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 $P(\text{bit } Y | \text{bit } X)$ ser simplesmente a probabilidade de um bit virar outro, poderia ser isso **CONJUGADO** à probabilidade de o receptor ter problemas técnicos/operacionais
 - $P(\text{bit } Y | \text{bit } X) = p_{\text{conversão_bit}}(Y | X) * p_{\text{problema_recepção}}(X)$
 $= c(Y | X) * r(X)$

MODELO *NOISY-CHANNEL*

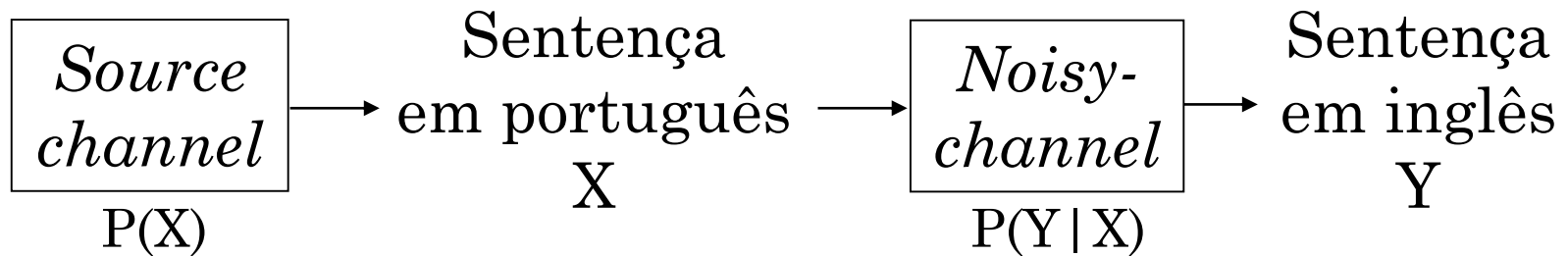
- Generalizando o modelo



Aplicação	Entrada (X)	Saída (Y)	P(X)	P(Y X)
Tradução Automática	Sequência de palavras	Sequência de palavras	Modelo de língua	Modelo de tradução
<i>Optical Character Recognition (OCR)</i>	Texto	Texto com erros	Prob. do texto	Modelo de erros de OCR
Reconhecimento de Fala	Sequência de palavras	Sinal acústico	Prob. de sequência de palavras	Modelo acústico

TRADUÇÃO AUTOMÁTICA

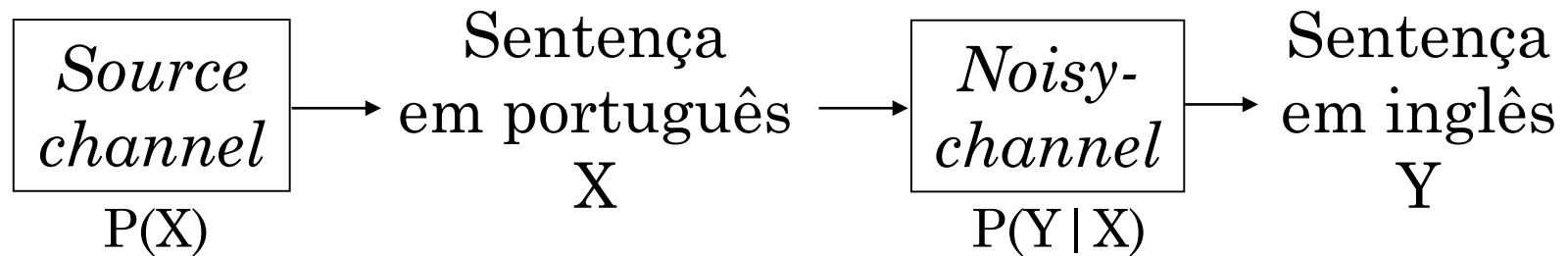
- Tradução de uma sentença em inglês para português



- Do que precisamos para saber $P(X|Y)$?

TRADUÇÃO AUTOMÁTICA

- Tradução de uma sentença em inglês para português



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

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.

$P(\text{tradução}) = t(\text{the} | \text{o}) \times t(\text{dog} | \text{cão}) \times t(\text{black} | \text{preto}) \times t(\text{died} | \text{morreu}) \times r(1 | 1) \times r(3 | 2) \times r(2 | 3) \times r(4 | 4)$

P(X)

○ Modelo de língua

- Como prever a probabilidade de uma sentença traduzida a partir de “*The boy fell.*”?

P(O menino caiu.)?

P(O menino colapsou.)?

P(O garoto caiu.)?

P(X)

○ Modelo de língua “clássico” 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

$P(\text{O menino caiu.}) =$

$$\begin{aligned} & \text{peso}_1 \times P(\text{O}) \times P(\text{menino}) \times P(\text{caiu}) \times P(\text{.}) + \\ & \text{peso}_2 \times P(\text{O, menino}) \times P(\text{menino, caiu}) \times P(\text{caiu, .}) + \\ & \text{peso}_3 \times P(\text{O, menino, caiu}) \times P(\text{menino, caiu, .}) + \\ & \text{peso}_4 \times P(\text{O, menino, caiu, .}) \end{aligned}$$

○ Uma possível alternativa (irreal): distribuição uniforme

- Toda sentença é igualmente provável

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

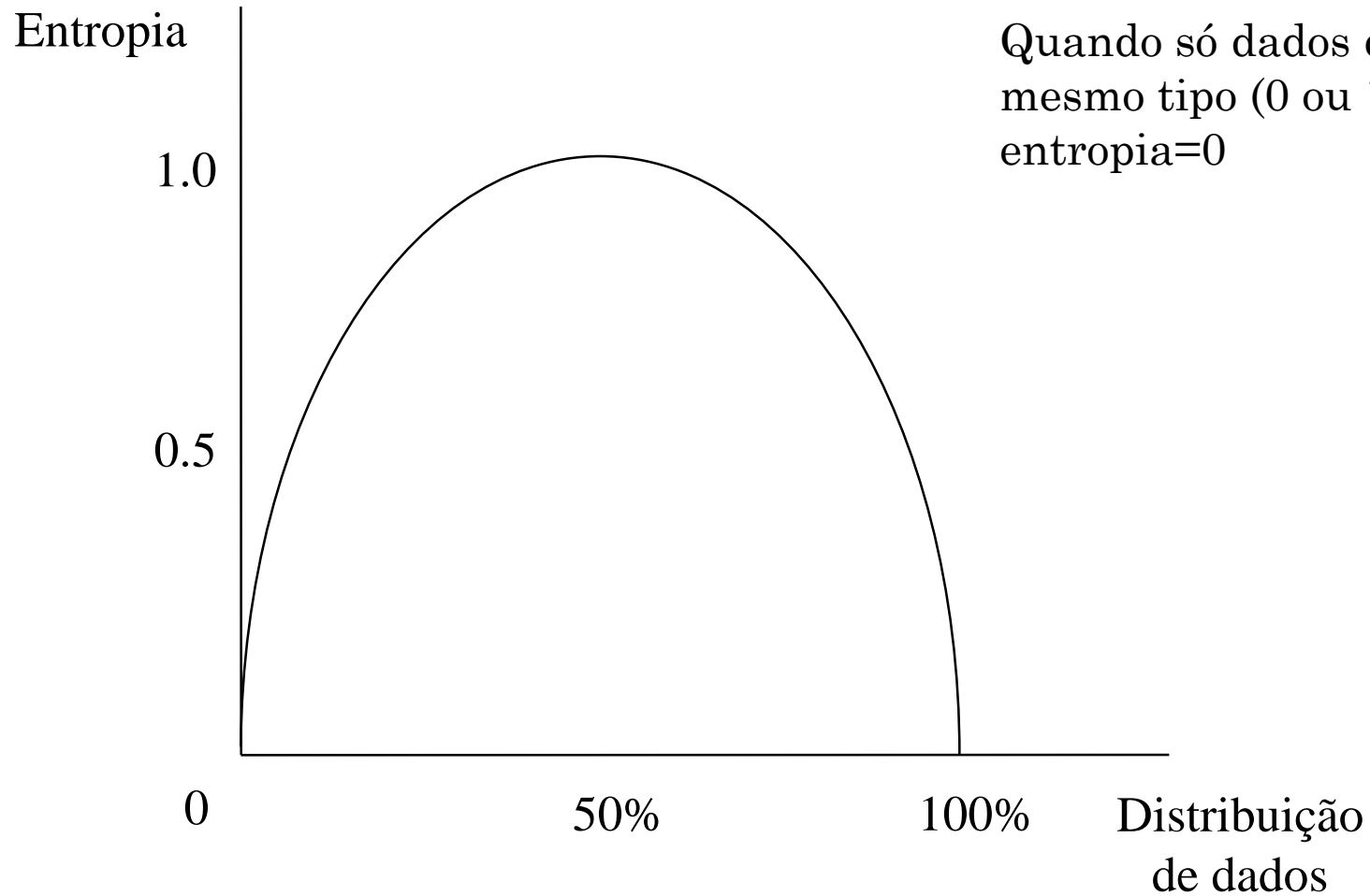
ENTROPIA

- **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!

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.

ENTROPIA



ENTROPIA

- A entropia é 0 se todos os exemplos são do mesmo tipo
 - Uma sequência 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 PLN, diferentes entropias podem indicar situações variadas
 - O que acontece com um fenômeno com alta entropia?
 - E em um com baixa entropia?

ENTROPIA

- Em um corpus em que **só há sentenças catafóricas**?
 - Entropia do fenômeno=0
- Em um corpus em que **metade das sentenças são catafóricas**?
 - Entropia do fenômeno=1
- Em um corpus em que **não há sentenças catafóricas**?
 - Entropia do fenômeno=0

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 frequências

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

- Entropia da língua

$$\begin{aligned} \text{Entropia}(S) = & -1/8 \cdot \log_2(1/8) - 1/4 \cdot \log_2(1/4) - 1/8 \cdot \log_2(1/8) \\ & - 1/4 \cdot \log_2(1/4) - 1/8 \cdot \log_2(1/8) - 1/8 \cdot \log_2(1/8) \end{aligned}$$

$$\text{Entropia}(S) = \mathbf{2,5 \text{ bits}}$$

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

ENTROPIA

○ Exemplo: língua polinésia simplificada

- Letras dessa língua e suas frequências

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

- Entropia da língua

$$\text{Entropia}(S) = -1/8 \cdot \log_2(1/8) - 1/4 \cdot \log_2(1/4) - 1/8 \cdot \log_2(1/8) \\ - 1/4 \cdot \log_2(1/4) - 1/8 \cdot \log_2(1/8)$$

$$\text{Entropia}(S) = \mathbf{2,5 \text{ bits}}$$

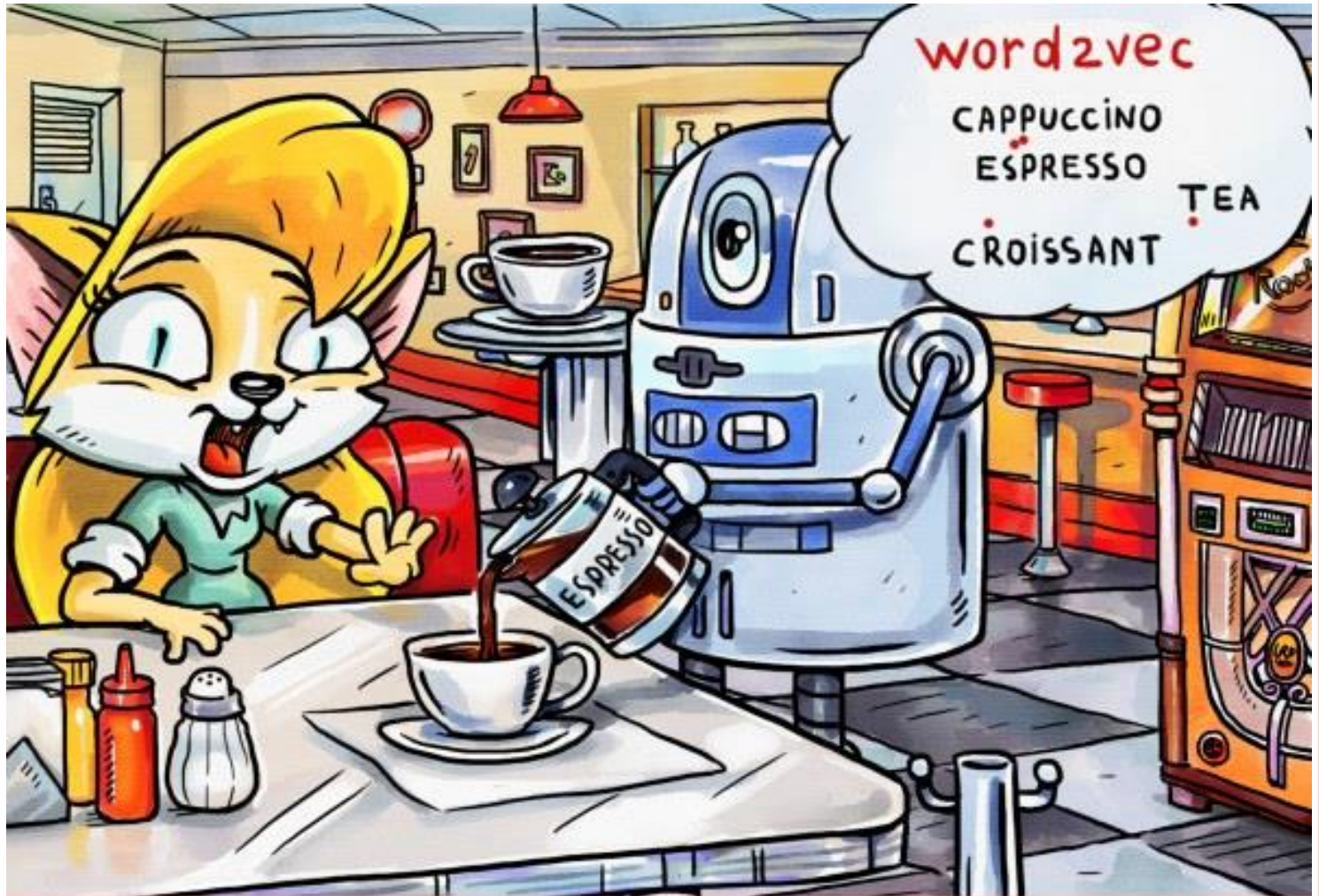
Menores códigos para letras
mais frequentes

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

ENTROPIA

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

Alguém consegue explicar essa figura?



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

MAIS RECENTEMENTE...

- O retorno dos **modelos distribucionais**
 - Inspiração no modelo do espaço vetorial de **Salton** (1971), originalmente aplicado para Recuperação de Informação, e em hipóteses linguísticas anteriores
 - Sofisticação e eficiência recentes
 - Grande poder computacional disponível
 - Grande volume de dados para “aprendizado”

INTUIÇÃO

- Palavras que ocorrem/se “distribuem” nos mesmos contextos tendem a ter o mesmo sentido
 - **Hipótese distribucional**, formulada por linguistas na década de 50 (Joos, 1950; Harris, 1954; Firth 1957)
 - Firth (1957)
 - *You shall know a word by the company it keeps!*

oculist and eye-doctor ... occur in almost the same environments

*A bottle of **tesgüino** is on the table
Everybody likes **tesgüino**
Tesgüino makes you drunk
We make **tesgüino** out of corn.
→ bebida alcóolica*

A IDEIA

- Osgood et al. (1957) e a tentativa de capturar o significado “afetivo” de cada palavra em um vetor de 3 dimensões
 - Valência (*valence*), intensidade (*arousal*) e controle (*dominance*) sobre o estímulo

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24
life	6.68	5.59	5.89

- Ideia revolucionária: “semântica vetorial”
 - O significado de *heartbreak* pode ser representado pelo vetor [2.45, 5.65, 3.58]

APLICAÇÕES

- Em várias frentes
 - Recuperação de Informação
 - Inferência Textual
 - Similaridade Lexical
 - Análise de Sentimentos
 - Tradução Automática
 - Etc.

MATRIZ TERMO-DOCUMENTO

- Ocorrência de palavras em 4 obras literárias
 - Matriz termo-documento de obras de Shakespeare

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- Cada documento é representado por um vetor
 - O espaço vetorial é, portanto, um conjunto de vetores

MATRIZ TERMO-DOCUMENTO

- Dois documentos são similares se seus vetores são similares

	<i>As You Like It</i>	<i>Twelfth Night</i>	<i>Julius Caesar</i>	<i>Henry V</i>
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

MATRIZ TERMO-DOCUMENTO

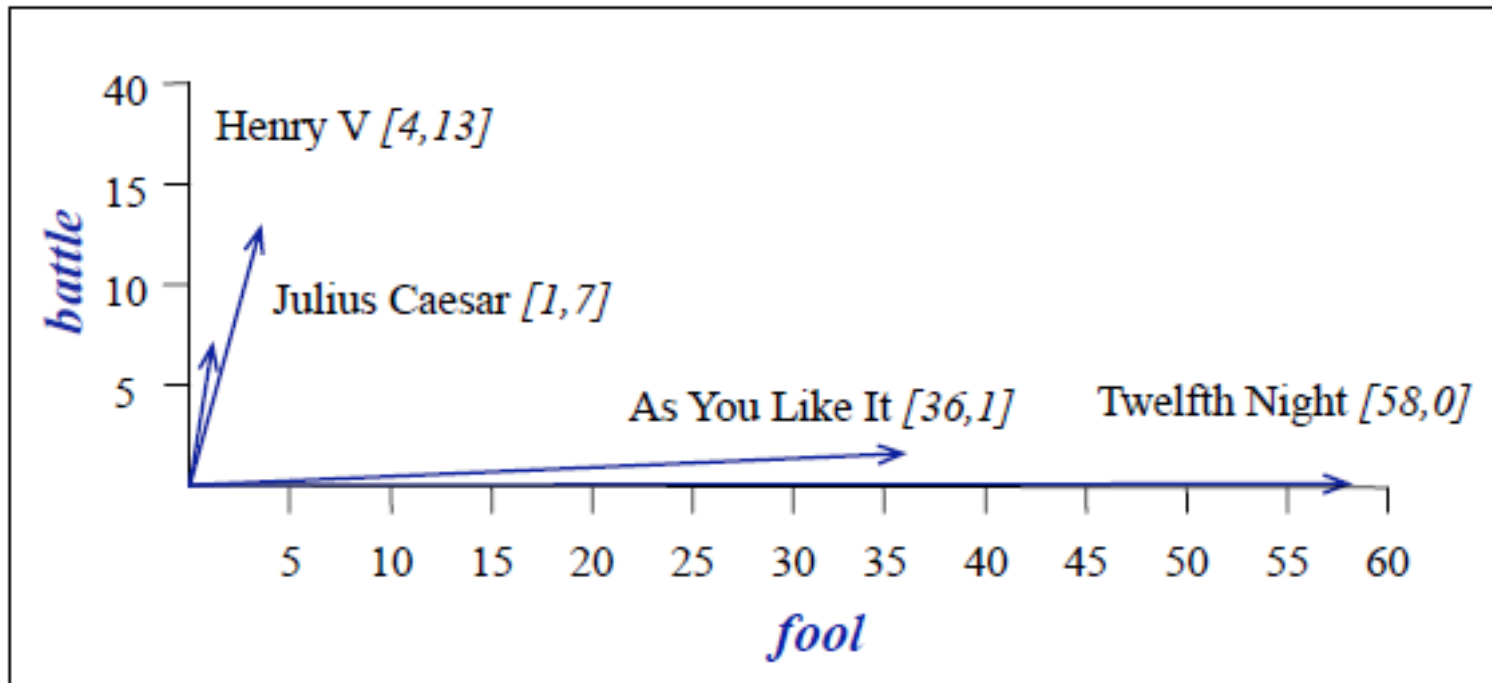
- Dois documentos são similares se seus vetores são similares

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Como essa ideia é usada em Recuperação de Informação?

PROJEÇÃO DAS OBRAS NO ESPAÇO VETORIAL

- Exemplo considerando apenas as dimensões “*fool*” e “*battle*”
 - Similaridade entre obras visualmente representada



CÔMPUTOS SOBRE AS MATRIZES

○ Células de uma matriz

- Frequência simples (de ocorrência de um termo no escopo considerado)
- Valor booleano (a palavra ocorreu ou não no escopo considerado)
- TF-IDF (*Term Frequency – Inverse Document Frequency*)
- Etc.

○ Usando ou não suavização

- Para que serve a suavização?

TF-IDF

- Alternativa mais interessante para a frequência simples, que pode ser enganosa
 - Termos muito comuns em todos os documentos não são discriminativos deles, por exemplo
- TF-IDF privilegia **termos frequentes em um documento** (TF) que **ocorrem relativamente pouco nos demais** (IDF), ou seja, que sejam discriminativos do documento de interesse

TF-IDF

- $\text{TF-IDF}(\text{termo } t, \text{ documento } d) = \text{TF}(t,d) * \text{IDF}(t)$
 - $\text{TF}(t,d)$ = frequência simples de t em d
 - Também é comum aplicar o log e suavizar (por quê?)
 - $\text{IDF}(t) = \log_{10}(N/df(t))$
 - N : número de documentos da coleção
 - $df(t)$: número de documentos da coleção em que t ocorreu
 - Quanto menor o número de documentos em que t ocorre, maior o $\text{IDF}(t)$
 - Exemplo: a palavra Romeu ocorre (113 vezes) somente em uma obra de Shakespeare (dentre 37 selecionadas)
 - $\text{TF}(\text{Romeu}) = \log_{10}(113+1) = 2,05$
 - $\text{IDF} = \log_{10}(37/1) = 1,56$
 - $\text{TF-IDF} = 2,05 * 1,57 = 3,22$

EXEMPLO

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3



De frequência
para TF-IDF

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Termos mais comuns em todas as obras (como “*good*”) perdem a importância, enquanto os mais discriminativos (como “*battle*”) ganham importância

MATRIZ TERMO-DOCUMENTO

- Base para a representação *Bag Of Words (BOW)*
 - Muito comum em Aprendizado de Máquina
 - Considerada *baseline* para muitas tarefas
 - E, por incrível que pareça, se sai muito bem em muitas tarefas
- Mas com claras limitações
 - Não considera a ordem e estruturação dos termos
 - Semântica limitada
 - Alta dimensionalidade
 - Estratégias para lidar: remoção de stopwords, normalização das palavras (por exemplo, lematização, radicalização e nominalização), corte por frequência

MATRIZ TERMO-CONTEXTO

- Para representar palavras, independentemente de obra, é mais usual ter matriz **termo-contexto**, ou **termo-termo**
 - Em vez de documentos inteiros, usam-se os contextos das palavras (em uma janela pré-especificada, com N palavras à esquerda e à direita da palavra em questão)
 - Palavras são similares se seus **contextos** são similares!

MATRIZ TERMO-CONTEXTO

- Exemplo: 4 palavras e suas co-ocorrências na Wikipédia, considerando uma janela de +- 4 palavras

is traditionally followed by **cherry** pie, a traditional dessert
often mixed, such as **strawberry** rhubarb pie. Apple pie
computer peripherals and personal **digital** assistants. These devices usually
a computer. This includes **information** available on the internet
...



Mapeamento para
matriz termo-contexto

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

Vetor de “digital” = [0, ..., 1670, 1683, 85, 5, 4, ...], mais próximo de “information” do que de “cherry”

MATRIZ TERMO-CONTEXT

Estamos capturando alguma semântica! A hipótese distribucional realmente ocorre!

- Exemplo: 4 palavras e suas descrições considerando uma janela de +/-

is traditionally followed by **cherry** pie, a traditional dessert
often mixed, such as **strawberry** rhubarb pie. Apple pie
computer peripherals and personal **digital** assistants. These devices usually
a computer. This includes **information** available on the internet
...



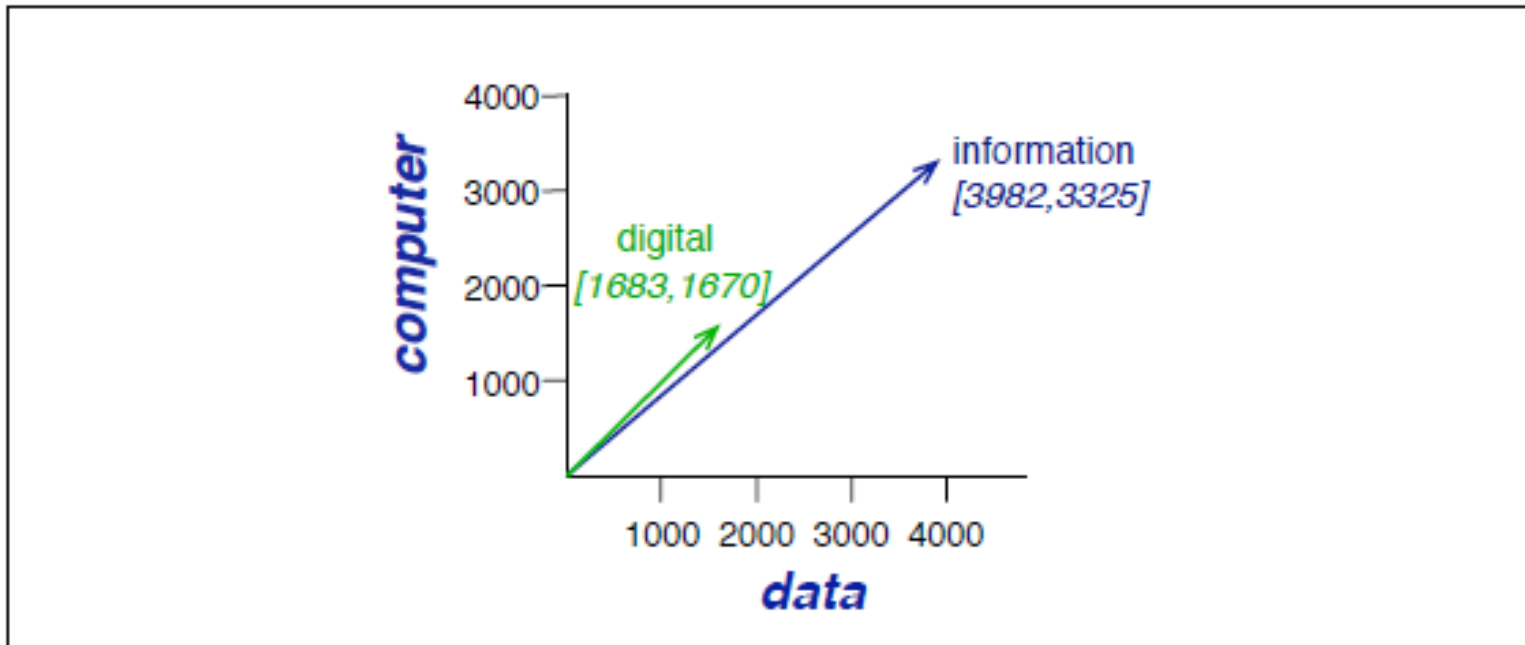
Mapeamento para matriz termo-contexto

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

Vetor de “digital” = [0, ..., 1670, 1683, 85, 5, 4, ...], mais próximo de “information” do que de “cherry”

PROJEÇÃO

- Considerando apenas as dimensões “*computer*” e “*data*”, vemos que as palavras são próximas



CÁLCULO COM VETORES

- Como calcular a distância (“similaridade”) entre os vetores, para descobrir, por exemplo, que “*digital*” é mais próxima de “*information*” do que de “*cherry*”?
 - Cosseno do ângulo entre os vetores!
 - Quanto menor o ângulo, maior o cosseno

MEDIDA DO COSSENO

- Produto escalar normalizado dos vetores
 - Quanto maiores os valores nas mesmas dimensões, maior a similaridade
 - Valor 1 se os vetores apontam na mesma direção; 0 se são ortogonais; -1 se apontam em direções opostas
- Cálculo entre dois vetores v e w

$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| |\mathbf{w}|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

MEDIDA DO COSSENO - EXEMPLO

- Considerando a pequena tabela termo-contexto abaixo

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

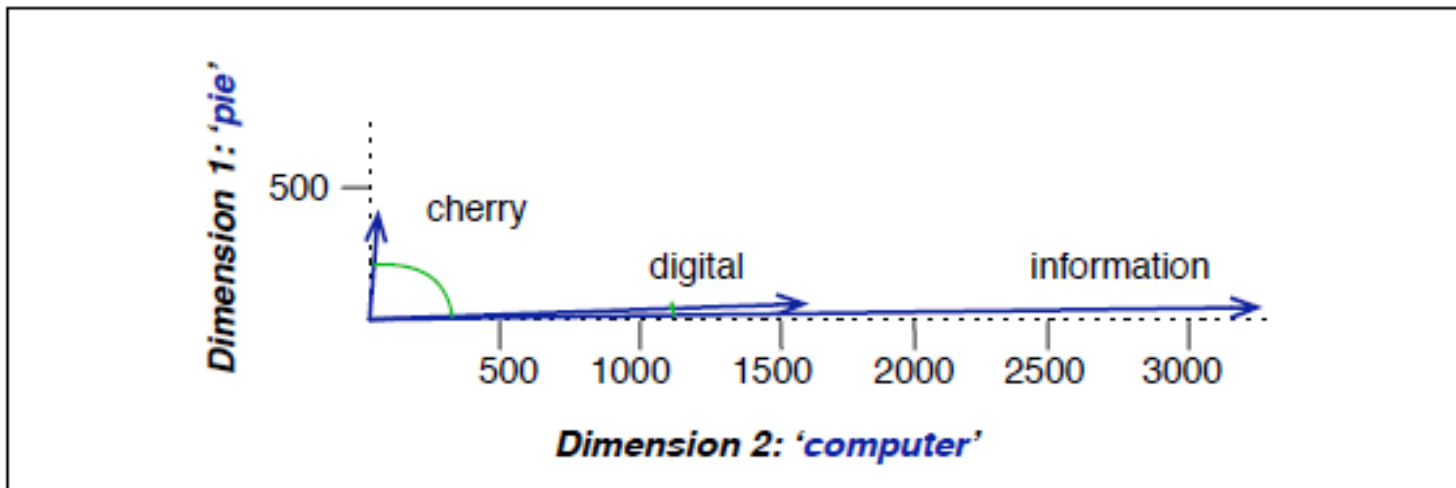
$$\cos(\text{cherry}, \text{information}) = \frac{442 * 5 + 8 * 3982 + 2 * 3325}{\sqrt{442^2 + 8^2 + 2^2} \sqrt{5^2 + 3982^2 + 3325^2}} = .017$$

$$\cos(\text{digital}, \text{information}) = \frac{5 * 5 + 1683 * 3982 + 1670 * 3325}{\sqrt{5^2 + 1683^2 + 1670^2} \sqrt{5^2 + 3982^2 + 3325^2}} = .996$$

- O termo “*information*” é muito mais similar a “*digital*” do que a “*cherry*”

MEDIDA DO COSSENO - EXEMPLO

- Representação gráfica da situação usando apenas 2 dimensões (“*pie*” e “*computer*”)



MATRIZ TERMO-CONTEXTO

- Exemplos anteriores para fins didáticos apenas
- Matrizes reais têm **milhares de linhas x milhares de colunas**
 - Altamente esparsas
 - Muitos zeros
- **Janela**
 - Tamanho variável, dependente do propósito
 - Se mais curta, mais sintática
 - Se mais longa, mais semântica
 - Linguisticamente motivada ou não
 - N-gramas, sintagmas, sentenças, parágrafos

VETORES

- Até então, vetores muito **esparso**
 - Muito grandes (20.000 a 50.000 elementos nas linhas e colunas... podendo haver muito mais)
- Tentativas de torna-los mais **denso**
 - Eficiência de representação: apenas “termos”/“dimensões” mais significativas
 - 50 a 1.000 termos, aproximadamente
 - Mais capacidade de capturar semântica (menos perceptível para modelos do estilo *bag of words*)
 - Eficiência computacional
 - Menos parâmetros para treinar em AM, melhor generalização e menos *overfitting*
- Surgimento do termo mais moderno: ***word embeddings***
 - A palavra está “embutida” (*embedded*) no espaço vetorial!

MÉTODOS

- Algumas abordagens já tradicionais (apesar de algumas serem bastante recentes)
 - **SVD** – *Singular Value Decomposition*
 - LSA (Deerwester et al., 1988)
 - **Redes neurais** (Bengio et al., 2003) e modelos preditivos
 - “Skip-grams” e “continuous bag of words” (Mikolov et al., 2013)
 - Métodos incorporados no pacote **word2vec**
 - **Métodos baseados em contagem**
 - GloVe (Pennington et al., 2014)
 - **BERT** (Devlin et al., 2019) e modelos contextuais
 - **BERTimbau** para o português (Souza et al., 2020)
- E muitas outras variações, para diferentes propósitos, inclusive
 - FastText, Wang2Vec, Doc2Vec, ELMo, RoBERTa, DeBERTa, Product2Vec, code2vec, etc.

TAREFAS

- Leitura da semana
 - Knight, K. (1999). A Statistical MT Tutorial Workbook.
 - No e-Disciplinas
- Provinha 5 disponível à tarde no e-Disciplinas