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

Aula 4: Representação de textos com técnicas tradicionais

Prof. Dennis Giovani Balreira

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

Conteúdo

- Introdução à representação de textos
- Modelos tradicionais de representação textual:
  - Bag of words
  - TF-IDF
- Medidas de similaridade entre textos
  - Distância de cosseno
- Problemas de representações tradicionais

Onde estamos em PLN?

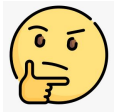
- Algoritmos tradicionais
  - Predominantes entre o final dos anos 1990 até ~2016
  - BoW features + Aprendizado de Máquina

- Embeddings fixas + Deep Learning
  - Predominates de ~2014 até ~2019
  - Word2vec, Glove, FastText + LSTM

- Embeddings contextuais + Large Language Models
  - Estado da arte em diversas tarefas
  - BERT, GPT, etc.

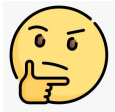
Representação de textos

- Como representar textos em computação?
  - No sentido de tipos de dados?



Representação de textos

- Como representar textos em computação?
  - No sentido de tipos de dados?
    - Caracteres
    - Strings
    - …
  - São na verdade símbolos codificados em “binários”
    - Mas qual relação eles possuem?



97	01100001	a
98	01100010	b
99	01100011	c
100	01100100	d
101	01100101	e
102	01100110	f
103	01100111	g

Representação de textos

- Como representar textos em computação?
  - No sentido de tipos de dados?
    - Caracteres
    - Strings
    - …
  - São na verdade símbolos codificados em “binários”
    - Mas qual relação eles possuem?
      - ‘a’ está antes de ‘b’ em ASCII
      - “amor” e “paixão”:
        - "amor": [97, 109, 111, 114]
        - "paixão": [112, 97, 105, 120, 227, 111]

97	01100001	a
98	01100010	b
99	01100011	c
100	01100100	d
101	01100101	e
102	01100110	f
103	01100111	g

Representação de textos

- Como representar textos em computação?
  - No sentido de tipos de dados?
    - Caracteres
    - Strings
    - …
  - São na verdade símbolos codificados em “binários”
    - Mas qual relação eles possuem?
      - ‘a’ está antes de ‘b’ em ASCII
      - “amor” e “paixão”:
        - "amor": [97, 109, 111, 114]
        - "paixão": [112, 97, 105, 120, 227, 111]

97	01100001	a
98	01100010	b
99	01100011	c
100	01100100	d
101	01100101	e
102	01100110	f
103	01100111	g

Legal, mas ‘a’ não tem nada a ver com ‘b’!

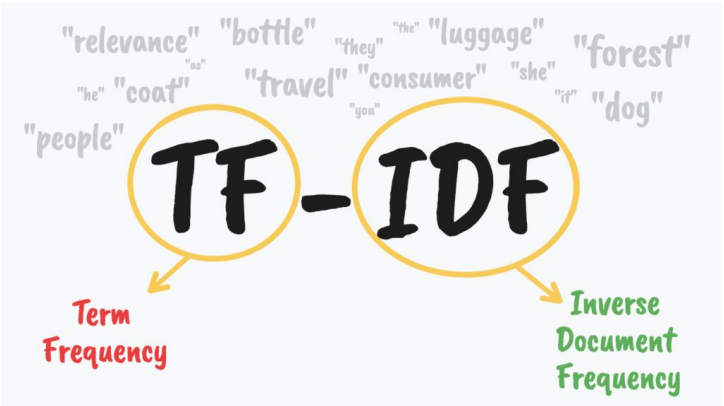
Números muito diferentes em palavras muito relacionadas!

Representação de textos

- Problema fundamental:
  - Máquinas processam números!
  - Mas textos são por natureza simbólicos…
    - Simbólico: palavras, expressões
    - Numérico: números, vetores, matrizes
  - Queremos poder processar ~~texto~~ números
    - Isso requer formas de representar os textos

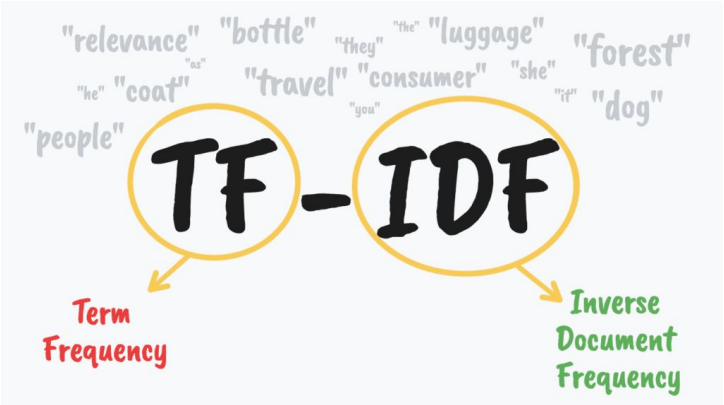
Modelos tradicionais de representação textual

- Modelos tradicionais para representar texto:
  - Bag of Words (BoW)
  - Term Frequency - Inverse Document Frequency (TF-IDF)



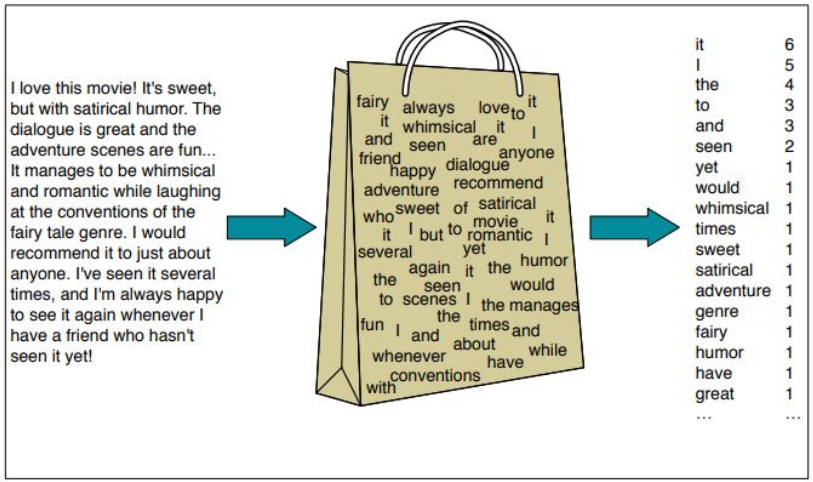
Modelos tradicionais de representação textual

- Modelos tradicionais para representar texto:
  - Bag of Words (BoW)
  - Term Frequency - Inverse Document Frequency (TF-IDF)



Modelos tradicionais: BoW

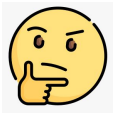
- Bag of Words (BoW)
  - Cada documento é transformado em um vetor, onde cada elemento representa a frequência de uma palavra específica
    - “Documento” neste caso é qualquer texto



	about	bird	heard	is	the	word	you
Doc1	About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
Doc2	You heard about the bird	1	1	1	0	1	0	1
Doc3	The bird is the word	0	1	0	1	2	1	0

Modelos tradicionais: BoW

- Bag of Words (BoW)
  - É uma das formas mais simples de representar texto
  - Qual o tamanho (dimensão) do vetor gerado?
    - Depende do que?



Modelos tradicionais: BoW

- Bag of Words (BoW)
  - É uma das formas mais simples de representar texto
  - Qual o tamanho (dimensão) do vetor gerado?
    - Depende do número de "termos" distintos que aparecem no vocabulário de todos os documentos!
  - Cada documento gera um vetor diferente, mas todos possuem a mesma dimensão!

Document D1	*The child makes the dog happy* the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	*The dog makes the child happy* the: 2, child: 1, makes: 1, dog: 1, happy: 1

↓

	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	[1,1,1,1,2]
D2	1	1	1	1	2	[1,1,1,1,2]

$$\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

dj: documento j. **wk,j:** frequência da palavra k (considerando todo o vocabulário) no documento j.

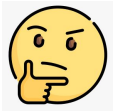
Modelos tradicionais: BoW

- Bag of Words (BoW)

  Exercício 1: Considere os seguintes documentos:

  Documento 1: "O gato correu rápido."
  Documento 2: "O cachorro correu devagar."
  Documento 3: "O gato e o cachorro correram."

  a) Qual o vocabulário dos documentos?
  b) Qual a dimensão (tamanho) dos vetores a serem gerados?
  c) Quais os vetores d1, d2 e d3?



Modelos tradicionais: BoW

- Bag of Words (BoW)

  Exercício 1: Considere os seguintes documentos:

  Documento 1: "O gato correu rápido."
  Documento 2: "O cachorro correu devagar."
  Documento 3: "O gato e o cachorro correram."

  a) Qual o vocabulário dos documentos?

  Vocabulário: ["O", "gato", "correu", "rápido", "cachorro", "devagar", "e", "correram"]

  b) Qual a dimensão (tamanho) dos vetores a serem gerados?

  8 (vocabulário contém oito palavras únicas)

  c) Quais os vetores d1, d2 e d3?

  d1 (Documento 1): [1, 1, 1, 1, 0, 0, 0, 0]
  d2 (Documento 2): [1, 0, 1, 0, 1, 1, 0, 0]
  d3 (Documento 3): [2, 1, 0, 0, 1, 0, 1, 1]

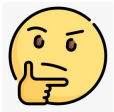
Modelos tradicionais: BoW

- Bag of Words (BoW)
  - Considere as frases:
    - d1: "O time venceu o jogo."
    - d2: "A equipe entendeu a estratégia."

  - BoW:
    - vocab: ['a', 'entendeu', 'equipe', 'estratégia', 'jogo', 'o', 'time', 'venceu']
    - d1: [0, 0, 0, 0, 1, 2, 1, 1]
    - d2: [2, 1, 1, 1, 0, 0, 0, 0]

Modelos tradicionais: BoW

- Bag of Words (BoW)
  - Considere as frases:
    - d1: "O time venceu o jogo."
    - d2: "A equipe entendeu a estratégia."
  - BoW:
    - vocab: ['a', 'entendeu', 'equipe', 'estratégia', 'jogo', 'o', 'time', 'venceu']
    - d1: [0, 0, 0, 0, 1, 2, 1, 1]
    - d2: [2, 1, 1, 1, 0, 0, 0, 0]
  - Qual problema principal temos usando BoW?



Modelos tradicionais: BoW

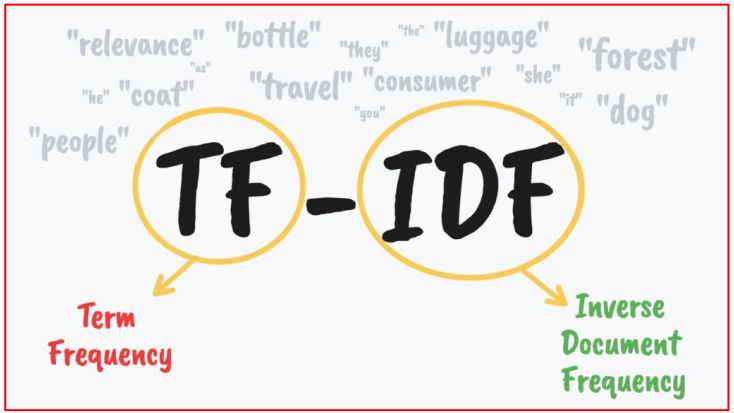
- Bag of Words (BoW)
  - Considere as frases:
    - d1: "O time venceu o jogo."
    - d2: "A equipe entendeu a estratégia."

  - BoW:
    - vocab: ['a', 'entendeu', 'equipe', 'estratégia', 'jogo', 'o', 'time', 'venceu']
    - d1: [0, 0, 0, 0, 1, 2, 1, 1]
    - d2: [2, 1, 1, 1, 0, 0, 0, 0]

  - Qual problema principal temos usando BoW?
    - Palavras “fracas” (não representam o contexto da frase) e palavras “fortes” (representam o contexto) possuem o mesmo peso

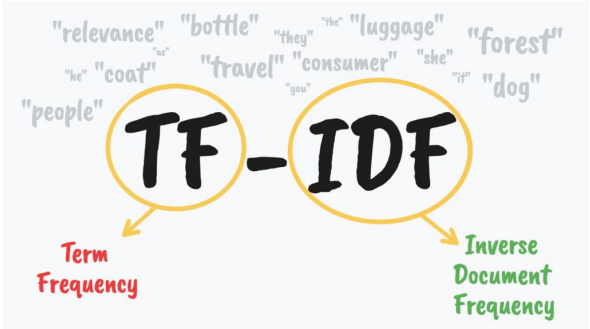
Modelos tradicionais: TF-IDF

- Modelos tradicionais para representar texto:
  - Bag of Words (BoW)
  - Term Frequency - Inverse Document Frequency (TF-IDF)



Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Mede a relevância de uma palavra em um documento em relação a um conjunto de documentos
  - Ajusta a contagem simples de palavras do BoW ponderando as palavras de acordo com sua importância (termos raros tem mais importância)
  - Proposto por Karen Spärck Jones (1972)
  - É formado pela multiplicação entre duas partes: TF x IDF
    - *Term Frequency* (TF): mede quantas vezes a palavra aparece no documento
    - *Inverse Document Frequency* (IDF): mede o quanto uma palavra é rara em todos os documentos do corpus



(26/08/1935 - 04/04/2007)

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)

termo documento número de vezes que a palavra t aparece no documento d

$$TF(t,d) = \frac{f_{t,d}}{N_d}$$

número total de palavras no documento d

número total de documentos do corpus

base 10

$$IDF(t) = \log\left(\frac{N}{1+df(t)}\right)$$

número de documentos que contêm o termo t

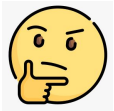
$$\text{TF-IDF}(t,d) = TF(t,d) \times IDF(t)$$

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Exemplo:
    - Documento 1: "O time venceu o jogo."
    - Documento 2: "A equipe entendeu a estratégia."
    - Documento 3: "O time entendeu a estratégia de jogo."

Qual o TF-IDF de todos os termos para cada documento?

(Considere após remoção das stopwords “o” e “a”)



$$TF(t,d) = \frac{f_{t,d}}{N_d}$$

$$IDF(t) = \log\left(\frac{N}{1+df(t)}\right)$$

$$\text{TF-IDF}(t,d) = TF(t,d) \times IDF(t)$$

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Exemplo:
    - Documento 1: "O time venceu o jogo."
    - Documento 2: "A equipe entendeu a estratégia."
    - Documento 3: "O time entendeu a estratégia de jogo."

$$TF(t,d) = \frac{f_{t,d}}{N_d}$$

$$IDF(t) = \log\left(\frac{N}{1+df(t)}\right)$$

$$\text{TF-IDF}(t,d) = TF(t,d) \times IDF(t)$$

Qual o TF-IDF de todos os termos para cada documento?

Vocabulário final (sem stopwords): ["time", "venceu", "jogo", "equipe", "entendeu", "estratégia"]

Documento 1: ["time", "venceu", "jogo"]

- "time": $TF(\text{"}time\text{"}) = 1/3 = 0.333$
- "venceu": $TF(\text{"}venceu\text{"}) = 1/3 = 0.333$
- "jogo": $TF(\text{"}jogo\text{"}) = 1/3 = 0.333$

Documento 2: ["equipe", "entendeu", "estratégia"]

- "equipe": $TF(\text{"}equipe\text{"}) = 1/3 = 0.333$
- "entendeu": $TF(\text{"}entendeu\text{"}) = 1/3 = 0.333$
- "estratégia": $TF(\text{"}estratégia\text{"}) = 1/3 = 0.333$

Documento 3: ["time", "entendeu", "estratégia", "jogo"]

- "time": $TF(\text{"}time\text{"}) = 1/4 = 0.25$
- "entendeu": $TF(\text{"}entendeu\text{"}) = 1/4 = 0.25$
- "estratégia": $TF(\text{"}estratégia\text{"}) = 1/4 = 0.25$
- "jogo": $TF(\text{"}jogo\text{"}) = 1/4 = 0.25$

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Exemplo:
    - Documento 1: "O time venceu o jogo."
    - Documento 2: "A equipe entendeu a estratégia."
    - Documento 3: "O time entendeu a estratégia de jogo."

$$TF(t,d) = \frac{f_{t,d}}{N_d}$$

$$IDF(t) = \log\left(\frac{N}{1+df(t)}\right)$$

$$\text{TF-IDF}(t,d) = TF(t,d) \times IDF(t)$$

Qual o TF-IDF de todos os termos para cada documento?

- "time" aparece em 2 documentos (1 e 3):

$$IDF(\text{"}time\text{"}) = \log\left(\frac{3}{1+2}\right) = \log(1) = 0$$

- "venceu" aparece em 1 documento (1):

$$IDF(\text{"}venceu\text{"}) = \log\left(\frac{3}{1+1}\right) = \log(1.5) \approx 0.176$$

- "jogo" aparece em 2 documentos (1 e 3):

$$IDF(\text{"}jogo\text{"}) = \log\left(\frac{3}{1+2}\right) = \log(1) = 0$$

- "equipe" aparece em 1 documento (2):

$$IDF(\text{"}equipe\text{"}) = \log\left(\frac{3}{1+1}\right) = \log(1.5) \approx 0.176$$

- "entendeu" aparece em 2 documentos (2 e 3):

$$IDF(\text{"}entendeu\text{"}) = \log\left(\frac{3}{1+2}\right) = \log(1) = 0$$

- "estratégia" aparece em 2 documentos (2 e 3):

$$IDF(\text{"}estratégia\text{"}) = \log\left(\frac{3}{1+2}\right) = \log(1) = 0$$

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Exemplo:
    - Documento 1: "O time venceu o jogo."
    - Documento 2: "A equipe entendeu a estratégia."
    - Documento 3: "O time entendeu a estratégia de jogo."

  Qual o TF-IDF de todos os termos para cada documento?

$$TF(t,d) = \frac{f_{t,d}}{N_d}$$

$$IDF(t) = \log\left(\frac{N}{1+df(t)}\right)$$

$$\text{TF-IDF}(t,d) = TF(t,d) \times IDF(t)$$

Documento 1:

- "time": $TF - IDF = 0.333 \times 0 = 0$
- "venceu": $TF - IDF = 0.333 \times 0.176 \approx 0.059$
- "jogo": $TF - IDF = 0.333 \times 0 = 0$

Documento 2:

- "equipe": $TF - IDF = 0.333 \times 0.176 \approx 0.059$
- "entendeu": $TF - IDF = 0.333 \times 0 = 0$
- "estratégia": $TF - IDF = 0.333 \times 0 = 0$

Documento 3:

- "time": $TF - IDF = 0.25 \times 0 = 0$
- "entendeu": $TF - IDF = 0.25 \times 0 = 0$
- "estratégia": $TF - IDF = 0.25 \times 0 = 0$
- "jogo": $TF - IDF = 0.25 \times 0 = 0$

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Análise com outro exemplo maior:
    Documento 1 (outros documentos omitidos):

"A **tecnologia** está mudando rapidamente o mundo em que vivemos. As empresas de **tecnologia** estão cada vez mais focadas em **inovação** para criar produtos que impactam positivamente a sociedade. A **inovação** em **inteligência** artificial (IA) é um dos principais fatores que impulsionam essas mudanças. As empresas estão investindo em pesquisa e desenvolvimento de IA para criar soluções inovadoras que transformem indústrias inteiras."

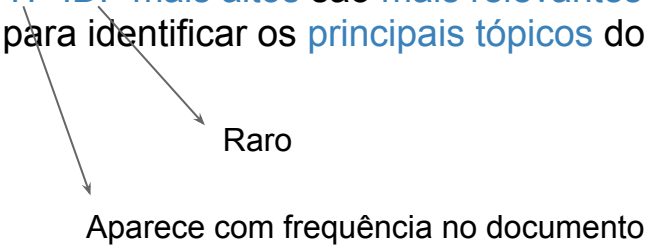
Assumindo os seguintes valores de TF-IDF:

TF-IDF("tecnologia") = $\frac{2}{50}$ × $\log\left(\frac{100}{70}\right)$ = 0.00616 → "Tecnologia" é uma palavra importante no documento, mas porque é comum no corpus tem valor de TF-IDF baixo.

TF-IDF("inteligência") = $\frac{1}{50}$ × $\log\left(\frac{100}{10}\right)$ = 0.02 → "Inteligência" é menos comum tanto no documento quanto no corpus, portanto, tem um valor de TF-IDF maior do que palavras muito comuns.

Modelos tradicionais: TF-IDF

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Destaca palavras que são frequentes em um documento específico e que não são comuns em outros documentos
    - Permite identificar termos que carregam mais significado em um contexto particular
  - Palavras com valores de TF-IDF mais altos são mais relevantes
    - Podem ser usadas para identificar os principais tópicos do documento

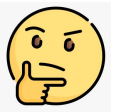


Similaridade de textos

- Considere os seguintes textos:

"Maria levou seu cachorro ao parque. Lá, o cachorro brincou com outros animais e correu feliz. No fim do dia, Maria deu ração ao seu cachorro."

"João levou seu gato ao veterinário. O gato ficou um pouco assustado com a presença de outros animais. Quando voltou para casa, João deu comida ao seu gato."

Quais semelhanças entre os textos? 

Similaridade de textos

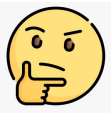
- Considere os seguintes textos:

"**Maria** levou seu **cachorro** ao parque. Lá, o **cachorro** brincou com outros **animais** e correu **feliz**. No fim do dia, **Maria** deu **ração** ao seu **cachorro**."

"**João** levou seu **gato** ao veterinário. O **gato** ficou **alegre** com a presença de outros **animais**. Quando voltou para casa, **João** deu **comida** ao seu **gato**."

Várias!! Algumas relações:
- Maria e João
- cachorro e gato
- animais
- alegre e feliz (sinônimo)
- ração e comida

Como achar similaridades sintáticas simples (com o que já vimos)?



Similaridade de textos

- Considere os seguintes textos:

"**Maria** levou seu **cachorro** ao parque. Lá, o **cachorro** brincou com outros **animais** e correu **feliz**. No fim do dia, **Maria** deu **ração** ao seu **cachorro**."

"**João** levou seu **gato** ao veterinário. O **gato** ficou **alegre** com a presença de outros **animais**. Quando voltou para casa, **João** deu **comida** ao seu **gato**."

Várias!! Algumas relações:
- Maria e João
- cachorro e gato
- animais
- alegre e feliz (sinônimo)
- ração e comida

Formas "sintáticas":
- 1. BoW: conta palavras comuns de dois textos
- 2. N-grams: conta a quantidade de palavras consecutivas iguais entre dois textos
- 3. Distância de *Levenshtein*: conta a quantidade de erros entre dois textos

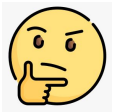
Similaridade de textos

- Considere os seguintes textos:

"**Maria** levou seu **cachorro** ao parque. Lá, o **cachorro** brincou com outros **animais** e correu **feliz**. No fim do dia, **Maria** deu **ração** ao seu **cachorro**."

"**João** levou seu **gato** ao veterinário. O **gato** ficou **alegre** com a presença de outros **animais**. Quando voltou para casa, **João** deu **comida** ao seu **gato**."

Mas e se quisermos tentar capturar similaridades semânticas (significado)?



Similaridade de textos

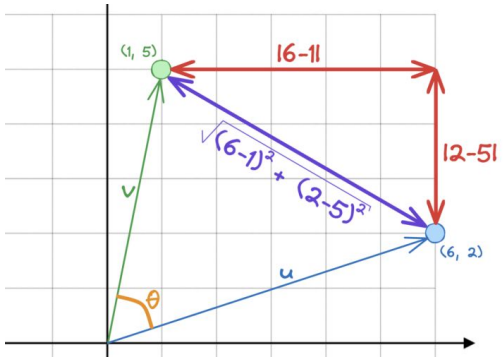
- Considere os seguintes textos:

"**Maria** levou seu **cachorro** ao parque. Lá, o **cachorro** brincou com outros **animais** e correu **feliz**. No fim do dia, **Maria** deu **ração** ao seu **cachorro**."

"**João** levou seu **gato** ao veterinário. O **gato** ficou **alegre** com a presença de outros **animais**. Quando voltou para casa, **João** deu **comida** ao seu **gato**."

Mas e se quisermos tentar capturar similaridades semânticas (significado)?

Podemos usar os modelos tradicionais para vetorizar palavras e utilizar distâncias para computar similaridade!



Similaridade de textos

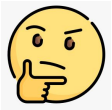
- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:

(Termos)

	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

(Documentos)

Quais textos são mais semelhantes?



Similaridade de textos

- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:

(Termos)

	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

(Documentos)

Quais textos são mais semelhantes?

- “fool” e “wit” são palavras que remetem a comédias

- “battle” remete a textos mais sérios

Similaridade de textos

- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:

(Termos)

	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

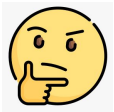
(Documentos)

Vetores:

As You Like It: [1,114,36,20]
Twelfth Night: [0,80,58,15]
Julius Caesar: [7,62,1,2]
Henry V: [13,89,4,3]

4D difícil de visualizar…

Como vê-los em 2D?



Similaridade de textos

- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:

(Termos)

	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

(Documentos)

Vetores:

As You Like It: [1,114,36,20]
Twelfth Night: [0,80,58,15]
Julius Caesar: [7,62,1,2]
Henry V: [13,89,4,3]

4D difícil de visualizar…

Como vê-los em 2D?

Utilizar técnicas de redução de dimensionalidade (transformar dados de alta dimensão em espaço de menor dimensão)
Ex: PCA, t-SNE

Similaridade de textos

- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:

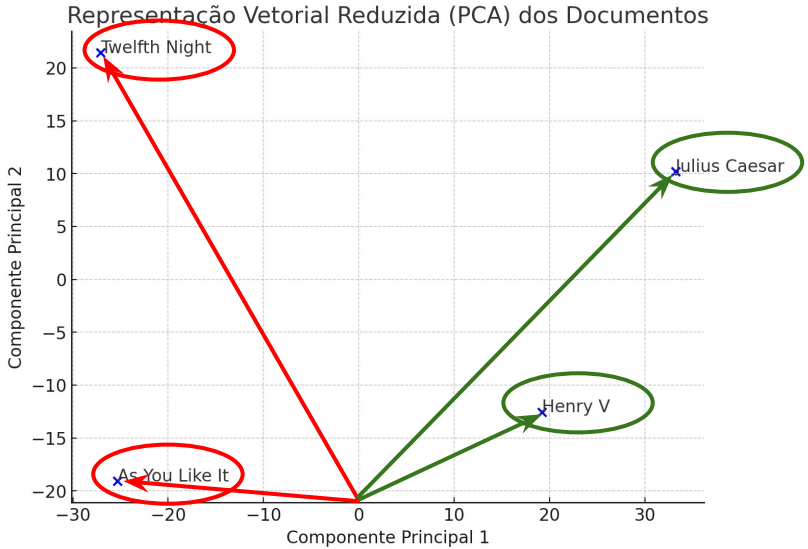
Vetores:

As You Like It: [1,114,36,20]
Twelfth Night: [0,80,58,15]
Julius Caesar: [7,62,1,2]
Henry V: [13,89,4,3]

Red. Dim

Vetores reduzidos em 2D:

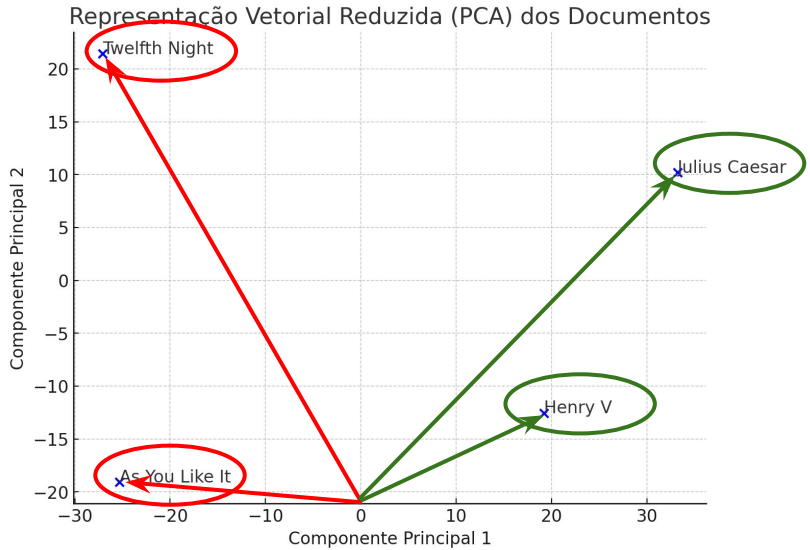
As You Like It: **[−25.35,−19.08]**
Twelfth Night: **[−27.10,21.45]**
Julius Caesar: **[33.23,10.20]**
Henry V: **[19.22,−12.56]**



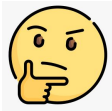
*PCA: *Principal Component Analysis*

Similaridade de textos

- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:

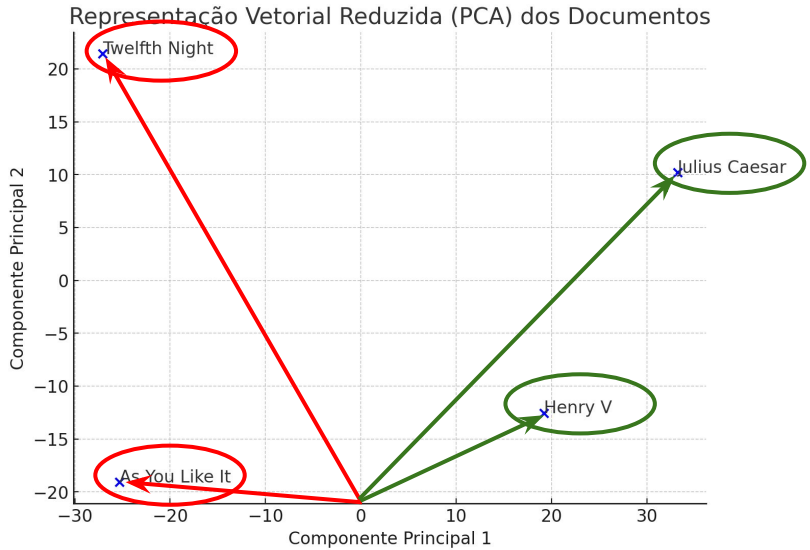


Qual distância mostra a similaridade de documentos?



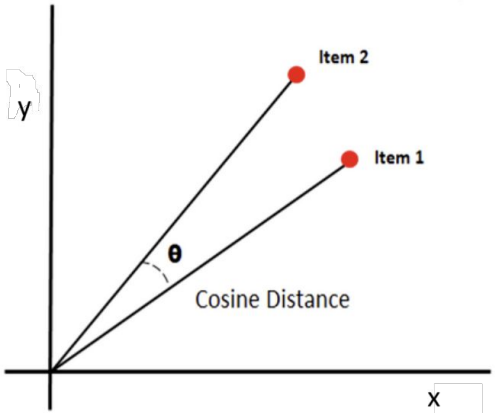
Similaridade de textos

- Ideia geral: documentos próximos no espaço vetorial têm conteúdo similar
  - Exemplo:



Qual distância mostra a similaridade de documentos?

- Euclidiana
- de Manhattan
- de Cosseno



Similaridade de textos: distância de cosseno

- Calcular o ângulo entre os vetores, em vez da magnitude
  - A similaridade do cosseno tem range [−1,1] e a distância [0,2], onde:
  - 0 indica que os vetores são idênticos
  - 1 indica que os vetores são completamente independentes (ortogonais)
  - 2 indica que os vetores estão em direções opostas

$$\text{s\_cos}(A, B) = \frac{A \cdot B}{||A||\,||B||}$$

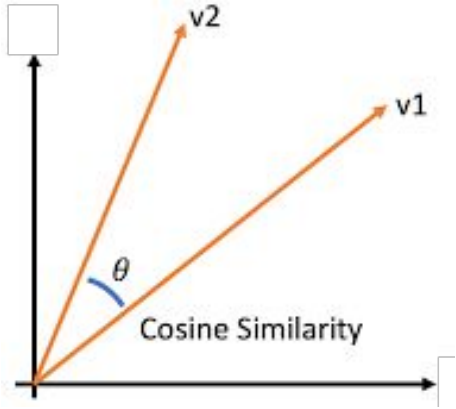
Produto escalar

Similaridade

Normas de A e B

$$\text{d\_cos}(A, B) = 1 - \text{s\_cos}(A, B)$$

Distância



Sobre os modelos tradicionais

- Ambas BoW e TF-IDF ajudam na representação numérica de palavras!
  - Abordagens simples e de fácil implementação
  - Interpretação intuitiva
    - Sabe-se exatamente o significado dos vetores gerados
  - Não precisam de treinamento prévio
    - São pouco custosos, sendo eficientes para pequenos conjuntos de textos

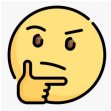
Questões fundamentais de “Linguagens Naturais”

- Ambiguidade: mesma palavra com significados distintos (contexto)
  Ex: "João foi ao banco para sacar dinheiro." vs.
  "João sentou no banco da praça para descansar."

- Sinônimos: diferentes palavras com mesmo significado,
  mas representadas de forma completamente diferente
  Ex: "O cachorro correu pelo parque." vs.
  "O canino correu pelo parque."

- Contexto: a ordem das palavras influencia o significado de uma frase
  Ex: "O cachorro mordeu o homem." vs. "O homem mordeu o cachorro."

  Como as abordagens BoW e TF-IDF lidam com estas questões?



Questões fundamentais de “Linguagens Naturais”

- Ambiguidade: mesma palavra com significados distintos (contexto)
  Ex: "João foi ao banco para sacar dinheiro." vs.
  "João sentou no banco da praça para descansar."

- Sinônimos: diferentes palavras com mesmo significado,
  mas representadas de forma completamente diferente
  Ex: "O cachorro correu pelo parque." vs.
  "O canino correu pelo parque."

- Contexto: a ordem das palavras influencia o significado de uma frase
  Ex: "O cachorro mordeu o homem." vs. "O homem mordeu o cachorro."

  Abordagens fundamentais (BoW e TF-IDF) permitem algum tipo de similaridade,
  mas não ajudam nestes pontos…

Material complementar

- Scikit-learn (https://scikit-learn.org/stable/)
  - TF-IDF: Fornece implementações eficientes para calcular TF-IDF com a classe TfidfVectorizer.
  - Bag of Words (BoW): A classe CountVectorizer permite gerar representações BoW.
  - Similaridade do cosseno: Pode ser calculada utilizando a função cosine_similarity.

Próximas aulas

- Aula prática (Laboratório 1)
- Aula teórica:
  - Aprendizado supervisionado para textos

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 Jurafsky)

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