

Classificação de textos

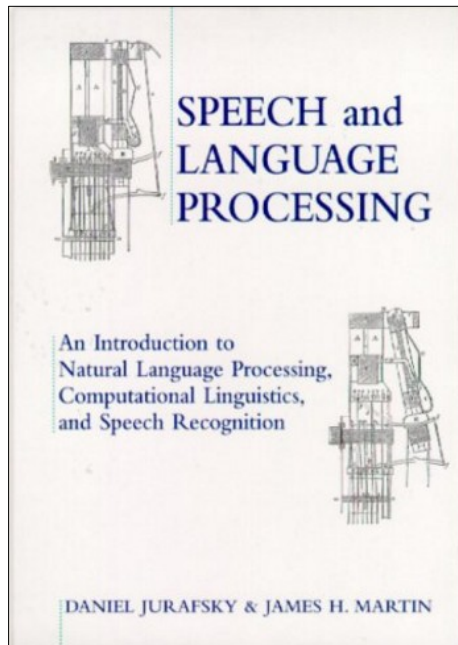
Prof. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

2Q-2019

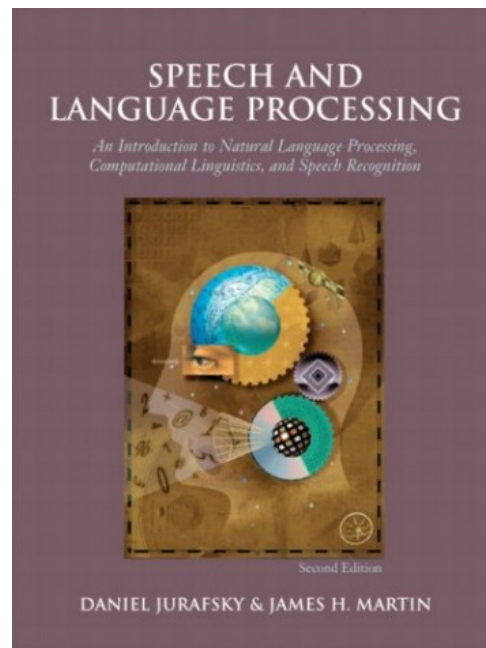
Bibliografia

Daniel Jurafsky & James H. Martin.

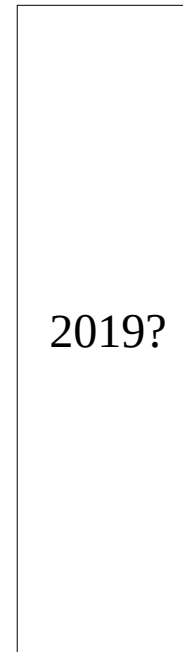
Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson/Prentice Hall.



2000



2009



2019?



Stanford University



University of Colorado, Boulder

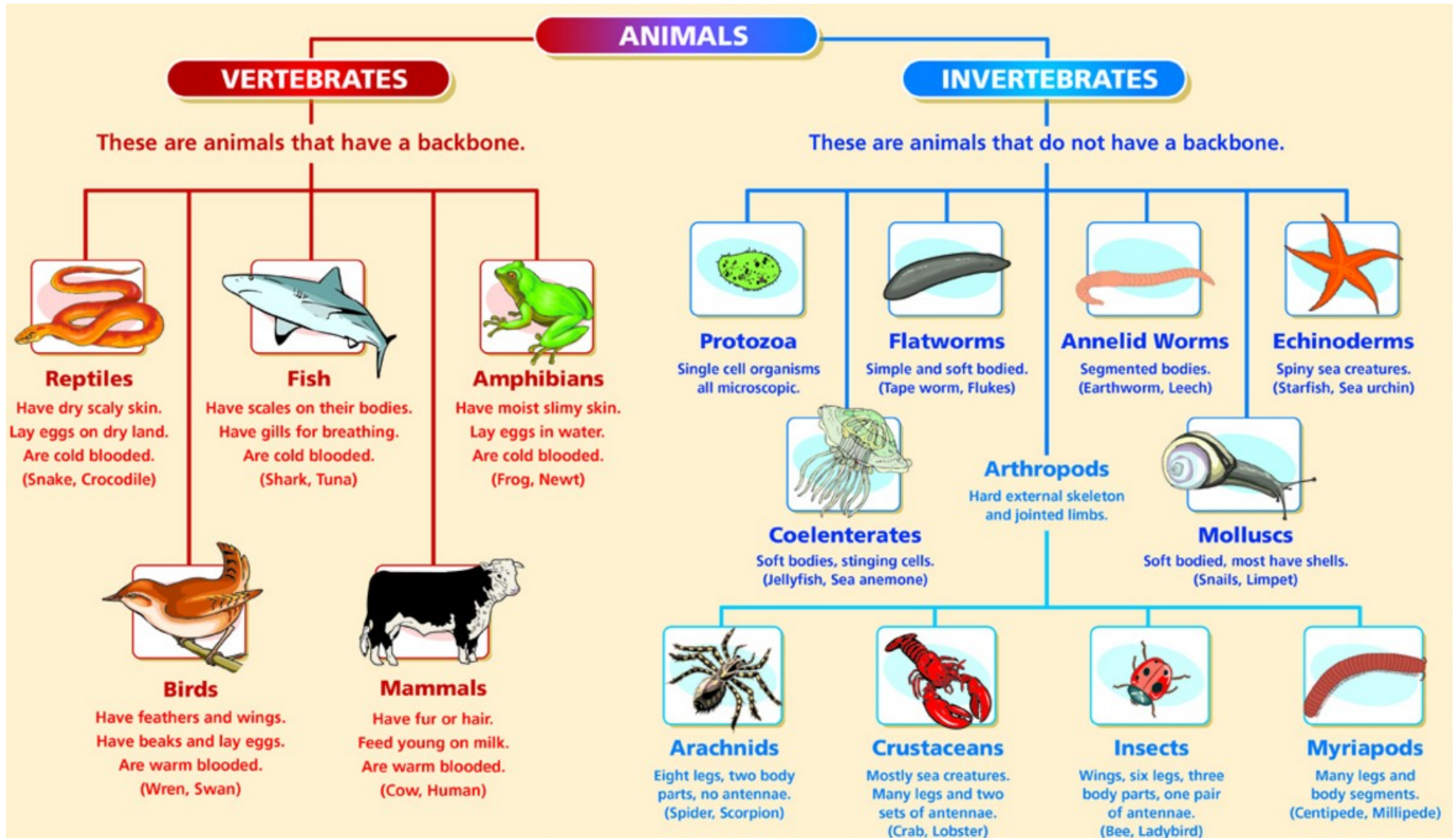
Bibliografía – Capítulo 6

Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

Chapter	Slides	Relation to 2nd ed.
1: Introduction		[Ch. 1 in 2nd ed.]
2: <u>Regular Expressions, Text Normalization, and Edit Distance</u>	Text [pptx] [pdf] Edit Distance [pptx] pdf]	[Ch. 2 and parts of Ch. 3 in 2nd ed.]
3: <u>Language Modeling with N-Grams</u>	LM [pptx] [pdf]	[Ch. 4 in 2nd ed.]
4: <u>Naive Bayes Classification and Sentiment</u>	NB [pptx] [pdf] Sentiment [pptx] [pdf]	[new in this edition]
5: <u>Logistic Regression</u>		
6: <u>Vector Semantics</u>	Vector1 [pptx] [pdf] Vector2 [pptx] [pdf]	

Classificação de animais

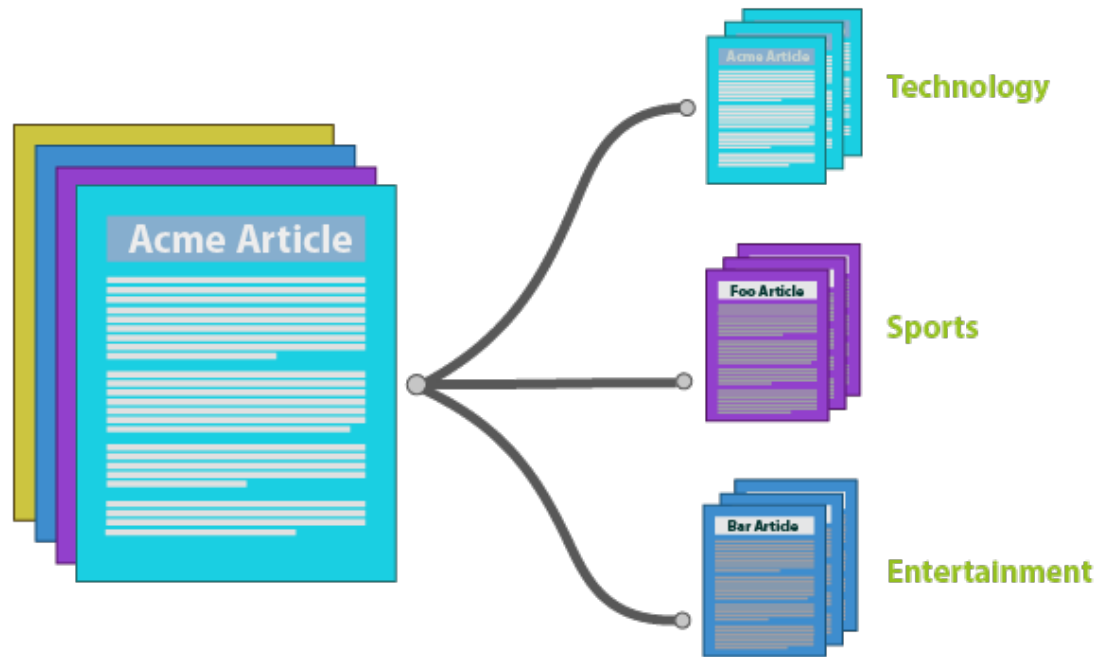


Classificação == distribuição por classes, categorias ou grupos com características semelhantes.



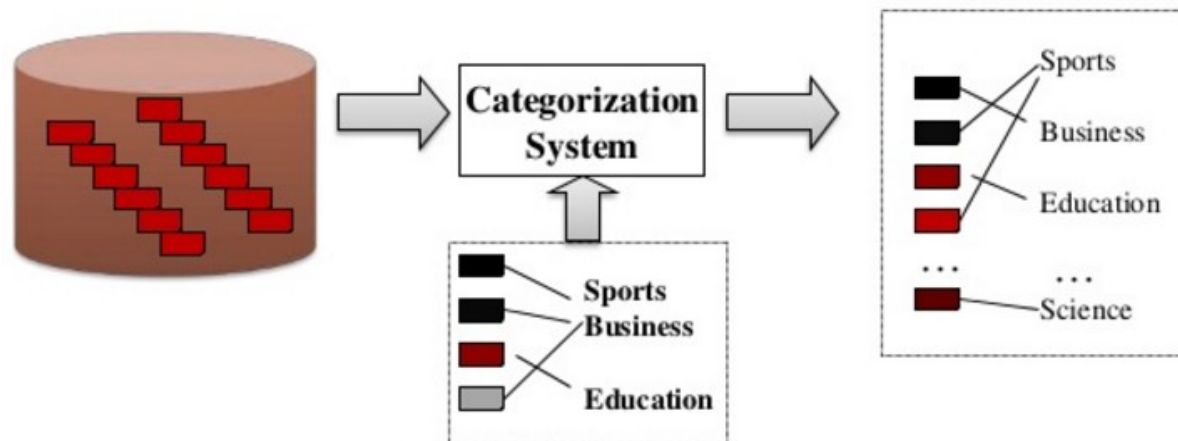
classificação de texto?

Text classification



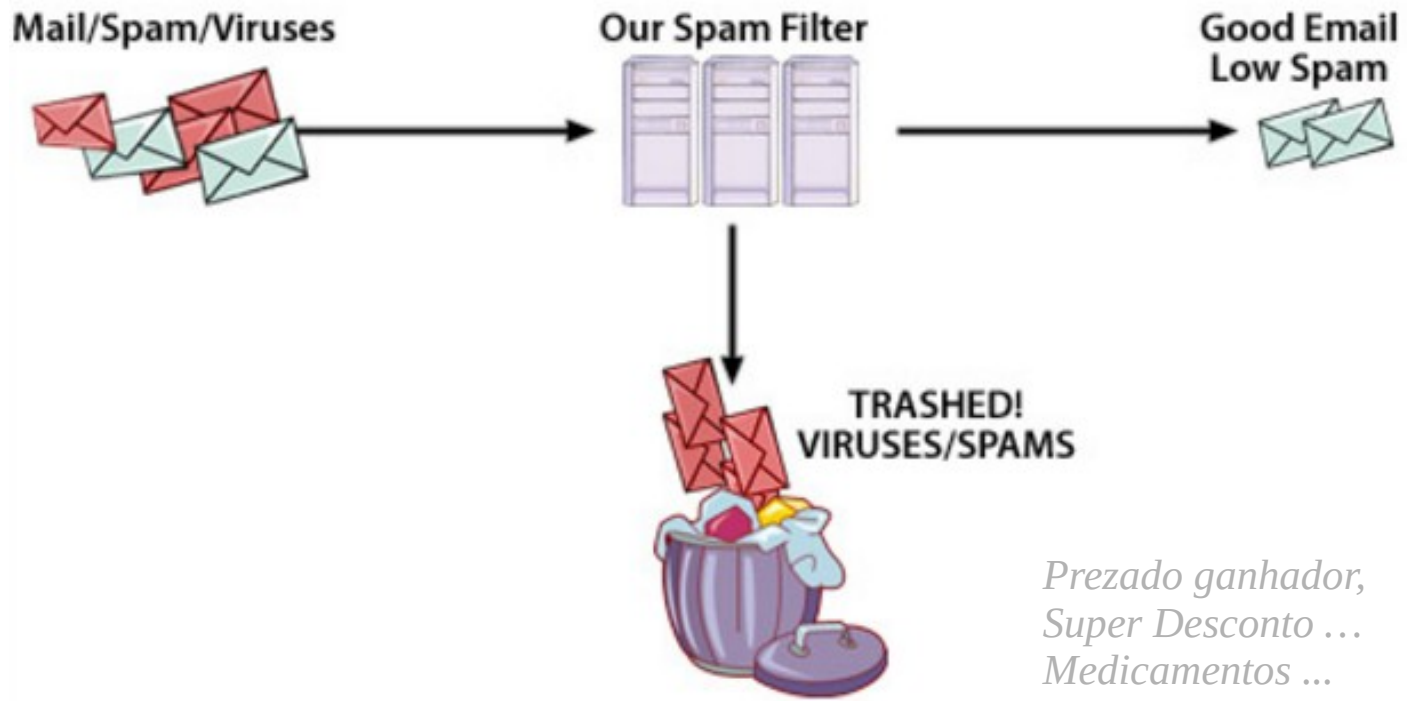
Categorização de textos:

Corresponde à tarefa de classificar textos inteiros atribuindo um rótulo (de um conjunto finito de rótulos).



Um sistema de categorização de textos:

- Permite classificar novos documentos
- Considera categorias pré-estabelecidas e documentos rotulados.



Um sistema de classificação de emails

Classificação

Formalmente, a tarefa de classificação pode ser definida como:

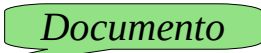

- Dada uma entrada \mathbf{x} , e
- Um conjunto finito de classes $Y = \{y_1, y_2, \dots, y_n\}$
- **Determinar**, para \mathbf{x} , uma classe \mathbf{y} que pertence a Y .

Classificação

Formalmente, a tarefa de classificação pode ser definida como:

- Dada uma entrada \mathbf{x} , e
- Um conjunto finito de classes $Y = \{y_1, y_2, \dots, y_n\}$
- **Determinar**, para \mathbf{x} , uma classe \mathbf{y} que pertence a Y .

Em Classificação de texto:

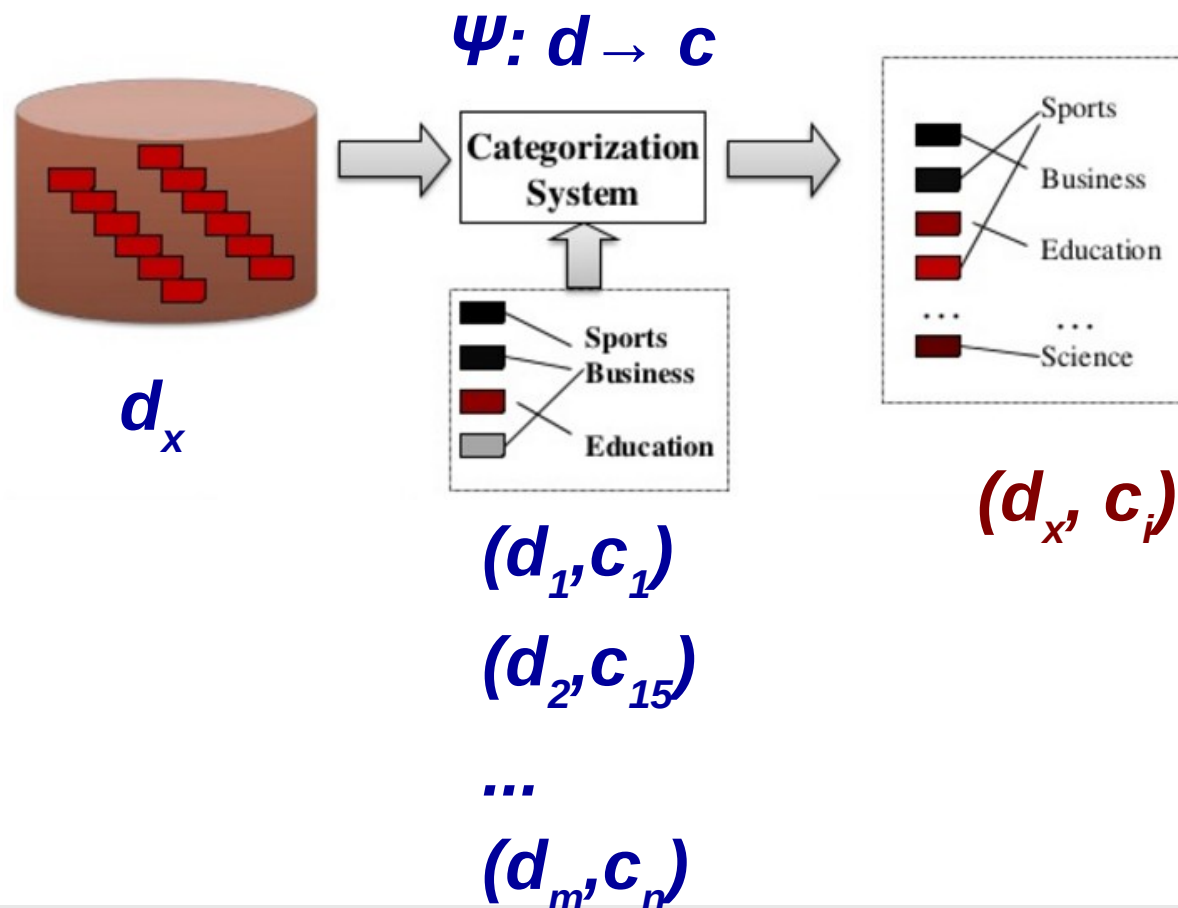
- Dada uma entrada \mathbf{d} , e 
- Um conjunto finito de classes $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ 
- **Determinar**, para \mathbf{d} , uma classe \mathbf{c} que pertence a \mathbf{C} .

Métodos de classificação: Usando regras

- Uso de regras baseadas em palavras ou combinação de palavras (ou outras características)
 - **Black-list-address OR (“dollars” AND “have been selected”)**
- A acurácia pode ser alta.
Se as regras são definidas por especialista(s)
- A manutenção/atualização das regras pode ser muito cara.

Métodos de classificação: Usando Aprendizado de Máquina

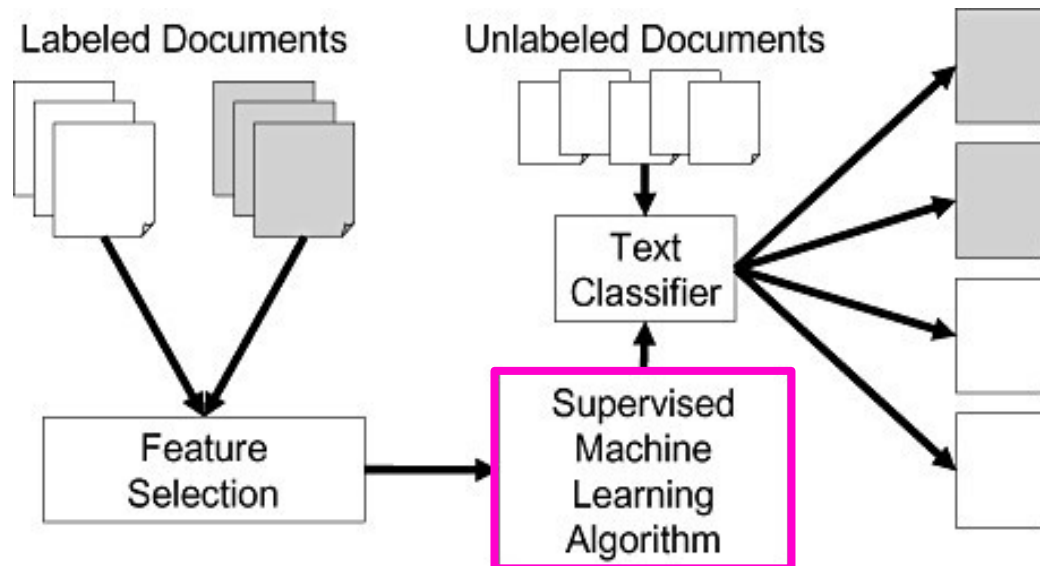
O **desafio**: Construir um classificador que seja capaz de **mapear** o novo documento d à sua classe correta c_i .



Métodos de classificação: Usando Aprendizado de Máquina

Existe uma quantidade grande classificadores:

- **Naive Bayes**
- Regressão logística
- Support-vector machines (SVM)
- K-Nearest Neighbors (KNN)





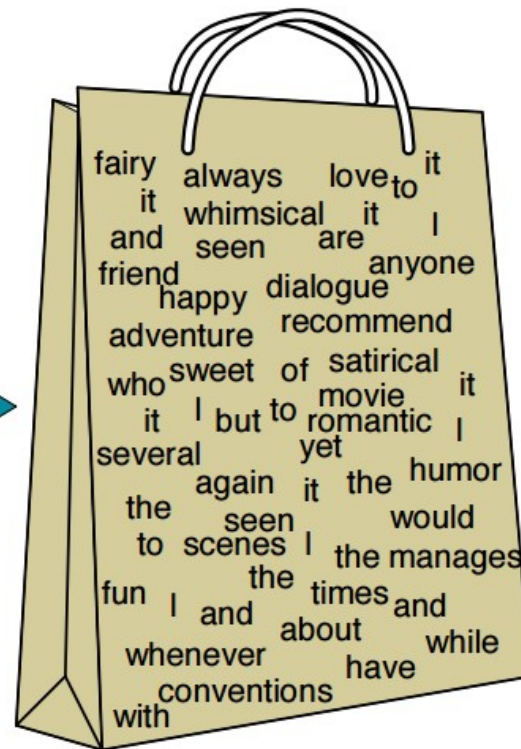
Classificação de texto usando *Naive Bayes*

Aprendizado supervisionado

Bag-of-words

Um documento pode ser representado como uma **bag-of-words**

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Bag-of-words: Conjunto não-ordenado de palavras (**desconsidera a gramática, mas mantendo a multiplicidade**).


Bag-of-words

Um documento pode ser representado como uma **bag-of-words**

Document 1	Term	Document 1	Document 2
The quick brown fox jumped over the lazy dog's back.	aid	0	1
	all	0	1
	back	1	0
	brown	1	0
	come	0	1
	dog	1	0
	fox	1	0
	good	0	1
	jump	1	0
	lazy	1	0
	men	0	1
	now	0	1
	over	1	0
	party	0	1
	quick	1	0
	their	0	1
	time	0	1

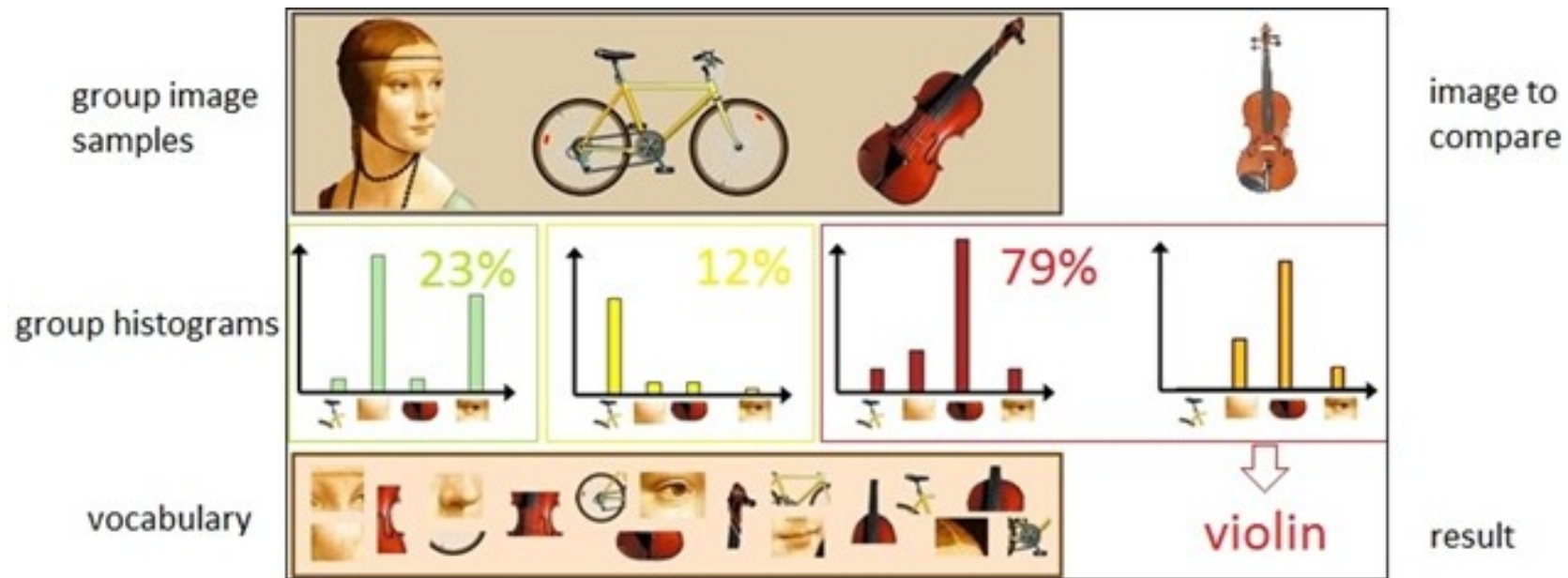
Stopword List

for
is
of
the
to



Bag-of-words: Conjunto não-ordenado de palavras (**desconsidera a gramática, mas mantendo a multiplicidade**).

Bag-of-visual-words (em imagens)



Bag-of-words em classificação de textos

Tópicos de ciência da computação

Test document

parser
language
label
translation
...

Machine Learning

learning
training
algorithm
shrinkage
network...

NLP

parser
tag
training
translation
language...

Garbage Collection

garbage
collection
memory
optimization
region...

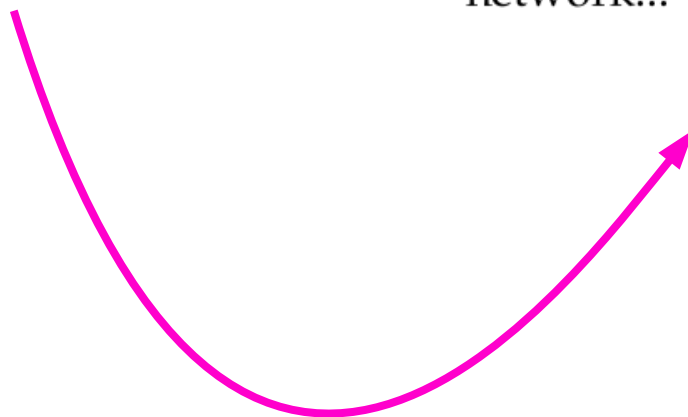
Planning

planning
temporal
reasoning
plan
language...

GUI

...

?



Classificador *Naive Bayes*

Conhecida também como **Classificador Bayesiano “simples”** ou “**ingênuo**”.

- É um classificador probabilístico.
- Usa a representação de textos como bag-of-words.
- É considerado **ingênuo** porque considera os atributos condicionalmente independentes (i.e., o valor de um atributo não está relacionado ao valor de um outro atributo).

A melhor classe

$$\hat{c} = \operatorname{argmax}_{c \in \mathcal{C}} P(c|d)$$

Documento

Probabilidade condicional dado um documento d

Classificador Naive Bayes

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|d)$$

MAP
Maximum a posteriori

$$= \operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)}$$

likelihood prior

$$= \operatorname{argmax}_{c \in C} \overbrace{P(d|c)} \quad \overbrace{P(c)}$$

Qual a probabilidade da classe c aparecer no corpus (treinamento)?

O documento d representado por um conjunto de características

$$= \operatorname{argmax}_{c \in C} P(f_1, f_2, \dots, f_n|c) P(c)$$

$$= \operatorname{argmax}_{c \in C} P(c) \prod_{f \in F} P(f|c)$$

Simplificação ingênua mas na prática permite resolver grandes problemas

Classificador *Naive Bayes*

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{f \in F} P(f|c)$$


Para aplicar o classificador para a sequência: $w_1, w_2, w_3, \dots, w_n$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{i \in \text{positions}} P(w_i|c)$$

Considerando bag-of-words
(a posição da palavra não importa)

$$= \operatorname{argmax}_{c \in C} \log P(c) + \sum_{i \in \text{positions}} \log P(w_i|c)$$

As probabilidades $P(w_i, c)$ são independentes



Classificador Bayesiano “ingênuo”: Aprendizado (treinamento)

$$c_{NB} = \operatorname{argmax}_{c \in \mathcal{C}} P(c) \prod_{i \in \text{positions}} P(w_i | c)$$

Treinamento

$$c_{NB} = \operatorname{argmax}_{c \in \mathcal{C}} P(c) \prod_{i \in \text{positions}} P(w_i | c)$$

Considerando um conjunto de dados de treinamento (**corpus rotulado**) composto de N_{doc} :

- Ex. $(doc_1, c_6), (doc_2, c_5), (doc_3, c_1), \dots (doc_N, c_2)$

Treinamento

$$c_{NB} = \operatorname{argmax}_{c \in \mathcal{C}} P(c) \prod_{i \in \text{positions}} P(w_i | c)$$

Considerando um conjunto de dados de treinamento (corpus rotulado) composto de N_{doc} :

- Ex. $(doc_1, c_6), (doc_2, c_5), (doc_3, c_1), \dots (doc_N, c_2)$

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

Número de documentos cuja classe é igual a c

Número de documentos totais (no treinamento)

Treinamento

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

Número de documentos
cuja classe é igual a c

Número de documentos
totais (no treinamento)

Corpus:



$$P(\text{classe_verde}) = 3/10$$

$$P(\text{classe_vermelha}) = 7/10$$

Treinamento

$$\hat{P}(w_i|c) = \frac{\text{count}(w_i, c)}{\sum_{w \in V} \text{count}(w, c)}$$

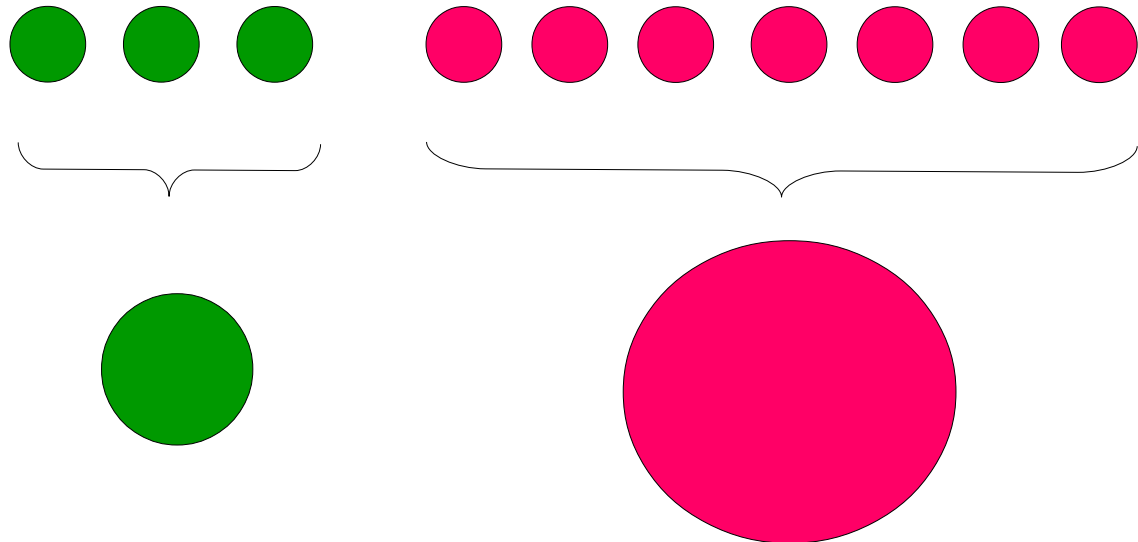
V é o vocabulário de todo o corpus (ie., de todas palavras de todos os documentos)

Treinamento

$$\hat{P}(w_i|c) = \frac{\text{count}(w_i, c)}{\sum_{w \in V} \text{count}(w, c)}$$

V é o vocabulário de todo o corpus (ie., de todas palavras de todos os documentos)

Corpus:



Treinamento

$$c_{NB} = \operatorname{argmax}_{c \in \mathcal{C}} P(c) \prod_{i \in \text{positions}} P(w_i | c)$$

Qual seria o valor de $P(w_x | c)$ quando w_x é palavra desconhecida no treinamento?

Treinamento

$$c_{NB} = \operatorname{argmax}_{c \in \mathcal{C}} P(c) \prod_{i \in \text{positions}} P(w_i | c)$$

Qual seria o valor de $P(w_x | c)$ quando w_x é palavra desconhecida no treinamento?

Zero! (não importando os outros termos)

Treinamento

$$c_{NB} = \operatorname{argmax}_{c \in \mathcal{C}} P(c) \prod_{i \in \text{positions}} P(w_i | c)$$

Qual seria o valor de $P(w_x | c)$ quando w_x é palavra desconhecida no treinamento?

Zero! (não importando os outros termos)

Laplace add-1 smoothing

Alternativa:

$$\begin{aligned} \hat{P}(w_i | c) &= \frac{\text{count}(w_i, c) + 1}{\sum_{w \in V} (\text{count}(w, c) + 1)} \\ &= \frac{\text{count}(w_i, c) + 1}{(\sum_{w \in V} \text{count}(w, c)) + |V|} \end{aligned}$$

function TRAIN NAIVE BAYES(D, C)

for each class $c \in C$

N_{doc} = number of documents in D

N_c = number of documents from D in class c

$logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$

$V \leftarrow$ vocabulary of D

$bigdoc[c] \leftarrow$ **append**(d) **for** $d \in D$ **with** class c

for each word w in V

$count(w, c) \leftarrow$ # of occurrences of w in $bigdoc[c]$

$loglikelihood[w, c] \leftarrow \log \frac{count(w, c) + 1}{\sum_{w' \text{ in } V} (count(w', c) + 1)}$

return $logprior, loglikelihood, V$

function TEST NAIVE BAYES(*testdoc*, *logprior*, *loglikelihood*, *C*, *V*)

for each class $c \in C$

$sum[c] \leftarrow logprior[c]$

for each position i in *testdoc*

$word \leftarrow testdoc[i]$

if $word \in V$

$sum[c] \leftarrow sum[c] + loglikelihood[word, c]$

return $argmax_c sum[c]$

Se uma palavra é desconhecida no treinamento, então será desconsiderada (solução padrão)



Atividade

Atividade 1



Considere o seguinte corpus (conjunto de treinamento) contendo duas classes (c_1 ='pos' e c_2 ='neg')

Classe	Texto
neg	just plain boring
neg	entirely predictable and lacks energy
neg	no surprises and very few laughs
pos	very powerful
pos	the most fun film of the summer

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

$$\hat{P}(w_i|c) = \frac{\text{count}(w_i, c) + 1}{\left(\sum_{w \in V} \text{count}(w, c)\right) + |V|}$$

Atividade 1

$P(c_1)$	2/5
$P(c_2)$	3/5
Vocabulário	and (2vezes) boring energy entirely few film fun just lacks laughs most no of plain powerful predictable summer surprises the (2 vezes) very (2 vezes)
Tamanho do Vocabulário	20

Atividade 1

S = "predictable with no fun"

$$P(\text{"predictable"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"no"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"fun"}|+) = \frac{1+1}{9+20}$$

$$P(\text{"predictable"}|-) = \frac{1+1}{14+20}$$

$$P(\text{"no"}|-) = \frac{1+1}{14+20}$$

$$P(\text{"fun"}|-) = \frac{0+1}{14+20}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$

S deve ser classificada como 'neg'

Prática 1: naiveBayes1.py

```
python3 naiveBayes1.py train1.txt
```

```
Total: classes=2 documentos=5 vocabulario=20
```

```
{'few', 'very', 'fun', 'no', 'energy', 'plain',  
'entirely', 'the', 'most', 'of', 'surprises',  
'boring', 'predictable', 'just', 'lacks',  
'powerful', 'film', 'summer', 'laughs', 'and'}
```

```
{'neg': ['just', 'plain', 'boring', 'entirely',  
'predictable', 'and', 'lacks', 'energy', 'no',  
'surprises', 'and', 'very', 'few', 'laughs'],  
'pos': ['very', 'powerful', 'the', 'most',  
'fun', 'film', 'of', 'the', 'summer']}
```

```
{'neg': 0.6, 'pos': 0.4}
```

```
('just', 'neg'): 0.058823529411764705,  
( 'entirely', 'neg'): 0.058823529411764705,  
( 'boring', 'neg'): 0.058823529411764705,  
( 'surprises', 'pos'): 0.034482758620689655,  
( 'film', 'pos'): 0.06896551724137931,  
( 'very', 'pos'): 0.06896551724137931,  
( 'energy', 'pos'): 0.034482758620689655,  
( 'no', 'pos'): 0.034482758620689655,  
( 'plain', 'neg'): 0.058823529411764705,  
( 'predictable', 'neg'): 0.058823529411764705,  
( 'fun', 'pos'): 0.06896551724137931,  
( 'few', 'pos'): 0.034482758620689655,  
( 'very', 'neg'): 0.058823529411764705,  
( 'the', 'neg'): 0.029411764705882353,  
( 'lacks', 'pos'): 0.034482758620689655,  
( 'and', 'pos'): 0.034482758620689655,  
( 'predictable', 'pos'): 0.034482758620689655,  
...
```

Testando: predictable with no fun

Teste 1: neg

Atividade 2

“I always like foreign films”

	Classe 'Pos'	Classe 'Neg'
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

Considere a mesma probabilidades à priori para cada classe.

$$P(\text{'Pos'}) = 0.00000585$$

$$P(\text{'Neg'}) = 0.00000950$$





Considerações finais

1) Aplicações

- Atribuição de categorias, tópicos ou generos.
- Detecção de SPAM.
- Identificação de autoria de escrita.
- Identificação de idade do autor do texto.
- Identificação de idioma.
- Análise de sentimentos em texto.

Opiniões em português?

pos eu amo esse sanduíche
pos este é um lugar incrível!
pos eu me sinto bem com essas cervejas
pos este é o meu melhor trabalho
pos que visão incrível

neg eu não gosto deste restaurante
neg estou cansado dessas coisas
neg não consigo lidar com isso
neg ele é meu inimigo jurado!
neg meu chefe é horrível

Opiniões em português?

TESTE

Neg "eu não gosto do meu trabalho"

Neg "eu não estou me sentindo bem hoje"

Pos "eu me sinto incrível"

Pos "Roberto é um amigo meu"

Neg "eu não posso acreditar que estou fazendo isso"

Identificar idioma?

pt	a capivara (nome científico: hydrochoerus hydrochaeris)
pt	é o maior roedor do mundo, pesando até 91 kg e medindo até 1,30 m de comprimento.
pt	a capivara também é chamada de carpincho, capim-cavalo ou capim-de-velho.
pt	a capivara foi descrita pela primeira vez por george georges schrebler em 1793.
pt	a característica mais chamativa na capivara é seu corpo cilíndrico e achatado.
pt	a cabeça é grande, com orelhas pequenas e sem pontas.
sp	el carpincho, capibara o chigüiro ² (hydrochoerus hydrochaeris)
sp	tiene un cuerpo pesado en forma de barril y una cabeza grande.
sp	la fórmula dental de este animal es de 1-0-1-3; lo que indica que es un roedor.
sp	la medida de los grupos y su estilo de vida depende de la disponibilidad de alimento.
eng	the capybara (hydrochoerus hydrochaeris) is a mammal native to south america.
eng	the capybara and the lesser capybara belong to the family chinchillidae.
eng	capybaras are herbivores, grazing mainly on grasses and aquatic plants.
eng	when in estrus, the female's scent changes subtly and she may become more aggressive.
eng	though quite agile on land (capable of running as fast as 30 km/h), they are excellent swimmers.

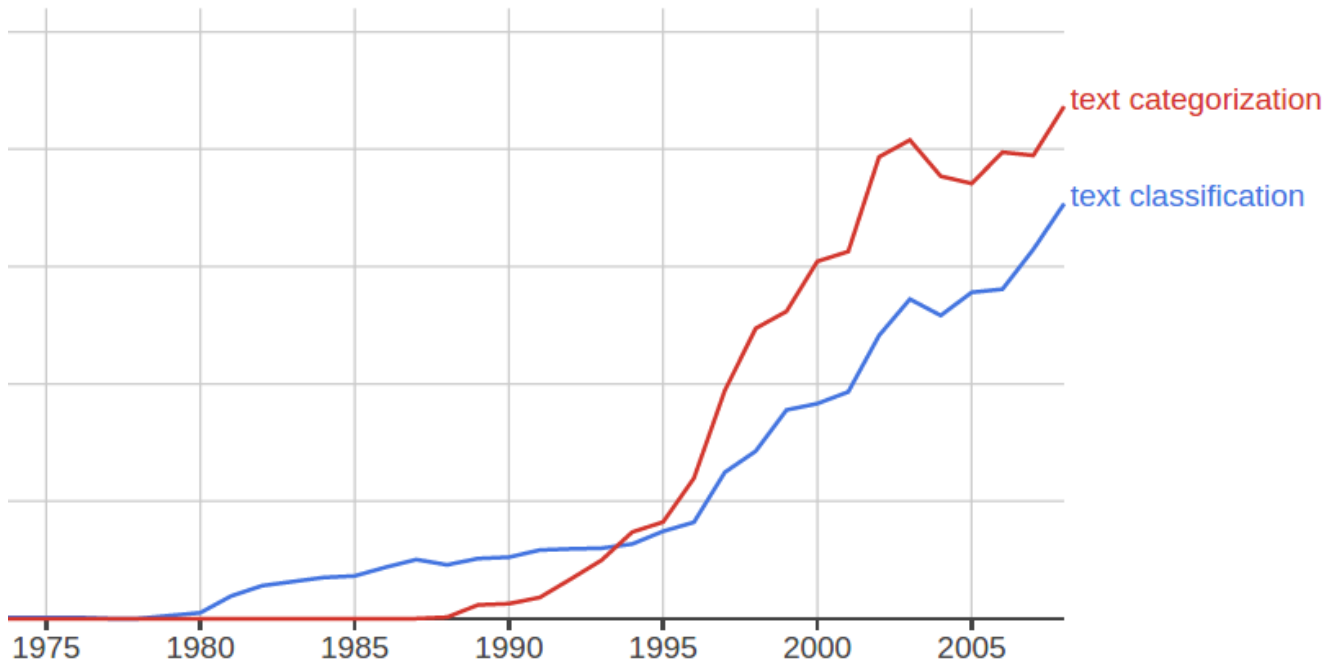
Classificar notícias?

<u>Noticia-Fapesp</u>	<u>atividades humanas já danificaram 75% d</u>
<u>Noticia-Fapesp</u>	<u>fapesp e finep apoiarão pesquisas em qu</u>
<u>Noticia-Fapesp</u>	<u>instituto oceanográfico da usp tem duas o</u>
<u>Noticia-Folha</u>	<u>esquerda critica netflix por causa de série</u>
<u>Noticia-Folha</u>	<u>uma defesa do facebook empresa fracass</u>
<u>Noticia-Sensacionalista</u>	<u>fifa pode suspender jogador que atribuir g</u>
<u>Noticia-Sensacionalista</u>	<u>brasil enfrenta epidemia de arrepios na es</u>

Identificar disciplina?

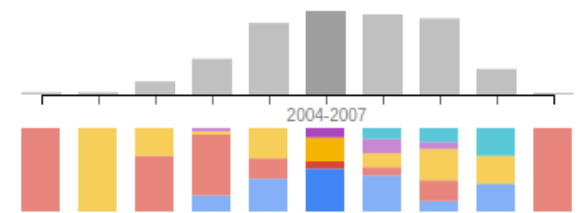
<u>abordagens tradicionais das relações internacionais</u>	<u>contextualização histórica da emergência</u>
<u>acionamentos elétricos</u>	<u>introdução aos sistemas de acionamento</u>
<u>acumuladores de energia</u>	<u>acumulação de energia por fotossíntese</u>
<u>administração pública e reforma do estado em perspectiva comparada</u>	<u>estado, política e administração pública</u>
<u>aeroacústica</u>	<u>fundamentos de acústica e propagação</u>
<u>aerodinâmica i</u>	<u>sustentação; teoria do perfil delgado; coeficiente</u>
<u>aerodinâmica ii</u>	<u>método da linha de sustentação. métodos</u>
<u>aeroelasticidade</u>	<u>comportamento aeroelástico de veículos</u>
<u>aeronáutica i-a</u>	<u>conhecimentos técnicos sobre aviões: projeto</u>
<u>aeronáutica i-b</u>	<u>conhecimentos técnicos sobre helicópteros</u>
<u>aeronáutica ii</u>	<u>regulamentação aeronáutica: regras de voo</u>
<u>álgebra linear</u>	<u>sistemas de equações lineares: sistemas</u>
<u>álgebra linear avançada i</u>	<u>corpos; espaço vetorial sobre um corpo</u>
<u>álgebra linear avançada ii</u>	<u>formas bilineares e sesquilineares: formas</u>
<u>algoritmos e estruturas de dados i</u>	<u>breve introdução à linguagem c. noções</u>
<u>algoritmos e estruturas de dados ii</u>	<u>hashing. introdução a arquivos. arquivos</u>

2) Text categorization



Google Ngram

Top 1000 results by filing date



Relative count of top 5 values

Assignees	Inventors	CPCs
<ul style="list-style-type: none"> Microsoft Corporation (7.6%) H04L G06N3/00 G06N3/02 G06N 		
<ul style="list-style-type: none"> International Business Machines Corporation (6.5%) Y10S707/99931 Y10S707/00 Y10S Y10S707/99933 		
<ul style="list-style-type: none"> Xerox Corporation (6%) G06F17/30722 G06K9/4676 G06F17/30011 G06K9/4671 		
<ul style="list-style-type: none"> Yahoo! Inc. (1.7%) G06K9/6269 G06K9/6268 G06K9/6267 G06N99/005 		
<ul style="list-style-type: none"> Google Inc. (1.3%) G06Q30/0274 G06Q30/0273 G06K9/00711 G06K9/00624 		

Patentes

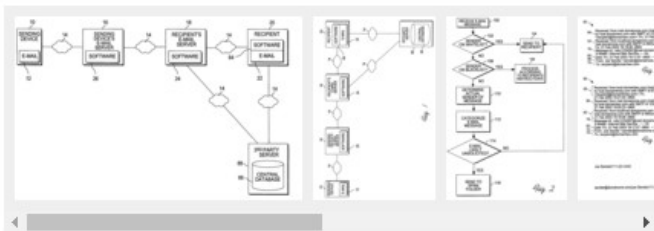
2) Categorization usando apenas texto?

Method and system for categorizing and processing e-mails

Abstract

An e-mail filtering method and system that categorize received e-mail messages based on information about the sender. Data about the sender is contained in the message and is used to identify the actual sender of the message using a signature combining pieces of information from the message header or derived from information in the message header. This and other information about the message is then sent by each member of an e-mail network to one or more central databases (in one embodiment, the information will also be stored at a database associated with the recipient's e-mail program and filtering software) which stores the information and compiles statistics about e-mails sent by the sender to indicate the likelihood that the e-mail is unsolicited and determine the reputation of the sender (a good reputation indicates the sender does not send unwanted messages while a bad reputation indicates the sender sends unsolicited e-mail messages). Information from the central database is then sent to recipients in order to determine the likelihood that a received e-mail message is spam (information may also be obtained from the local database associated with the recipient's e-mail program and filtering software).

Images (9)



US7206814B2
US Grant

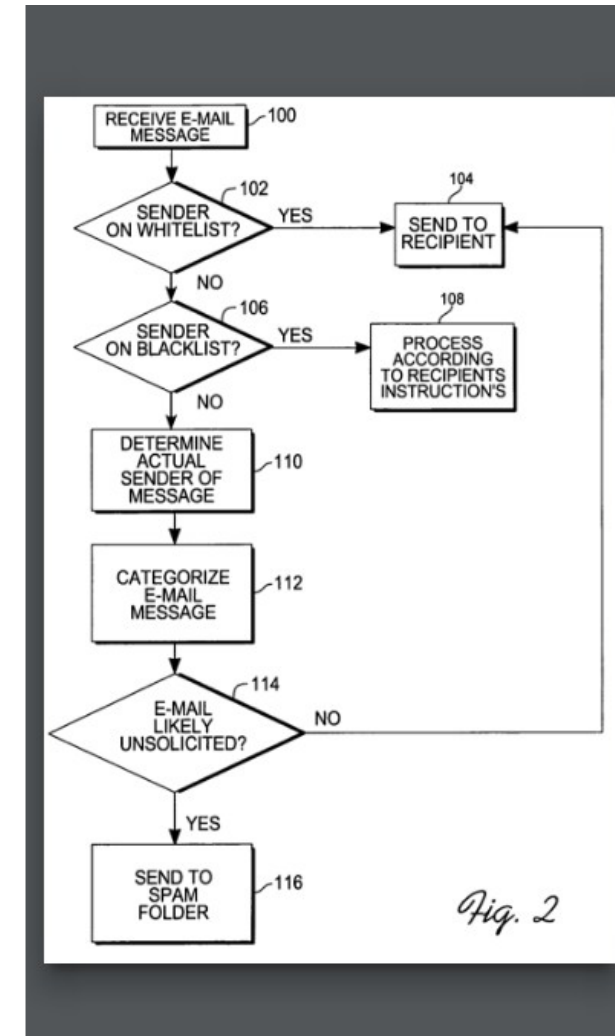
Download PDF Find Prior Art Similar

Inventor: Steven T. Kirsch
Current Assignee: PROOFPOINT Inc
Original Assignee: Propel Software Corp
Priority date: 2003-10-09

Family: US (1)

Date	App/Pub Number	Status
2003-10-09	US10683598	Active
2005-04-14	US20050080856A1	Application
2007-04-17	US7206814B2	Grant

Info: Patent citations (31), Non-patent citations (3), Cited by (192), Legal events, Similar documents, Priority and Related Applications





Sobre o projeto

PLN-UFABC: Entrega Número 1

Este formulário será utilizado para registrar: (i) os membros do grupo de trabalho, e (ii) o artigo científico a ser replicado para a disciplina de PLN. Note que esta entrega corresponde a 10% da nota do projeto.

Observações:

- Apenas um membro do grupo deve entregar o formulário.
 - A implementação do artigo não precisa ser de forma idêntica ao apresentado pelos seus autores. O importante é compreender a lógica da proposta e fazer uma rerepresentação sobre os conceitos de PLN. A rerepresentação pode ser limitada, isto é, não precisa implementar tudo. Pode implementar parte desde que se justifique o motivo da limitação.
 - Sobre o conjunto de dados utilizado no artigo: Para sua implementação não precisa ser o mesmo conjunto. Pode usar um outro similar, um outro coletado pelo seu grupo de trabalho, ou um conjunto extraído de outra fonte ou de outro artigo.
 - Sobre a avaliação ou estudo comparativo: Não precisa implementar todos os algoritmos considerados para comparar a proposta com outras abordagens. O importante é implementar principalmente a proposta do artigo.
 - Deadline para preenchimento do formulário: 04/julho (23h55)
-

Email address *

Valid email address

This form is collecting email addresses. [Change settings](#)

*****SOBRE O GRUPO DE ALUNOS*****

O grupo deve ser composto por 4 alunos. Caso o grupo for composto por 2 ou 3 alunos deve ser apresentado uma justificativa concreta

SOBRE O ARTIGO SELECIONADO

Description (optional)

Título do artigo *

Short answer text

Nome da revista ou evento onde foi publicado o artigo *

Short answer text

Ano de publicação *

Short answer text

Número de páginas *

Short answer text

Número de citações *

Short answer text

Fonte consultada sobre o número de citações *

Short answer text

URL da fonte consultada sobre o número de citações *

Short answer text

PDF do artigo científico *

ADD FILE

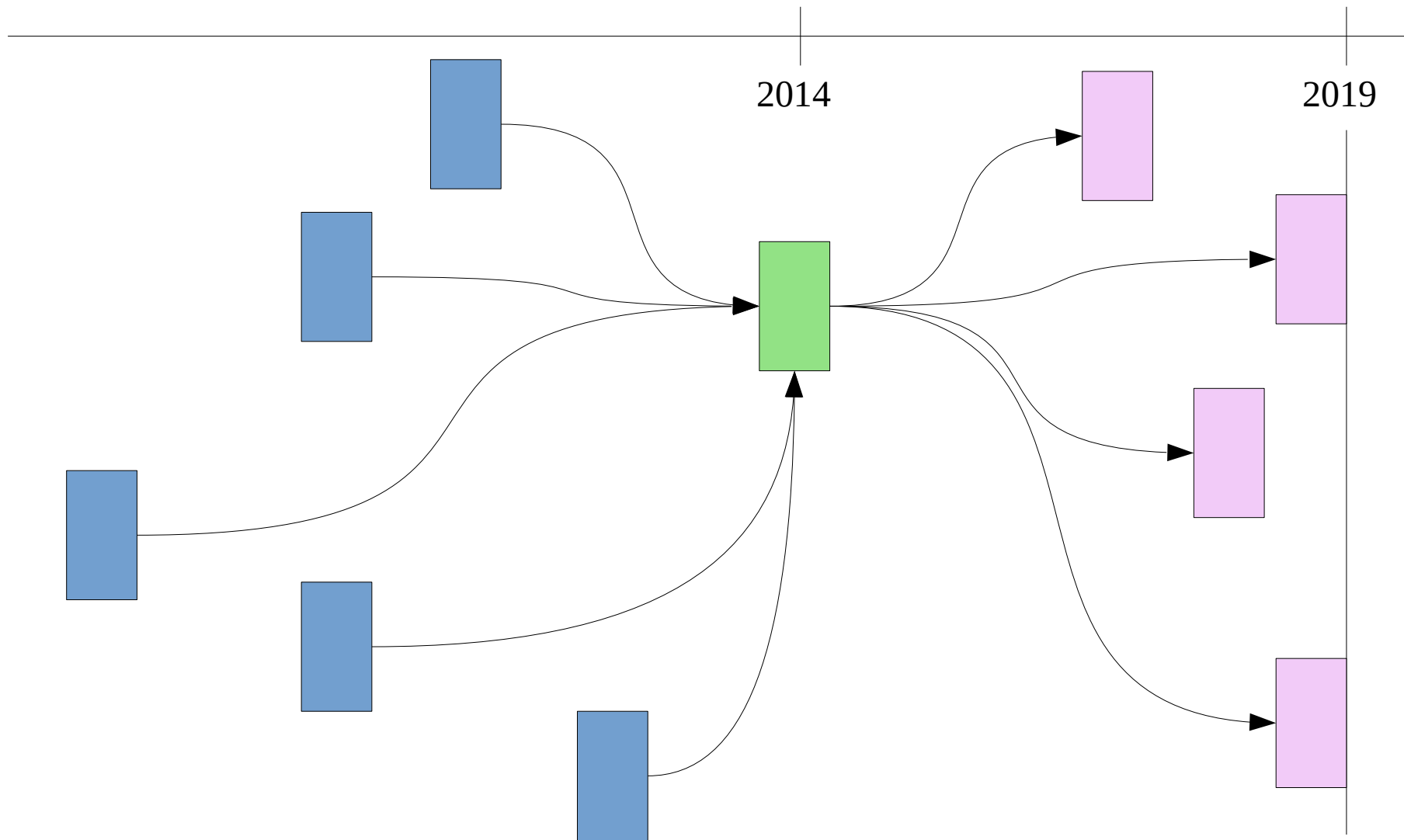
Qual é o objetivo do artigo? (descreva brevemente) *

Long answer text

Qual é a técnica de PLN que é explorada no artigo? *

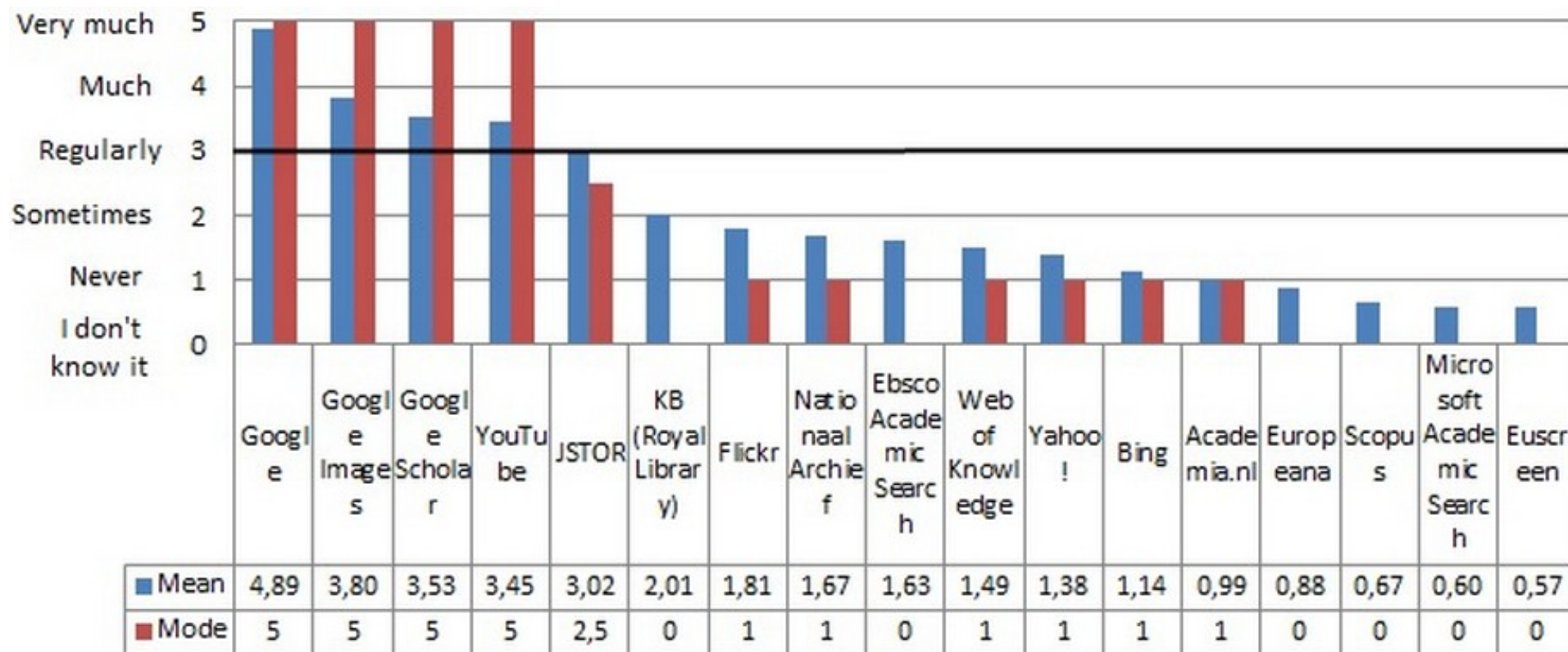
Long answer text

Sobre revisão bibliográfica



Just Google it?

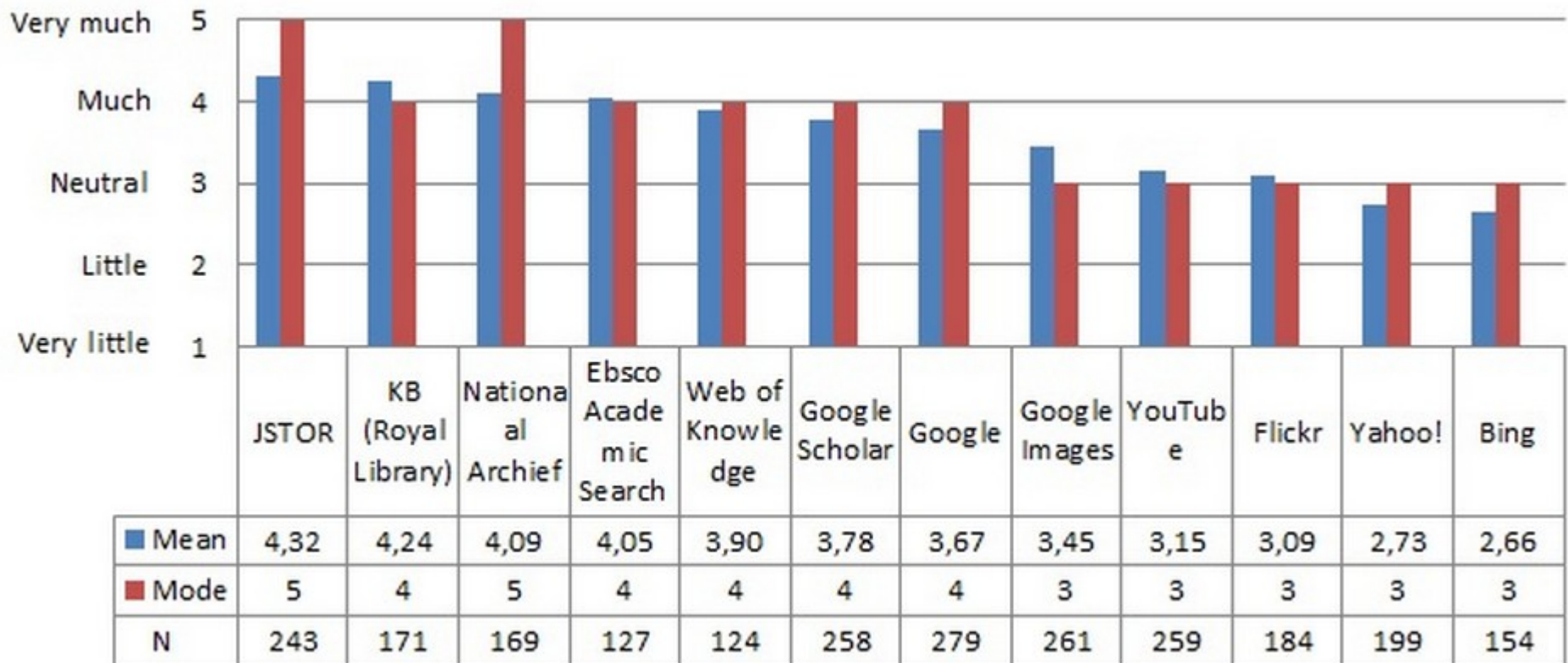
“Which of the following search engines, websites or databases do you use?”



Fonte: Kemman, M., Kleppe, M. and Scagliola, S., 2014. **Just Google It.** In Proceedings of the Digital Humanities Congress 2012. HRI Online Publications.

Just Google it?

“How much do you trust the following search engines”



Fonte: Kemman, M., Kleppe, M. and Scagliola, S., 2014. **Just Google It.** In Proceedings of the Digital Humanities Congress 2012. HRI Online Publications.

Sobre os ombros de gigantes?

- Google, pode não cobrir todas as fontes relevantes (cobre a maioria).
- **Apenas evidências:** Com o google (a partir de 2004)
 - O impacto de revistas que **não** são da “**elite**” aumentou.
 - O impacto de artigos **antigos** aumentou.
- Artigos são indexados pelo seu título (e search snippets) **dando menor ênfase para a revista onde foram publicados**
- Artigos com maior número de citações apresentam maior ranking. → “Efeito Mateus” / “Rico fica mais rico”.

Principais plataformas e bases de dados

Para computação:

- DBLP (computer science bibliography): <http://dblp.uni-trier.de>
 - ACM Digital library: <http://dl.acm.org/dl.cfm>
 - IEEE Computer society digital library: <https://www.computer.org/csdl>
 - Arxiv: Computing research repository: <https://arxiv.org/corr>
-
- Portal de Periódicos CAPES: <http://www.periodicos.capes.gov.br>
 - Scielo: <http://www.scielo.br>
-
- Semantic scholar: <https://www.semanticscholar.org>
 - CiteSeerX: <http://csxstatic.ist.psu.edu>

Semantic scholar

Cut through the clutter

Find peer-reviewed research from the world's most trusted sources

All Fields

Try: [Yoshua Bengio](#) [Cyber Security](#) [Diabetes Insipidus](#)

Semantic Scholar is a free, nonprofit, academic search engine from [AI2](#).

Supplement your research with Semantic Scholar

Get Up to Speed with Videos Describing the Paper

Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks

[Tim Salimans](#), [Diederik P. Kingma](#)

We present weight normalization: a reparameterization of the weight vectors in a neural network that decouples the length of those weight vectors from their direction.

By... [CONTINUE READING](#)



VIDEO

Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Ne

Tad Zona
30 March 2018



26:06

Baidu scholar

百度学术 高级搜索 登录 注册

Natural Language Processing (Almost) from Scratch

来自 OALib | 喜欢 1 阅读量: 1320

作者: Collobert, Ronan, Weston, Jason, Bottou, Leon...

摘要: We propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. This versatility is achieved by trying to avoid task-specific engineering and therefore disregarding a lot of prior knowledge. Instead of exploiting man-made input features carefully optimized for each task, our system learns internal representations on the basis of vast amounts of mostly unlabeled training data. This work is then used as a basis for...

关键词: natural language processing neural networks

DOI: 10.1016/j.chemolab.2011.03.009

被引量: 2082

收藏 引用 批量引用 报错 分享

全部来源 免费下载 求助全文

- OALib
- EBSCO
- arXiv.org
- ResearchGate
- ACM
- 查看更多

相似文献 参考文献 引证文献

Natural language processing (almost) from scratch.

We propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-...

R Collobert, J Weston, L Bottou, ... - 《Journal of Machine Learning Research》

被引量: 2798 发表: 2011年

The Homemade Alternative: Teaching Human Neurophysiology with Instrumentation Made (Almost) from Scratch

来源期刊

Journal of Machine Learning Research
August 2011

引用走势

累加量 2018年被引量

1966 192

研究点分析

- Language Processing
- neural networks
- natural language processing

学术很专业, 搜索很简单, 给你一个无广告搜索APP

更有实时翻译、页内查找、长截屏等多种实用工具

< 简单搜索 > 立刻扫码体验