Análise de Sentimento em Textos Curtos Baseada em Processamento de Linguagem Natural e Aprendizado de Máquina

Breno Arosa

Natanael N. de Moura Junior
Luiz Pereira Calôba
Felipe Fink Grael

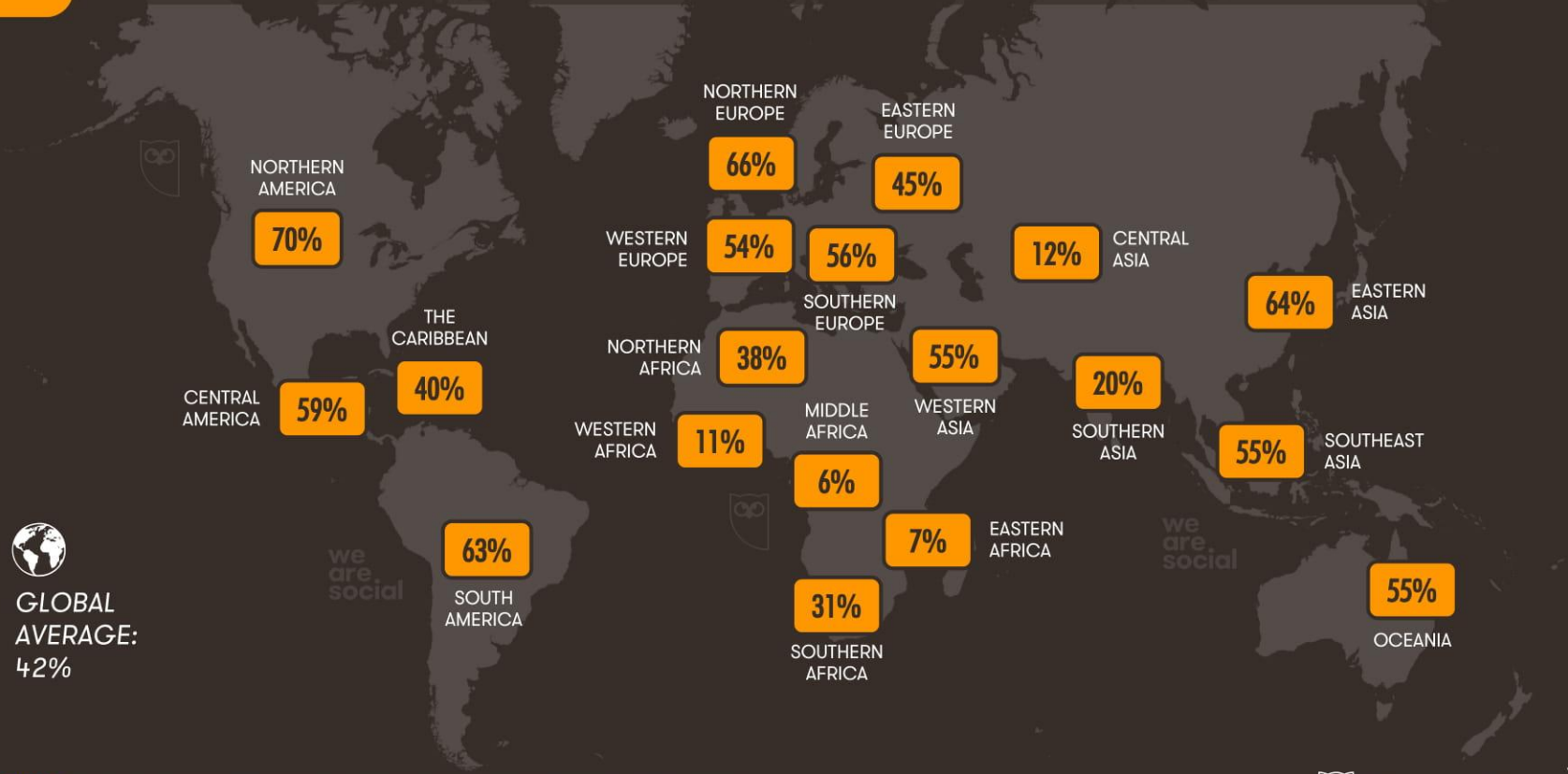
Agenda

- Motivação
- Processamento de Linguagem Natural
- Classificadores
- Método
- Resultados
- Conclusão

JAN 2018

SOCIAL MEDIA PENETRATION BY REGION

TOTAL ACTIVE ACCOUNTS ON THE MOST ACTIVE SOCIAL NETWORK IN EACH COUNTRY, COMPARED TO POPULATION

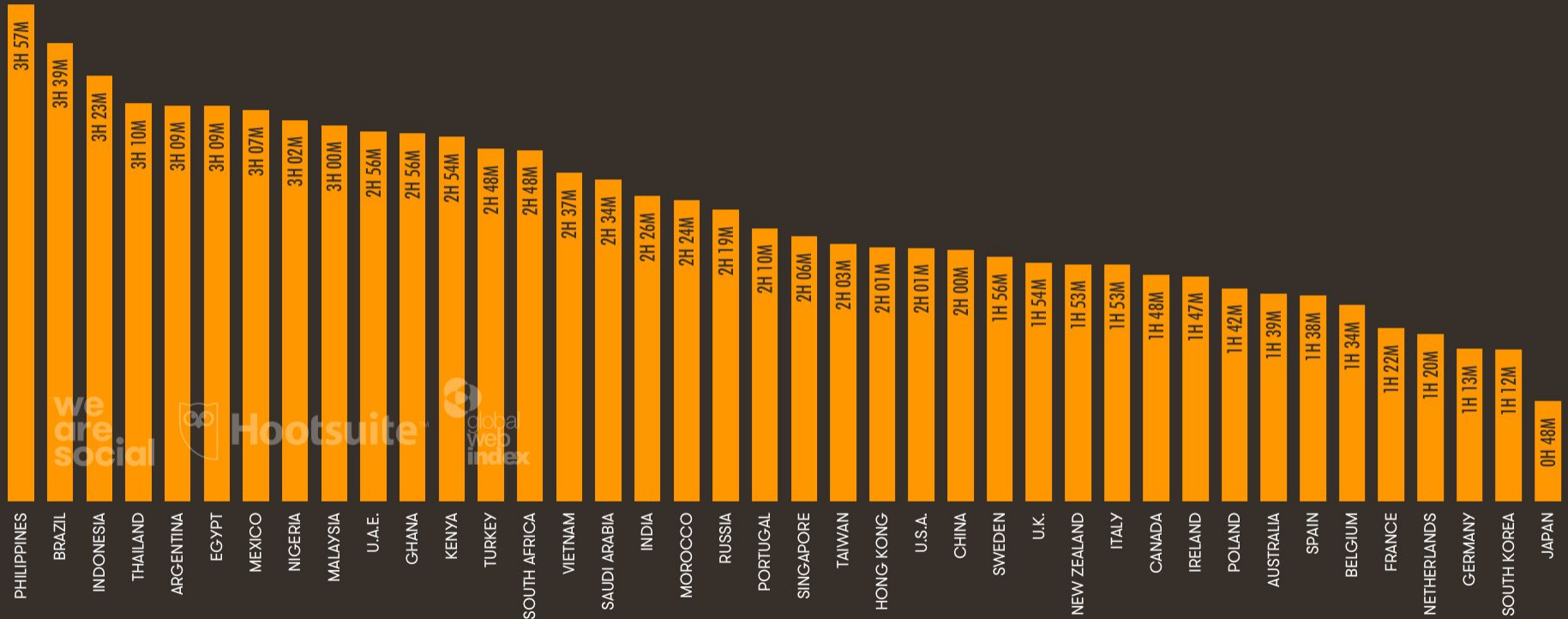


SOURCES: FACEBOOK; TENCENT; VKONTAKTE; KAKAO; NAVER; DING; TECHRASA; SIMILARWEB; KEPIOS ANALYSIS.
NOTE: PENETRATION FIGURES ARE FOR TOTAL POPULATION, REGARDLESS OF AGE.

JAN 2018

TIME SPENT ON SOCIAL MEDIA

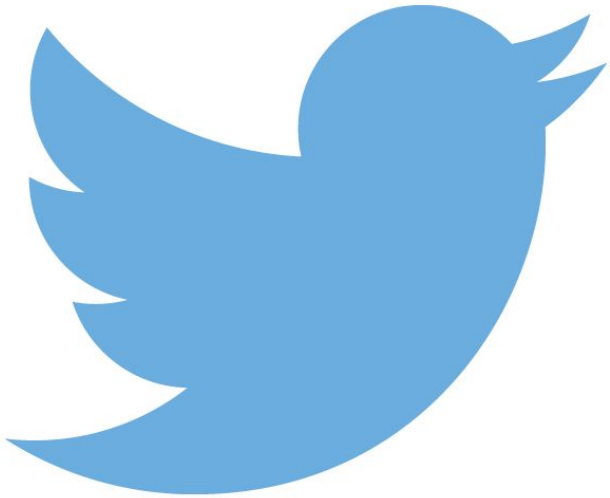
AVERAGE NUMBER OF HOURS THAT SOCIAL MEDIA USERS SPEND USING SOCIAL MEDIA EACH DAY VIA ANY DEVICE [SURVEY BASED]



SOURCE: GLOBALWEBINDEX, Q2 & Q3 2017. BASED ON A SURVEY OF INTERNET USERS AGED 16-64.

310M
Milhões de usuário ativos

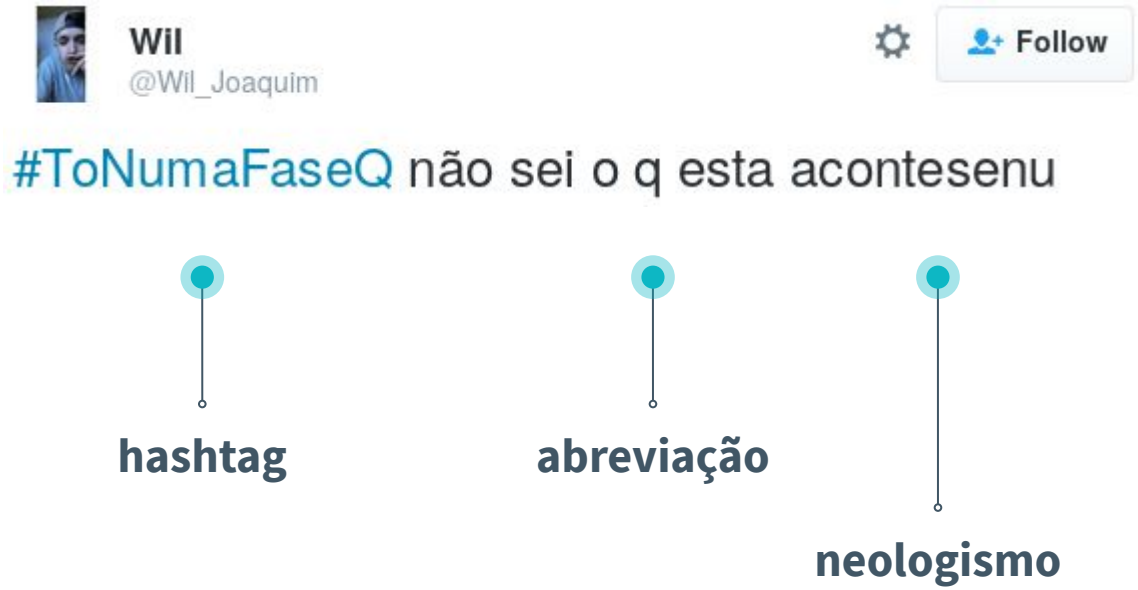
500M
Tweets por dia



Twitter


Donald J. Trump
@realDonaldTrump
Follow
Just heard Foreign Minister of North Korea speak at U.N. If he echoes thoughts of Little Rocket Man, they won't be around much longer!
Louis Tomlinson
@Louis_Tomlinson
Follow
Thoughts go out to everyone in Paris .
#prayforparis

Tweets


Wil
@Wil_Joaquim
Follow
#ToNumaFaseQ não sei o q esta acontesenu
hashtag
abreviação
neologismo

Tweets

Mídias

Data

Re-publicações e Curtidas

Análise de sentimento

Análise de sentimento pode ser considerada como a extração de ***polaridade*** de mensagem. Esta consiste em classificar a mensagem entre positiva, neutra e negativa.

Polaridade de mensagem

Multiplicidade de idiomas

Sentimento definido pela forma

Ambiguidade

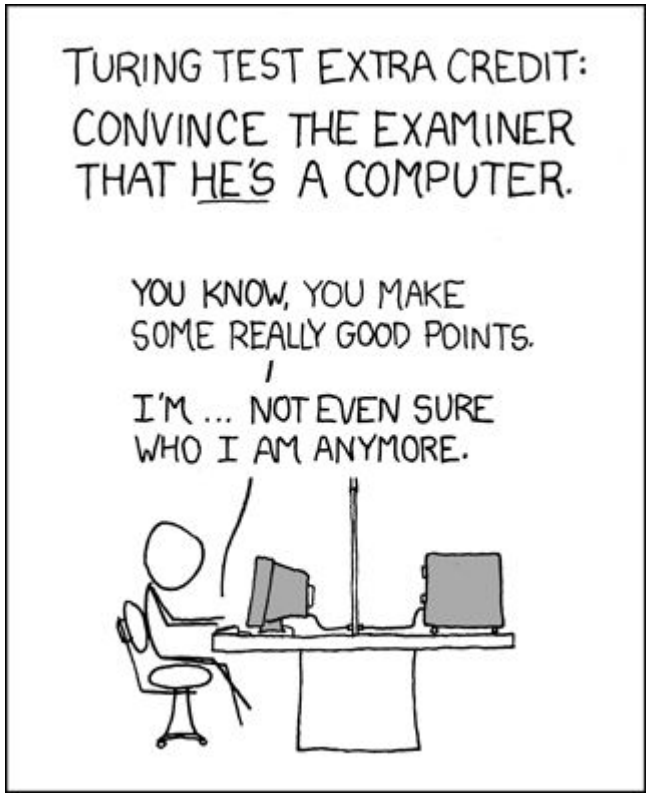
Ironia

Objetivos

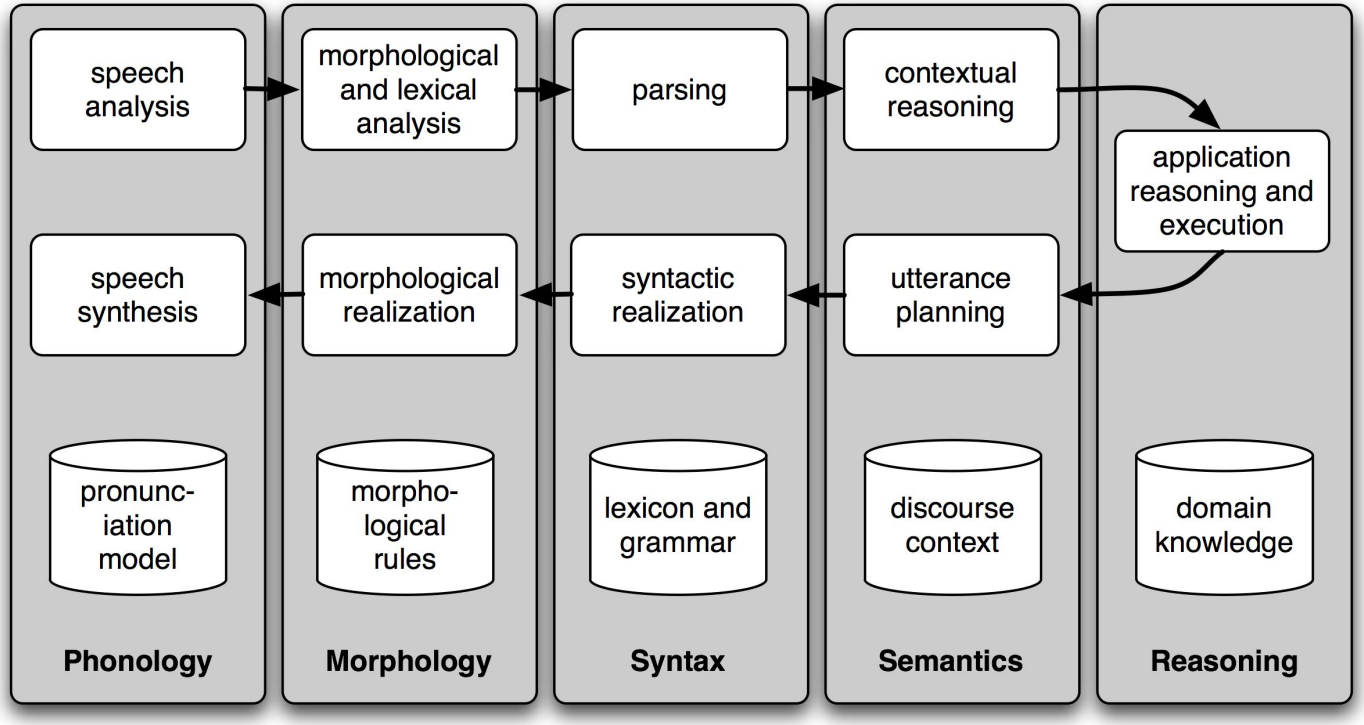
- Método de produção de classificadores de análise de sentimento.
- Avaliar técnicas de Deep Learning.

Processamento de linguagem natural

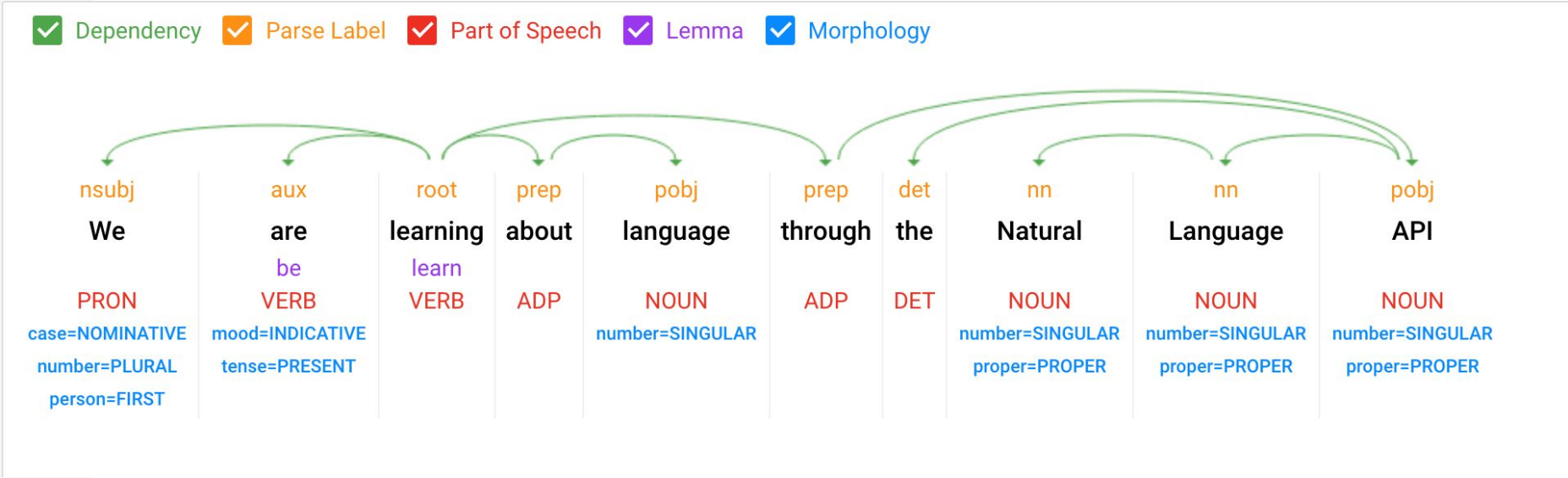
NLP é o ramo da inteligência artificial que trata da **análise**, **compreensão** e **geração** de linguagem que os seres humanos usam para comunicação em contextos escritos e/ou falados.



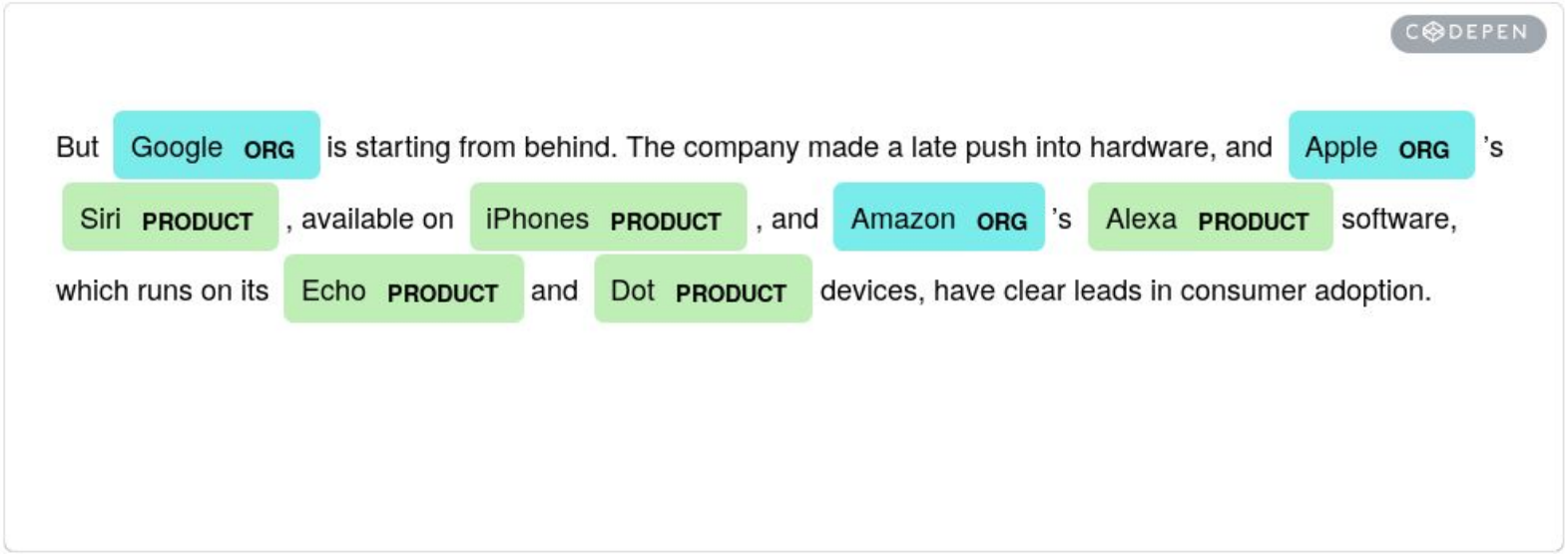
Níveis de Processamento



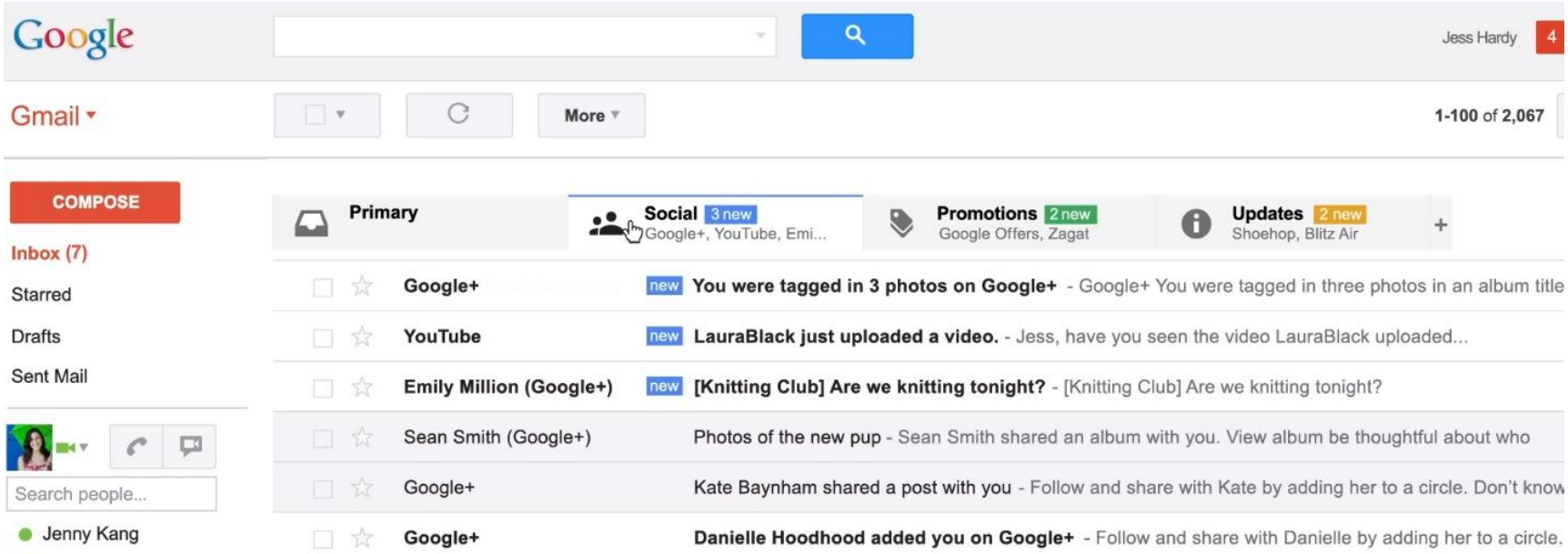
Análise Sintática



Reconhecimento de Entidade



Categorização



Resumo

Julian Assange

UK prosecutors admit destroying key emails in Julian Assange case

Correspondence between CPS and its Swedish counterparts about WikiLeaks founder deleted after lawyer retired in 2014

[–] autotldr [score hidden] 11 hours ago

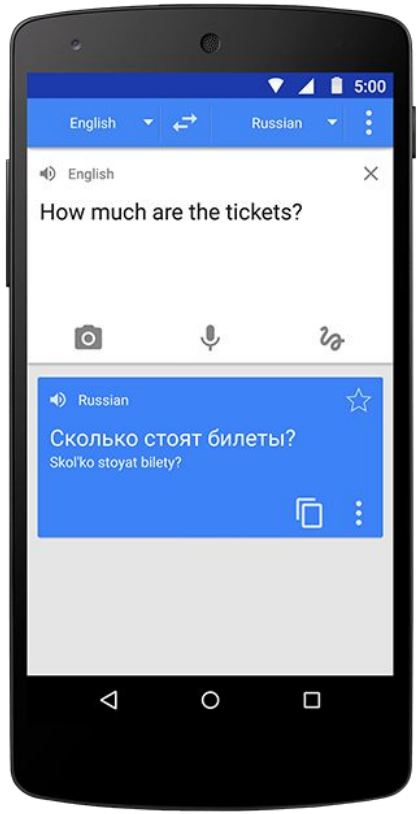
This is the best tl;dr I could make, original reduced by 91%. (I'm a bot)

> The Crown Prosecution Service is facing embarrassment after admitting it destroyed key emails relating to the WikiLeaks founder Julian Assange, who is holed up in Ecuador's London embassy fighting extradition.
>
> Email exchanges between the CPS and its Swedish counterparts over the high-profile case were deleted after the lawyer at the UK end retired in 2014.
>
> The CPS, responding to questions from the Guardian, denied there were any legal implications of the data loss for an Assange case if it were to come to court in the future.

Extended Summary | FAQ | Feedback | *Top keywords*: **CPS**#1 **Assange**#2 **case**#3 **Swedish**#4 **Email**#5

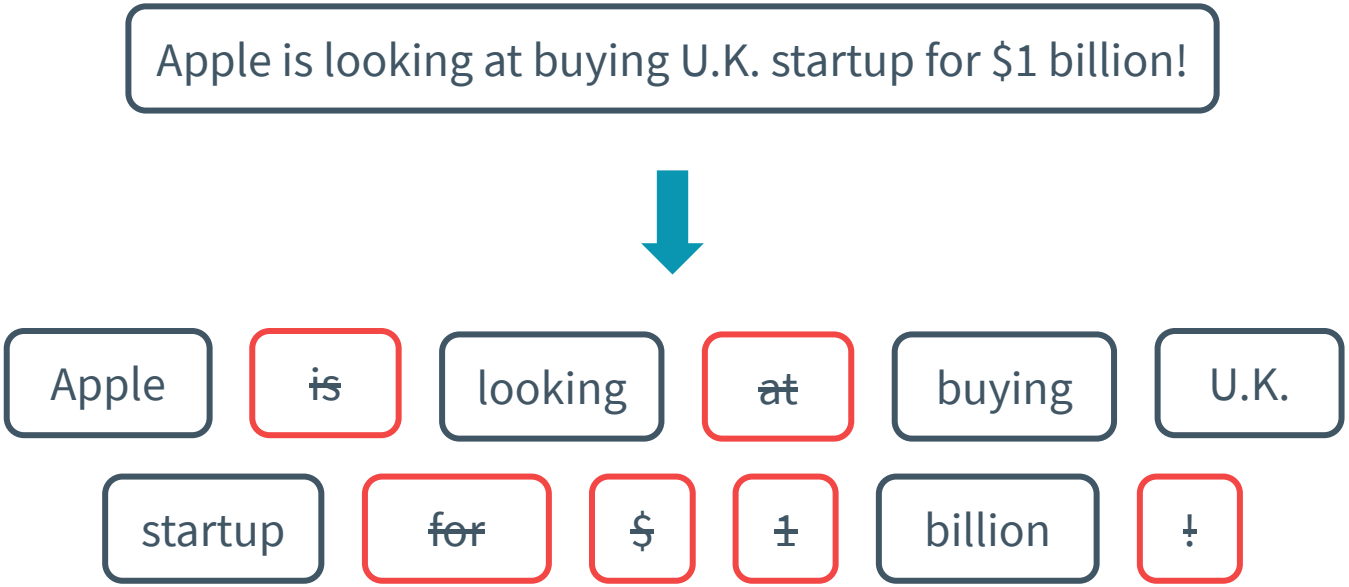
permalink embed save report give gold REPLY

Tradução


5:00
English
Russian
English
How much are the tickets?
Russian
Сколько стоят билеты?
Skol'ko stoyat bilety?

Conversação Chatbot

Tokenização



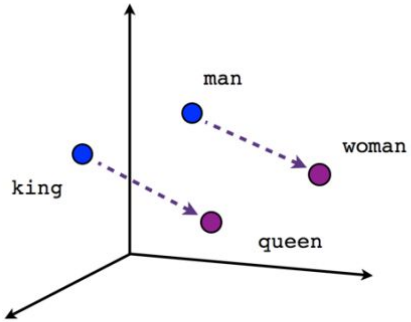
Bag of words

Representa-se uma mensagem por um vetor binário referente a todas suas palavras contidas.

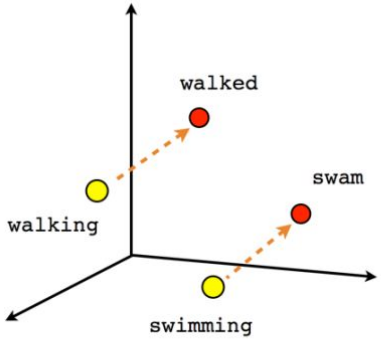
Apple will buy.

google	0
buy	**1**
apple	0
walk	0
shoes	0
apple	**1**
startup	0
will	**1**

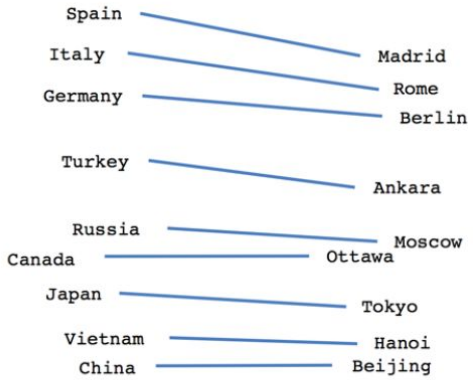
Word2vec



Male-Female



Verb tense



Country-Capital

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Outros Embeddings

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

fastText

Library for efficient text classification and representation learning

Naïve Bayes

Assume-se independência entre as probabilidades dos pares de palavras.

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval.*

Naïve Bayes

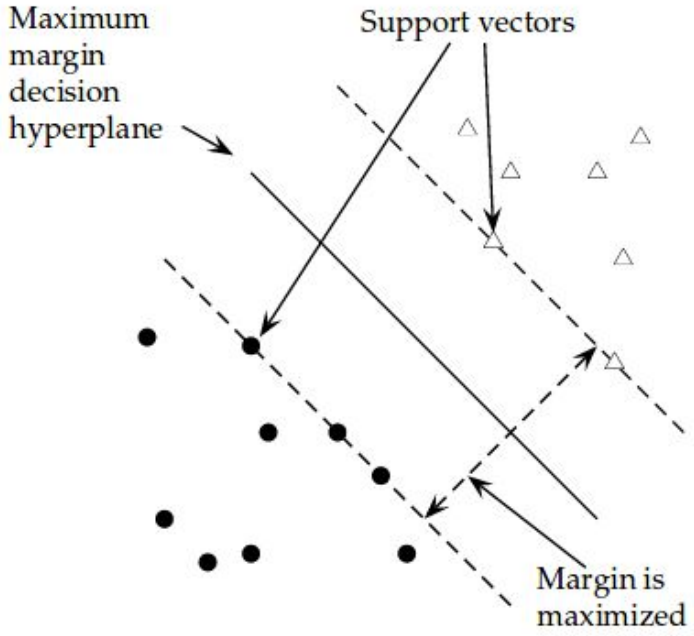
$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

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

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

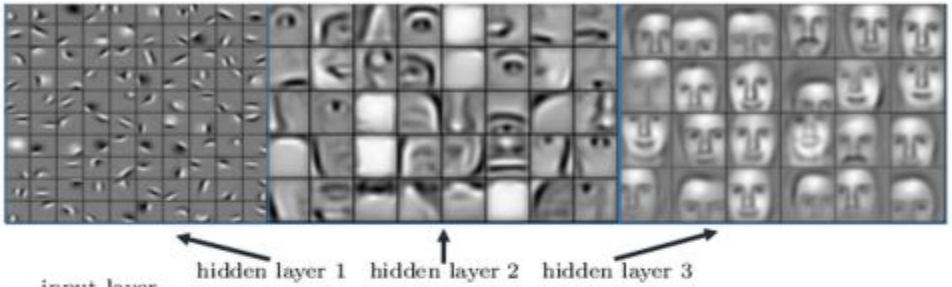
C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval.*

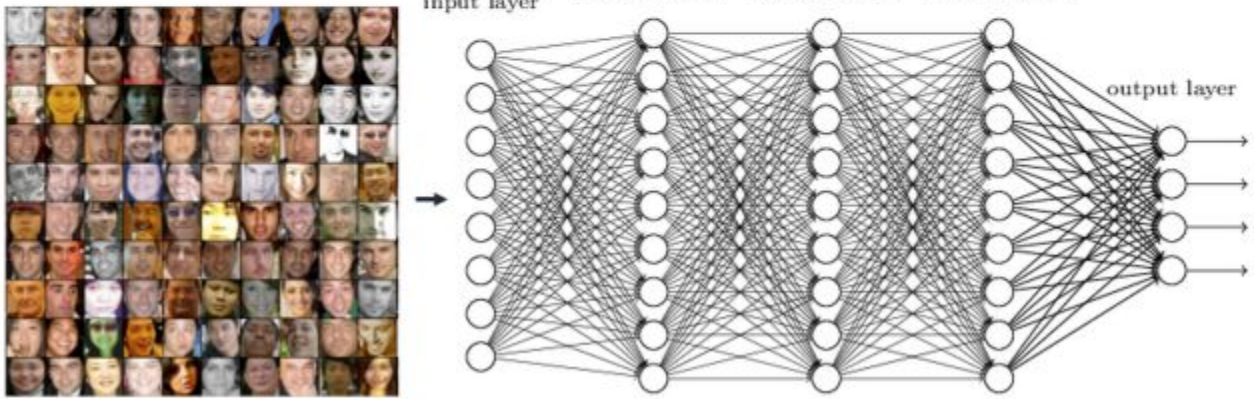
Support Vector Machine



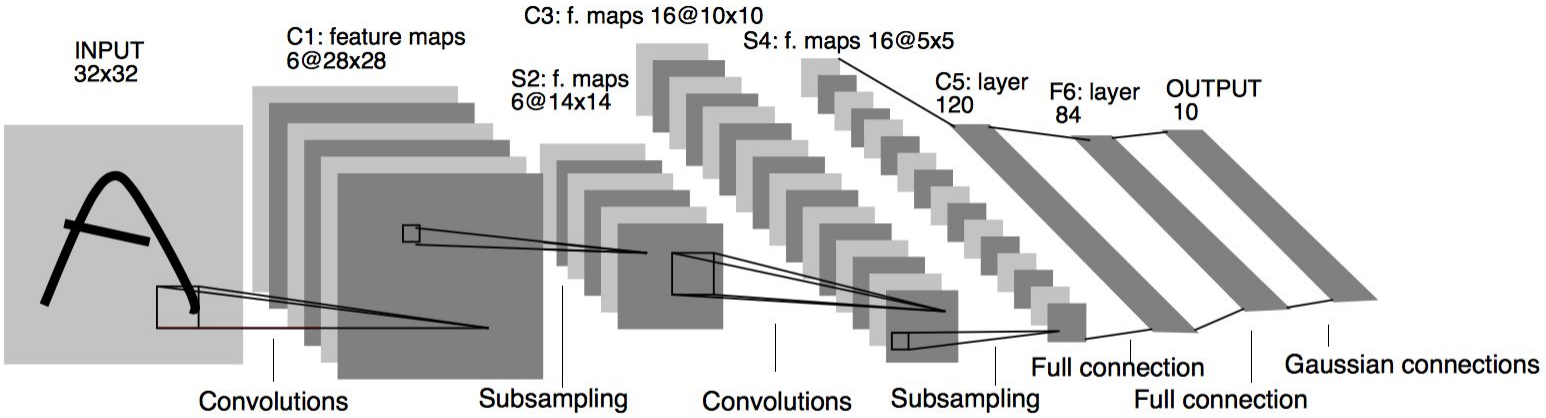
Deep Learning

Deep neural networks learn hierarchical feature representations



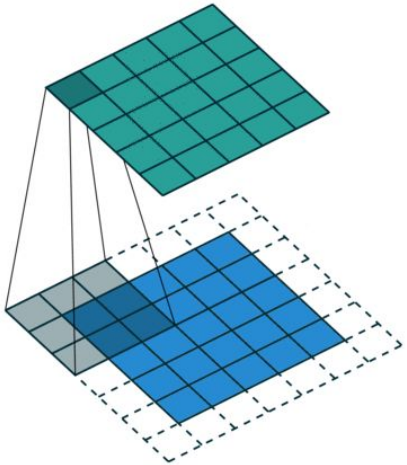


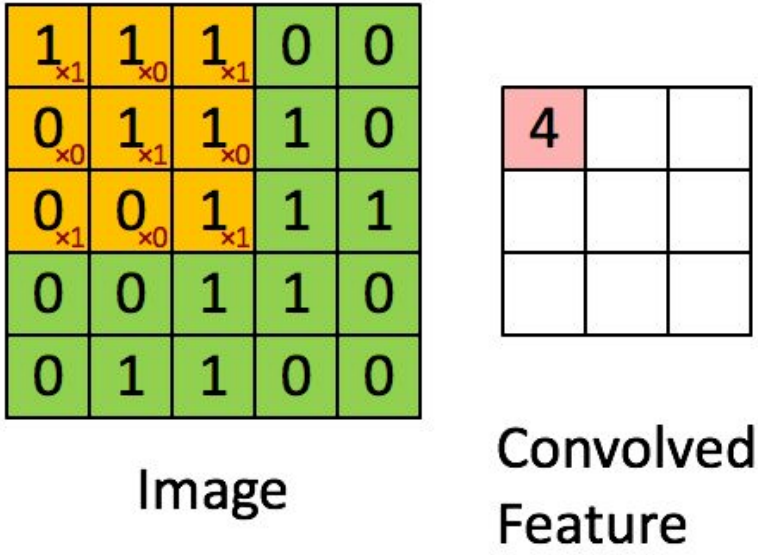
Redes Neurais Convolucionais



Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. *Gradient-based learning applied to document recognition.*

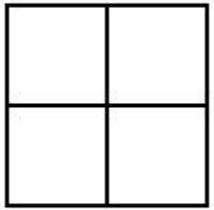
CNN
Convolução



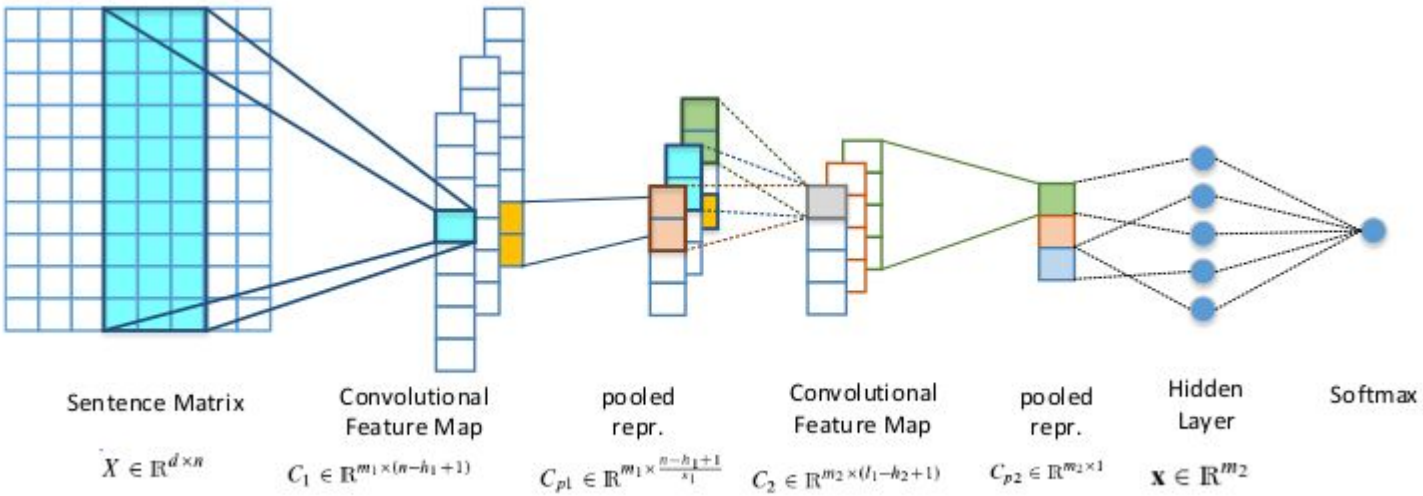


CNN
Max-Pooling

1	3	2	9
7	4	1	5
8	5	2	3
4	2	1	4



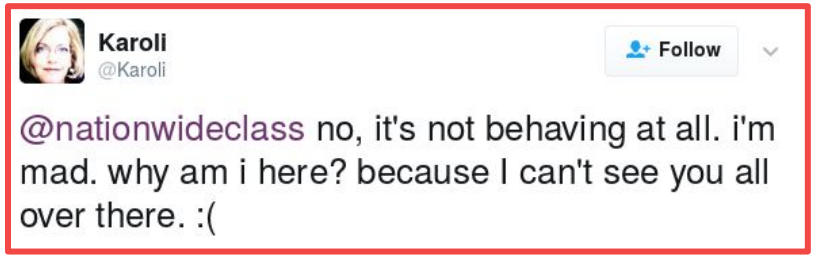
CNN
Aplicadas em Texto



SwissCheese at SemEval-2016 Task 4: Sentiment Classification Using an Ensemble of Convolutional Neural Networks with Distant Supervision

Supervisão distante

Emoticons positivos	Emoticons negativos
:)	:(
:-)	:-(
:D	: (



A. Go, R. Bhayani, L. Huang. *Twitter Sentiment Classification using Distant Supervision.*

Resultados de referência

Acurácia dos classificadores do Sentiment140.

	Máxima Entropia	Naive Bayes	SVM
Unigrama	80,5%	81,3%	82,2%
Bigrama	79,1%	81,6%	78,8%
Unigrama + Bigrama	83,0%	82,7%	81,6%
Unigrama + POS	79,9%	79,9%	81,9%

A. Go, R. Bhayani, L. Huang. *Twitter Sentiment Classification using Distant Supervision.*

Semantic Evaluation

- Conferência anual de análise semântica computacional.
- Realiza competições em análise de sentimento de tweets desde 2013.
- Participantes disponibilizam artigos com metodologia, técnicas e resultados obtidos.

Método

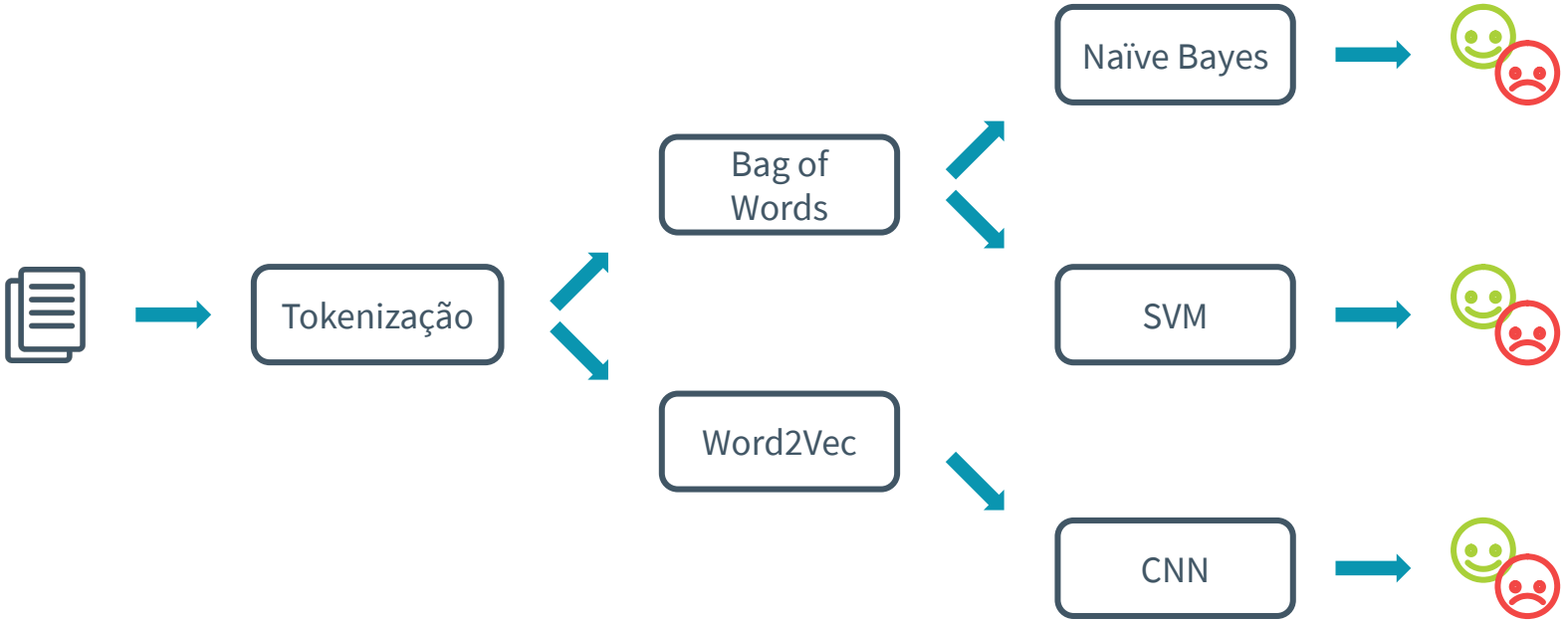
Bases de dados

	Tamanho	Anotação	Ano
Sentiment140 Treino	**1,6M**	**supervisão distante**	**2009**
Sentiment140 Teste	**360**	**manual**	**2009**
Base Própria	**40M**	**supervisão distante**	**2015 - 2017**
SemEval	**60K**	**manual**	**2013 - 2017**

Desenvolvimento

	Objetivo	Treinamento	Teste
Etapa 1	**Replicar classificadores do Sentiment140.**	**Sentiment140 Treino**	**Sentiment140 Teste**
Etapa 2	**Treinar os classificadores a partir de base de dados própria.**	**Base Própria**	**SemEval**
Etapa 3	**Aplicar técnicas de Deep Learning para classificação de sentimento.**	**Base Própria**	**SemEval**

Classificadores
Visão Geral



Etapa 3
Word2Vec

- Vocabulário de **3M termos**
- Embedding de **300 dimensões**
- Treinado em 100B palavras de notícias.
- Contém **89k** dos **1,1M** termos únicos dos datasets de treino e teste.

Etapa 3
Hiperparâmetros

- ReLU
- Entropia Cruzada
- Adam
- L2
- Dropout
- Early-Stopping

- Nº Camadas: 1 e 2
- Nº Filtros Conv.: 100 e 200
- Tam. Filtros Conv.: 2 e 3
- Tam. Pooling: 2, 3 e 5

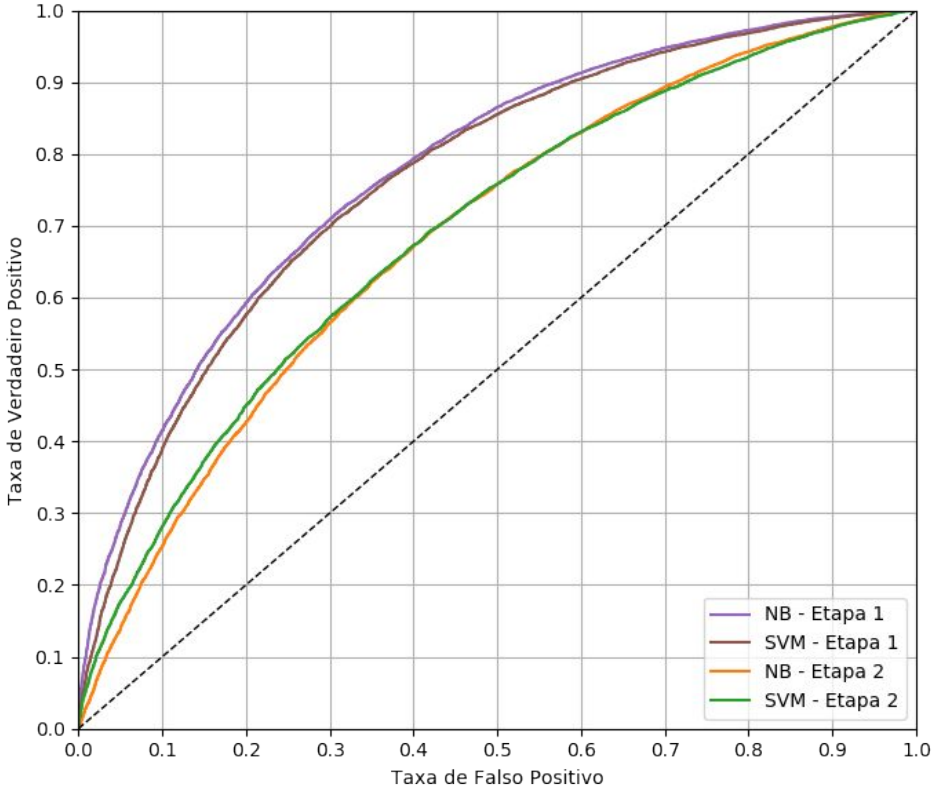
Resultados

Etapa 1
Resultados

Acurácia

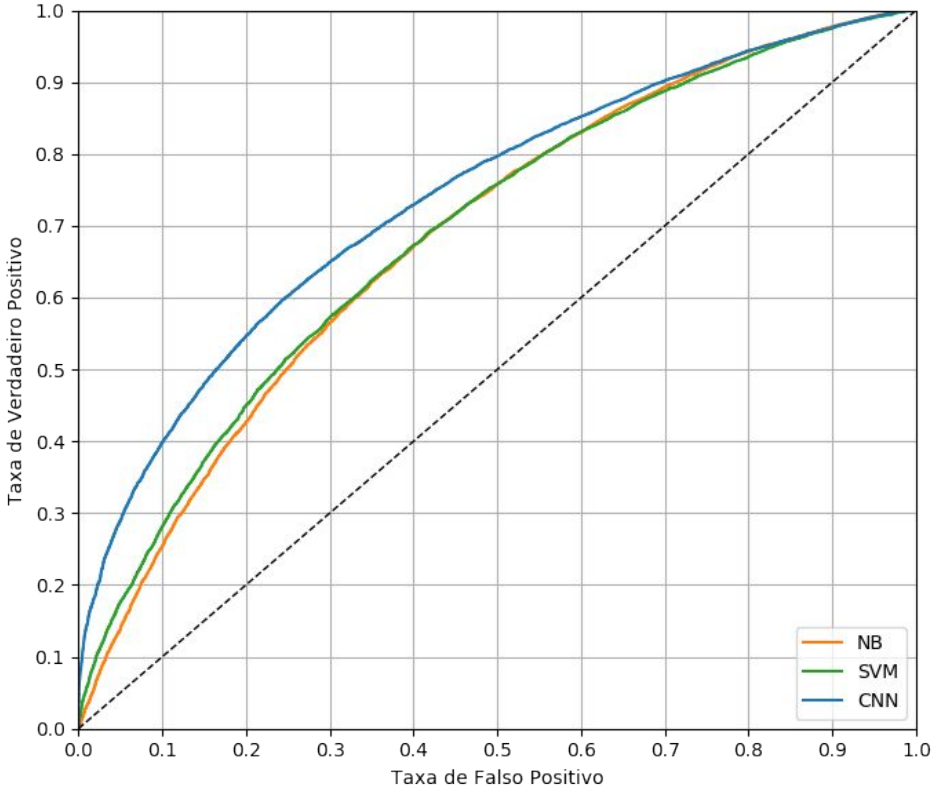
	Referência	Reprodução
Naive Bayes	**81,3%**	**83,3%**
SVM	**82,2%**	**83,0%**

Etapa 2
Resultados



	AUC	SP
Naïve Bayes	**0,688**	**0,639**
SVM	**0,693**	**0,640**

Etapa 3
Resultados



	AUC	SP
Naïve Bayes	**0,688**	**0,639**
SVM	**0,693**	**0,640**
CNN	**0,738**	**0,675**

Conclusões

1. Reproduzimos com sucesso classificadores de sentimento de redes sociais.
2. Se treinaram classificadores eficientes a partir de bases anotadas por supervisão distante.
3. Deep Learning obteve desempenho superior a técnicas clássicas.

Trabalhos Futuros

- Aplicar o processo para português.
- Adicionar sentimento neutro.
- Avaliar outras técnicas de Deep Learning.

Obrigado!

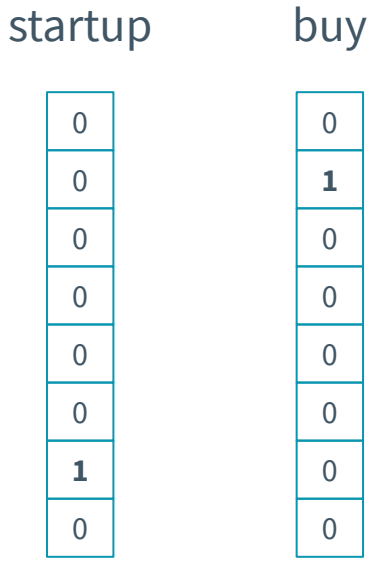
Perguntas?

X.
Extras

Polaridade de mensagem

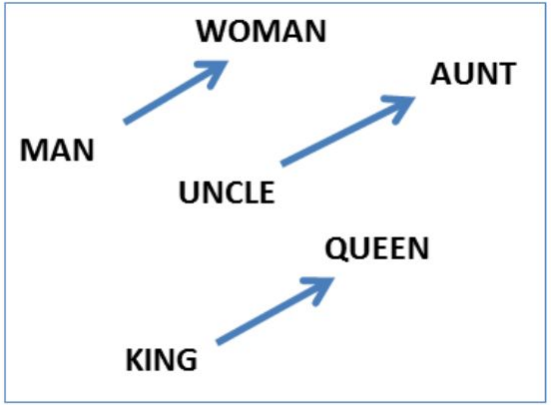
Codificação One-Hot

É criado um espaço vetorial no qual cada palavra do vocabulário corresponde a uma dimensão.



Word2vec

- Embedding que representa uma palavra em um vetor real de tamanho arbitrário.
- É produzido a partir de janelas de contexto ao redor das palavras.



T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. *Distributed Representations of Words and Phrases and their Compositionality.*

Word2vec
Janela de Contexto

In this paper we present several extensions that improve both the quality of the vectors and the training speed.

Word2vec
Arquiteturas

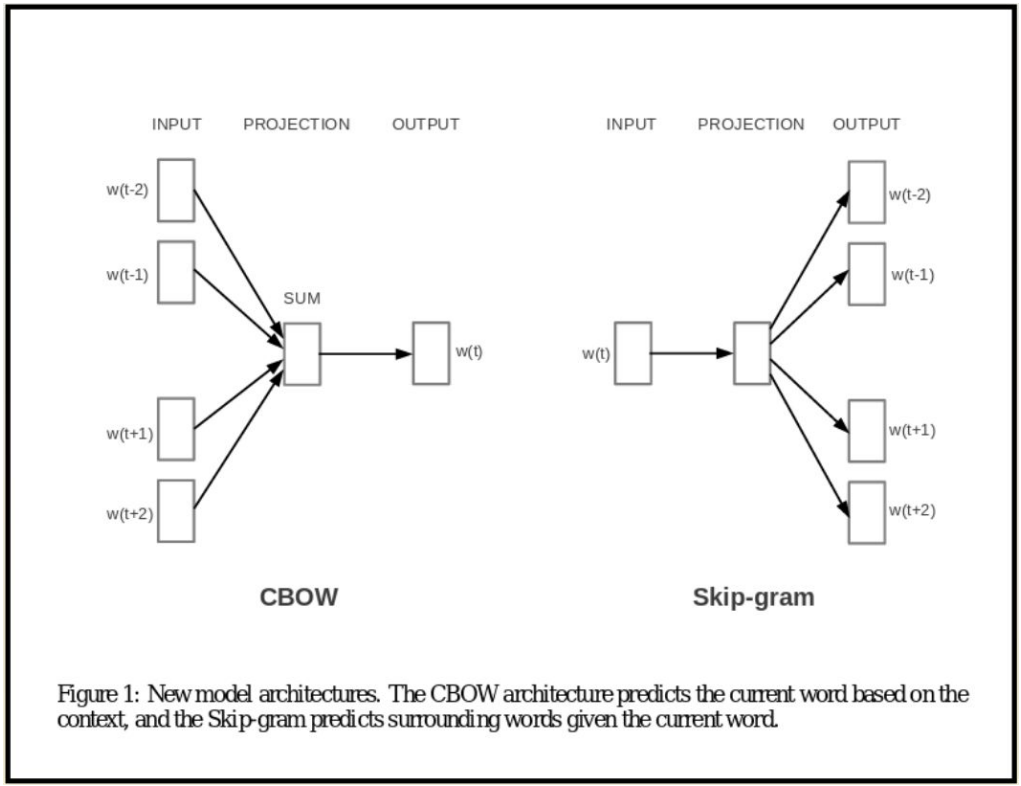
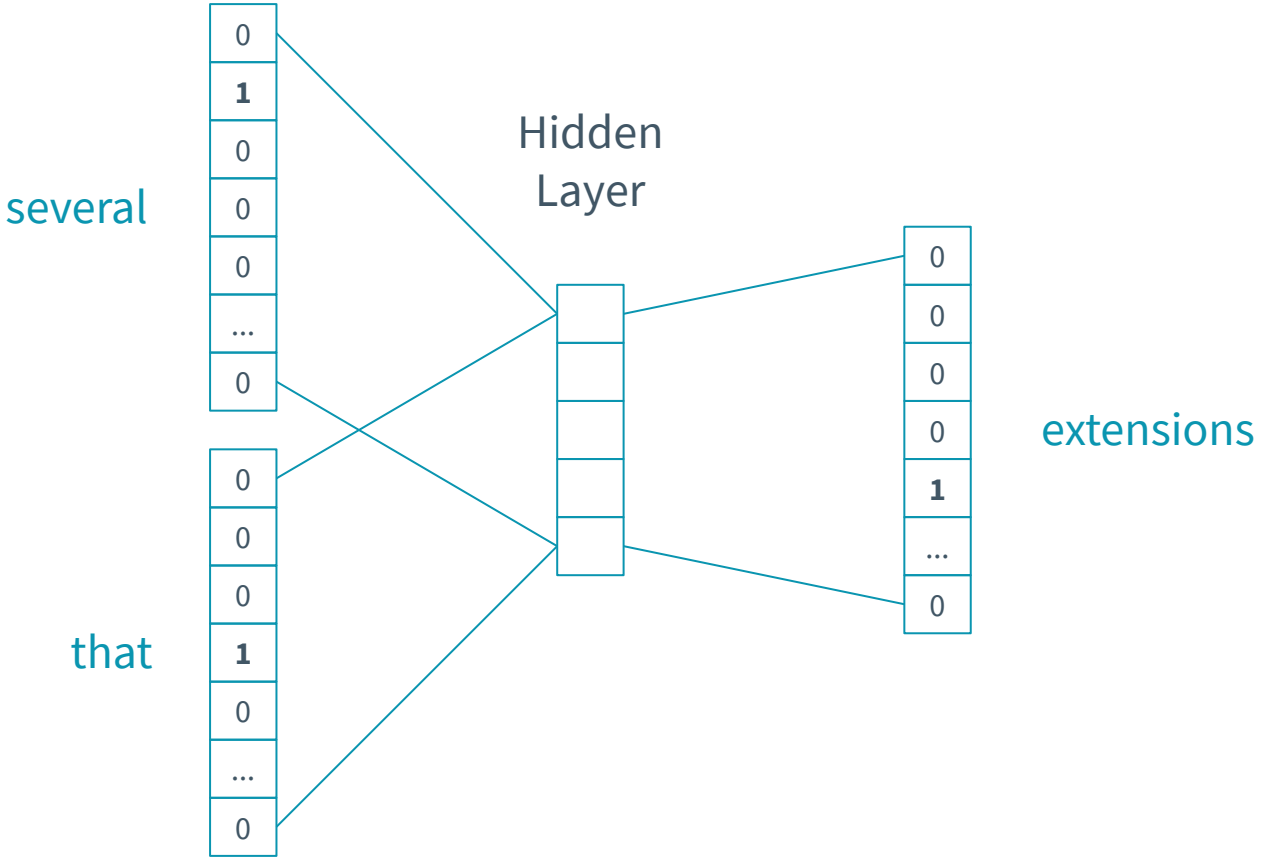


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

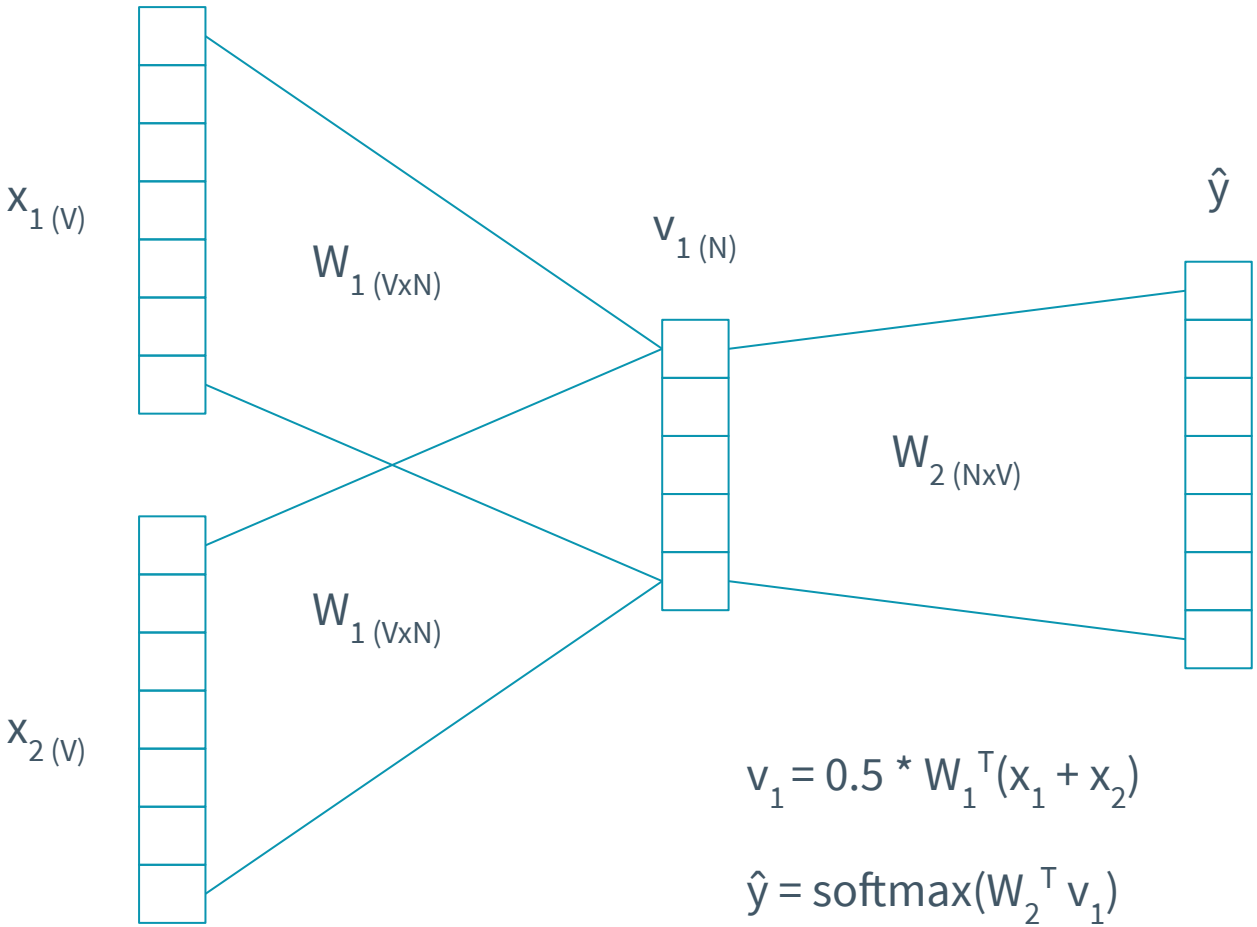
Treinamento Word2vec

CBOW



Treinamento Word2vec

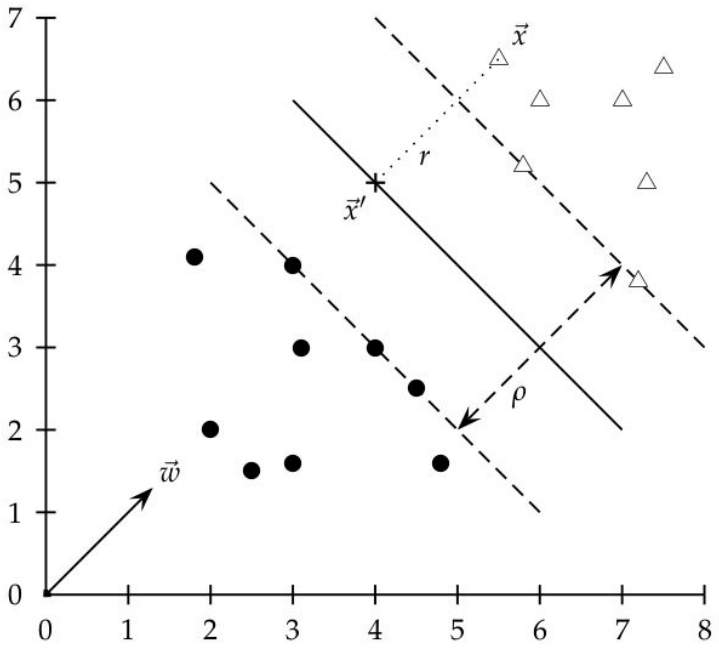
CBOW



$$v_1 = 0.5 * W_1^T(x_1 + x_2)$$

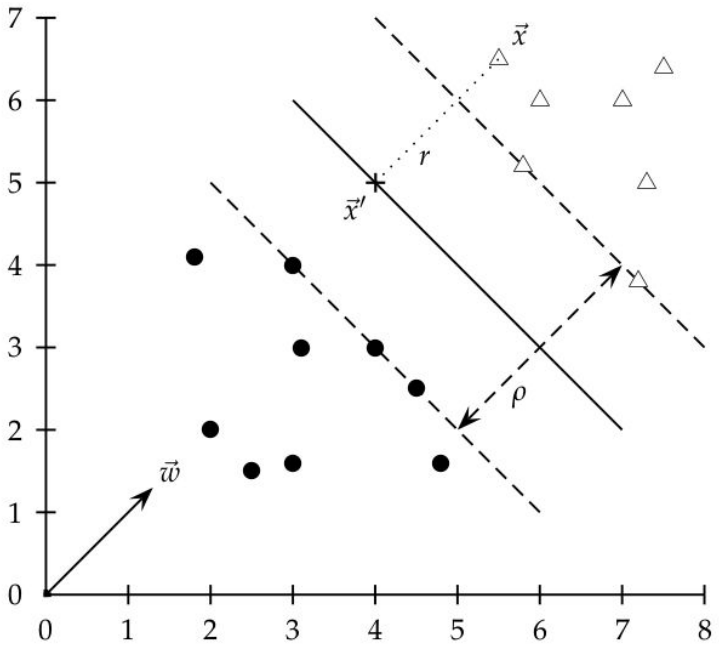
$$\hat{y} = \text{softmax}(W_2^T v_1)$$

SVM



$$f(\vec{x}) = \text{sign}(\vec{w}^{\mathrm{T}}\vec{x} + b)$$

SVM



$$\vec{x}' = \vec{x} - yr\frac{\vec{w}}{|\vec{w}|}$$

$$\vec{w}^{\mathrm{T}}\left(\vec{x} - yr\frac{\vec{w}}{|\vec{w}|}\right) + b = 0$$

$$r = y\frac{\vec{w}^{\mathrm{T}}\vec{x} + b}{|\vec{w}|}$$

$\rho = 2/|\vec{w}|$ is maximized

For all $(\vec{x}_i, y_i) \in \mathbb{D}$, $y_i(\vec{w}^{\mathrm{T}}\vec{x}_i + b) \geq 1$

Inicialização dos neurônios

Neurônios: inicialização Glorot uniforme[1]

Bias: inicialização em 0

$$^{1}limites = \pm ganho \cdot \sqrt{\frac{2}{entradas + saidas}} \cdot \sqrt{3}$$

$$ganho = \begin{cases} \sqrt{2} & \text{se } \varphi = \text{ReLU} \\ 1 & \text{se } \varphi = \text{Sigmoid} \end{cases}$$

X. Glorot, Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks.*

Regra de atualização
Otimização Adam

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1$^{\text{st}}$ moment vector)
$v_0 \leftarrow 0$ (Initialize 2$^{\text{nd}}$ moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

D. Kingma, J. Lei Ba. *Adam: A Method for Stochastic Optimization*.

Função custo

- Função custo: Entropia cruzada

  Permite interpretar ŷ como probabilidade

$$\frac{-\sum[(\mathbf{y} \cdot log(\hat{\mathbf{y}}) + (1 - \mathbf{y}) \cdot log(1 - \hat{\mathbf{y}})) + \frac{\lambda}{2} \cdot \mathbf{W}^2]}{n}$$

Hiperparâmetros

Otimização

- $\alpha = 0{,}001$
- $\beta_1 = 0{,}9$
- $\beta_2 = 0{,}999$

Regularização

- L2: $\lambda = 0{,}001$
- Dropout: 0,5

D. Kingma, J. Lei Ba. *Adam: A Method for Stochastic Optimization*.

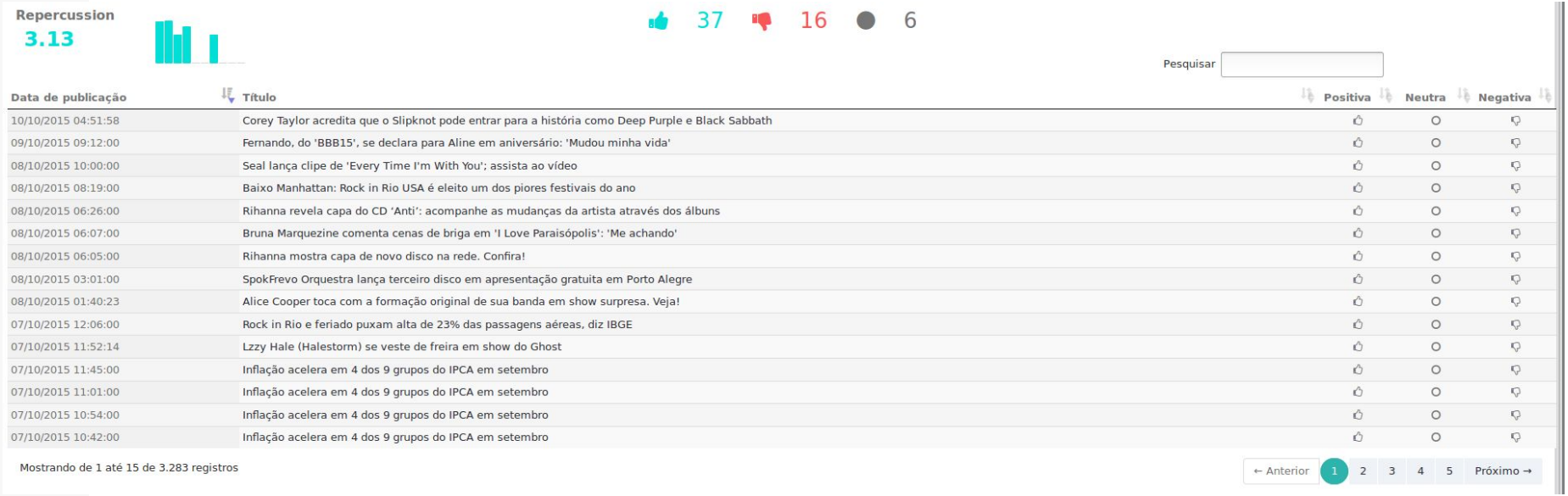
Aquisiçao de dados

Tweets podem ser facilmente coletados através de API.

Podem ser coletados tweets de um contexto específico, a partir de filtro por hashtag ou palavras chave, ou não.

9. Aplicações

Aplicações


Repercussion
3.13
37
16
6
Pesquisar
Data de publicação
Título
Positiva
Neutra
Negativa
10/10/2015 04:51:58 Corey Taylor acredita que o Slipknot pode entrar para a história como Deep Purple e Black Sabbath
09/10/2015 09:12:00 Fernando, do 'BBB15', se declara para Aline em aniversário: 'Mudou minha vida'
08/10/2015 10:00:00 Seal lança clipe de 'Every Time I'm With You'; assista ao vídeo
08/10/2015 08:19:00 Baixo Manhattan: Rock in Rio USA é eleito um dos piores festivais do ano
08/10/2015 06:26:00 Rihanna revela capa do CD 'Anti': acompanhe as mudanças da artista através dos álbuns
08/10/2015 06:07:00 Bruna Marquezine comenta cenas de briga em 'I Love Paraisópolis': 'Me achando'
08/10/2015 06:05:00 Rihanna mostra capa de novo disco na rede. Confira!
08/10/2015 03:01:00 SpokFrevo Orquestra lança terceiro disco em apresentação gratuita em Porto Alegre
08/10/2015 01:40:23 Alice Cooper toca com a formação original de sua banda em show surpresa. Veja!
07/10/2015 12:06:00 Rock in Rio e feriado puxam alta de 23% das passagens aéreas, diz IBGE
07/10/2015 11:52:14 Lzzy Hale (Halestorm) se veste de freira em show do Ghost
07/10/2015 11:45:00 Inflação acelera em 4 dos 9 grupos do IPCA em setembro
07/10/2015 11:01:00 Inflação acelera em 4 dos 9 grupos do IPCA em setembro
07/10/2015 10:54:00 Inflação acelera em 4 dos 9 grupos do IPCA em setembro
07/10/2015 10:42:00 Inflação acelera em 4 dos 9 grupos do IPCA em setembro
Mostrando de 1 até 15 de 3.283 registros
← Anterior
1
2
3
4
5
Próximo →

Aplicações


jogar
fiscalizacao
irregularidades
chuva
pouca
barrada
encontra
trocam
homem
procon
briga
encontram
guarda
sao
escravidao
irregular
trabalhadores
pm
r7
faith
publico
saida
civil
condicoes
tecnico
pega
analogas
trabalho
resgata
municipal
festa
russa
tiros
problema
fiscais
namorado
plateia
more
pompa
policial
rihanna
produtos
dormiam
militar
apreende
noticias
alojamento
brinquedo
sunset
logo


nostalgia
festival
setlist
ganha
cidade
depp
katy
perry
slipknot
fas
smith
fecha
system
rihanna
palco
veja
fa
noticias
medina
canta
dowa
nao
marquezine
casamento
primeiro
bruce
chuva
vampires
ultimo
levam
tocar
musica
classicos
camerata
encerra
cantora
brasil
metal
beijos
more
pompa