Introdução a Processamento de Linguagem Natural

Jonatas Grosman
Luiz Schirmer
William Fernandes

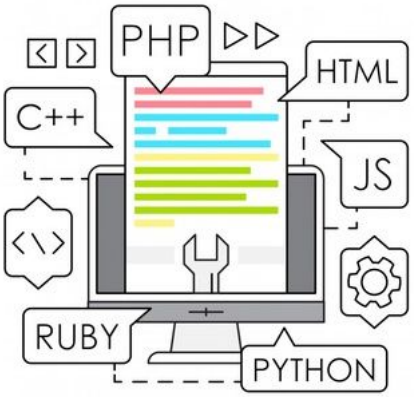
PUC RIO

DEPARTAMENTO DE INFORMÁTICA
PUC·RIO

O que é Linguagem Natural?

Linguagem natural

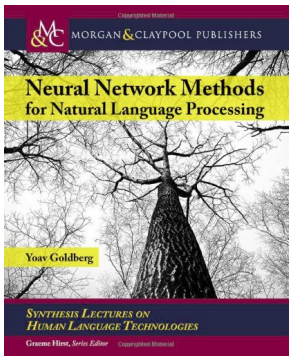
Linguagem artificial



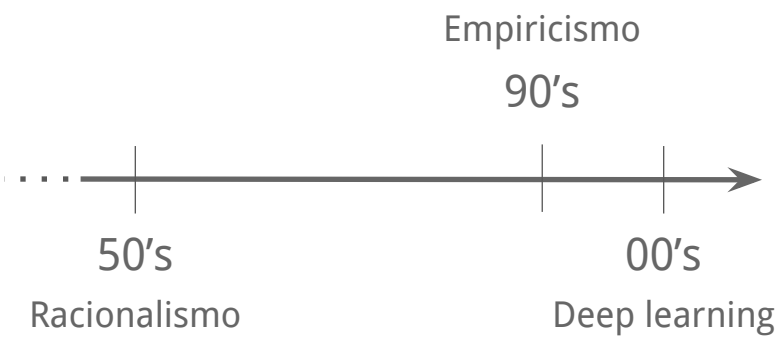
O que é Processamento de Linguagem Natural?

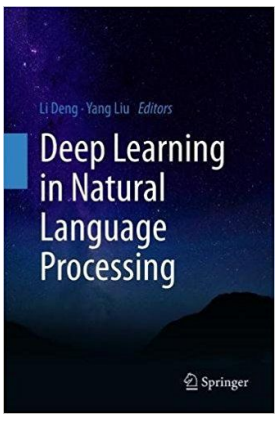
"Processamento de linguagem natural (PLN ou NLP do inglês *natural language processing*) é o campo que projeta métodos e algoritmos que recebem como entrada ou produzem como saída dados em linguagem natural não estruturados"

Goldberg, Y. (2017)



As ondas de NLP





Li, D., & Yang, L. (2014)

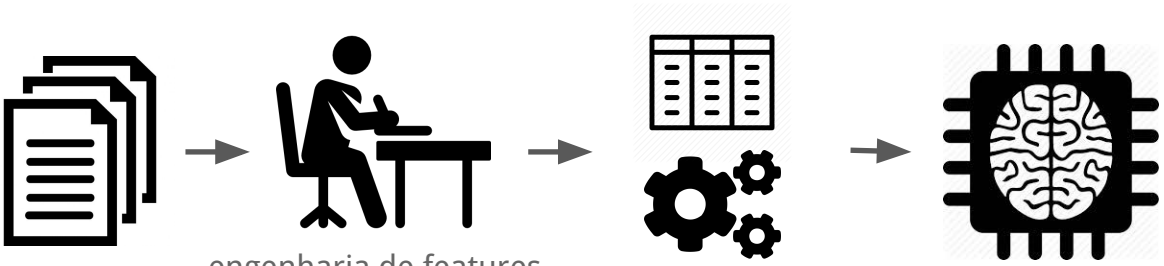
Onda racionalista

- Conhecimento da linguagem na mente humana é fixo
- Baseado em regras

criação de regras

Onda empiricista

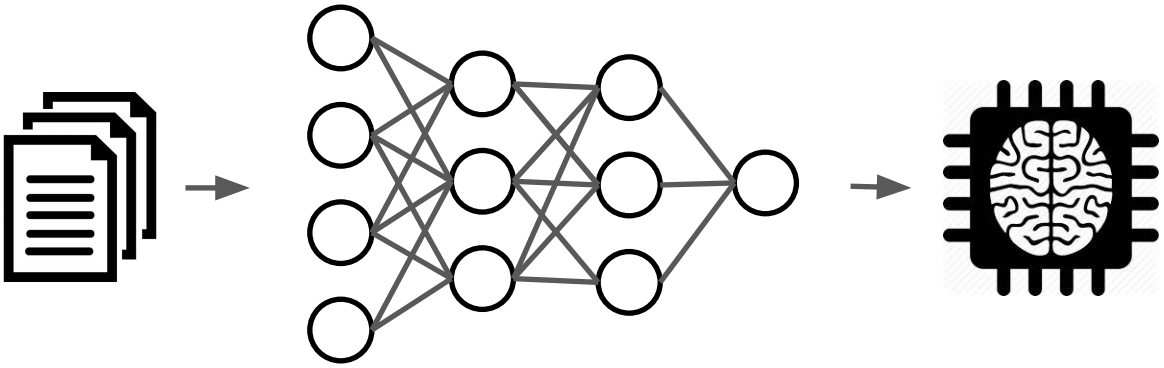
- A mente humana somente começa com operações gerais de associação, reconhecimento de padrões e generalização
- Baseada em corpora



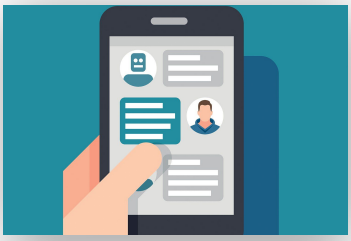
engenharia de features

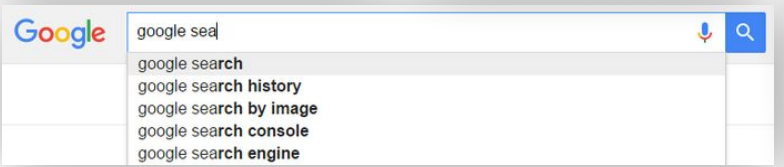
Onda de deep learning

- Independente de especialista de domínio
- Baseado em “grandes corpora”



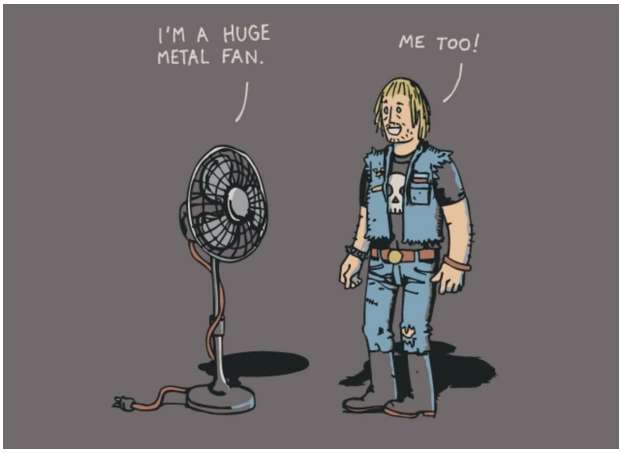
Exemplos de aplicação de NLP





Por que as tarefas de NLP são tão difíceis?

A linguagem natural é altamente ambígua



Por que as tarefas de NLP são tão difíceis?

A linguagem natural é dependente do contexto


DAILY Mirror
FALKLANDS ARE TAKEN
VICTORY
Argentine troops flee task force
Generals meet for ceasefire
Mirror comment

Por que as tarefas de NLP são tão difíceis?

A linguagem natural está sempre evoluindo

Tarefas de NLP

Brasil

tarefa sintática

substantivo

6 letras

2 vogais

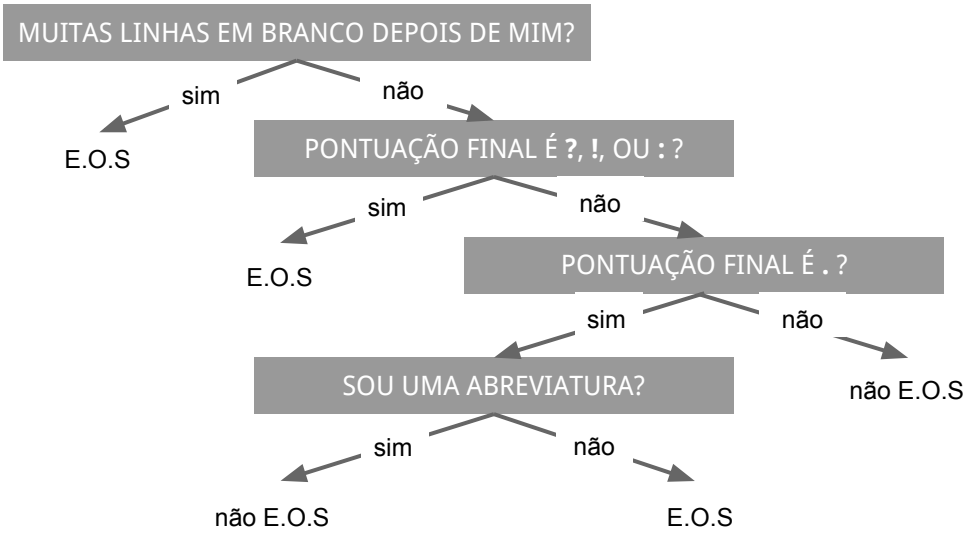
...

tarefa semântica

Segmentação de sentenças

- !,? não são muito ambíguos
- A pontuação "." é muito ambígua
  - Números .04 or 5.0 …
  - Limite da sentença
  - Abreviações (Inc. , Mr.)
- Construção de um classificador binário!
  - Olhe um "."
  - Decide EndOfSentence/NotEndOfSentence
  - Classificadores: regras escritas manualmente, expressões regulares, machine learning

Segmentação de sentenças



Segmentação de palavras

- Separar um trecho de texto contínuo em palavras separadas

  ogatonochapéu -> o | gato | no | chapéu

- Em Português é uma tarefa trivial -> espaços
- Em Chinês, Japonês etc as palavras não são limitadas desta maneira

Lematização

- Duas palavras têm a mesma raiz?

  sou, são, é -> ser
  carro, carros -> carro

  Os carros de os meninos são de cores diferentes -> ???

Stemização

- Reduzir palavras flexionadas a sua base

  voto, vota, votei -> vot

  provimento, proveu, provido -> prov

  Voto em o sentido de dar provimento a o recurso -> ???

Etiquetagem de classe gramatical (POS tagging)

- Determinar a classe gramatical de cada palavra em uma sentença

Palavra:	O	carro	é	preto
POS:	DET	N	V	ADJ

Etiquetagem de classe gramatical (POS tagging)

Exemplo de conjunto de tags:

Nome	**Classe gramatical**
N	Substantivo
PROP	Nome próprio
SPEC	Especificador (pronome ou adjectivo)
DET	Determinante (artigo, pronome ou adjectivo)
PERS	Pronome pessoal
ADJ	adjectivo
ADV	advérbio
V	verbo
NUM	numeral
KS	conjunção subordinativa
KC	conjunção coordenativa
PRP	preposição
IN	interjeição
EC	prefixo

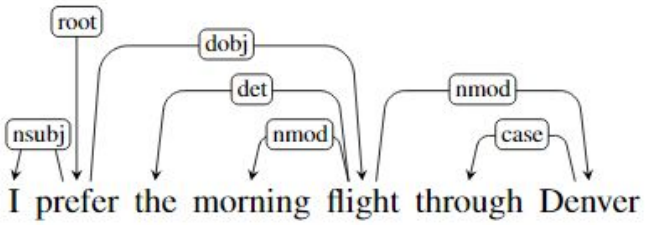
Etiquetagem de classe gramatical (POS tagging)

- Componente importante para análise sintática
- Predição de palavras em reconhecimento de fala
  - – Pronomes possessivos (meu, seu, dela) são geralmente seguidos por substantivos
  - – Pronomes pessoais (eu, você, ele) são geralmente seguidos por verbos
- Tradução automática

Dependency parsing

- Idéia básica:

  - A estrutura sintática consiste de relações binárias, assimétricas entre palavras de uma sentença



Dependency parsing

Exemplo de conjunto de tags:

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Figure 14.2 Selected dependency relations from the Universal Dependency set. (de Marneffe et al., 2014)

Predição de palavra

- Calcular a probabilidade de uma palavra vir a seguir

Geração de linguagem natural

bathrooms	bedrooms	year_built	num_floors	house_style
4	3.5	2007	1	Craftsman
3	2	1961	1	Bungalow
4	2	2004	1	Ranch
4	2	1930	2	Craftsman
3	2.5	1900	1	Ranch
2	2	1942	2	Bungalow
3	2	2008	2	Cape Cod
2	1.5	1946	2	Modern
3	1.5	1937	2	Craftsman
4	3	1900	1	Bungalow

Classificação de documentos

- SPAM?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

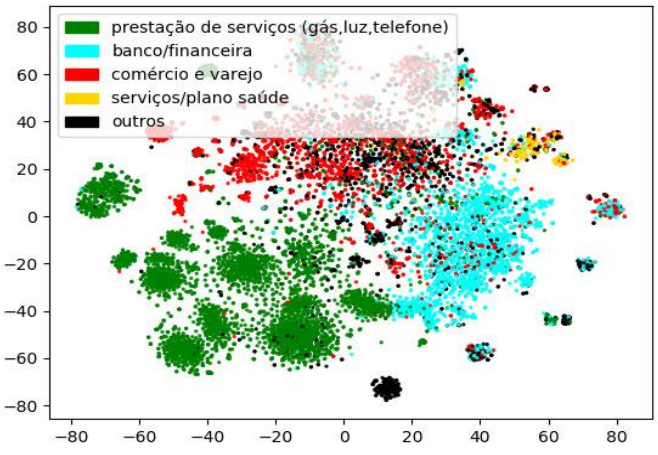
You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Classificação de documentos

Qual é o contexto destes processos judiciais?



Classificação de documentos

- Atribuir assunto, tópico ou gênero
- Detecção de spam
- Identificação de autoria
- Identificação de idade/gênero
- Identificação de língua
- Análise de sentimento
- ...

Entrada:
um documento d
um conjunto fixo de classes C = {c1, c2,…, cj}

Saída: uma classe predita c de C

Classificação de documentos

Regras escritas manualmente

- Regras baseadas em combinações de palavras ou outros atributos
  - spam: endereço em lista negra, US$
- Qualidade é alta -> especialista
- Construir e manter regras é caro

Classificação de documentos

Aprendizado de máquina

Entrada:

- um documento d
- um conjunto fixo de classes C = {c1, c2,…, cj}
- um conjunto de documentos anotados manualmente

Saída: um classificador treinado

Classificadores

- Naïve Bayes
- Regressão logística
- Support-vector machines
- k-Nearest Neighbors

Reconhecimento de entidades nomeadas

- Encontrar e classificar nomes em um texto

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Reconhecimento de entidades nomeadas

- Encontrar e classificar nomes em um texto

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Reconhecimento de entidades nomeadas

- Encontrar e classificar nomes em um texto

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Reconhecimento de entidades nomeadas

- Encontrar e classificar nomes em um texto

Person
Date
Location
Organization

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Extração de relações

- Texto:

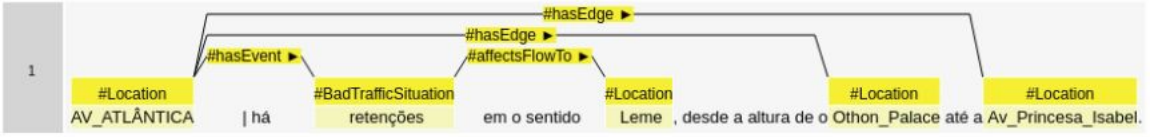
  "A professora Simone Barbosa trabalha para a PUC-Rio".

- Entidades nomeadas extraídas:
  - Pessoa Simone Barbosa
  - Organização PUC-Rio

- Relação extraída:
  - Trabalha-para Simone Barbosa -> PUC-Rio

Extração de relações

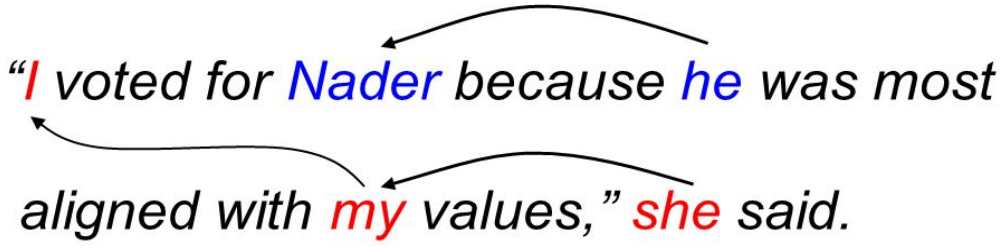
(a)



(b)

Resolução de correferência

Resolução de correferência é a tarefa de encontrar todas as expressões que referem a mesma entidade em um texto



Desambiguação

APPLE

Respostas a perguntas

Sumariamento

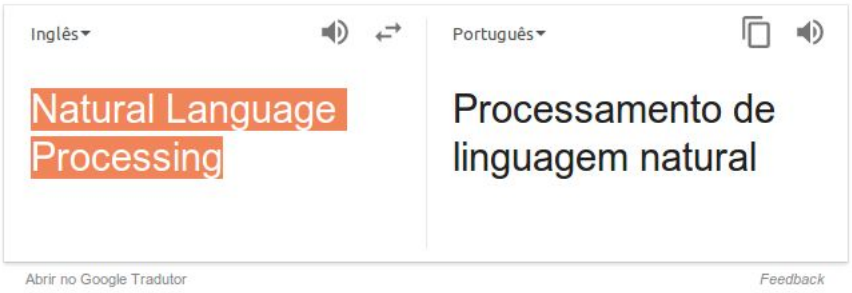
cia documents reveal iot-specific televisions can be used to secretly record conversations .
cybercriminals who initiated the attack managed to commandeer a large number of internet-connected devices in current use .
cia documents revealed that microwave ovens can spy on you - maybe if you personally don't suffer the consequences of the sub-par security of the iot .

Internet of Things (IoT) security breaches have been dominating the headlines lately . WikiLeaks's trove of CIA documents revealed that internet-connected televisions can be used to secretly record conversations . Trump's advisor Kellyanne Conway believes that microwave ovens can spy on you - maybe she was referring to microwave cameras which indeed can be used for surveillance . And don't delude yourself that you are immune to IoT attacks , with 96 % of security professionals responding to a new survey expecting an increase in IoT breaches this year . Even if you personally don't suffer the consequences of the sub-par security of the IoT , your connected gadgets may well be unwittingly cooperating with criminals . Last October , Internet service provider Dyn came under an attack that disrupted access to popular websites . The cybercriminals who initiated the attack managed to commandeer a large number of internet-connected devices (mostly DVRs and cameras) to serve as their helpers . As a result , cybersecurity expert Bruce Schneier has called for government regulation of the IoT , concluding that both IoT manufacturers and their customers don't care about the security of the 8.4 billion internet-connected devices in current use . Whether because of government regulation or good old-fashioned self-interest , we can expect increased investment in IoT security technologies . In its recently-released TechRadar report for security and risk professionals , Forrester Research discusses the outlook for the 13 most relevant and important IoT security technologies , warning that " there is no single , magic security bullet that can easily fix all IoT security issues . " Based on Forrester's analysis , here's my list of the 6 hottest technologies for IoT security : IoT network security : Protecting and securing the network connecting IoT devices to back-end systems on the internet . IoT network security is a bit more challenging than traditional network security because there is a wider range of communication protocols , standards , and device capabilities , all of which pose significant issues and increased complexity . Key capabilities include traditional endpoint security features such as antivirus and antimalware as well as other features such as firewalls and intrusion prevention and detection systems . Sample vendors : Bayshore Networks , Cisco , Darktrace , and Senrio . IoT authentication : Providing the ability for users to authenticate an IoT device , including managing multiple users of a single device (such as a connected car) , ranging from simple static password/pins to more robust authentication mechanisms such as two-factor

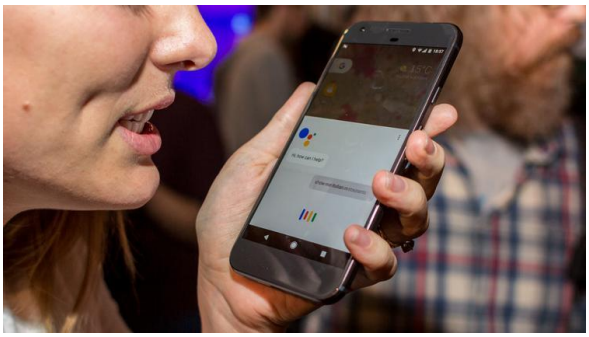
Sumariamento

- Aplicações
  - esboços ou resumos de qualquer documento
  - sumários de threads de email
  - itens de ação de uma reunião
  - simplificação de texto pela compressão de sentenças

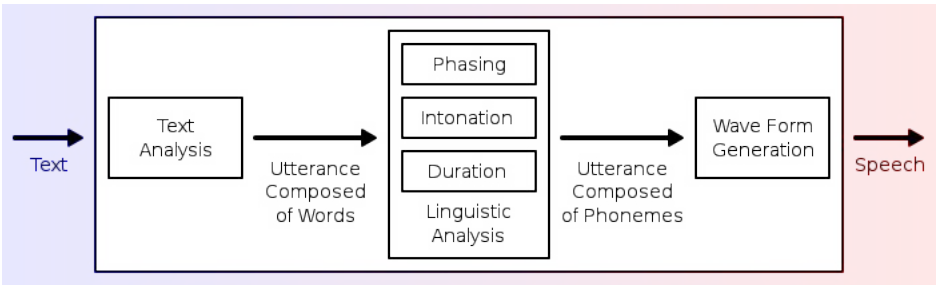
Tradução



Reconhecimento de fala



Sintetização de fala



Regular expression

- A formal language for specifying text strings
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

Regular expression

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Regular expression

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	I have no exquisite reason"
[^e^]	Neither e nor ^	Look here
a^b	The pattern a carat b	Look up a^b now

Regular expression

Pattern	Matches
groundhog|woodchuck	
yours|mine	yours mine
a|b|c	= [abc]
[gG]roundhog|[Ww]oodchuck	

Regular expression

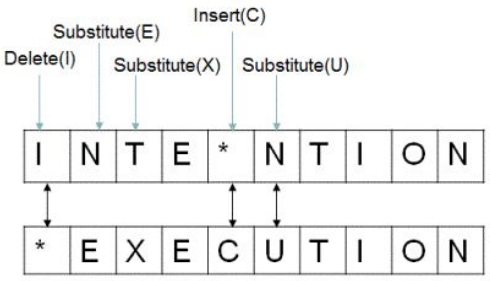
Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun begun beg3n

Regular expression

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
\.$	The end.
.$	The end? The end!

Edit distance

- How similar are two strings?
- Spell correction:
  - The user typed “graffe”
- What is the closest?
  - graffite
  - graf
  - grail
  - giraffe
- Machine translation, Speech Recognition..



Bag of words

Also known as vector space models (VSMs)

Sentence 1 (S1): The car is blue.

Sentence 2 (S2): The color of her eyes is blue.

	the	of	is	her	blue	color	eyes	car
S1:	1	0	1	0	1	0	0	1
S2:	1	1	1	1	1	1	1	0

Bag of words

$$\mathbf{D} = \begin{pmatrix} \mathrm{tf}\,(t_1, d_1) & \cdots & \mathrm{tf}\,(t_N, d_1) \\ \vdots & \ddots & \vdots \\ \mathrm{tf}\,(t_1, d_\ell) & \cdots & \mathrm{tf}\,(t_N, d_\ell) \end{pmatrix}$$

Where:

l is the number of documents

N is the number of terms

$tf(t_i,d_j)$ is the frequency of term t_i in document d_j

Bag of words > input matrix

		the	of	is	her	blue	color	eyes	car
D	**S1**	1	0	1	0	1	0	0	1
	S2	1	1	1	1	1	1	1	0

		S1	S2
D'	**the**	1	1
	of	0	1
	is	1	1
	her	0	1
	blue	1	1
	color	0	1
	eyes	0	1
	car	1	0

Bag of words > input matrix

Term-by-term matrix D'D

Document-by-document matrix DD'

Where

$D' = D^T$

Bag of words > semantic issues

VSM representation ignores semantic relations

Word position, context, synonyms, homonyms, etc.

Bag of words > improving the quality

Associate different weights (w_i) to each term t_i

Remove uninformative terms (*and*, *or*, *the*, etc.)

Stemming (structural and structured to structur)

Remove the effect of the length of the document

Operations can be performed in sequence

Bag of words > kernel

K = DD'

Uses the document-based representation

$$\kappa(d_1, d_2) = \langle \phi(d_1), \phi(d_2) \rangle = \sum_{j=1}^{N} \mathrm{tf}(t_j, d_1)\, \mathrm{tf}(t_j, d_2)$$

Bag of words > term weighting

Not all terms have the same importance

Stop words (*and*, *or*, *the*, etc.) are removed before the analysis

The frequency of a word across the documents → amount of information

Bag of words > term weighting

We consider *inverse document frequency* (*idf*)

$$w(t) = \ln\left(\frac{\ell}{\mathrm{df}(t)}\right)$$

Where

l: number of documents in the corpus

df(*t*): number of documents containing term t

Stop words can be treated here (*w(t) = 0*), but for efficiency it is better to remove them earlier

Bag of words > term weighting

Given a term weighing, we define a new VSM

We choose the diagonal matrix **R** as

$$\mathbf{R}_{tt} = w(t)$$

The kernel computes

$$\tilde{\kappa}(d_1, d_2) = \boldsymbol{\phi}(d_1)\,\mathbf{R}\mathbf{R}'\boldsymbol{\phi}(d_2)' = \sum_t w(t)^2\,\mathrm{tf}(t, d_1)\,\mathrm{tf}(t, d_2)$$

The evaluation of this kernel involves

Term frequencies and inverse document frequencies (tf-idf)

Bag of words > term proximity

tf-idf is not capable of recognizing two terms that are semantically related

Two documents that share no terms, but share synonyms are not connected

Therefore we need to establish semantic similarities between terms

We need a non-zero off-diagonal matrix **P**

Where

$\mathbf{P}_{ij} > 0$, if term *i* is semantically related to term *j*

Bag of words > term proximity

Given **P**, the vector space kernel

$$\tilde{\kappa}(d_1, d_2) = \phi(d_1)\,\mathbf{P}\mathbf{P}'\phi(d_2)'$$

Related to an IR technique known as *query expansion*

We can view **PP'** as encoding semantic strength between terms

We will present some methods for obtaining semantic relationships

Bag of words > term proximity > explicit construction

Use a semantic network such as Wordnet

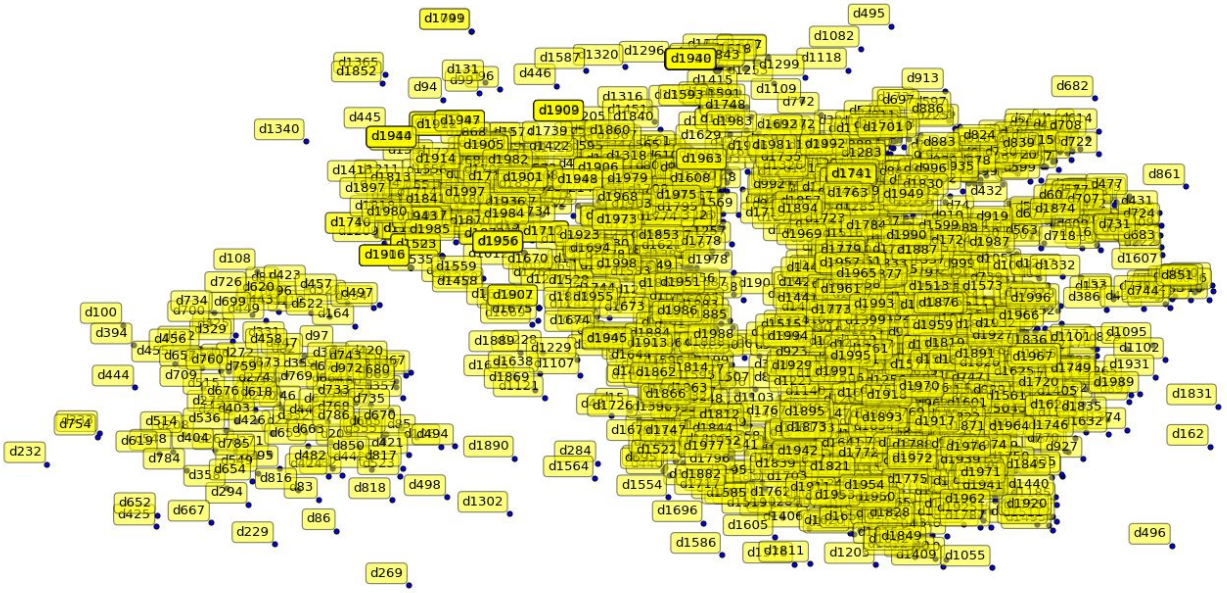
S: (n) **wife**, married woman
- *direct hyponym* / *full hyponym*
- *has instance*
- ***direct hypernym*** / *inherited hypernym* / *sister term*
  - S: (n) woman, adult female
  - S: (n) spouse, partner, married person, mate, better half
- *antonym*
- *derivationally related form*

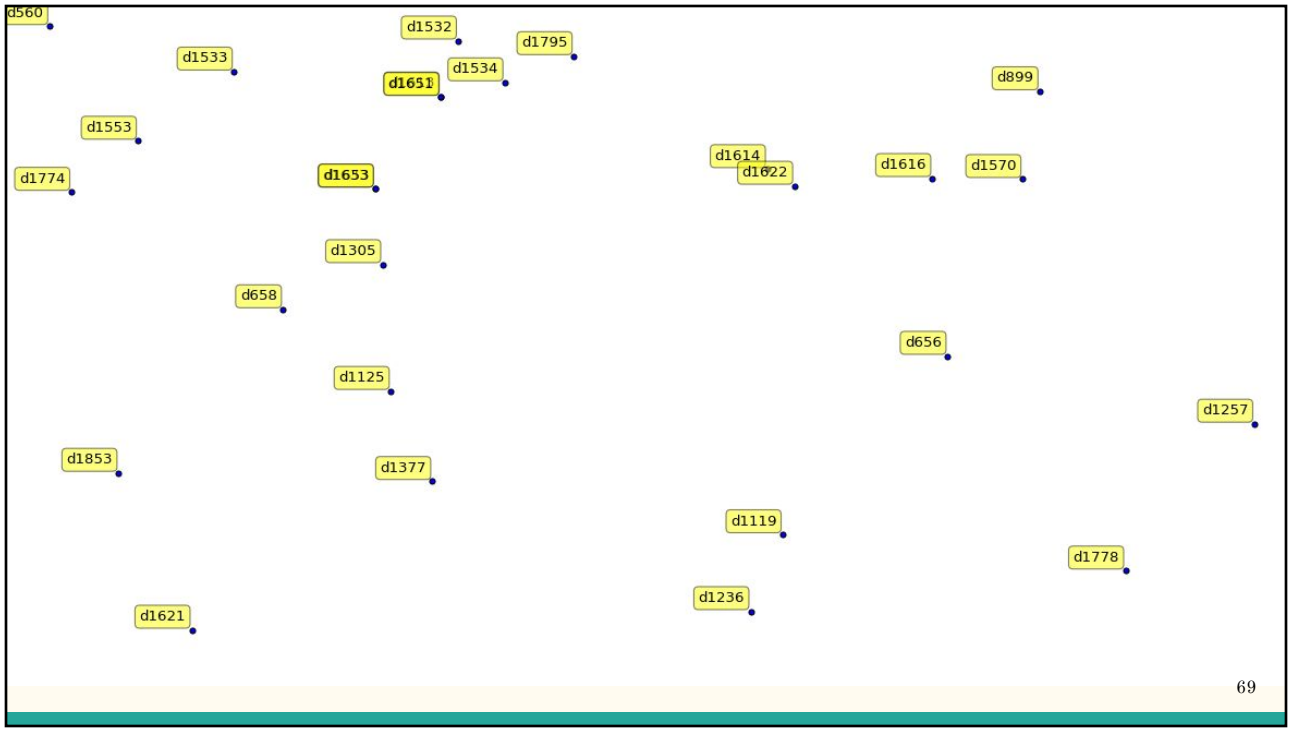
We handcraft matrix **P**

Bag of words > application

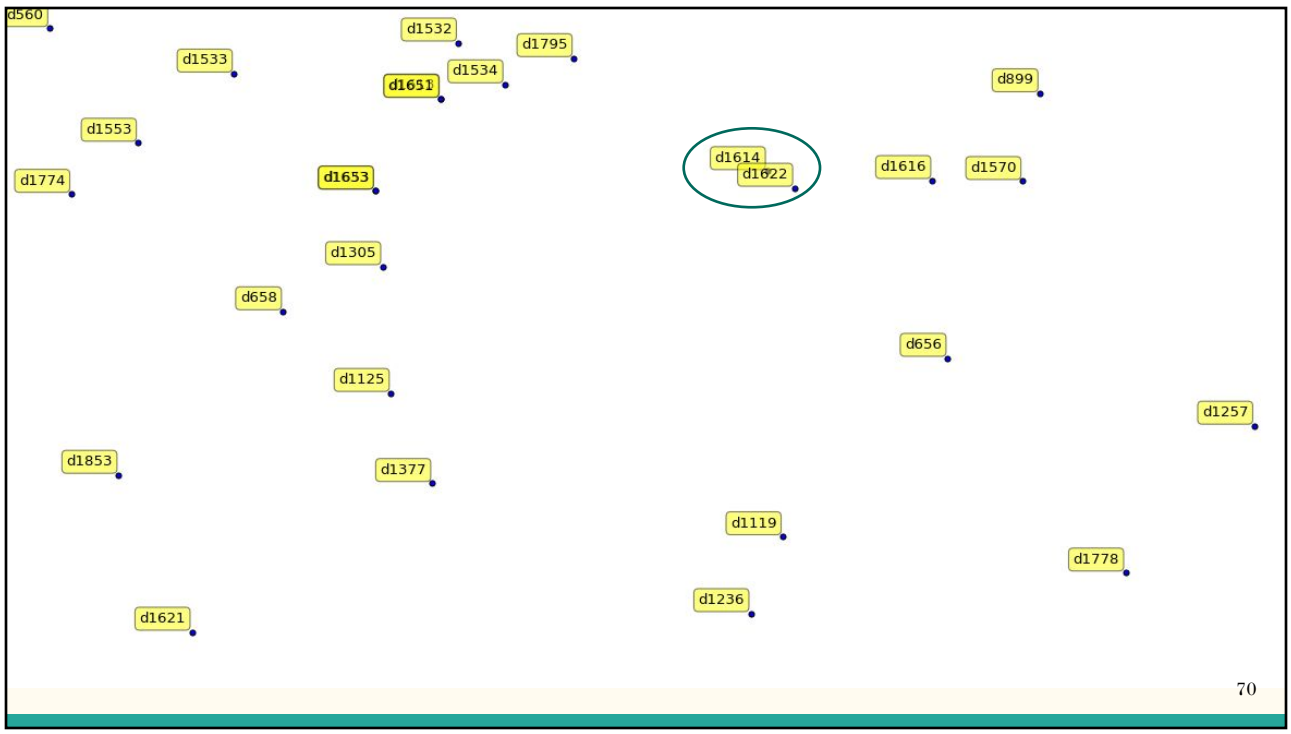
Document-based kernel

K = DD’

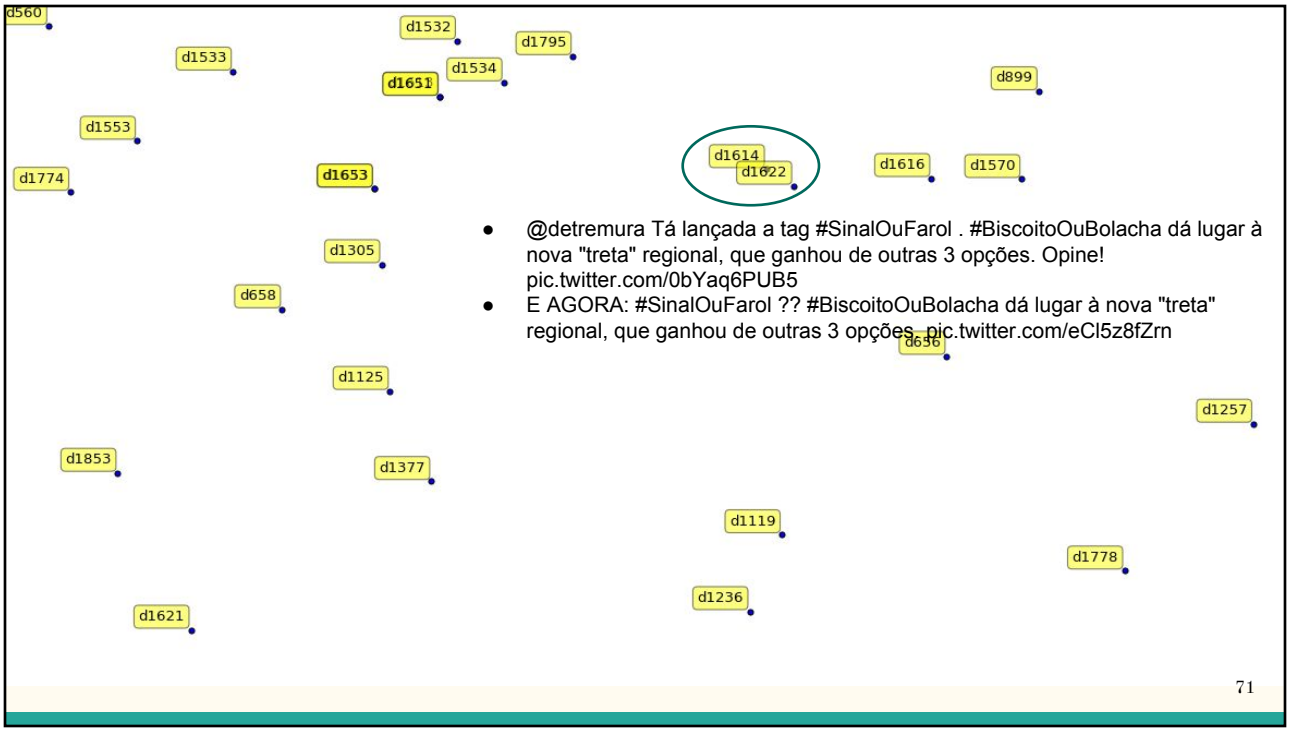



d560
d1532
d1795
d1533
d1534
d1651
d899
d1553
d1614
d1616
d1570
d1774
d1653
d1622
d1305
d658
d656
d1125
d1257
d1853
d1377
d1119
d1778
d1236
d1621

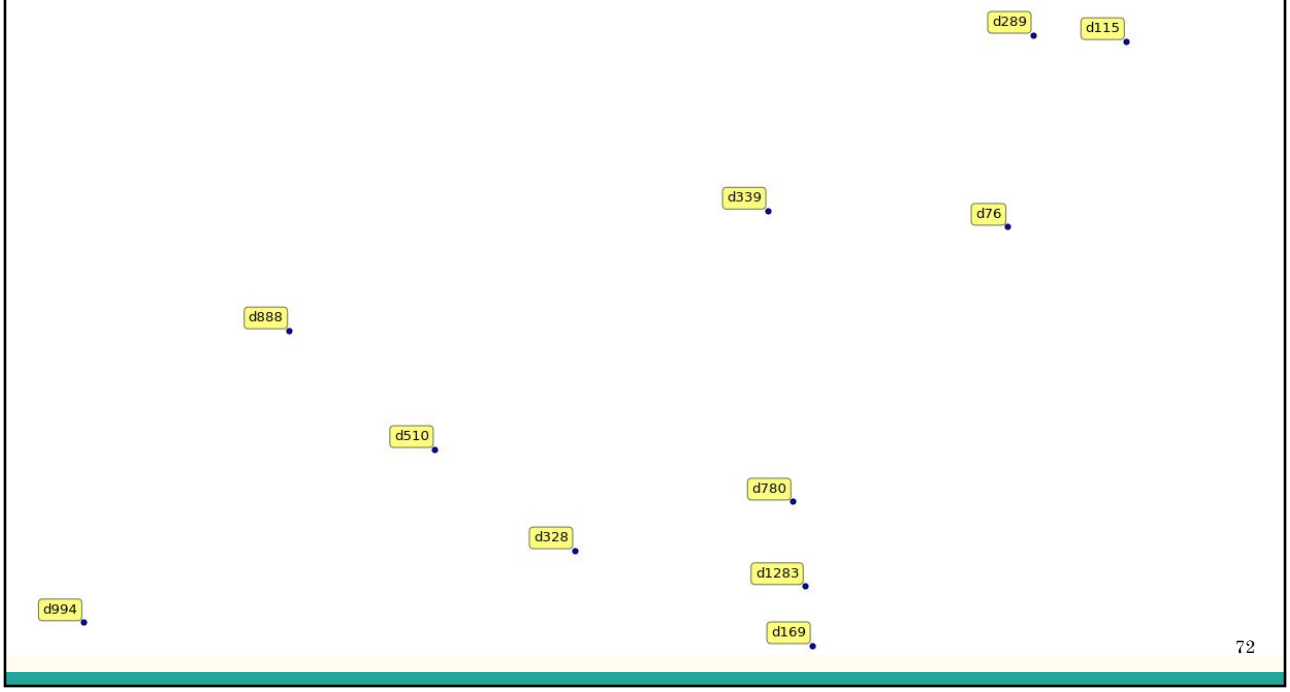
69

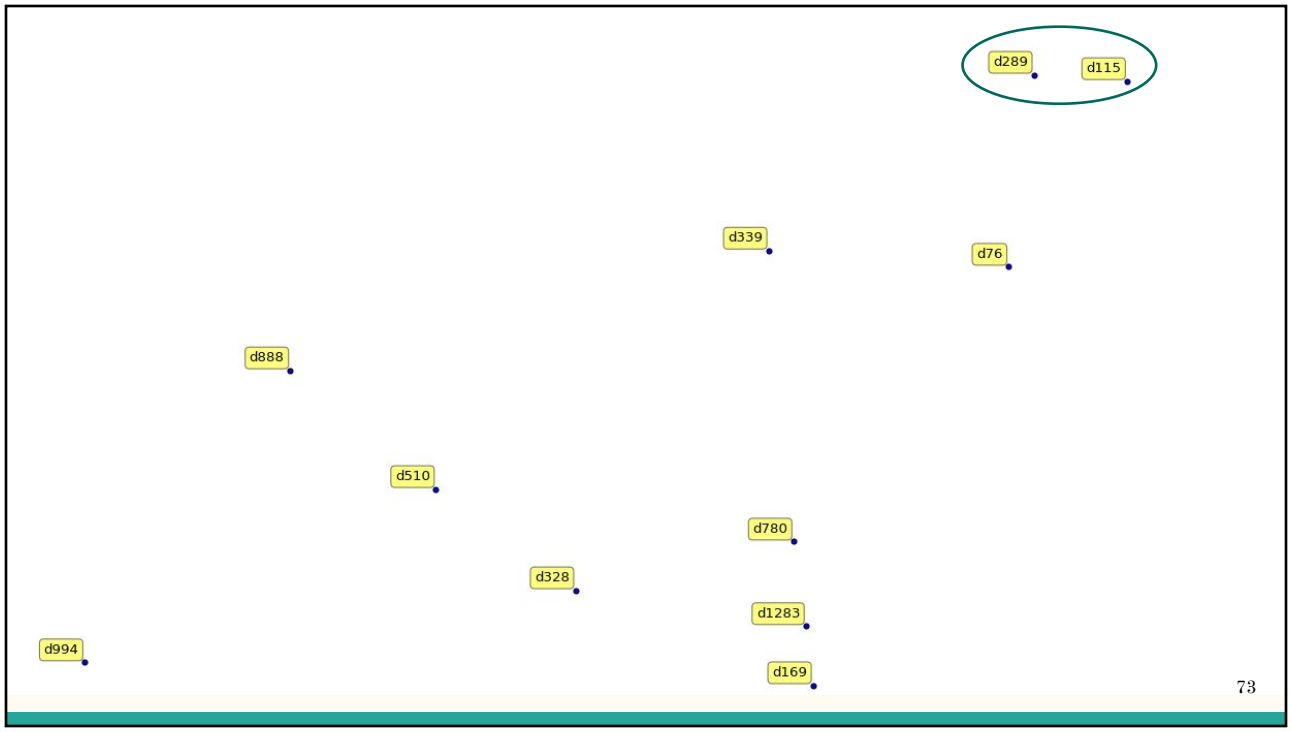

d560
d1532
d1795
d1533
d1534
d1651
d899
d1553
d1614
d1616
d1570
d1774
d1653
d1622
d1305
d658
d656
d1125
d1257
d1853
d1377
d1119
d1778
d1236
d1621

70

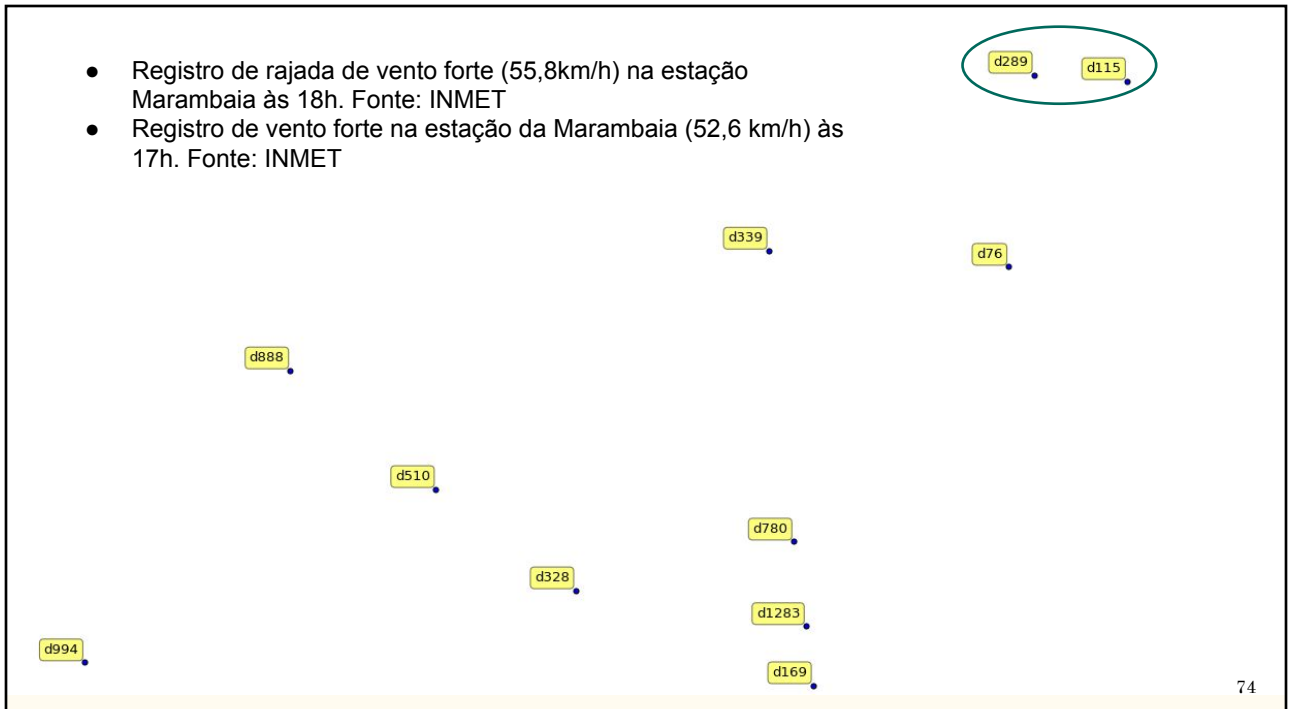


- @detremura Tá lançada a tag #SinalOuFarol . #BiscoitoOuBolacha dá lugar à nova "treta" regional, que ganhou de outras 3 opções. Opine! pic.twitter.com/0bYaq6PUB5
- E AGORA: #SinalOuFarol ?? #BiscoitoOuBolacha dá lugar à nova "treta" regional, que ganhou de outras 3 opções. pic.twitter.com/eCl5z8fZrn





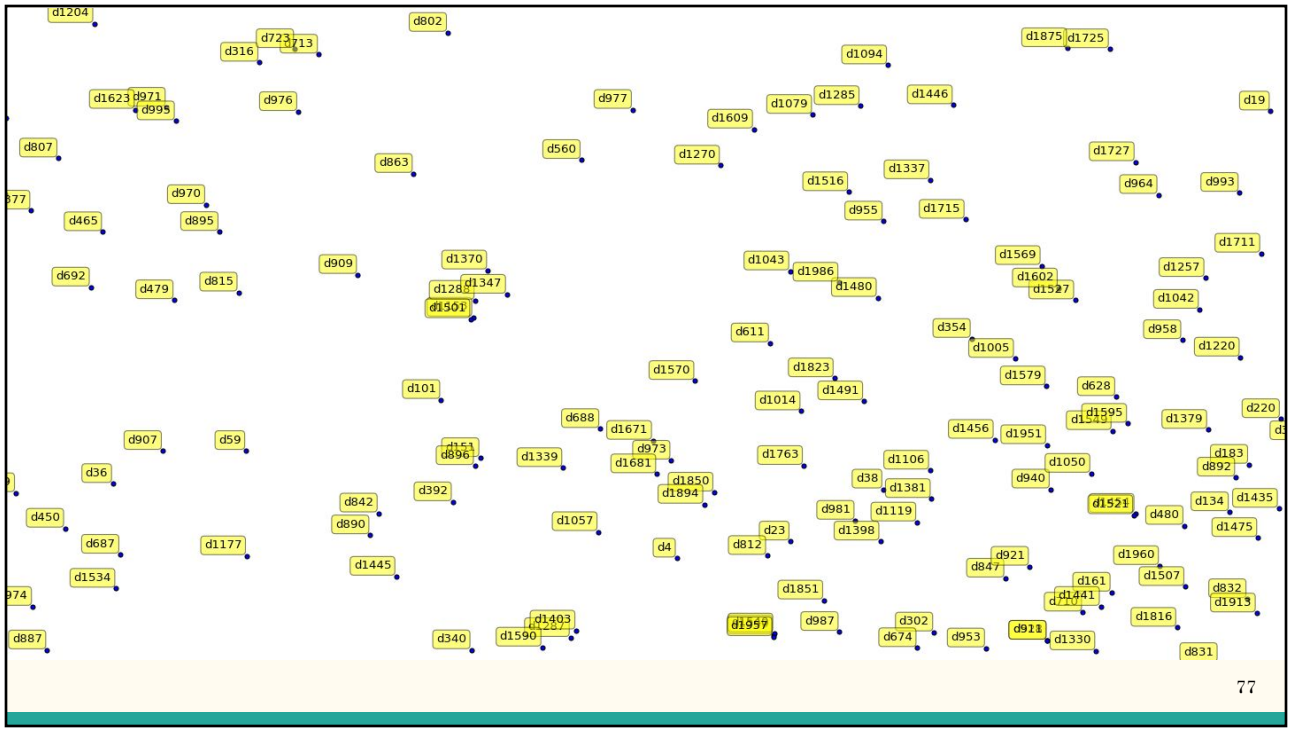
- Registro de rajada de vento forte (55,8km/h) na estação Marambaia às 18h. Fonte: INMET
- Registro de vento forte na estação da Marambaia (52,6 km/h) às 17h. Fonte: INMET

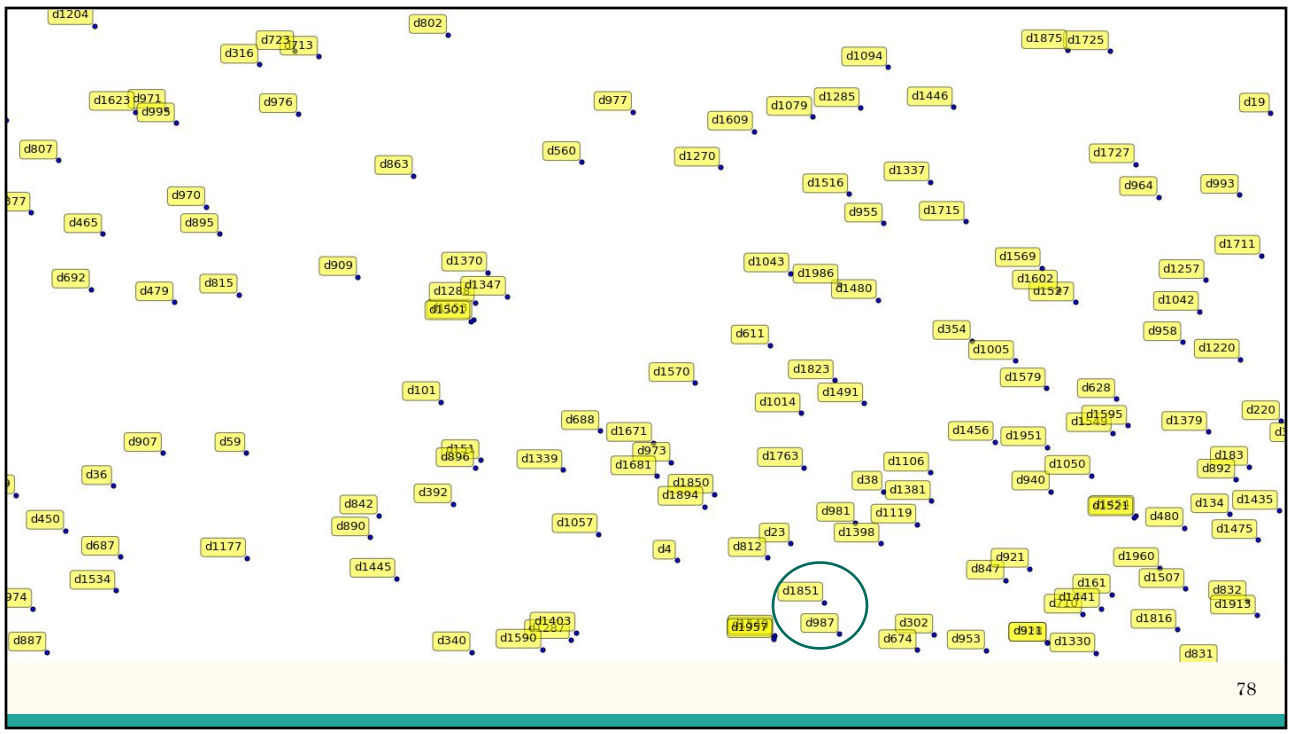


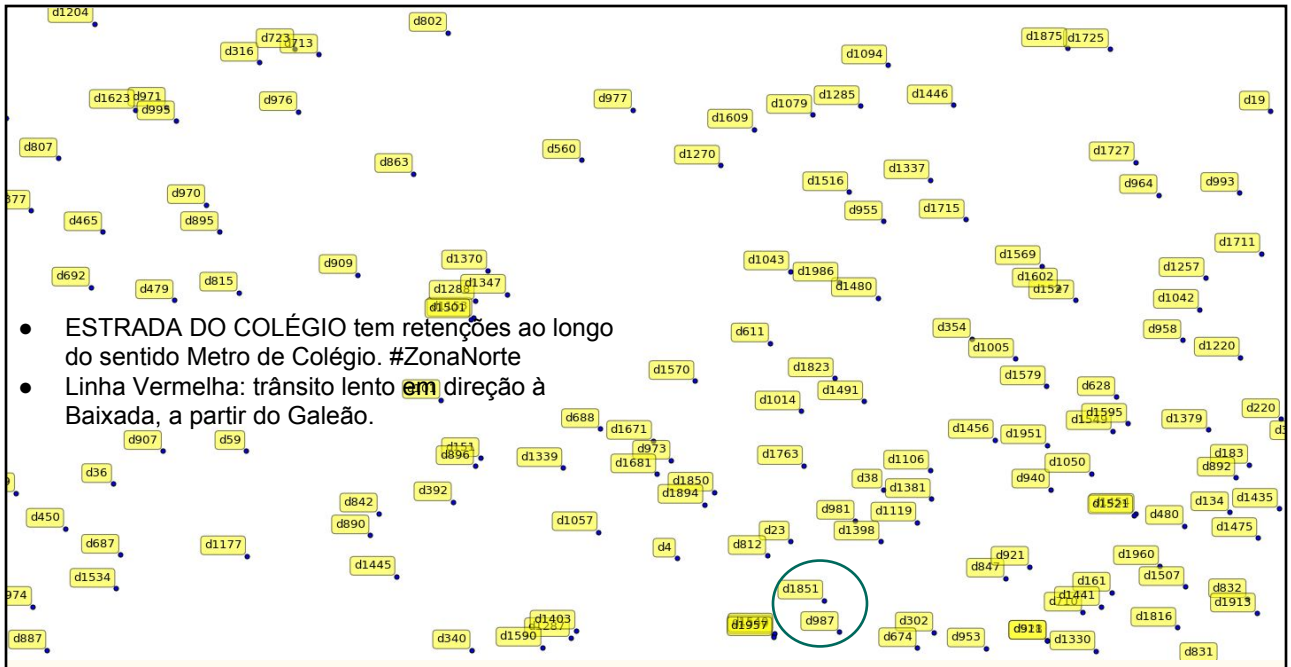
Bag of words > application

Document-based tf-idf kernel

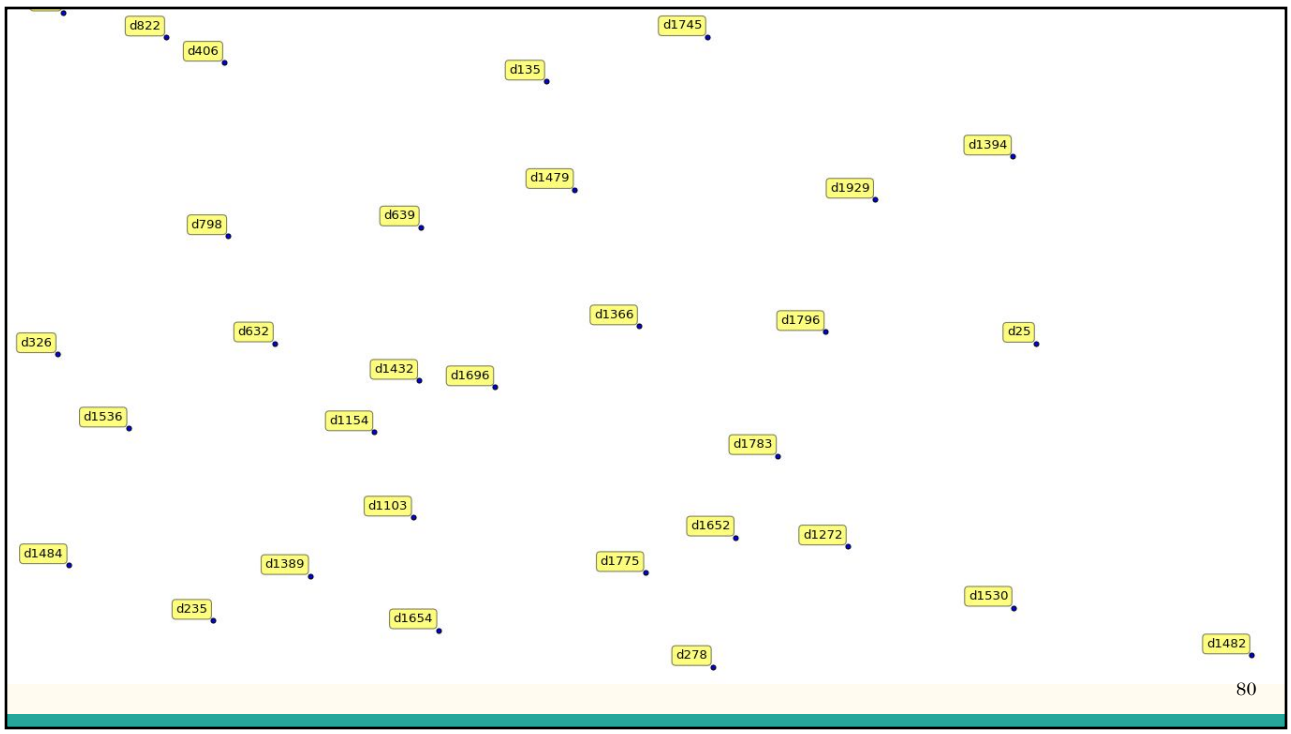
K = DRR'D'

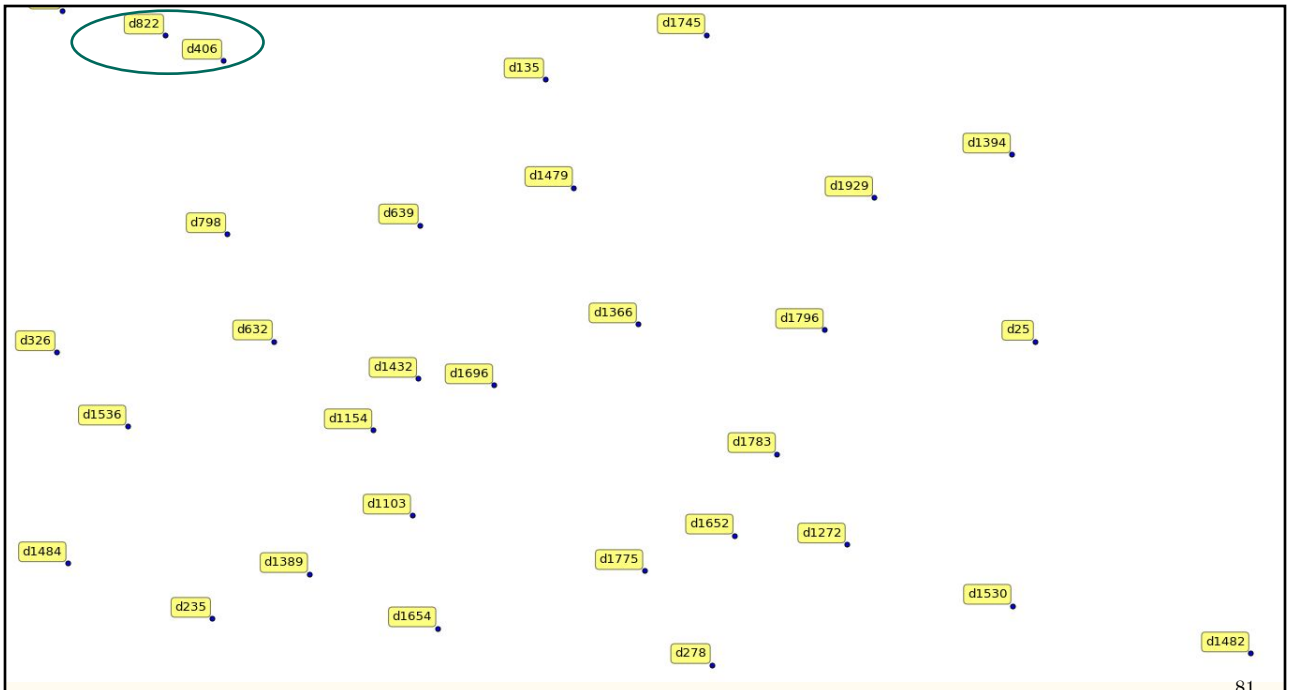

d1204
d802
d723
d316
d1875
d1725
d1094
d1623
d976
d977
d1079
d1285
d1446
d19
d1609
d807
d560
d1270
d1727
d863
d1516
d1337
d964
d993
d970
d955
d1715
d465
d895
d1711
d1370
d1043
d1569
d909
d1986
d1257
d692
d815
d1602
d1480
d1042
d1347
d611
d354
d958
d1005
d1220
d1570
d1823
d1579
d101
d1491
d628
d1014
d688
d1671
d1456
d1951
d907
d59
d973
d1763
d1106
d1339
d1681
d1050
d36
d1850
d38
d1381
d940
d842
d392
d1894
d981
d1119
d450
d1057
d1398
d1475
d687
d1177
d4
d812
d1960
d1445
d921
d847
d1507
d1534
d1851
d1441
d1403
d987
d302
d1816
d887
d1590
d340
d674
d953
d1330
d831

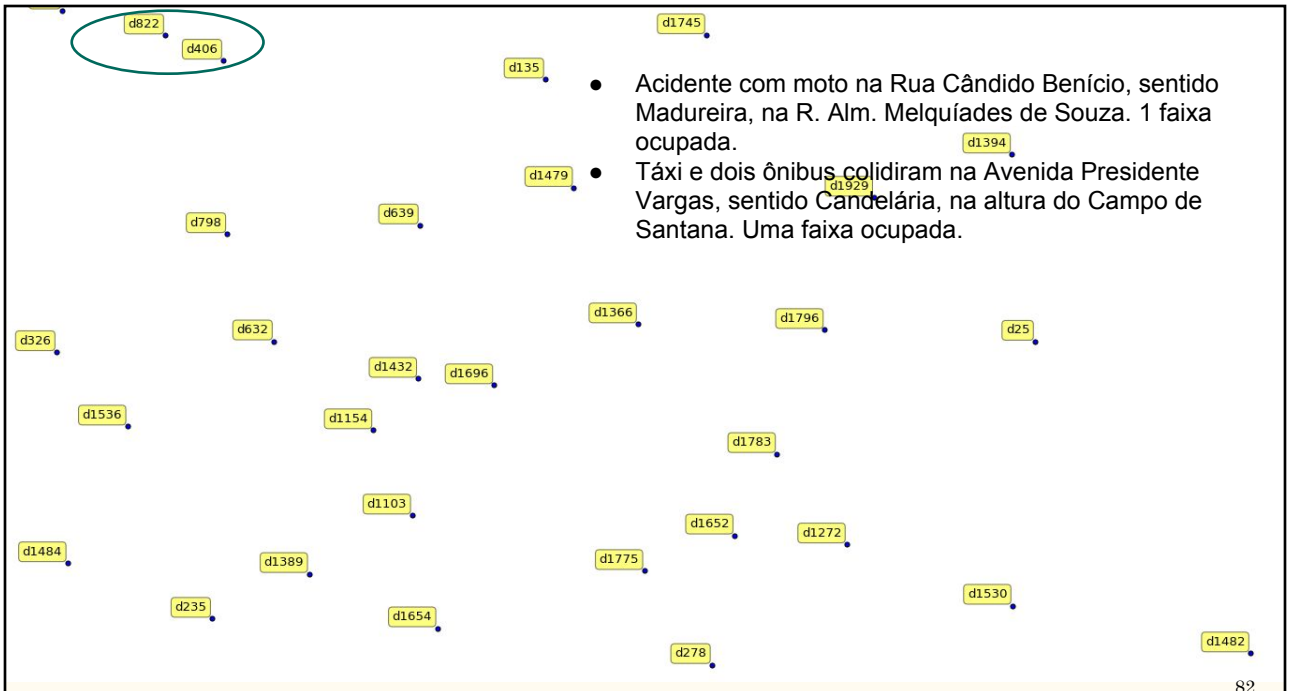

d1204
d802
d723
d316
d1875
d1725
d1094
d1623
d976
d977
d1079
d1285
d1446
d19
d1609
d807
d560
d1270
d1727
d863
d1516
d1337
d964
d993
d970
d955
d1715
d465
d895
d1711
d1370
d1043
d1569
d909
d1986
d1257
d692
d815
d1602
d1480
d1042
d1347
d611
d354
d958
d1005
d1220
d1570
d1823
d1579
d101
d1491
d628
d1014
d688
d1671
d1456
d1951
d907
d59
d973
d1763
d1106
d1339
d1681
d1050
d36
d1850
d38
d1381
d940
d842
d392
d1894
d981
d1119
d450
d1057
d1398
d1475
d687
d1177
d4
d812
d1960
d1445
d921
d847
d1507
d1534
d1851
d1441
d1403
d987
d302
d1816
d887
d1590
d340
d674
d953
d1330
d831



- ESTRADA DO COLÉGIO tem retenções ao longo do sentido Metro de Colégio. #ZonaNorte
- Linha Vermelha: trânsito lento em direção à Baixada, a partir do Galeão.





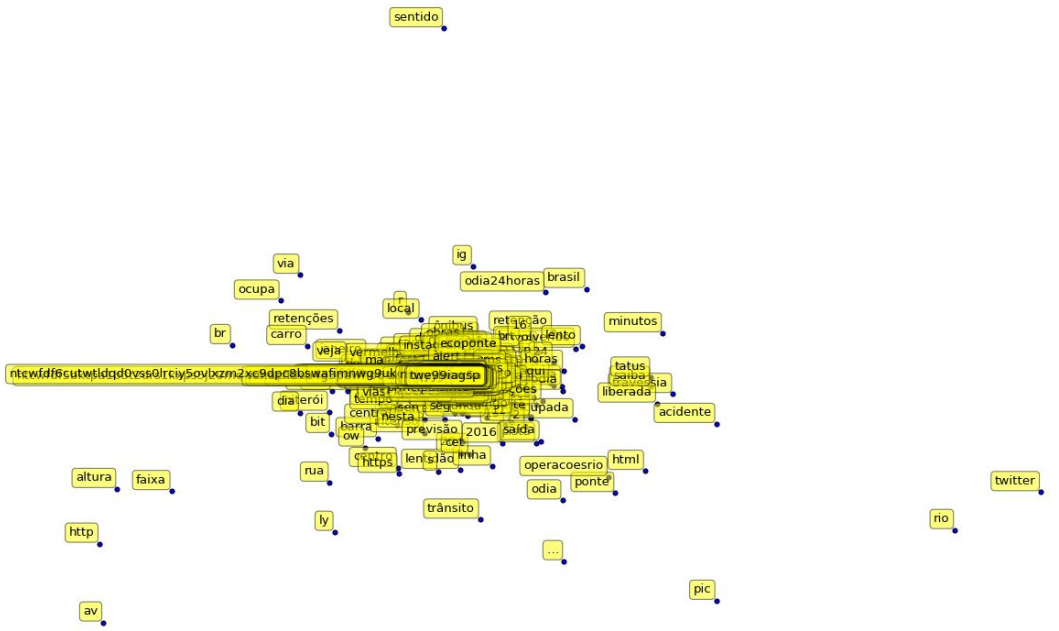


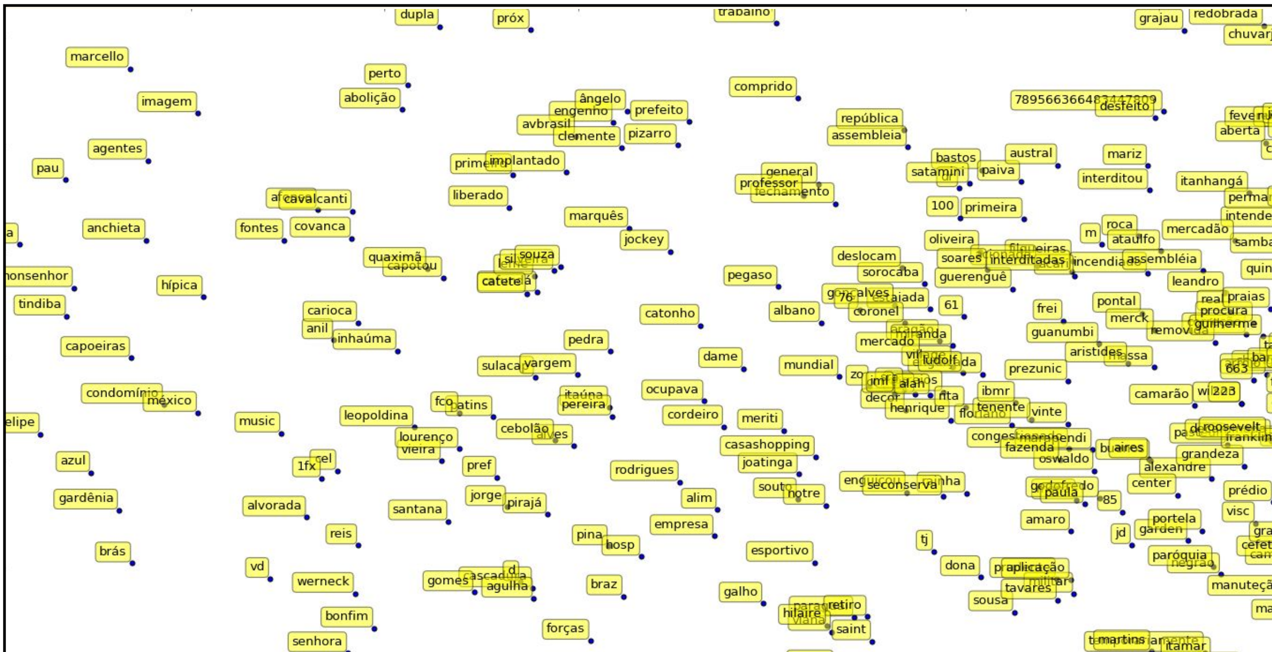
- Acidente com moto na Rua Cândido Benício, sentido Madureira, na R. Alm. Melquíades de Souza. 1 faixa ocupada.
- Táxi e dois ônibus colidiram na Avenida Presidente Vargas, sentido Candelária, na altura do Campo de Santana. Uma faixa ocupada.

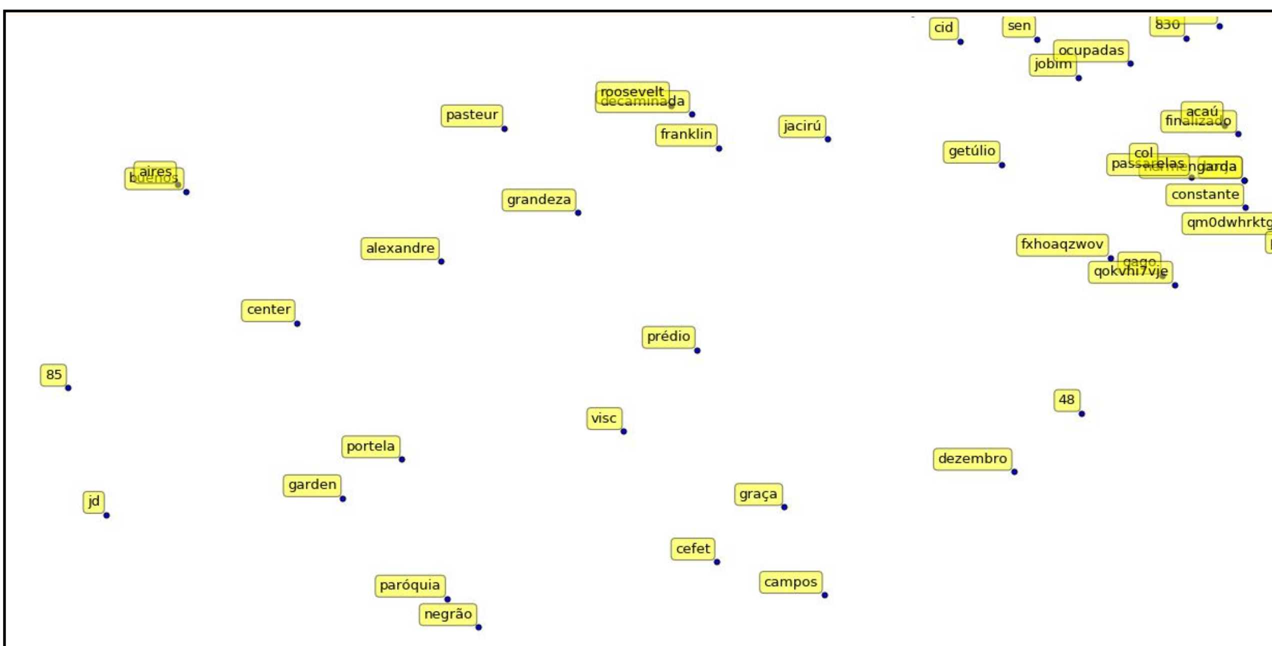
Bag of words > application

Term-based kernel

K = D'D




dupla
próx
trabalho
grajau
redobrada
chuvarj
marcello
perto
imagem
abolição
comprido
ângelo
engenho
prefeito
avbrasil
clemente
pizarro
república
assembleia
desfeito
agentes
implantado
bastos
austral
mariz
pau
general
satamini
paiva
interditou
itanhangá
professor
fechamento
cavalcanti
liberado
100
primeira
anchieta
fontes
covanca
marquês
jockey
oliveira
roca
ataulfo
mercadão
samba
deslocam
soares
interditadas
incendia
assembléia
quaximã
capotou
souza
sorocaba
guerenguê
quinta
hípica
catete
pegaso
leandro
tindiba
carioca
praias
pontal
procura
61
frei
merck
guilherme
catonho
albano
coronel
anil
inhaúma
guanumbi
capoeiras
pedra
mercado
dame
mundial
aristides
sulaca
vargem
ludolf
prezunic
663
condomínio
méxico
ocupava
itaúna
ibmr
camarão
223
patins
pereira
rita
music
leopoldina
cordeiro
meriti
henrique
tenente
vinte
cebolão
roosevelt
lourenço
vieira
fazenda
casashopping
aires
azul
cel
1fx
pref
rodrigues
joatinga
oswaldo
grandeza
alexandre
gardênia
jorge
pirajá
alim
notre
seconserva
paula
center
prédio
alvorada
santana
85
visc
empresa
amaro
portela
garden
jd
tj
pina
hosp
reis
brás
esportivo
vd
paróquia
werneck
gomes
agulha
braz
dona
aplicação
tavares
manuteção
galho
retiro
hilaire
sousa
bonfim
saint
forças
senhora
martins
itamar


cid
sen
830
ocupadas
jobim
roosevelt
decaminada
pasteur
franklin
jacirú
acaú
finalizado
getúlio
col
passarelas
garda
aires
buenos
grandeza
constante
qm0dwhrktg
alexandre
fxhoaqzwov
gago
qokvhi7vje
center
prédio
85
48
visc
portela
dezembro
garden
jd
graça
cefet
paróquia
campos
negrão

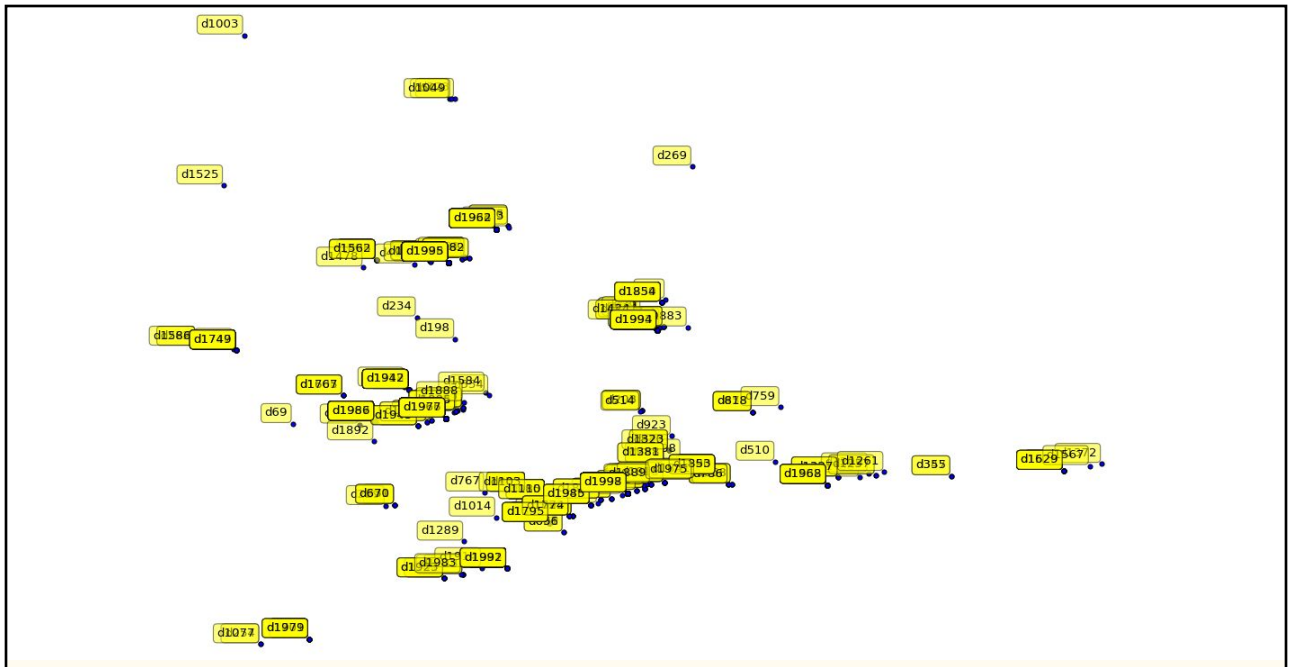
Bag of words > application

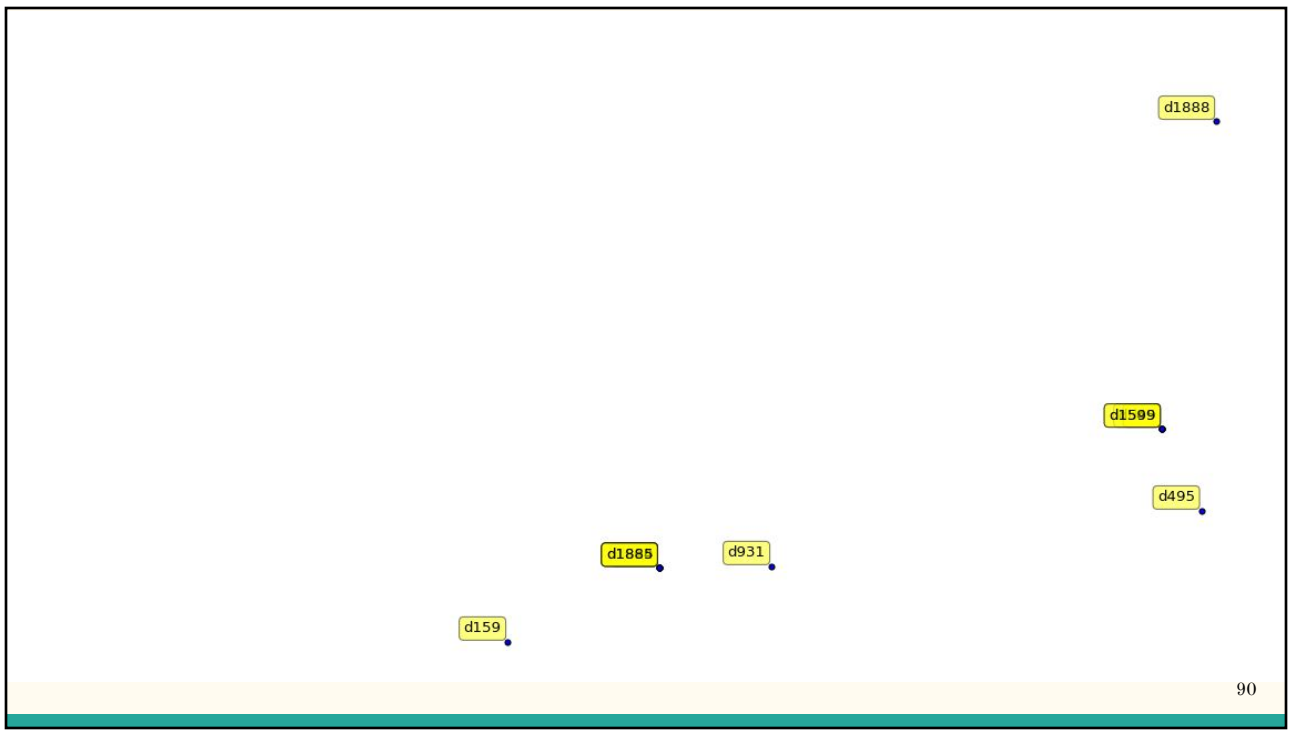
Document-based kernel with proximity

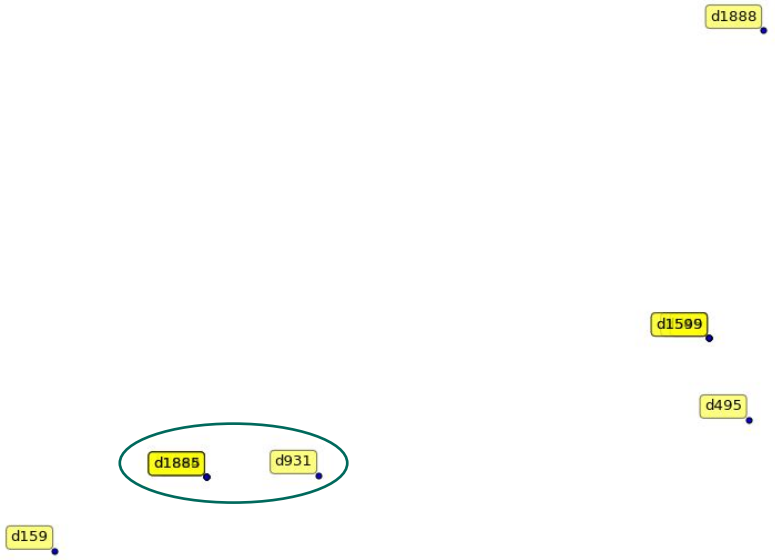
K = DPP'D'

Bag of words > application > synonyms

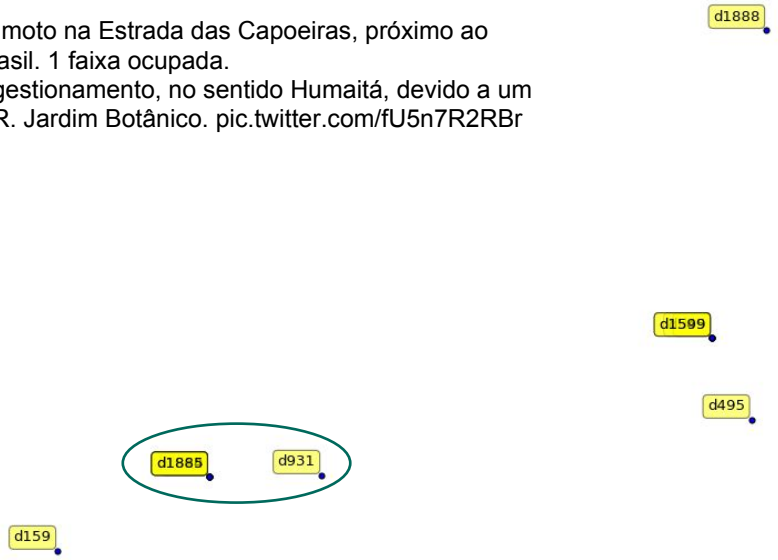
- interdicao - fechada
- colisao - engavetamento - choque - capotamento
- queda - atropelamento
- enguicado - pane
- trafego - transito
- chuva - alagamento - bolsao d agua
- nas duas direcoes - nos dois sentidos
- direcao - sentido
- altura - perto de - proximo de
- neste momento - agora - em 1 hora
- carro - onibus - motocicleta
- bom - livre
- intenso - comretencoes
- lento - lentidao - parado - congestionamento - retencao
- causou - gerou - ocasionou - complica


d1003
d1049
d269
d1525
d1962
d1562
d1995
d1854
d1994
d234
d198
d1749
d1942
d1767
d1986
d1977
d69
d1892
d514
d818
d923
d1381
d510
d1968
d355
d1998
d1985
d1795
d1014
d1289
d1992
d1983
d1077
d1979


d1888
d1599
d495
d1885
d931
d159



- Atropelamento envolvendo moto na Estrada das Capoeiras, próximo ao McDonald's, sentido Av. Brasil. 1 faixa ocupada.
- PRAÇA SIBÉLIUS: há congestionamento, no sentido Humaitá, devido a um acidente mais a frente, na R. Jardim Botânico. pic.twitter.com/fU5n7R2RBr



Pareamento de palavras

Similar to VSMs

Sentence 1: New York is very crowded.

Sentence 2: Washington Post is a leading newspaper.

	new	york	is	very	crowded	washington	post	...
S1	1	1	1	1	1			
S2						1	1	

Pareamento de palavras

Similar to VSMs

Sentence 1: New York is very crowded.

Sentence 2: Washington Post is a leading newspaper.

	new	...	new york	york is	is very	very crowded	washington post	post is	is a	a leading	leading newspaper
S1	1		1	1	1	1					
S2							1	1	1	1	1

Pareamento de palavras

Similar to VSMs

Sentence 1: New York is very crowded.

Sentence 2: Washington Post is a leading newspaper.

	new	...	new york	york is	is very	very crowded	washington post	post is	is a	a leading	leading newspaper
S1	1		1	1	1	1					
S2							1	1	1	1	1

Pareamento de palavras

Bi-gram: Santa Maria, Best Buy, Universidade Federal

Tri-gram: Rio de Janeiro, Juiz de Fora, New York Times, Pontifícia Universidade Católica

4-gram: United States of America, Folha de São Paulo

Machine learning

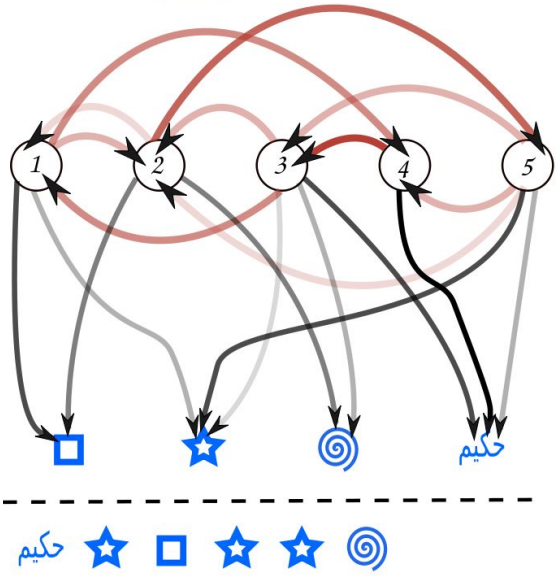
Naïve Bayes

Logistic Regression

Hidden Markov Models (HMM's)

Long Short Term Memory Networks (LSTM's)

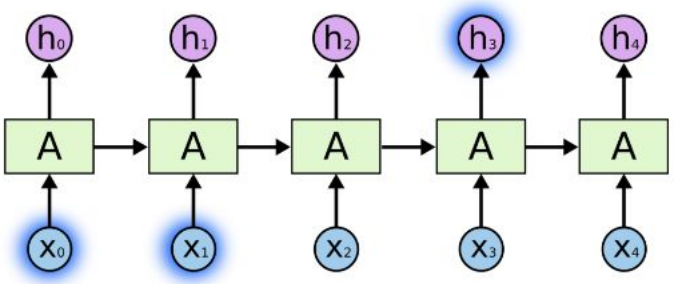
Machine learning > HMM's



Machine learning > LSTM's

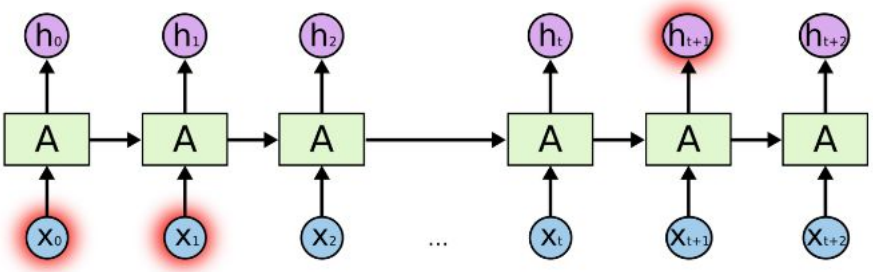
Long Short Term Memory Networks are a special kind of Recurrent Neural Networks (RNN's)

The idea of RNN's is to use past information to the present task



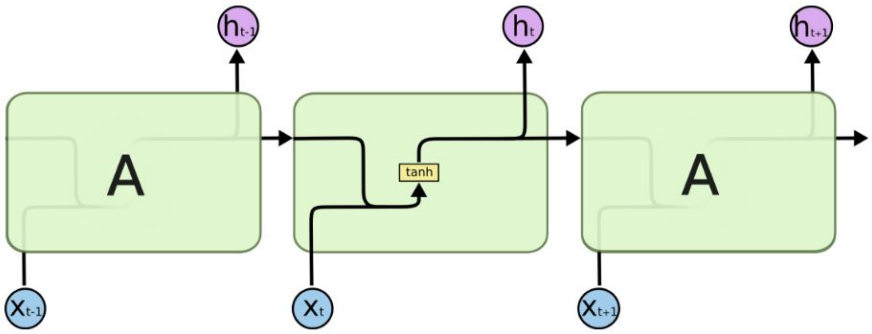
Machine learning > LSTM's

However RNN's do not deal with long distances



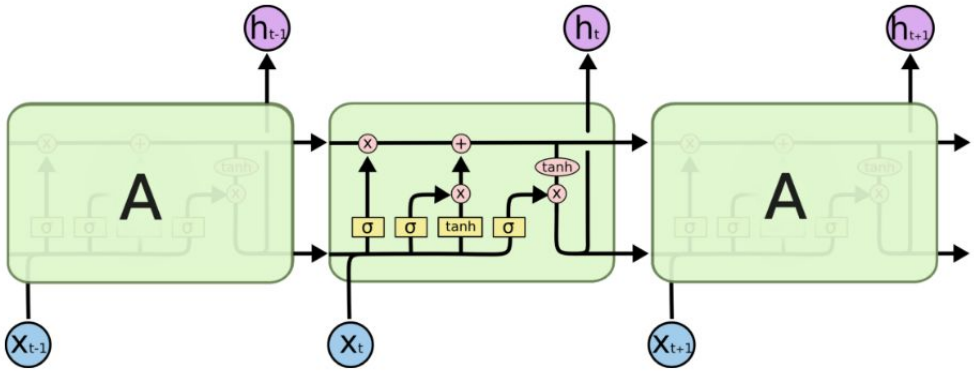
Machine learning > LSTM's

The repeating module in a RNN has one layer



Machine learning > LSTM's

The repeating module in a LSTM has four layers



word2vec

word2vec is a group of related models used to produce word embeddings

These models are used to reconstruct linguistic contexts of words

Takes as input a large corpus of text

Produces a vector space

For each unique word is assigned a vector

It was created at Google by a team of researchers led by Tomas Mikolov

word2vec

Simple neural network

One hidden layer

Train the neural network to execute a task

Take only the hidden layer

word2vec

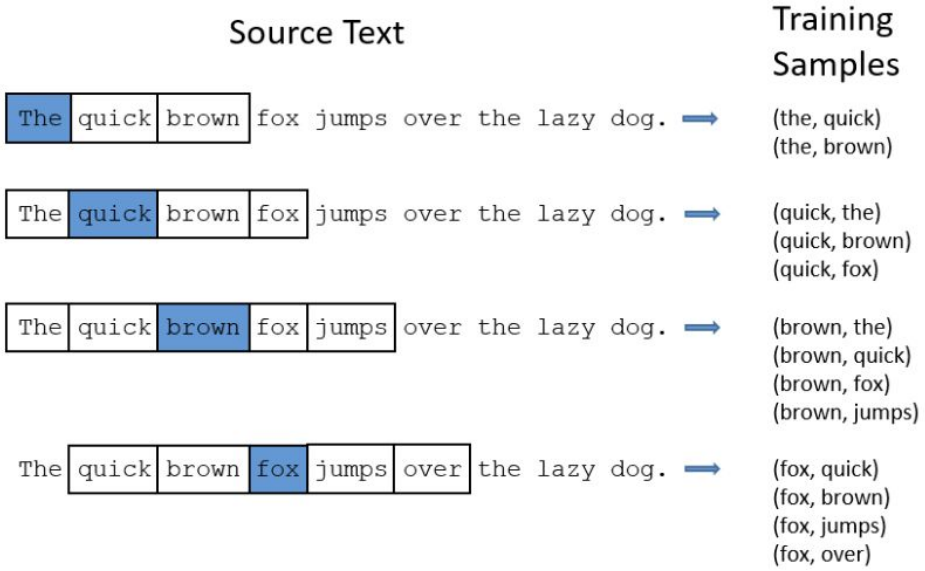
Given a specific word in the middle of a sentence

Pick up a word at random

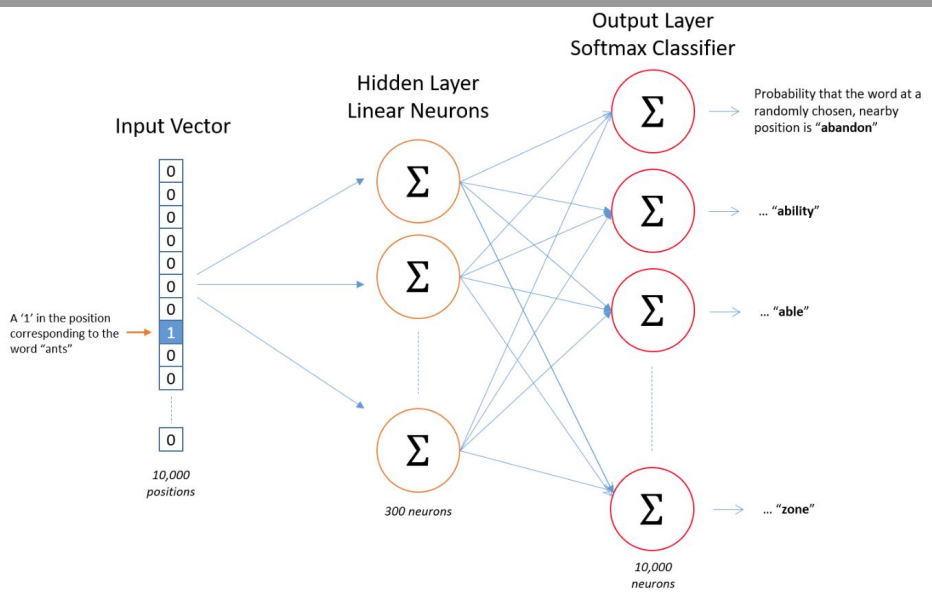
The network will tell the probability for every word in the vocabulary of being the "nearby word"

Input word *Soviet* → output probability of *Union* and *Russia* is higher than *watermelon* and *bottle*

word2vec > dataset



word2vec > model



word2vec > training

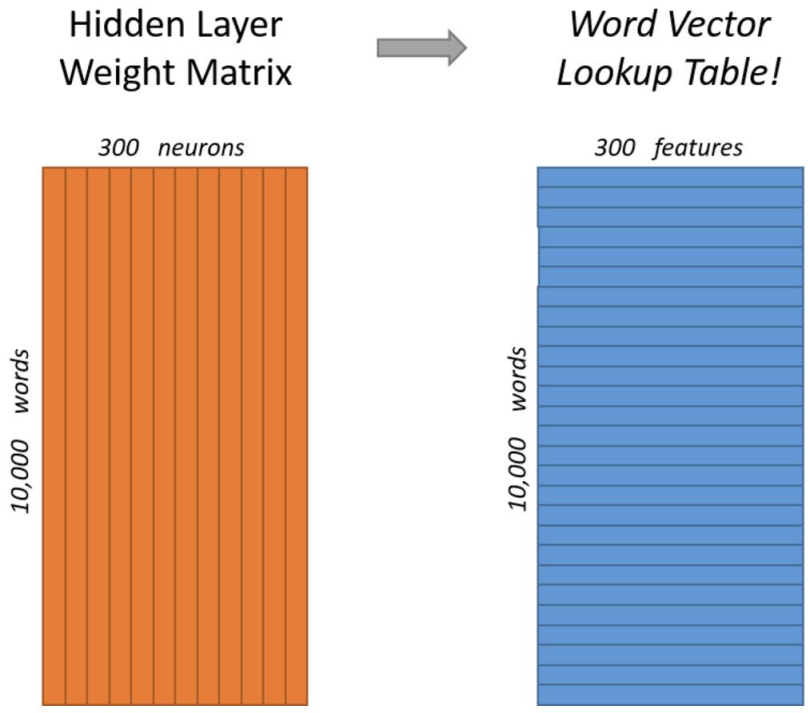
Word pairs

(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

Input is a one-hot vector

True output is a one-hot vector

Network output is a probability distribution



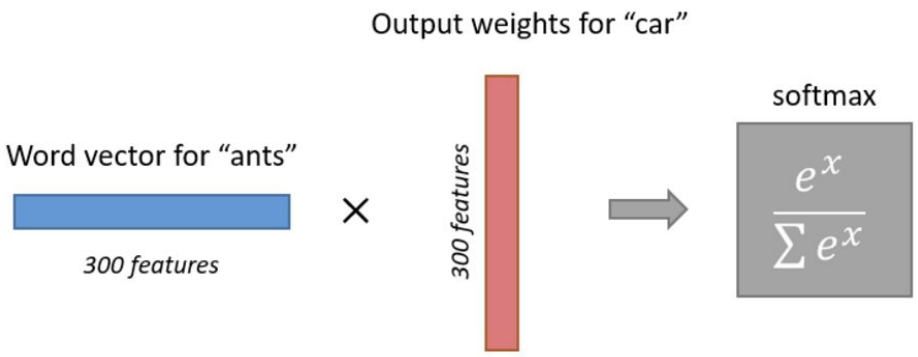
word2vec > hidden layer

The one-hot vector is almost all zeros

What is the effect of that?

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Word embedding > word2vec > output layer



Word embedding > word2vec > intuition

Two different words with similar contexts

The model needs to output very similar results

Two word vectors are very similar

Synonyms like intelligent and smart will have similar contexts

Related words like engine and transmission will have similar contexts

Ant and ants will have similar contexts

Word embedding > word2vec > modifications

Skip-gram model for word2vec is a huge neural network

With 300 components and 10,000 words

300 x 10,000 = 3 million weights in each layer

Running gradient descent will be slow

Then the authors of word2vec proposed three innovations in their second paper

Word embedding > word2vec > innovations

Treating common word pairs as single words

Subsampling frequent words to decrease the number of training samples

Using a technique called *Negative Sampling* to update only a small percentage of model weights

Thanks for your attention :)

Jonatas Grosman jgrosman@inf.puc-rio.br
Luiz Schirmer lschirmer@inf.puc-rio.br
William Fernandes wfernandes@inf.puc-rio.br