MCZA017-13
Processamento de Linguagem Natural

NLTK através de exemplos: Sumarização

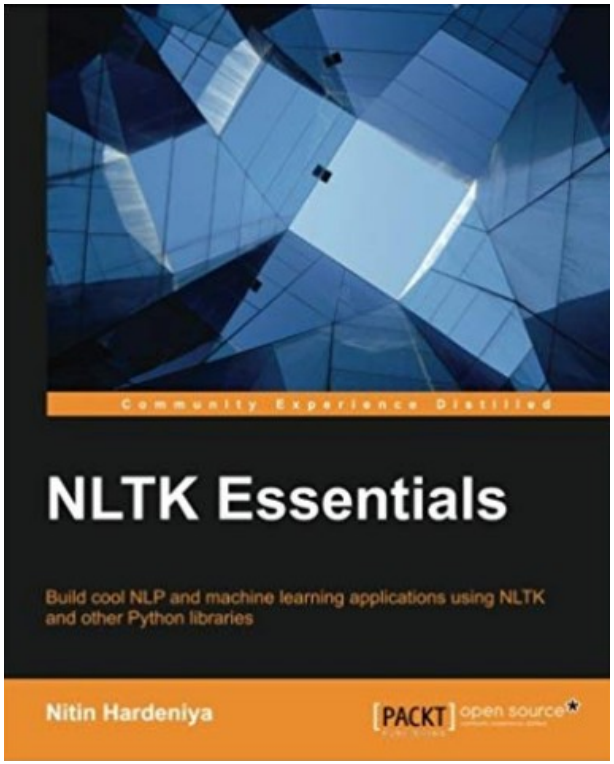
Prof. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

2Q-2019

Bibliografia (introdutória)

Nitin Hardeniya.

NLTK Essentials: Build cool NLP and machine learning applications using NLTK and other Python libraries.
PACKT.



Texto de exemplo

Federal University of ABC (Portuguese: Universidade Federal do ABC, UFABC) is a Brazilian institution of higher learning based in Santo André, with operations in several municipalities in the ABC region, all in the state of São Paulo. The chairman of the committee that formulated the proposal of the university was Luiz Bevilacqua, who became its second rector.

UFABC is the only federal university in Brazil with 100% of its professors holding Ph.D.s and, for the second consecutive year in 2011, emerged as the only university in Brazil with impact factor in scientific publications above the world average according to SCImago Institutions Rankings.

In 2004, the Ministry of Education sent to Congress the Bill 3962/2004, which created the Federal University of ABC. This law was enacted by the president and published in the Official Gazette of 27 July 2005, under No. 11,145 and dated 26 July 2005.

nltk.sent_tokenize(*string*)

```
0 : >>Federal University of ABC (Portuguese: Universidade Federal do ABC, UFABC) is
a Brazilian institution of higher learning based in Santo André, with
operations in several municipalities in the ABC region, all in the state of
São Paulo.<<

1 : >>The chairman of the committee that formulated the proposal of the
university was Luiz Bevilacqua, who became its second rector.<<

2 : >>UFABC is the only federal university in Brazil with 100% of its professors
holding Ph.D.s and, for the second consecutive year in 2011, emerged as the
only university in Brazil with impact factor in scientific publications above
the world average according to SCImago Institutions Rankings.<<

3 : >>In 2004, the Ministry of Education sent to Congress the Bill 3962/2004, which
created the Federal University of ABC.<<

4 : >>This law was enacted by the president
and published in the Official Gazette of 27 July 2005, under No.<<

5 : >>11,145 and
dated 26 July 2005.<<
```

nltk.word_tokenize(*sentence*)

sentence:
Federal University of ABC (Portuguese: Universidade Federal do ABC, UFABC) is a Brazilian institution of higher learning based in Santo André, with operations in several municipalities in the ABC region, all in the state of São Paulo.

```
['Federal', 'University', 'of', 'ABC', '(', 'Portuguese', ':', 'Universidade',
'Federal', 'do', 'ABC', ',', 'UFABC', ')', 'is', 'a', 'Brazilian', 'institution',
'of', 'higher', 'learning', 'based', 'in', 'Santo', 'André', ',', 'with',
'operations', 'in', 'several', 'municipalities', 'in', 'the', 'ABC',
'region', ',', 'all', 'in', 'the', 'state', 'of', 'São', 'Paulo', '.']
```

sentence:
The chairman of the committee that formulated the proposal of the university was Luiz Bevilacqua, who became its second rector.

```
['The', 'chairman', 'of', 'the', 'committee', 'that', 'formulated', 'the',
'proposal', 'of', 'the', 'university', 'was', 'Luiz', 'Bevilacqua', ',',
'who', 'became', 'its', 'second', 'rector', '.']
```

nltk.pos_tag(*tokens*)

tokens:

```
['Federal', 'University', 'of', 'ABC', '(', 'Portuguese', ':', 'Universidade',
'Federal', 'do', 'ABC', ',', 'UFABC', ')', 'is', 'a', 'Brazilian', 'institution',
'of', 'higher', 'learning', 'based', 'in', 'Santo', 'André', ',', 'with',
'operations', 'in', 'several', 'municipalities', 'in', 'the', 'ABC',
'region', ',', 'all', 'in', 'the', 'state', 'of', 'São', 'Paulo', '.']
```

```
[('Federal', 'NNP'),
('University', 'NNP'),
('of', 'IN'),
('ABC', 'NNP'),
('(', '('),
('Portuguese', 'JJ'),
(':', ':'),
('Universidade', 'JJ'),
('Federal', 'NNP'),
('do', 'VBP'),
('ABC', 'NNP'),
(',', ','),
('UFABC', 'NNP'),
(')', ')'),
('is', 'VBZ'),
```

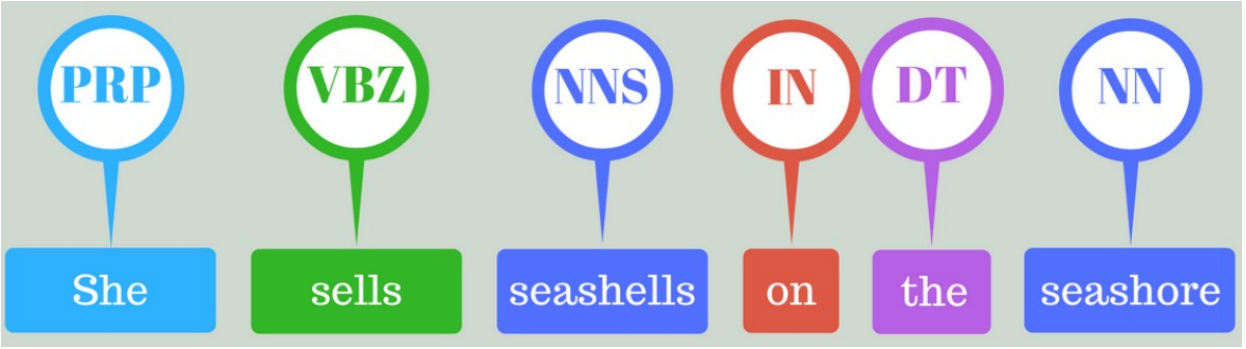
```
('a', 'DT'),
('Brazilian', 'JJ'),
('institution', 'NN'),
('of', 'IN'),
('higher', 'JJR'),
('learning', 'NN'),
('based', 'VBN'),
('in', 'IN'),
('Santo', 'NNP'),
('André', 'NNP'),
(',', ','),
('with', 'IN'),
('operations', 'NNS'),
('in', 'IN'),
('several', 'JJ'),
```

```
('municipalities', 'NNS'),
('in', 'IN'),
('the', 'DT'),
('ABC', 'NNP'),
('region', 'NN'),
(',', ','),
('all', 'DT'),
('in', 'IN'),
('the', 'DT'),
('state', 'NN'),
('of', 'IN'),
('São', 'NNP'),
('Paulo', 'NNP'),
('.', '.')]
```

pos_tag: Part of speech tagging

A POS-tagging é um dos principais componentes de quase todas as análises de PLN.

A tarefa de etiquetagem de POS implica simplesmente em **rotular palavras com sua característica de linguagem** (i.e., substantivo, verbo, adjetivo, advérbio, pronome).



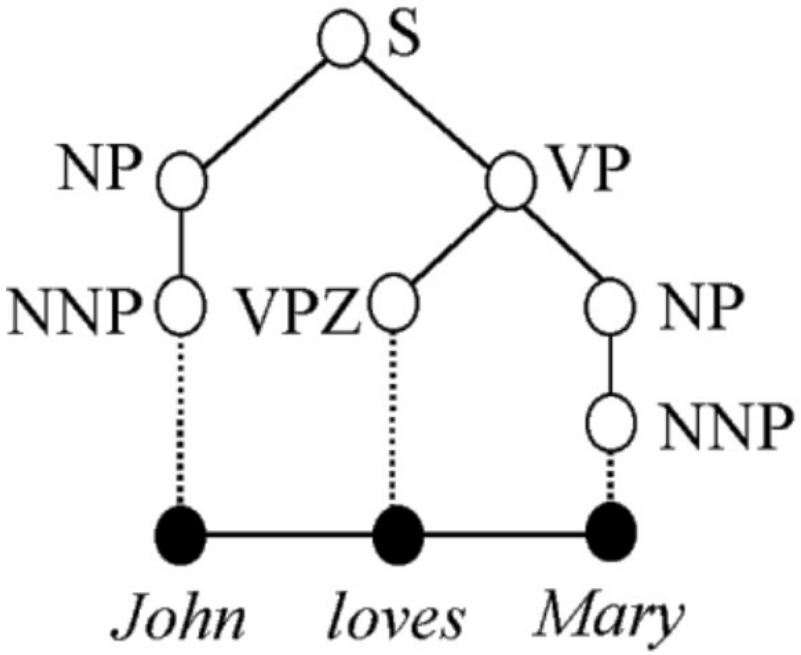
O conjunto de rótulos mais popular é o treebank (U. Penn)

pos_tag: Part of speech tagging

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

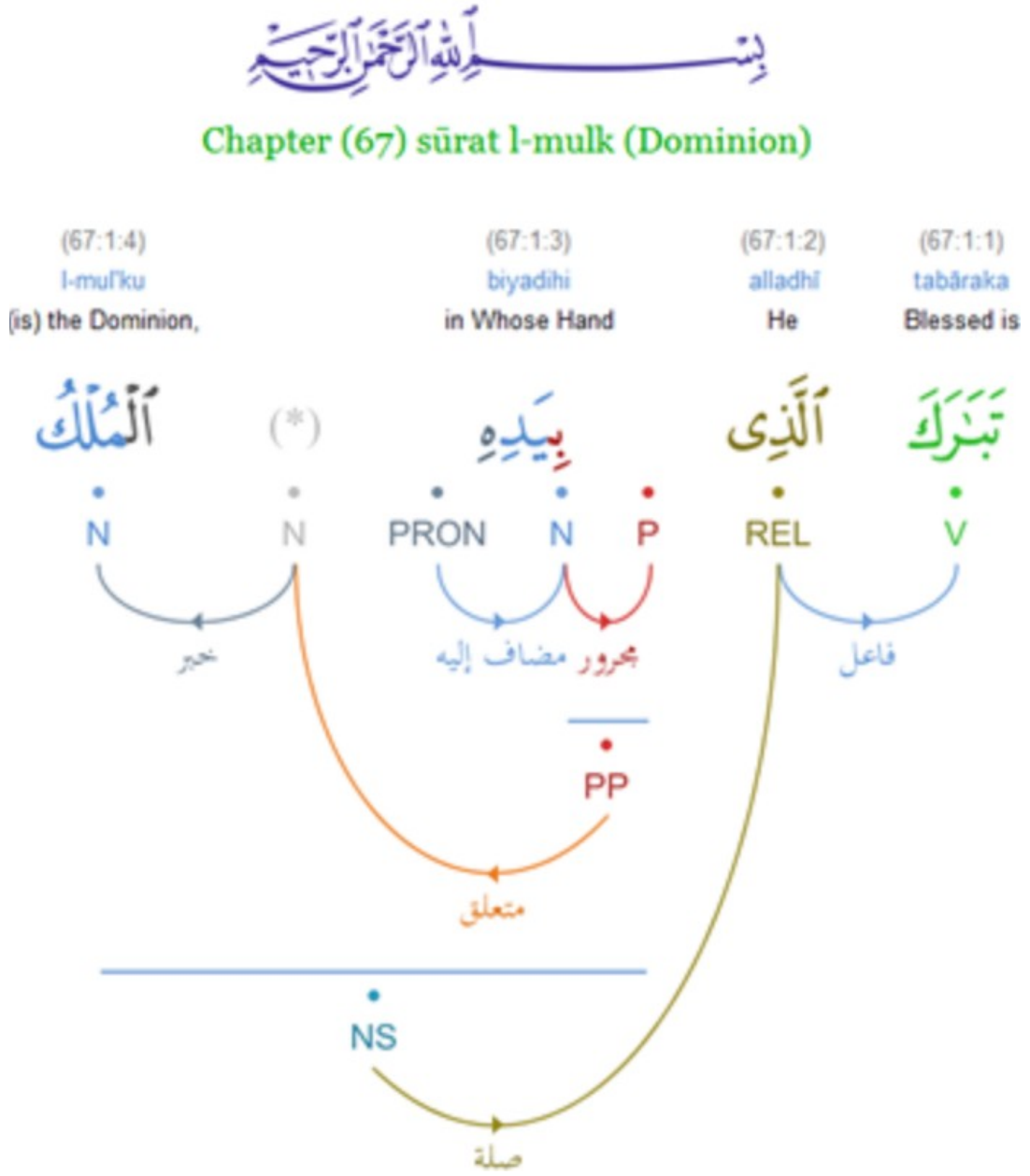
Treebank

In linguistics, a **treebank** is a parsed text corpus that annotates syntactic or semantic sentence structure. The construction of parsed corpora in the early 1990s revolutionized computational linguistics, which benefitted from large-scale empirical data.[1] The exploitation of treebank data has been important ever since the first large-scale treebank, The Penn Treebank, was published. However, although originating in computational linguistics, the value of treebanks is becoming more widely appreciated in linguistics research as a whole. For example, annotated treebank data has been crucial in syntactic research to test linguistic theories of sentence structure against large quantities of naturally occurring examples.




```
(S (NP (NNP John))
   (VP (VPZ loves)
       (NP (NNP Mary)))
   (. .))
```

Treebank



Existe uma forte dependência ao idioma que está sendo tratado.

Não é uma tarefa fácil a criação de um conjunto de rótulos. Veja o caso para o árabe.

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural

VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

JJ	Adjective

IN	Preposition or subordinating conjunction

DT	Determiner

```
[('Federal', 'NNP'),
('University', 'NNP'),
('of', 'IN'),
('ABC', 'NNP'),
('(', '('),
('Portuguese', 'JJ'),
(':', ':'),
('Universidade', 'JJ'),
('Federal', 'NNP'),
('do', 'VBP'),
('ABC', 'NNP'),
(',', ','),
('UFABC', 'NNP'),
(')', ')'),
('is', 'VBZ'),
```

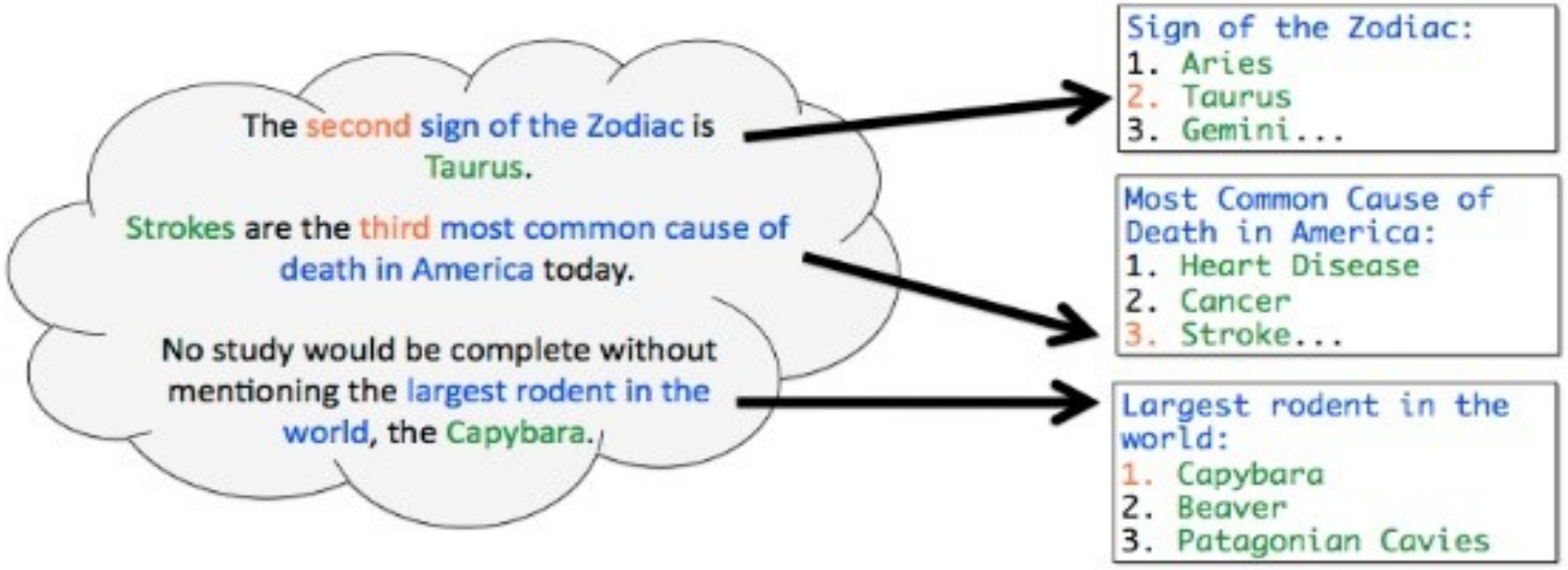
```
('a', 'DT'),
('Brazilian', 'JJ'),
('institution', 'NN'),
('of', 'IN'),
('higher', 'JJR'),
('learning', 'NN'),
('based', 'VBN'),
('in', 'IN'),
('Santo', 'NNP'),
('André', 'NNP'),
(',', ','),
('with', 'IN'),
('operations', 'NNS'),
('in', 'IN'),
('several', 'JJ'),
```

```
('municipalities', 'NNS'),
('in', 'IN'),
('the', 'DT'),
('ABC', 'NNP'),
('region', 'NN'),
(',', ','),
('all', 'DT'),
('in', 'IN'),
('the', 'DT'),
('state', 'NN'),
('of', 'IN'),
('São', 'NNP'),
('Paulo', 'NNP'),
('.', '.')]
```

Reconhecedor de entidades nomeadas

O reconhecimento de entidades nomeadas (**NER**, *Named Entity Recognition*) permite identificar “entidades nomeadas”.

Comumente são nomes próprios, locais e são utilizados para etapas de PLN que permita identificar **entidades-chave**.



nltk.ne_chunk(*tagged*, binary=True)

```
('Federal', 'NNP'),
('University', 'NNP'),
('of', 'IN'),
('ABC', 'NNP'),
('(', '('),
('Portuguese', 'JJ'),
(':', ':'),
('Universidade', 'JJ'),
('Federal', 'NNP'),
('do', 'VBP'),
('ABC', 'NNP'),
(',', ','),
('UFABC', 'NNP'),

…

('in', 'IN'),
('the', 'DT'),
('state', 'NN'),
('of', 'IN'),
('São', 'NNP'),
('Paulo', 'NNP'),
('.', '.')
```

```
(S
  (NE Federal/NNP University/NNP)
  of/IN
  (NE ABC/NNP)
  (/(
  (NE Portuguese/JJ)
  :/:
  Universidade/JJ
  Federal/NNP
  do/VBP
  (NE ABC/NNP)

  …

  in/IN
  the/DT
  state/NN
  of/IN
  (NE São/NNP Paulo/NNP)
  ./.)
```

nltk.ne_chunk(*tagged*, binary=false)

```
('Federal', 'NNP'),
('University', 'NNP'),
('of', 'IN'),
('ABC', 'NNP'),
('(', '('),
('Portuguese', 'JJ'),
(':', ':'),
('Universidade', 'JJ'),
('Federal', 'NNP'),
('do', 'VBP'),
('ABC', 'NNP'),
(',', ','),
('UFABC', 'NNP'),

…

('in', 'IN'),
('the', 'DT'),
('state', 'NN'),
('of', 'IN'),
('São', 'NNP'),
('Paulo', 'NNP'),
('.', '.')
```

```
(S
  (ORGANIZATION Federal/NNP University/NNP)
  of/IN
  (ORGANIZATION ABC/NNP)
  (/(
  (GPE Portuguese/JJ)
  :/:
  (PERSON Universidade/JJ Federal/NNP)
  do/VBP
  (ORGANIZATION ABC/NNP)
  ,/,
  (ORGANIZATION UFABC/NNP)
…
  in/IN
  the/DT
  state/NN
  of/IN
  (ORGANIZATION São/NNP Paulo/NNP)
  ./.)
```

Tipos de EN mais comuns

Dependerá muito da aplicação, mas na seguinte tabela temos uma lista das 6 entidades nomeadas mais comuns.

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Tappan Zee Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Sumarização

Sumarização de informação

É um processo de geração automática de um resumo associado a um objeto (e.g., texto, artigo, história).

A sumarização permite não somente entender a estrutura de uma frase, mas também a estrutura do texto inteiro.

A sumarização permitiria evidenciar o tema tratado.

Abordagem intuitiva ...

Assumiremos que a sumarização é nada mais do que a **classificação das frases** baseadas em sua importância.

Podemos também assumir que: Normalmente uma frase que tem **mais entidades e substantivos** tem maior importância que outras sentenças.

Usando essa lógica vamos ver um exemplo "simples” para calculo de **pontuação de importância.**

Podemos considerar um limiar para selecionar apenas as primeiras frases mais importantes.

sumarizacao.py

```
1   #!/usr/bin/env python
2   import sys
3   import nltk
4
5   if __name__ == '__main__':
6       fileName = sys.argv[1]
7       document = open(fileName,'r')
8       content  = document.read()
9       ranking  = []
10
11      for (i, sentence) in enumerate(nltk.sent_tokenize(content)):
12          tokens         = nltk.word_tokenize(sentence)
13          tagged         = nltk.pos_tag(tokens)
14          named_entities = nltk.ne_chunk(tagged, binary=True)
15          nTokens        = len(tokens)
16
17          print (f"\n{i} : >>{sentence}<<\n")
18
19          # Contagem dos substantivos: NN e NNP.
20          nNouns = 0
21          for (word, pos) in tagged:
22              if pos in ["NN", "NNP"]:
23                  nNouns += 1
24                  print (word, pos)
```

sumarizacao.py

```
25
26          # Contagem de entidades nomeadas
27          nNes = 0
28          for ne in named_entities:
29              if hasattr(ne, 'label') and ne.label() == "NE":
30                  nNes += 1
31                  print (ne)
32
33          # Cálculo do score ingenuo
34          score = (nNes + nNouns)/float(nTokens)
35          ranking.append( (i, score, nNes, nNouns, nTokens, sentence) )
36
37      # Imprimindo as frases em ordem invertida pelo score.
38      for r in sorted( ranking, key=lambda x: x[1], reverse=True):
39          print (r)
```

Texto de exemplo

Federal University of ABC (Portuguese: Universidade Federal do ABC, UFABC) is a Brazilian institution of higher learning based in Santo André, with operations in several municipalities in the ABC region, all in the state of São Paulo. The chairman of the committee that formulated the proposal of the university was Luiz Bevilacqua, who became its second rector.

UFABC is the only federal university in Brazil with 100% of its professors holding Ph.D.s and, for the second consecutive year in 2011, emerged as the only university in Brazil with impact factor in scientific publications above the world average according to SCImago Institutions Rankings.

In 2004, the Ministry of Education sent to Congress the Bill 3962/2004, which created the Federal University of ABC. This law was enacted by the president and published in the Official Gazette of 27 July 2005, under No. 11,145 and dated 26 July 2005.

Texto de exemplo

(0, **0.54**, 9, 15, 44, 'Federal University of ABC (Portuguese: Universidade Federal do ABC, UFABC) is\na Brazilian institution of higher learning based in Santo André, with\ noperations in several municipalities in the ABC region, all in the state of\nSão Paulo.')

(3, **0.45**, 3, 7, 22, 'In 2004, the Ministry of Education sent to Congress the Bill 3962/2004, which\ncreated the Federal University of ABC.')

(2, **0.36**, 4, 14, 49, 'UFABC is the only federal university in Brazil with 100% of its professors\nholding Ph.D.s and, for the second consecutive year in 2011, emerged as the\nonly university in Brazil with impact factor in scientific publications above\nthe world average according to SCImago Institutions Rankings.')

(1, **0.36**, 1, 7, 22, 'The chairman of the committee that formulated the proposal of the\ nuniversity was Luiz Bevilacqua, who became its second rector.')

(4, **0.23**, 0, 5, 21, 'This law was enacted by the president\nand published in the Official Gazette of 27 July 2005, under No.')

(5, **0.14**, 0, 1, 7, '11,145 and\ndated 26 July 2005.')

```
# Cálculo do score ingenuo
score = (nNes + nNouns)/float(nTokens)
ranking.append( (i, score, nNes, nNouns, nTokens, sentence) )
```

(0, **0.54**, 9, 15, 44, 'Federal University of ABC (Portuguese: Universidade Federal do ABC, UFABC) is\na Brazilian institution of higher learning based in Santo André, with\ noperations in several municipalities in the ABC region, all in the state of\nSão Paulo.')

(3, **0.45**, 3, 7, 22, 'In 2004, the Ministry of Education sent to Congress the Bill 3962/2004, which\ncreated the Federal University of ABC.')

(2, **0.36**, 4, 14, 49, 'UFABC is the only federal university in Brazil with 100% of its professors\nholding Ph.D.s and, for the second consecutive year in 2011, emerged as the\nonly university in Brazil with impact factor in scientific publications above\nthe world average according to SCImago Institutions Rankings.')

(1, **0.36**, 1, 7, 22, 'The chairman of the committee that formulated the proposal of the\ nuniversity was Luiz Bevilacqua, who became its second rector.')

(4, **0.23**, 0, 5, 21, 'This law was enacted by the president\nand published in the Official Gazette of 27 July 2005, under No.')

(5, **0.14**, 0, 1, 7, '11,145 and\ndated 26 July 2005.')

Table of Contents

DEPARTMENTS

From the president

Cerf's up

Letters to the editor

BLOG@CACM

SECTIONS

Practice

Contributed articles

Review articles

Research highlights

COLUMNS

News

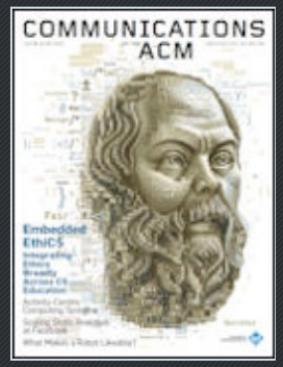
Education

Economic and business dimensions

Kode vicious

Viewpoint

Last byte



Previous Issue | Next Issue

View the Digital Library

View the Digital Edition

Download a PDF of this issue

SECTION: REVIEW ARTICLES

Activity-Centric Computing Systems

The ability to build a construct that organizes work from different devices and information resources is as complex as it is invaluable.

Jakob E. Bardram, Steven Jeuris, Paolo Tell, Steven Houben, Stephen Voida

Pages 72-81

The History of Digital Spam

Tracing the tangled web of unsolicited and undesired email and possible strategies for its demise.

Emilio Ferrara

Pages 82-91

SECTION: RESEARCH HIGHLIGHTS

Technical Perspective: The True Cost of Popularity

The main focus of "Heavy Hitters via Cluster-Preserving Clustering" by Kasper Green Larsen et al. is on building up sufficient information to allow a more effective search process.

Graham Cormode

Page 94

Heavy Hitters via Cluster-Preserving Clustering

We develop a new algorithm for the turnstile heavy hitters problem in general turnstile streams, the EXPANDERSKETCH, which finds the approximate top-k items in a universe of size n using the same asymptotic O(k log n) words of …

Kasper Green Larsen, Jelani Nelson, Huy L. Nguyễn, Mikkel Thorup

Pages 95-100

CACM: SPAM!

review articles

DOI:10.1145/3299768

Tracing the tangled web of unsolicited and undesired email and possible strategies for its demise.

BY EMILIO FERRARA

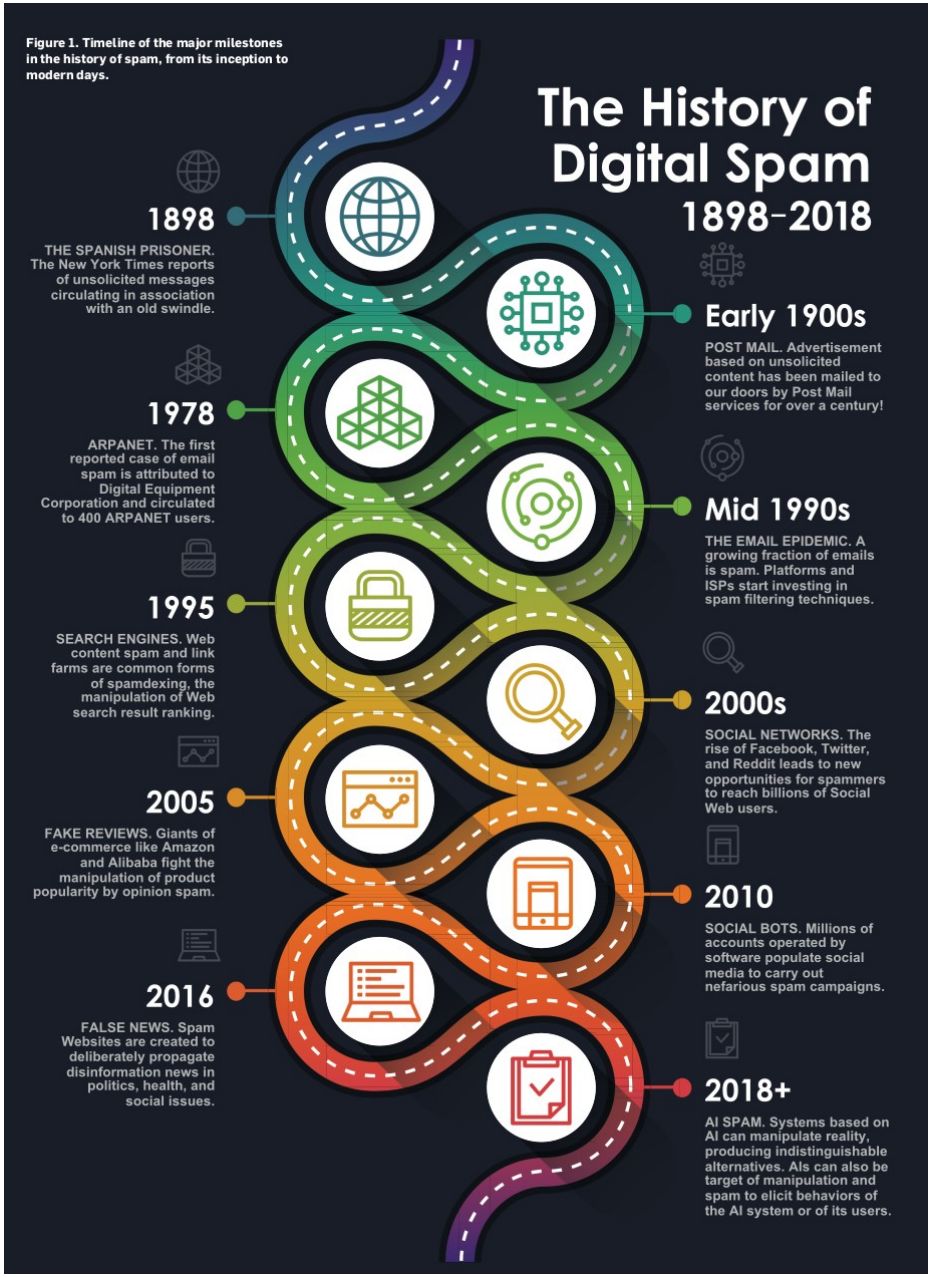
The History of Digital Spam

SPAM! THAT'S WHAT Lorrie Faith Cranor and Brian LaMacchia exclaimed in the title of a popular call-to-action article that appeared 20 years ago in *Communications*.[10] And yet, despite the tremendous efforts of the research community over the last two decades to mitigate this problem, the sense of urgency remains unchanged, as emerging technologies have brought new dangerous forms of digital spam under the spotlight. Furthermore, when spam is carried out with the intent to deceive or influence at scale, it can alter the very fabric of society and our behavior. In this article, I will briefly review the history of digital spam: starting from its quintessential incarnation, spam emails, to modern-days forms of spam affecting the Web and social media, the survey will close by depicting future risks associated with spam and abuse of new technologies, including artificial intelligence (AI), for example, digital humans. After providing a taxonomy of spam, and its most popular applications emerged throughout the last two decades, I will review technological and regulatory approaches proposed in the literature, and suggest some possible solutions to tackle this ubiquitous digital epidemic moving forward.

An omni-comprehensive, universally acknowledged definition of digital spam is hard to formalize. Laws and regulation attempted to define particular forms of spam, for example, email (see 2003's Controlling the Assault of Non-Solicited Pornography and Marketing Act.) However, nowadays, spam occurs in a variety of forms, and across different techno-social systems. Each domain may warrant a slight different definition that suits what spam is in that precise context: some features of spam in a domain, for example, volume in mass spam campaigns, may not apply to

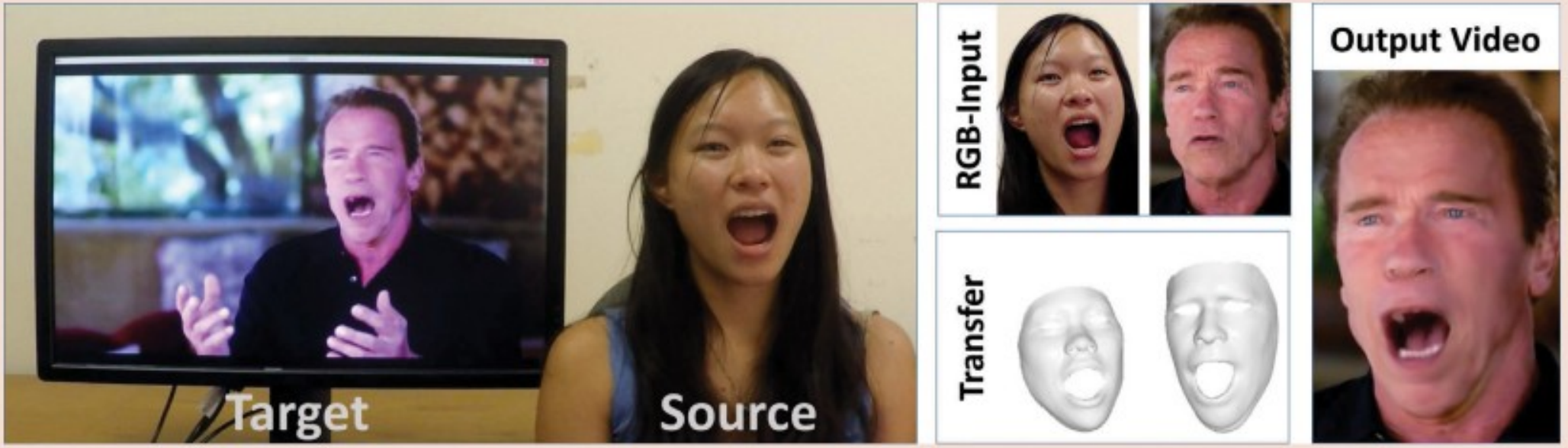
» key insights

- **Throughout the Internet's history, digital spam has pervaded all techno-social platforms and it is constantly evolving. This article provides a taxonomy of digital spam, from its inception to current spam techniques.**
- **Since the email spam epidemic of the early 1990s, new forms of spam have emerged, including search engine spam, fake reviews, spam bots, and false news. In its latest incarnation, spam threats to pollute AI systems making them biased and ultimately dangerous for our society.**
- **By illustrating some of the risks posed by digital spam in all its forms, including AI spam, we provide policy recommendations and technical insights to tackle old and new forms of spam.**



CACM: SPAM!

Figure 2. Video sequence real-time reenactment using AI.[34] This proof-of-concept technology could be abused to create AI-fueled multimedia spam.



SPAM!

(118, **1.0**, 1, 5, 6, 'Social Spam Applications\nPolitical manipulation.')

(126, **1.0**, 1, 2, 3, 'Public health.')

(55, **0.5**, 0, 2, 4, 'Fighting email spam.')

(10, **0.46**, 1, 5, 13, 'SMTP is the protocol at the foundation of the email exchange\
ninfrastructure.')

(129, **0.46**, 1, 5, 13, 'Automatic trading algorithms leverage information from social\
nmedia to predict stock prices.')

(39, **0.44**, 4, 12, 36, 'Flooded By Junk Email\nThe 1998 article by Cranor and LaMacchia10 in Communications, characterized the problem of junk email messages, or email spam, as one of the\nearliest forms of digital spam.')

(48, **0.42**, 1, 2, 7, 'Email spam: Risks and challenges.')

...

SPAM!

(80, **0.05**, 0, 2, 35, 'Furthermore,\nfakes reviews can be posted by individual users or even groups of users.27,30 Spammers can deliberately\nuse fake accounts on e-commerce\nplatforms, created only with the scope\nof posting fake reviews.')

(57, **0.03**, 0, 1, 30, 'However, it has become increasingly more obvious that\nsolutions based exclusively on regulatory affairs are ineffective: spam operations can move to countries with less\nrestrictive Internet regulations.')

(109, **0.03**, 0, 1, 30, 'Although isolated examples exist of such bots being used for nefarious purposes, I am unaware of any reports of systematic abuse carried out\nby bots in those contexts.')

(9, **0.0**, 0, 0, 7, 'Some\nmilestones include:\nSMTP solutions.')

(27, **0.0**, 0, 0, 7, 'Networking companies are developing anti-spam appliances.')

(65, **0.0**, 0, 0, 6, 'Web 2.0 or Spam 2.0?')

(107, **0.0**, 0, 0, 4, 'social bots).')

Notícia FAPESP

Atividades humanas já danificaram 75% da superfície terrestre

26 de março de 2018

EN ES

Número deve chegar a 90% até 2050, segundo o novo relatório sobre degradação e restauração de áreas degradadas divulgado pela Plataforma Intergovernamental sobre Biodiversidade e Serviços Ecossistêmicos (IPBES) (*foto: Dudarev Mikhail / Shutterstock.com*)

Karina Toledo, de Medellín | Agência FAPESP – Apenas 25% da superfície terrestre permanece livre de impactos substanciais causados por atividades humanas. E o índice deve cair para meros 10% até 2050, segundo projeções da Plataforma Intergovernamental sobre Biodiversidade e Serviços Ecossistêmicos (IPBES).

"Apenas algumas regiões nos polos, desertos e as partes mais inacessíveis das florestas tropicais permanecem intactas", afirmou o sul-africano Robert Scholes, um dos coordenadores do relatório temático sobre Degradação e Restauração de Terras Degradadas divulgado pela IPBES nesta segunda-feira (26/3), em Medellín, na Colômbia.

O documento na íntegra e um sumário para tomadores de decisão foram aprovados pelos 129 países-membros da entidade durante a 6ª Reunião Plenária, que ocorreu entre os dias 17 e 24 de março.

Segundo o texto, até o ano de 2014, mais de 1,5 bilhão de hectares de ecossistemas naturais foram convertidos em áreas agrícolas. Plantações e pastagens cobrem atualmente mais de um terço da superfície do planeta. "Os processos mais recentes de desmatamento estão ocorrendo nas regiões do globo mais ricas em biodiversidade", afirmaram os autores no texto.

Notícia FAPESP

(34, **0.85**, 0, 6, 7, 'avançar em direção das boas práticas.')

(45, **0.79**, 0, 19, 24, 'esta última corresponde a iniciativas voltadas a recuperar algumas das funções críticas da terra e criar condições para que talvez ela seja recuperada.')

(25, **0.66**, 0, 12, 18, '“é possível aumentar a produção sem avançar sobre áreas naturais e sem abusar de produtos químicos.')

(44, **0.66**, 0, 6, 9, '“fizemos uma diferenciação entre restauração e reabilitação.')

(60, **0.63**, 0, 14, 22, 'poucos turistas que visitam o parque nacional da tijuca sabem que estão caminhando em uma área restaurada”, disse joly.')

(27, **0.63**, 0, 24, 38, 'para scholes, o brasil está em uma posição favorável para lidar com essas questões por ter fortalecido ao longo dos últimos anos sua capacidade de realizar pesquisas científicas e por ter especialistas capazes de orientar soluções.')

...

Desafio 3: Bônus +0.5 na MF

Resumo de uma artigo da Magazine Communication of the ACM (CACM) https://cacm.acm.org/

O artigo selecionado deve ter sido publicado ou em 2018 ou em 2019. O artigo deve conter, no mínimo, 2 páginas.

Resumo nos mesmos moldes dos resumos de aula.

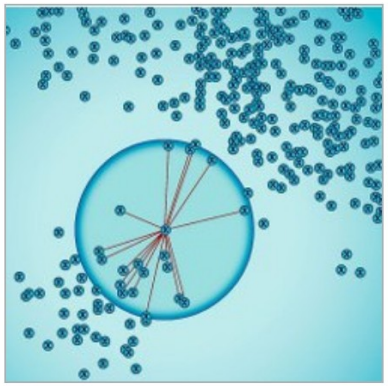
Envio pelo tidia (seção atividades).

Resumo de um artigo da CACM: Dos anos 2018 ou 2019

NEWS

Good Algorithms Make Good Neighbors

By Erica Klarreich
Communications of the ACM, July 2019, Vol. 62 No. 7, Pages 11-13
10.1145/3329712
Comments



Credit: Andrij Borys Associates

A host of different tasks—such as identifying the song in a database most similar to your favorite song, or the drug most likely to interact with a given molecule—have the same basic problem at their core: finding the point in a dataset that is closest to a given point. This "nearest neighbor" problem shows up all over the place in machine learning, pattern recognition, and data analysis, as well as many other fields.

Yet the nearest neighbor problem is not really a single problem. Instead, it has as many different manifestations as there are different notions of what it means for data points to be similar. In recent decades, computer scientists have devised efficient nearest neighbor algorithms for a handful of different definitions of similarity: the ordinary Euclidean distance between points, and a few other distance measures.

However, "every time you needed to work with a new space or distance measure, you would kind of have to start from scratch" in designing a nearest neighbor algorithm, said Rasmus Pagh, a computer scientist at the IT University of Copenhagen. "Each space required some kind of craftsmanship."

Because distance measures are so varied, many computer scientists doubted these ad hoc methods would ever give way to a more general approach that could cover many different distance measures at once. Now, however, a team of five computer scientists has proven the doubters—who originally included themselves—were wrong.

In a pair of papers published last year (in the *Proceedings of the ACM Symposium on Theory of Computing* and the IEEE Annual Symposium on Foundations of Computer Science, respectively), the researchers set forth an efficient approximation algorithm for nearest neighbor search that covers a wide class of distance functions. Their algorithm finds, if not the very closest neighbor, then one that's almost as close, which is good enough for many applications.

The distance functions covered by the new algorithm, called norms, "encompass the majority of interesting distance functions," said Piotr Indyk, a computer scientist at the Massachusetts Institute of Technology.

The new algorithm is a big leap forward, Pagh said, who added, "I wouldn't have guessed such a general result was possible."

Resumo de um artigo da CACM: requisito (ser maior ou igual a 2 páginas)

CONTRIBUTED ARTICLES

Ways of Thinking in Informatics

By Christopher Frauenberger, Peter Purgathofer
Communications of the ACM, July 2019, Vol. 62 No. 7, Pages 58-64
10.1145/3329674
Comments

VIEW AS: SHARE: AddThis

Credit: Andrij Borys Associates /

With the intimate entanglement of digital technology with humans and their social way of being, computer science has changed. While some of the problems we deal with are (still) well defined and mostly computationally solvable, many problems are now found to be wicked and ill defined. People and technologies are now part of an interwoven socio-material web in which humans are not the only actors anymore. This pervasive complexity raises challenges for computer scientists and technologists that go well beyond of what could be addressed by a traditional understanding of engineering the most efficient computational tools. It requires us to rethink what must be the core competencies of future computer scientists. New skills stemming from the social sciences or philosophy need to complement engineering skills to create digital technologies within lived experiences. With it comes a major shift in responsibility. In a *New York Times* article, Farhad