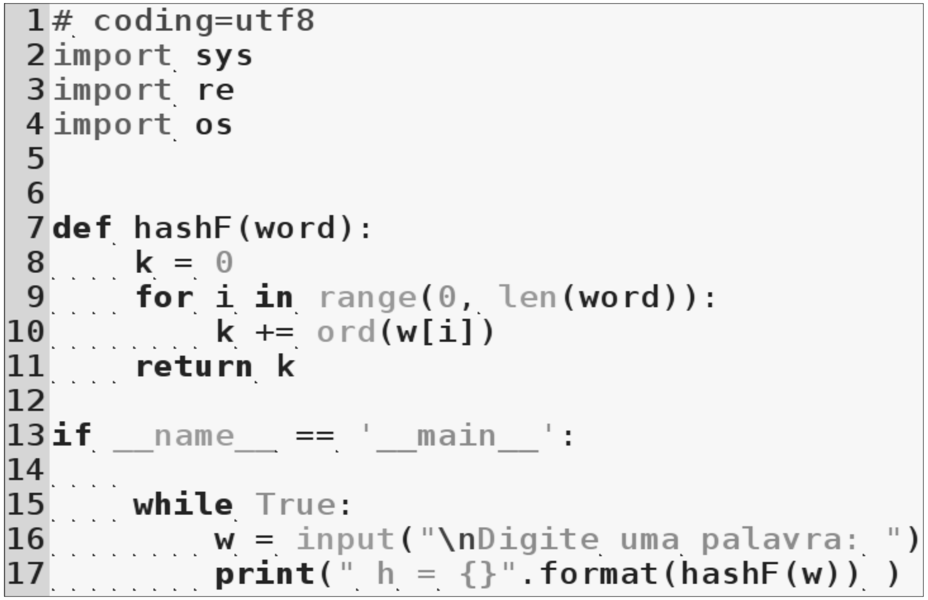
MCZA017-13
Processamento de Linguagem Natural

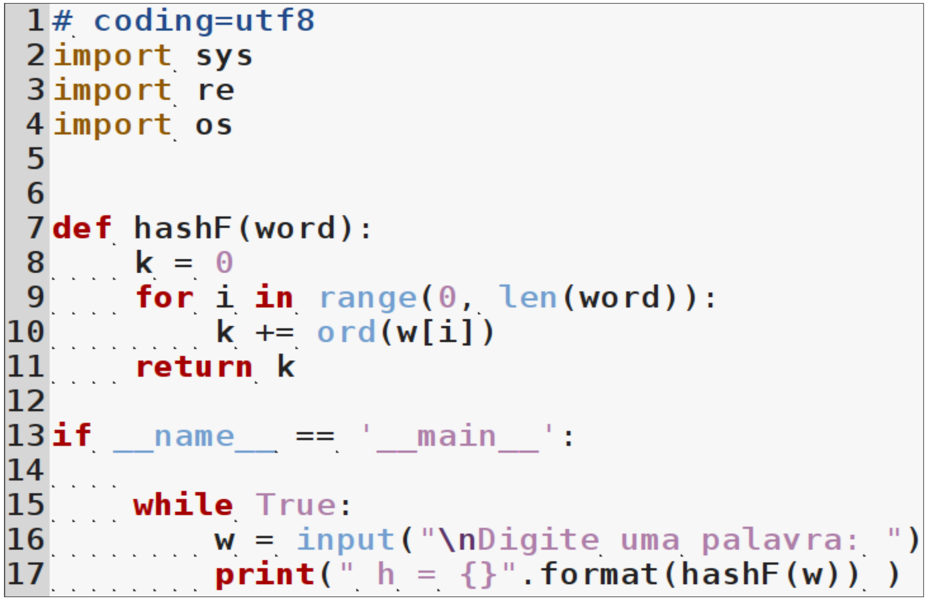
Reconhecimento de entidades nomeadas

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

2Q-2019



```python
# coding=utf8
import sys
import re
import os


def hashF(word):
    k = 0
    for i in range(0, len(word)):
        k += ord(w[i])
    return k

if __name__ == '__main__':
    
    while True:
        w = input("\nDigite uma palavra: ")
        print(" h = {}".format(hashF(w)) )
```



```python
# coding=utf8
import sys
import re
import os


def hashF(word):
    k = 0
    for i in range(0, len(word)):
        k += ord(w[i])
    return k

if __name__ == '__main__':
    
    while True:
        w = input("\nDigite uma palavra: ")
        print(" h = {}".format(hashF(w)) )
```

Empirical Software Engineering
pp 1–34 | Cite as

Does syntax highlighting help programming novices?

Authors | Authors and affiliations

Christoph Hannebauer ✉ , Marc Hesenius, Volker Gruhn

Article
First Online: 28 February 2018

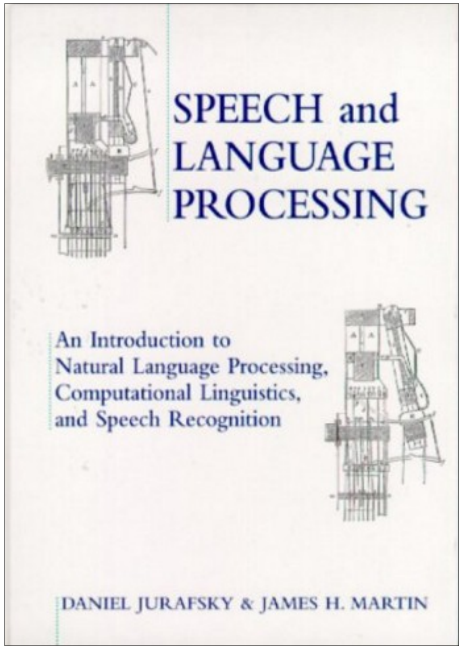
Abstract

Program comprehension is an important skill for programmers – extending and debugging existing source code is part of the daily routine. Syntax highlighting is one of the most common tools used to support developers in understanding algorithms. However, most research in this area originates from a time when programmers used a completely different tool chain. We examined the influence of syntax highlighting on novices' ability to comprehend source code. Additional analyses cover the influence of task type and programming experience on the code comprehension ability itself and its relation to syntax highlighting. We conducted a controlled experiment with 390 undergraduate students in an introductory Java programming course. We measured the correctness with which they solved small coding tasks. Each test subject received some tasks with syntax highlighting and some without. The data provided no evidence that syntax highlighting improves novices' ability to comprehend source code. There are very few similar experiments and it is unclear as of yet which factors impact the effectiveness of syntax highlighting. One major limitation may be the types of tasks chosen for this experiment. The results suggest that syntax highlighting squanders a feedback channel from the IDE to the programmer that can be used more effectively.

Bibliografia

Daniel Jurafsky & James H. Martin.

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

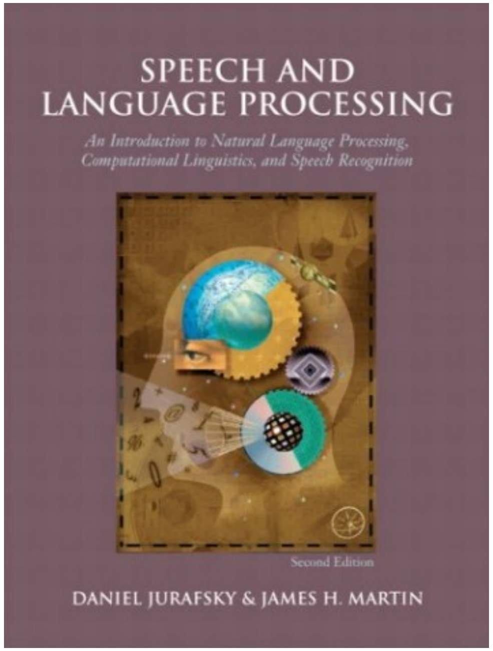


SPEECH and LANGUAGE PROCESSING

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

DANIEL JURAFSKY & JAMES H. MARTIN

2000



SPEECH AND LANGUAGE PROCESSING

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Second Edition

DANIEL JURAFSKY & JAMES H. MARTIN

2009

2019?

Stanford University

University of Colorado, Boulder

Extração de informação

Extração de informação

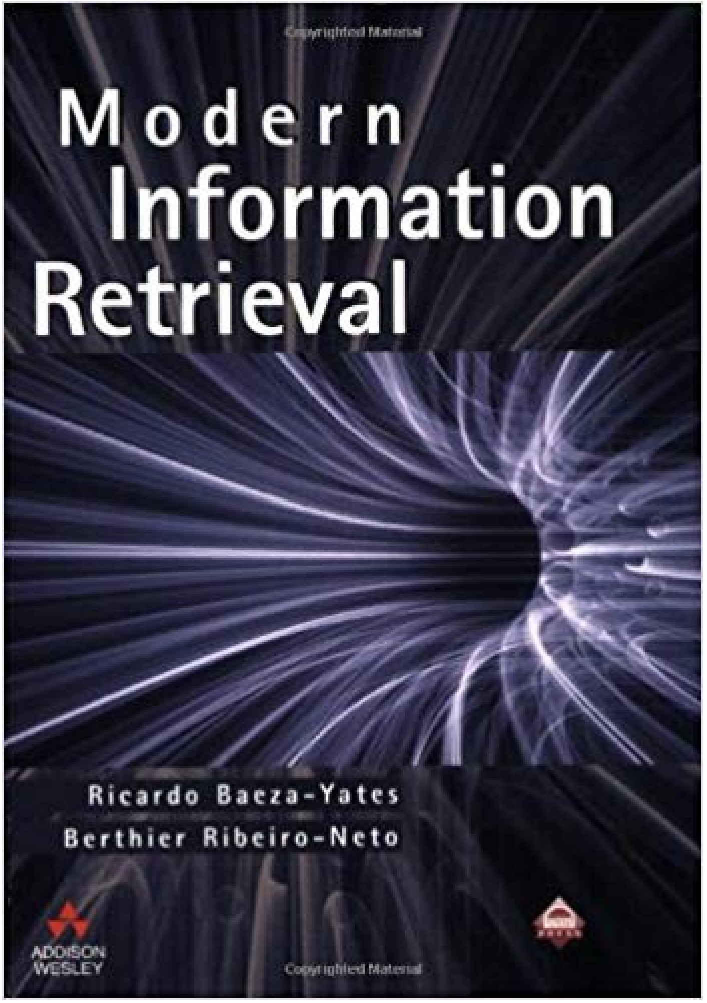
Os sistemas de extração de informação permitem:

- Transformar os dados **não estruturados** (incorporados em textos) em **dados estruturados**.
- **Encontrar** partes relevantes do texto.
- **Obter informação** de trechos de texto.
- **Produzir uma representação estruturada** de informação relevante.

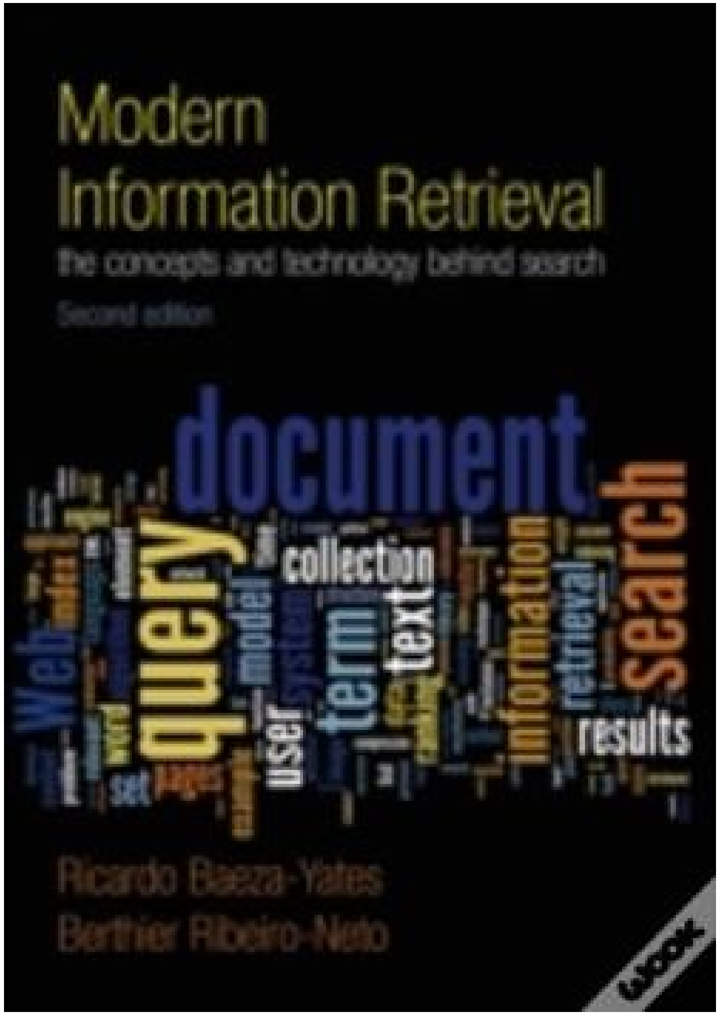
Objetivos:

- Organizar informação que seja útil para as pessoas.
- Colocar informações de forma clara que sejam úteis para inferencias realizadas por algoritmos computacionais.

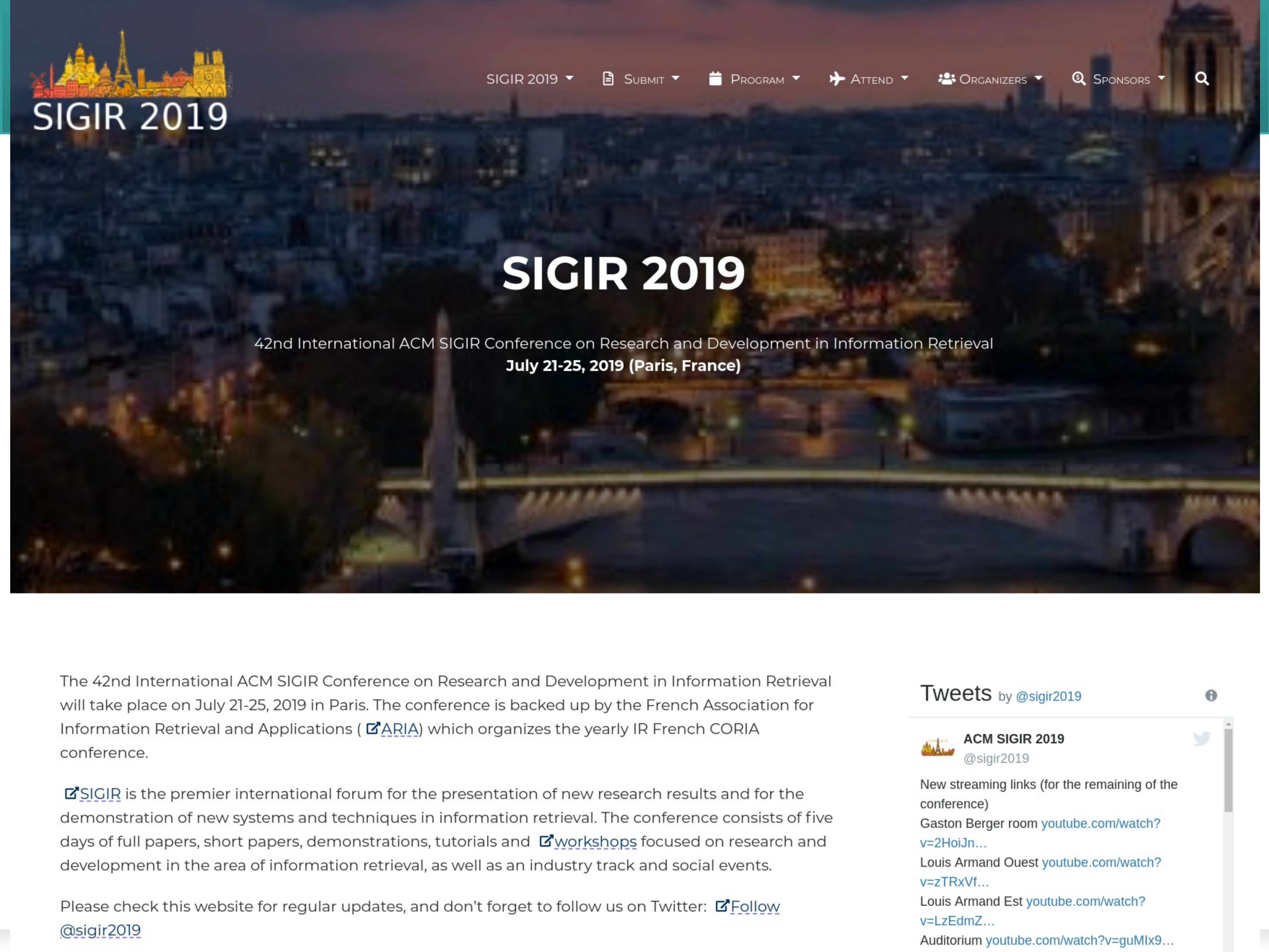
Information retrieval



1999



2010



SIGIR 2019

42nd International ACM SIGIR Conference on Research and Development in Information Retrieval
July 21-25, 2019 (Paris, France)

The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval will take place on July 21-25, 2019 in Paris. The conference is backed up by the French Association for Information Retrieval and Applications (ARIA) which organizes the yearly IR French CORIA conference.

SIGIR is the premier international forum for the presentation of new research results and for the demonstration of new systems and techniques in information retrieval. The conference consists of five days of full papers, short papers, demonstrations, tutorials and workshops focused on research and development in the area of information retrieval, as well as an industry track and social events.

Please check this website for regular updates, and don't forget to follow us on Twitter: Follow @sigir2019

Tweets by @sigir2019

ACM SIGIR 2019
@sigir2019

New streaming links (for the remaining of the conference)
Gaston Berger room youtube.com/watch?v=2HoiJn…
Louis Armand Ouest youtube.com/watch?v=zTRxVf…
Louis Armand Est youtube.com/watch?v=LzEdmZ…
Auditorium youtube.com/watch?v=guMIx9…

Generic Intent Representation in Web Search

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul Bennett, Nick Craswell and Saurabh Tiwary

Generic Intent Encoding | Query Embedding | User Intent Understanding

Harvesting Drug Effectiveness from Social Media

Zi Chai, Xiaojun Wan, Zhao Zhang and Minjie Li

Drug effectiveness discovery | Relation extraction | Social media mining | Graph-based Information Transfers Over Time

Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation

Aymé Arango, Jorge Perez and Barbara Poblete

hate speech detection | hate speech classification | experimental validation | benchmark datasets | deep learning | social media

Health Cards for Consumer Health Search

Jimmy Jimmy, Guido Zuccon, Bevan Koopman and Gianluca Demartini

Health cards | Consumer health search | User study

Hierarchical Matching Network for Crime Classification

Pengfei Wang, Yu Fan, Yongfeng Zhang, Shuzi Niu, Ze Yang and Jiafeng Guo

Hierarchical multi-label classification | Crime Classification | Hierarchical Matching Network

Hot Topic-Aware Retweet Prediction with Masked Self-attentive Model

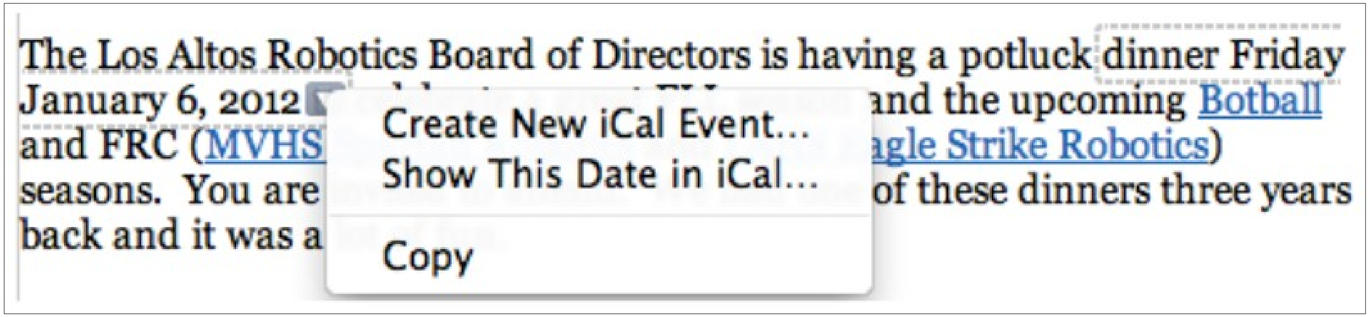
Renfeng Ma, Qi Zhang, Xiangkun Hu, Xuanjing Huang and Yu-Gang Jiang

Retweet prediction | Hot topics | Social Medias

Exemplos

Vários sistemas atuais permitem identificar regiões textuais de interese para o usuário.

Email: identificar uma data para associar com a agenda.

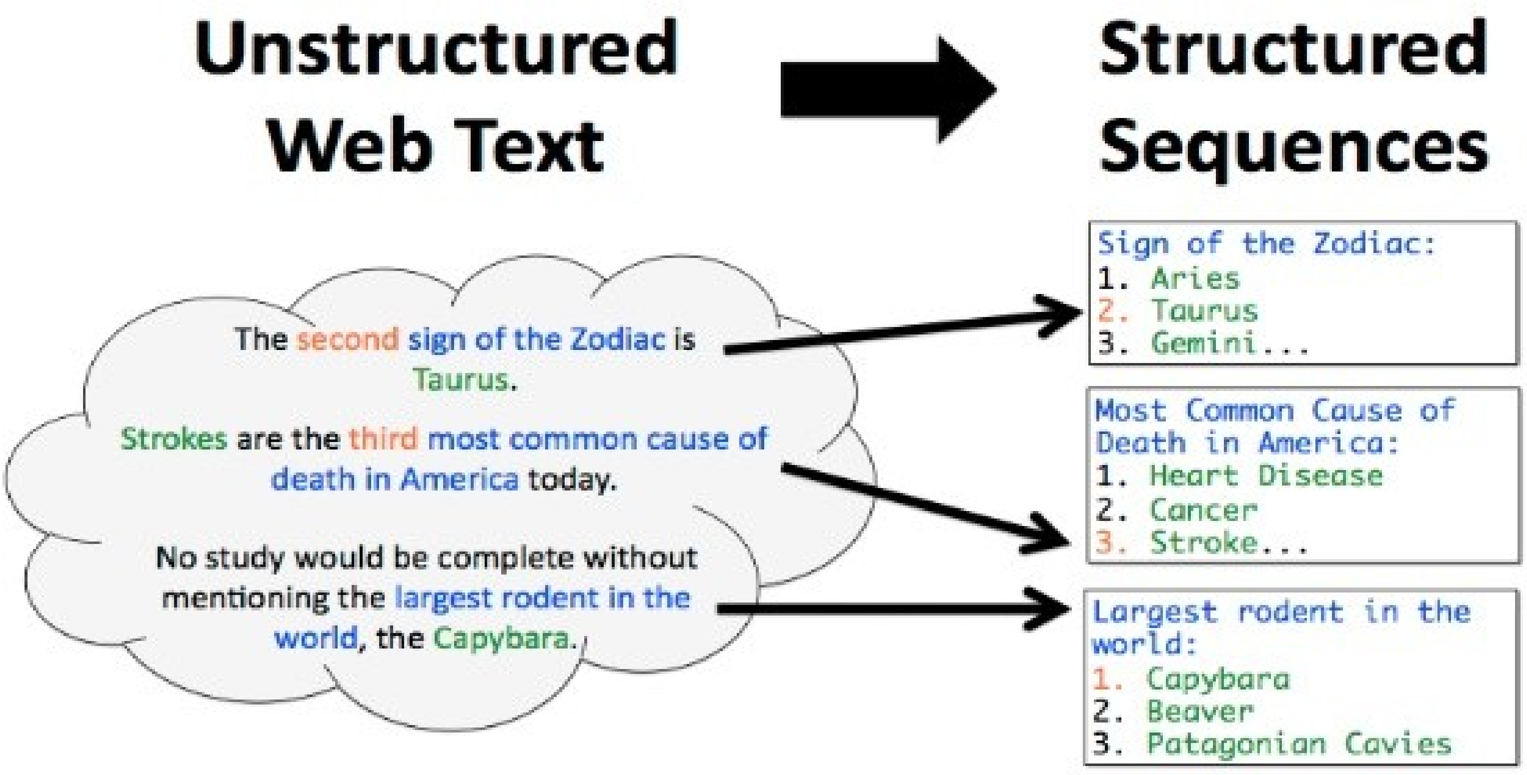


The Los Altos Robotics Board of Directors is having a potluck dinner Friday
January 6, 2012 and the upcoming Botball
and FRC (MVHS agle Strike Robotics)
seasons. You are of these dinners three years
back and it was a

Apple

Frequentemente são utilizadas expressões regulares ou lista de nomes.

Exemplos



The second sign of the Zodiac is Taurus.

Strokes are the third most common cause of death in America today.

No study would be complete without mentioning the largest rodent in the world, the Capybara.

```
Sign of the Zodiac:
1. Aries
2. Taurus
3. Gemini...
```

```
Most Common Cause of
Death in America:
1. Heart Disease
2. Cancer
3. Stroke...
```

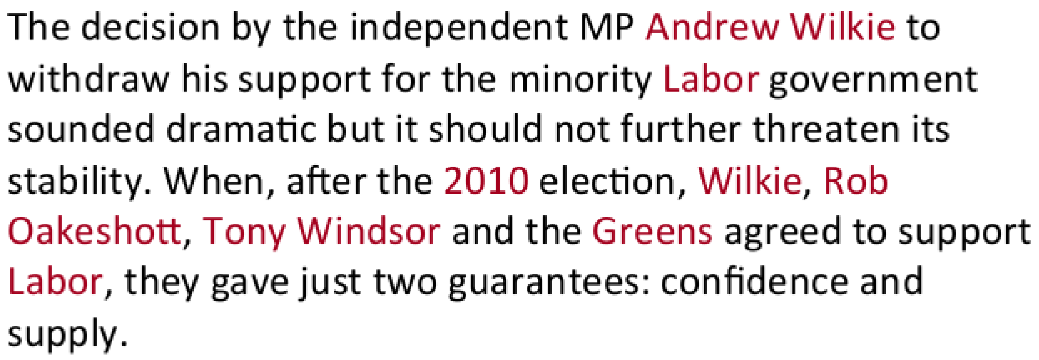
```
Largest rodent in the
world:
1. Capybara
2. Beaver
3. Patagonian Cavies
```

Reconhecimento de Entidades nomeadas
NER – Named Entity Recognition

Reconhecimento de entidades nomeadas

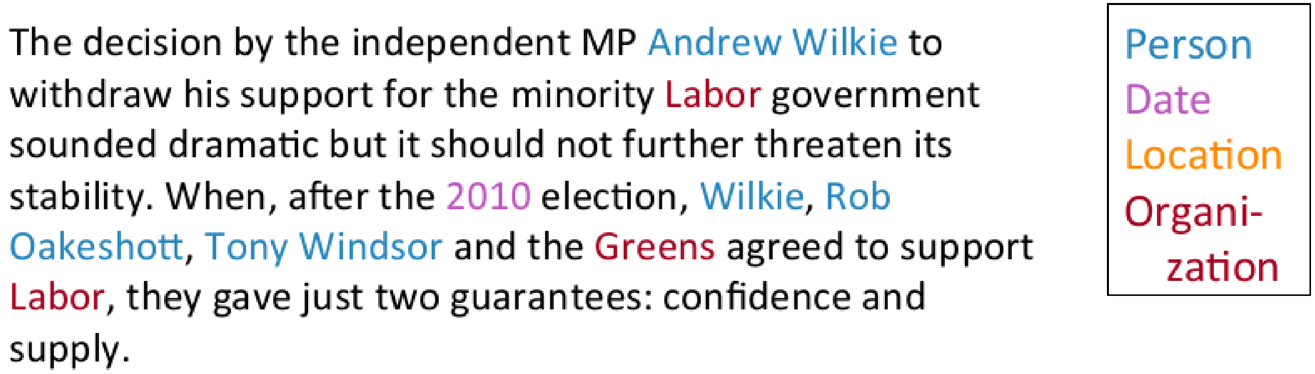
Um reconhecedor permite **identificar** e **classificar** as EN em um texto escrito em linguagem natural.

Identificação



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.

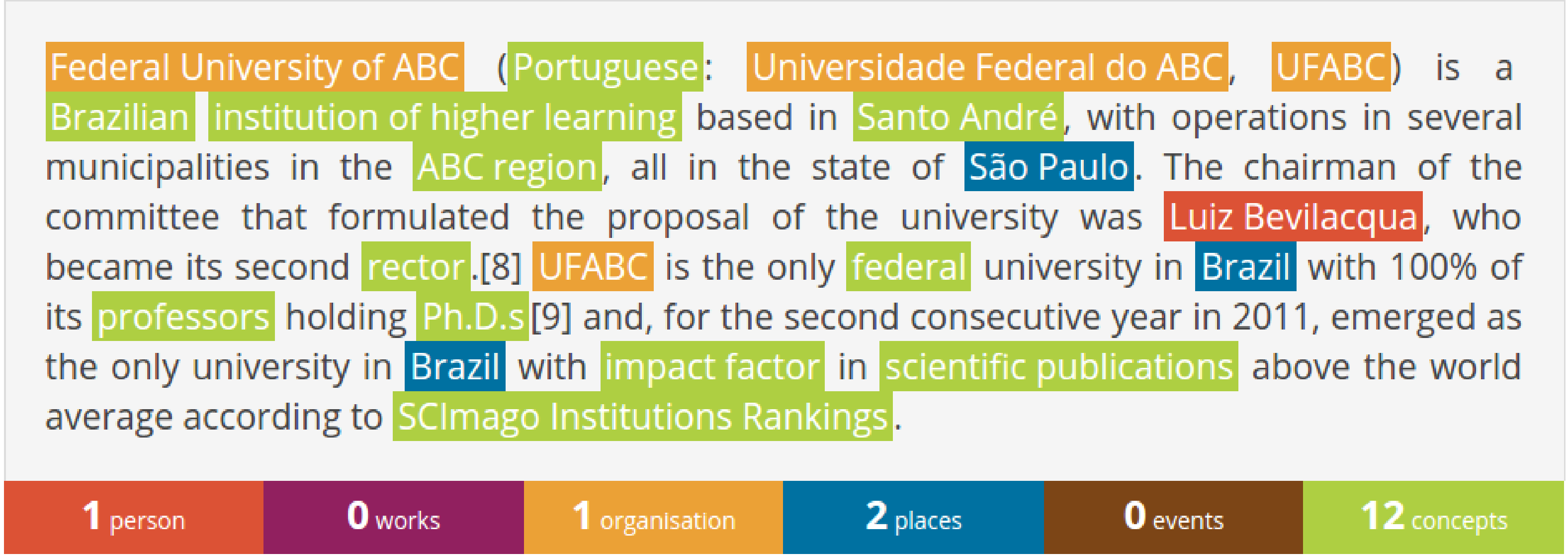
Classificação



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.

Person
Date
Location
Organi-
zation

Reconhecimento de entidades nomeadas



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.[8] UFABC is the only federal university in Brazil with 100% of its professors holding Ph.D.s[9] 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.

1 person	0 works	1 organisation	2 places	0 events	12 concepts

Usos

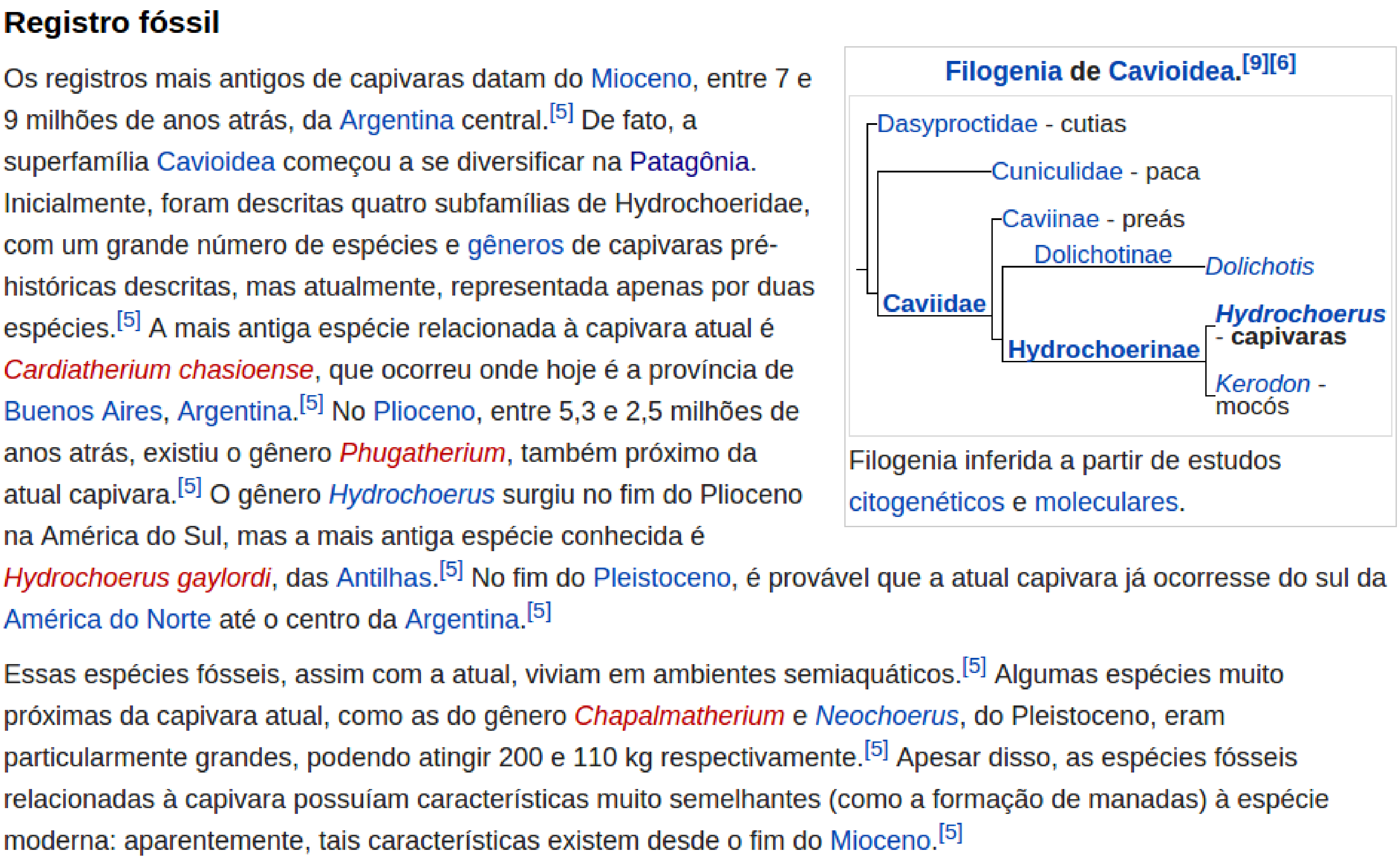
As **EN** podem ser **índices** para:

  - Conceitos.
  - Novas relações / associações entre outras entidades.

Na web:

- Às entidades nomeadas são associadas Links com maiores informações.

Usos: Dados estruturados na wikipedia



Registro fóssil

Os registros mais antigos de capivaras datam do Mioceno, entre 7 e 9 milhões de anos atrás, da Argentina central.[5] De fato, a superfamília Cavioidea começou a se diversificar na Patagônia. Inicialmente, foram descritas quatro subfamílias de Hydrochoeridae, com um grande número de espécies e gêneros de capivaras pré-históricas descritas, mas atualmente, representada apenas por duas espécies.[5] A mais antiga espécie relacionada à capivara atual é *Cardiatherium chasioense*, que ocorreu onde hoje é a província de Buenos Aires, Argentina.[5] No Plioceno, entre 5,3 e 2,5 milhões de anos atrás, existiu o gênero *Phugatherium*, também próximo da atual capivara.[5] O gênero *Hydrochoerus* surgiu no fim do Plioceno na América do Sul, mas a mais antiga espécie conhecida é *Hydrochoerus gaylordi*, das Antilhas.[5] No fim do Pleistoceno, é provável que a atual capivara já ocorresse do sul da América do Norte até o centro da Argentina.[5]

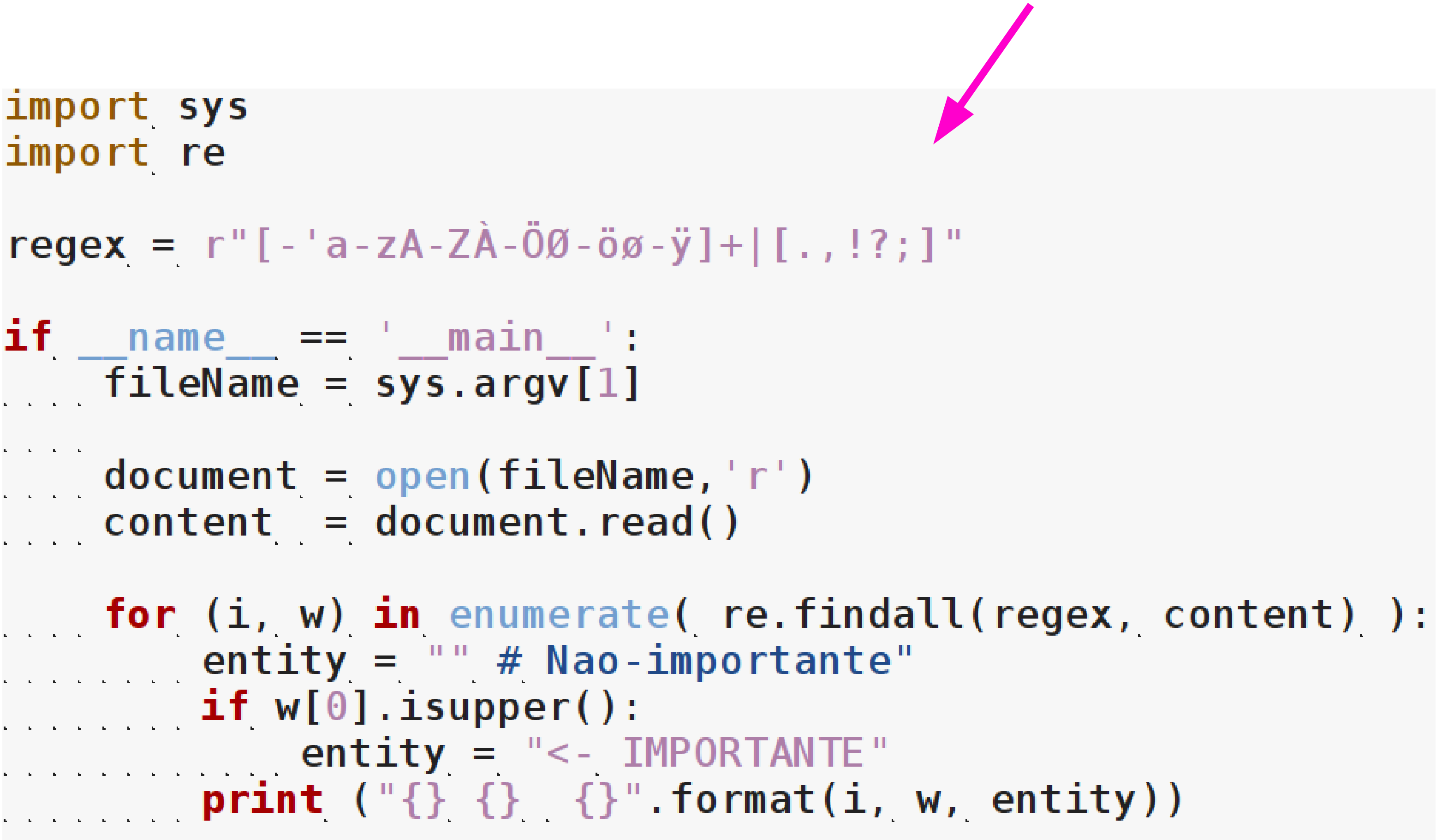
Filogenia de Cavioidea.[9][6]

Filogenia inferida a partir de estudos citogenéticos e moleculares.

Essas espécies fósseis, assim com a atual, viviam em ambientes semiaquáticos.[5] Algumas espécies muito próximas da capivara atual, como as do gênero *Chapalmatherium* e *Neochoerus*, do Pleistoceno, eram particularmente grandes, podendo atingir 200 e 110 kg respectivamente.[5] Apesar disso, as espécies fósseis relacionadas à capivara possuíam características muito semelhantes (como a formação de manadas) à espécie moderna: aparentemente, tais características existem desde o fim do Mioceno.[5]

Primeira abordagem
teste1.py

teste1.py



```python
import sys
import re

regex = r"[-'a-zA-ZÀ-ÖØ-öø-ÿ]+|[.,!?;]"

if __name__ == '__main__':
    fileName = sys.argv[1]

    document = open(fileName,'r')
    content  = document.read()

    for (i, w) in enumerate( re.findall(regex, content) ):
        entity = "" # Nao-importante"
        if w[0].isupper():
            entity = "<- IMPORTANTE"
        print ("{} {}  {}".format(i, w, entity))
```

teste1.py

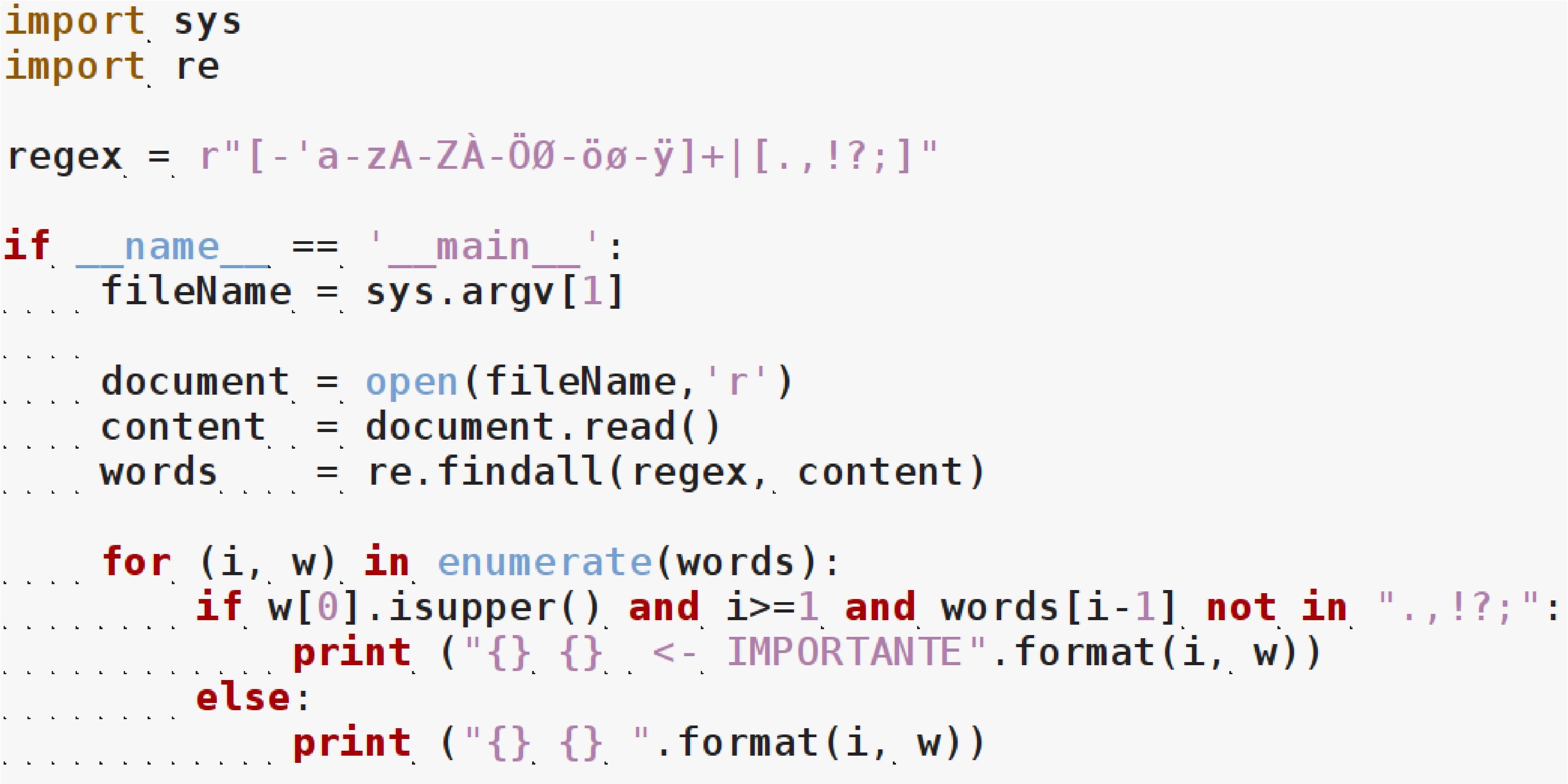
```
python3 teste1.py  capivara-pt.txt
0 A  <- IMPORTANTE
1 Capivara  <- IMPORTANTE
2 nome
3 científico
4 Hydrochoerus  <- IMPORTANTE
5 hydrochaeris
6 é
7 uma

...


13 família
14 Caviidae  <- IMPORTANTE
15 e
16 subfamília
17 Hydrochoerinae  <- IMPORTANTE
18 .
19 Alguns  <- IMPORTANTE
20 autores
```

O script é uma versão **muito simples** de identificação de palavras importantes de um texto.

Note que algumas não deveriam ser consideradas importantes.

Modifique o programa!

teste2.py



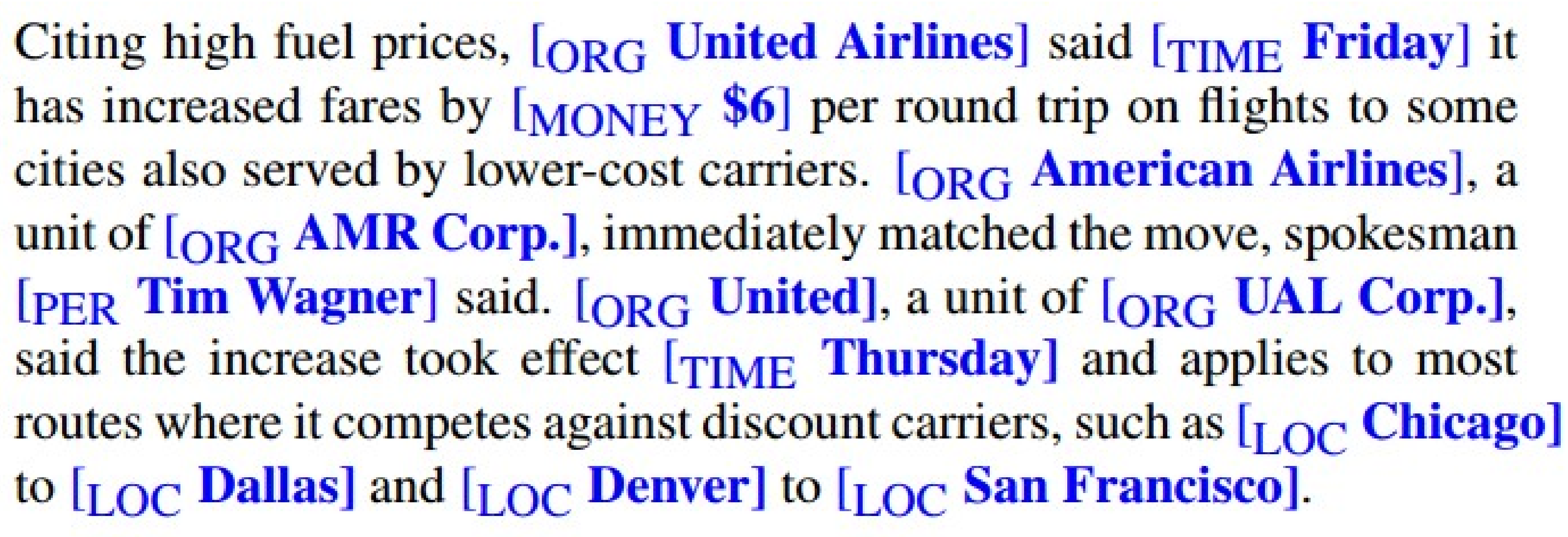
```python
import sys
import re

regex = r"[-'a-zA-ZÀ-ÖØ-öø-ÿ]+|[.,!?;]"

if __name__ == '__main__':
    fileName = sys.argv[1]
    
    document = open(fileName,'r')
    content  = document.read()
    words    = re.findall(regex, content)

    for (i, w) in enumerate(words):
        if w[0].isupper() and i>=1 and words[i-1] not in ".,!?;":
            print ("{} {}  <- IMPORTANTE".format(i, w))
        else:
            print ("{} {} ".format(i, w))
```

teste2.py

```
python3 teste2.py  capivara-pt.txt
0 A
1 Capivara  <- IMPORTANTE
2 nome
3 científico
4 Hydrochoerus  <- IMPORTANTE
5 hydrochaeris
6 é
7 uma

…

13 família
14 Caviidae  <- IMPORTANTE
15 e
16 subfamília
17 Hydrochoerinae  <- IMPORTANTE
18 .
19 Alguns
20 autores
```

Tipos de Entidades Nomeadas

Usos

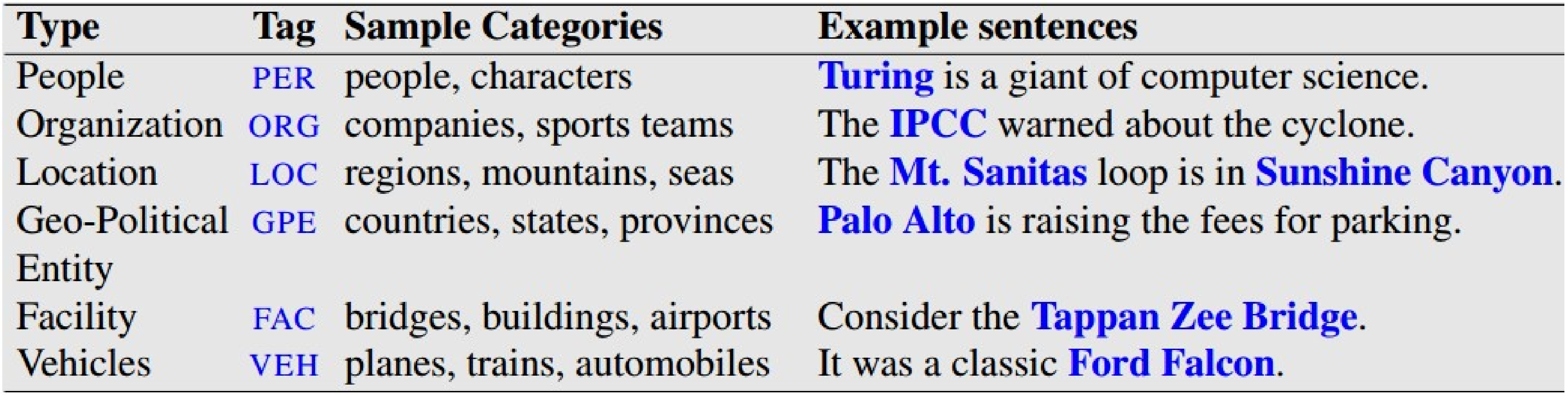


Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

ORG	Organization
TIME	Time period
MONEY	Currency
PER	Person
LOC	Location

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**.

Um reconhecedor de EN

- Permite **encontrar** **segementos** **de texto** que constituam nomes próprios e, em seguida, classificar seu tipo da entidade.

- O reconhecimento é difícil, em parte, devido **à ambiguidade da segmentação.** Precisamos decidir o que é uma entidade e o que não é, e quais são os limites.

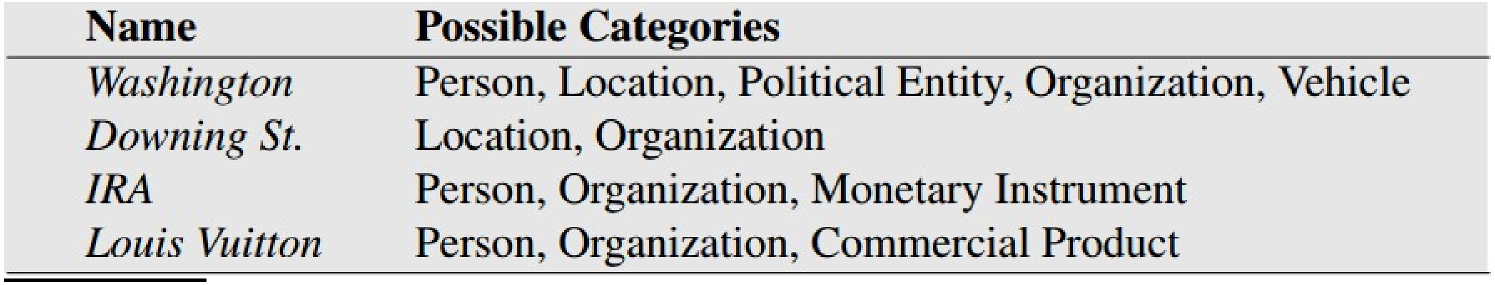
  Exemplo:

  JK (Juscelino Kubitschek)

  - Escola?
  - Avenida?
  - Pessoa?
  - Governo?

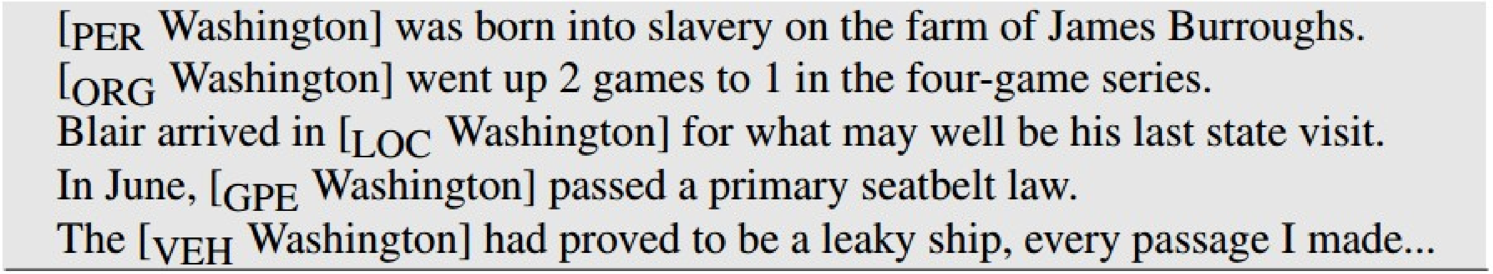
Um reconhecedor de EN

- Outro exemplo de possível multiple categoria para segmentos de texto.



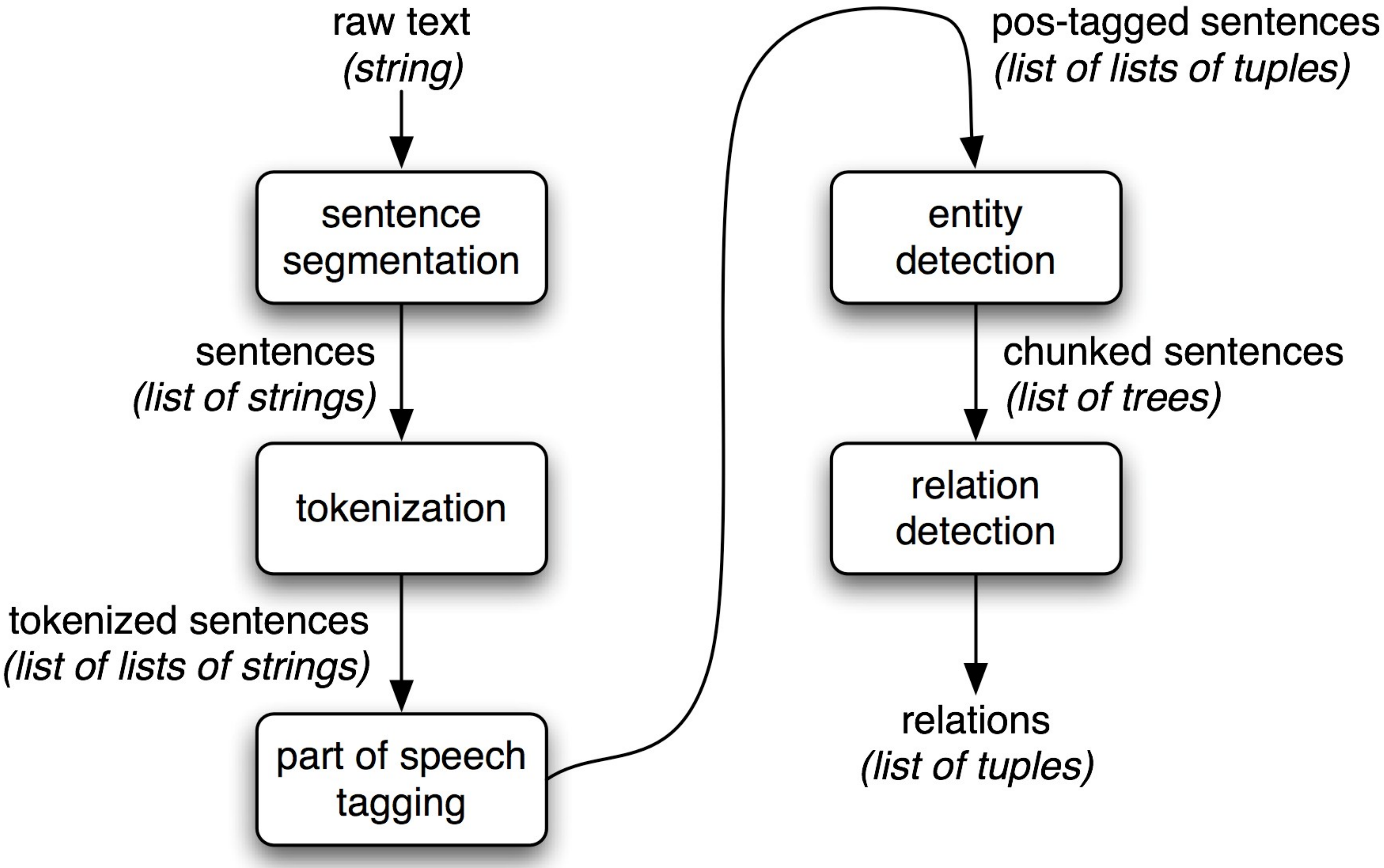
Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Vehicle
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

Por falar de Washington:



[$_{\text{PER}}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{\text{ORG}}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{\text{LOC}}$ Washington] for what may well be his last state visit.
In June, [$_{\text{GPE}}$ Washington] passed a primary seatbelt law.
The [$_{\text{VEH}}$ Washington] had proved to be a leaky ship, every passage I made...

Pipeline comum: reconhecimento de EN



raw text
(string)

sentence segmentation

sentences
(list of strings)

tokenization

tokenized sentences
(list of lists of strings)

part of speech tagging

pos-tagged sentences
(list of lists of tuples)

entity detection

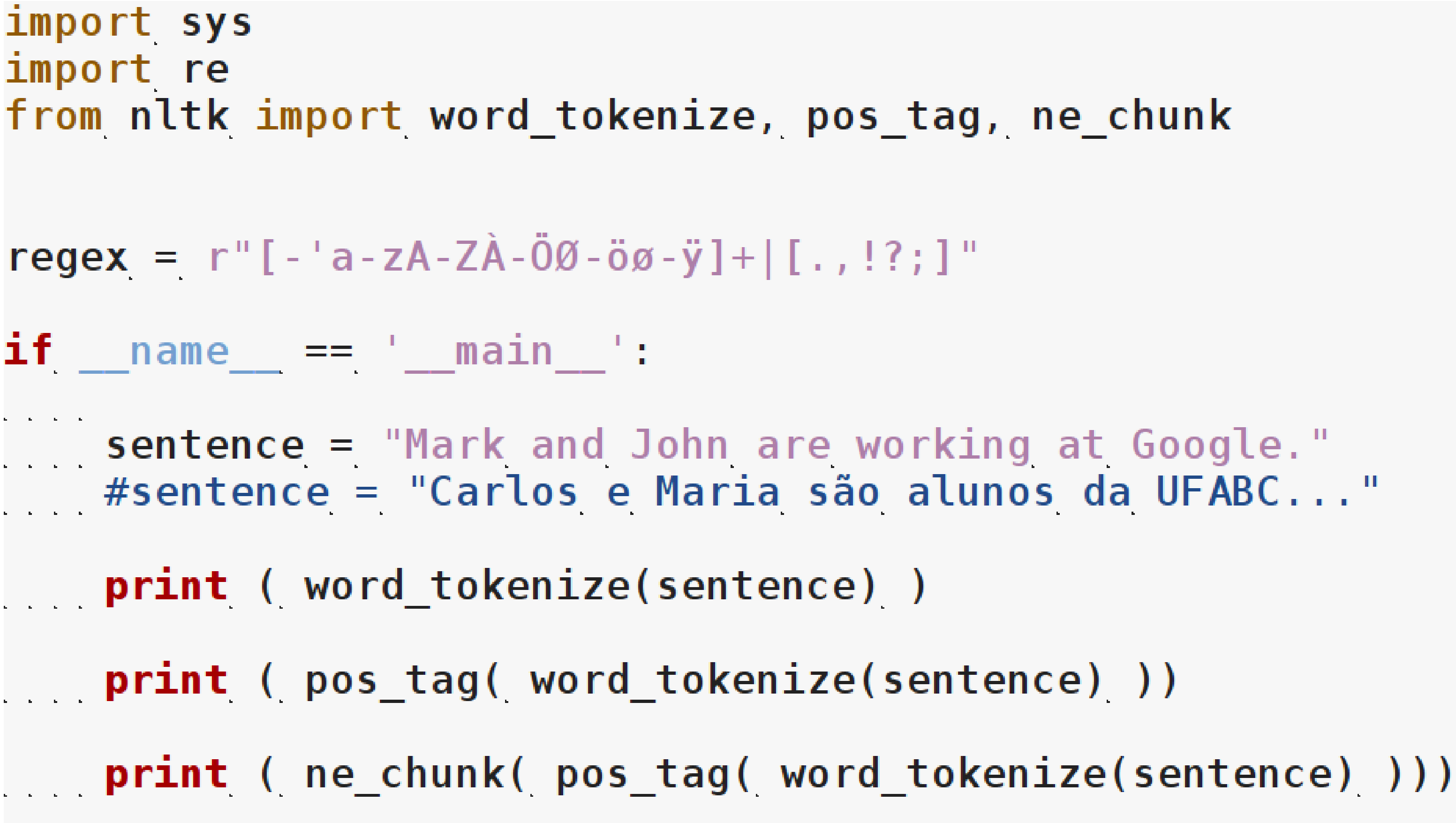
chunked sentences
(list of trees)

relation detection

relations
(list of tuples)

teste3.py

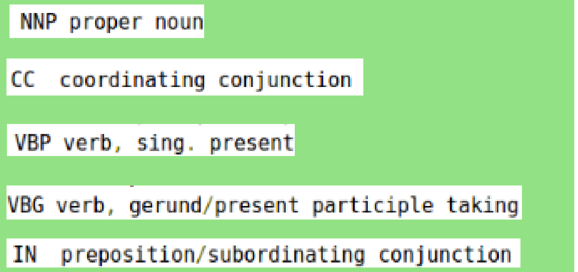
pos_tag = part of speech tagger
ne_chuck = named entity (tree)



```python
import sys
import re
from nltk import word_tokenize, pos_tag, ne_chunk


regex = r"[-'a-zA-ZÀ-ÖØ-öø-ÿ]+|[.,!?;]"

if __name__ == '__main__':

    sentence = "Mark and John are working at Google."
    #sentence = "Carlos e Maria são alunos da UFABC..."

    print ( word_tokenize(sentence) )

    print ( pos_tag( word_tokenize(sentence) ))

    print ( ne_chunk( pos_tag( word_tokenize(sentence) )))
```

Pipeline comum: reconhecimento de EN

```
python3 teste3.py

['Mark', 'and', 'John', 'are', 'working', 'at', 'Google', '.']


[('Mark', 'NNP'), ('and', 'CC'), ('John', 'NNP'),
('are', 'VBP'), ('working', 'VBG'), ('at', 'IN'),
('Google', 'NNP'), ('.', '.')]



(S
  (PERSON Mark/NNP)
  and/CC
  (PERSON John/NNP)
  are/VBP
  working/VBG
  at/IN
  (ORGANIZATION Google/NNP)
  ./.)
```



NNP proper noun

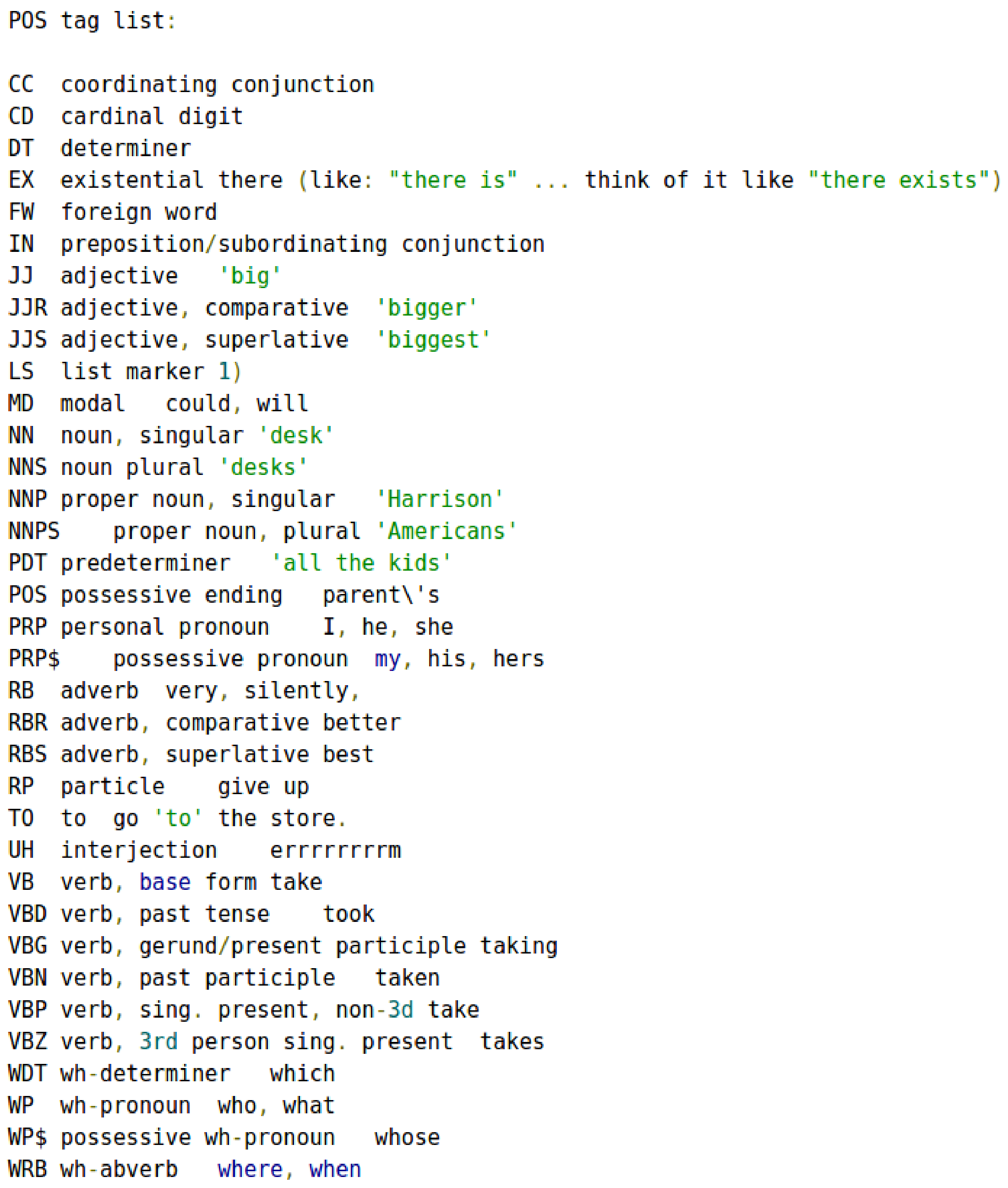
CC coordinating conjunction

VBP verb, sing. present

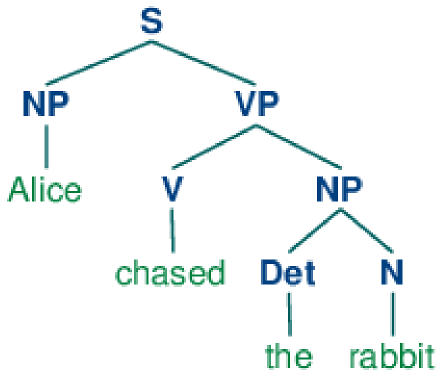
VBG verb, gerund/present participle taking

IN preposition/subordinating conjunction

Part of speech TAG

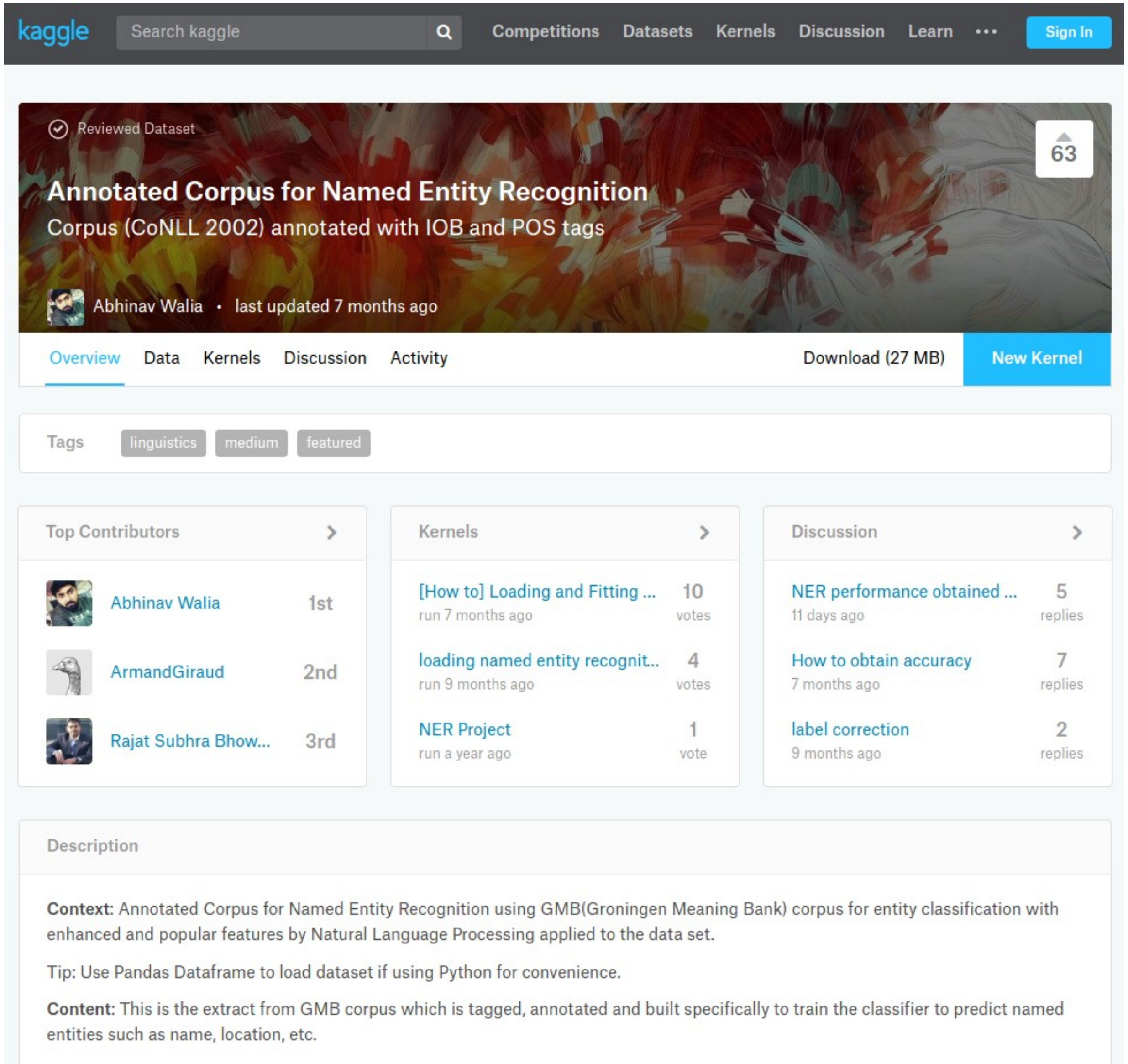


```
POS tag list:

CC  coordinating conjunction
CD  cardinal digit
DT  determiner
EX  existential there (like: "there is" ... think of it like "there exists")
FW  foreign word
IN  preposition/subordinating conjunction
JJ  adjective   'big'
JJR adjective, comparative  'bigger'
JJS adjective, superlative  'biggest'
LS  list marker 1)
MD  modal   could, will
NN  noun, singular 'desk'
NNS noun plural 'desks'
NNP proper noun, singular   'Harrison'
NNPS    proper noun, plural 'Americans'
PDT predeterminer   'all the kids'
POS possessive ending   parent\'s
PRP personal pronoun    I, he, she
PRP$    possessive pronoun  my, his, hers
RB  adverb  very, silently,
RBR adverb, comparative better
RBS adverb, superlative best
RP  particle    give up
TO  to  go 'to' the store.
UH  interjection    errrrrrrrm
VB  verb, base form take
VBD verb, past tense    took
VBG verb, gerund/present participle taking
VBN verb, past participle   taken
VBP verb, sing. present, non-3d take
VBZ verb, 3rd person sing. present  takes
WDT wh-determiner   which
WP  wh-pronoun  who, what
WP$ possessive wh-pronoun   whose
WRB wh-abverb   where, when
```



Recursos disponíveis

Corpus para treinamento de um reconhecedor de EN



Annotated Corpus for Named Entity Recognition

Corpus (CoNLL 2002) annotated with IOB and POS tags

Abhinav Walia • last updated 7 months ago

Tags: linguistics, medium, featured

Top Contributors

Abhinav Walia	1st
ArmandGiraud	2nd
Rajat Subhra Bhow...	3rd

Kernels

[How to] Loading and Fitting ... run 7 months ago	10 votes
loading named entity recognit.. run 9 months ago	4 votes
NER Project run a year ago	1 vote

Discussion

NER performance obtained ... 11 days ago	5 replies
How to obtain accuracy 7 months ago	7 replies
label correction 9 months ago	2 replies

Description

Context: Annotated Corpus for Named Entity Recognition using GMB(Groningen Meaning Bank) corpus for entity classification with enhanced and popular features by Natural Language Processing applied to the data set.

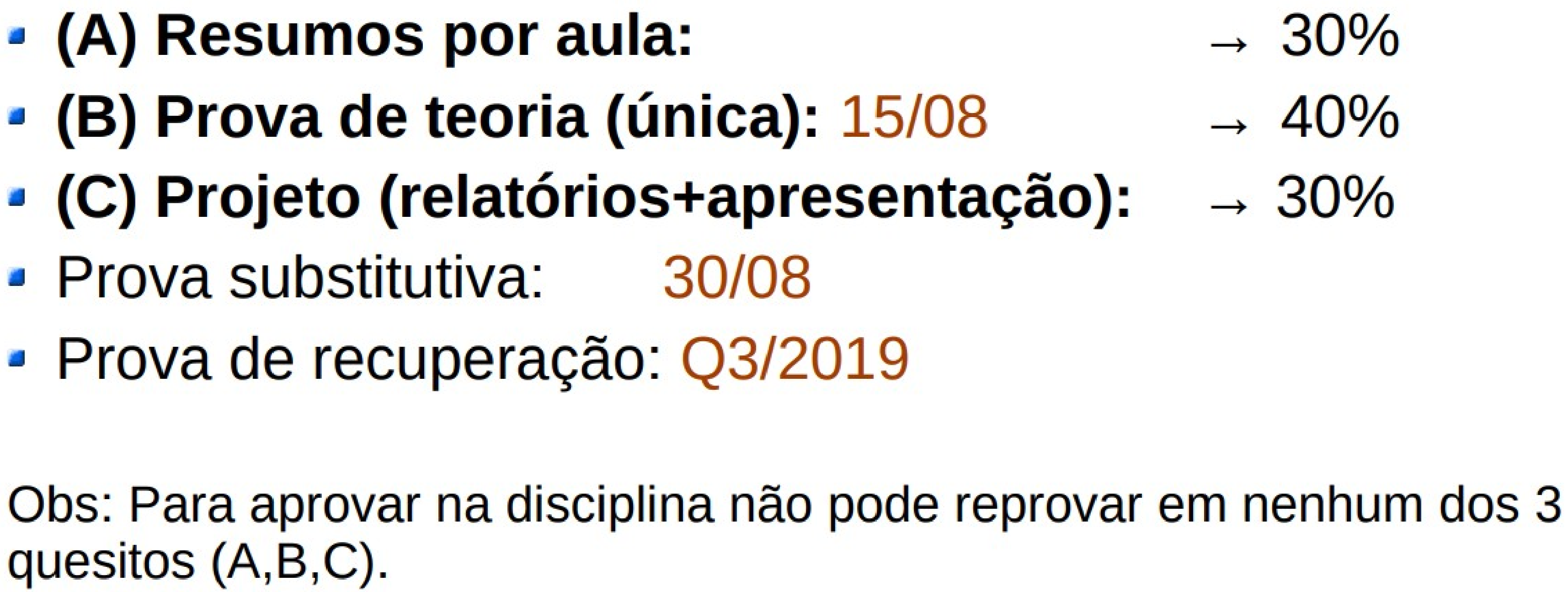
Tip: Use Pandas Dataframe to load dataset if using Python for convenience.

Content: This is the extract from GMB corpus which is tagged, annotated and built specifically to train the classifier to predict named entities such as name, location, etc.

https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus

Sobre a entrega 2: Projeto

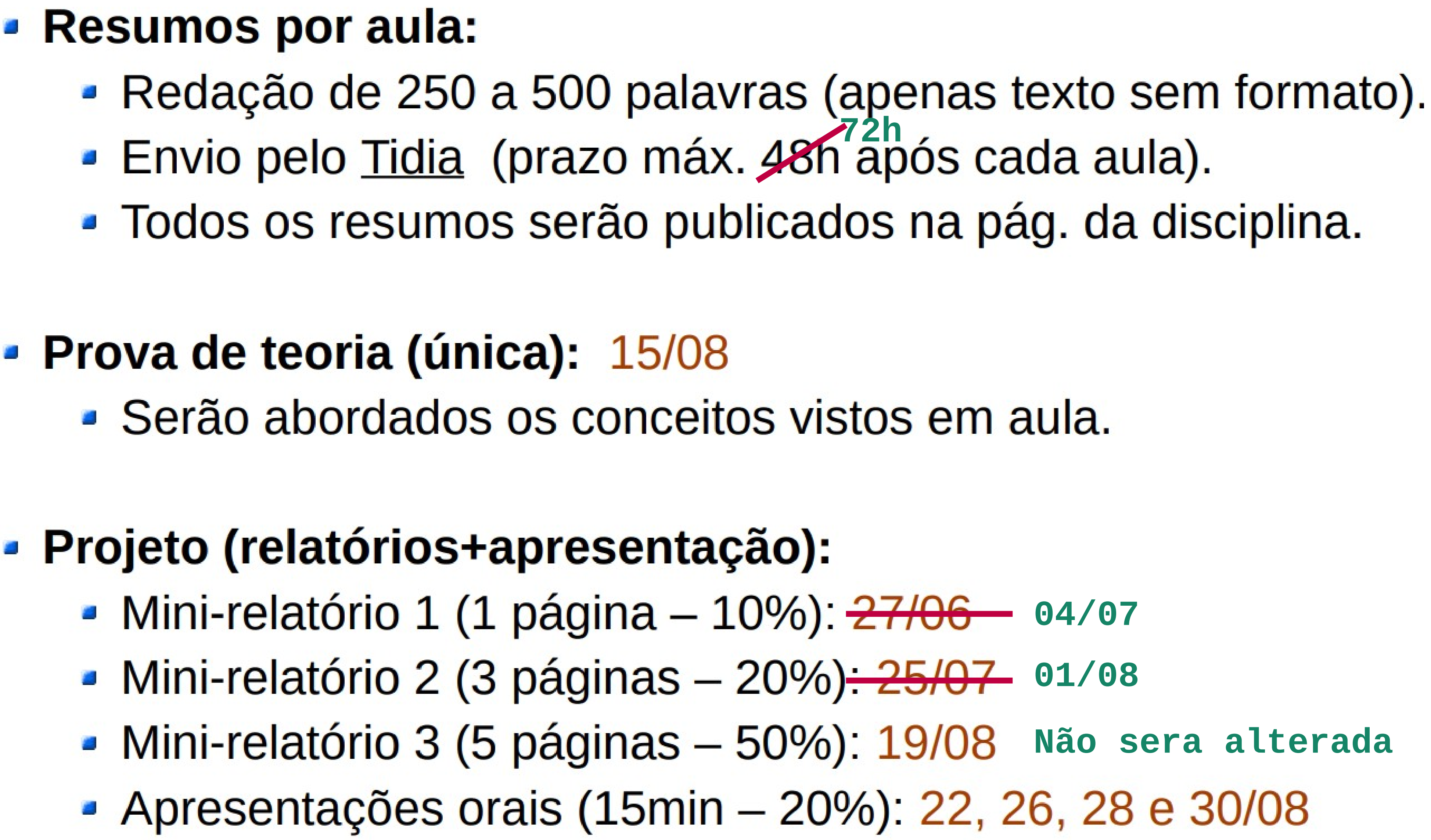
Sobre a avaliação



- **(A) Resumos por aula:** → 30%
- **(B) Prova de teoria (única):** 15/08 → 40%
- **(C) Projeto (relatórios+apresentação):** → 30%
- Prova substitutiva: 30/08
- Prova de recuperação: Q3/2019

Obs: Para aprovar na disciplina não pode reprovar em nenhum dos 3 quesitos (A,B,C).

Sobre a avaliaç]ao



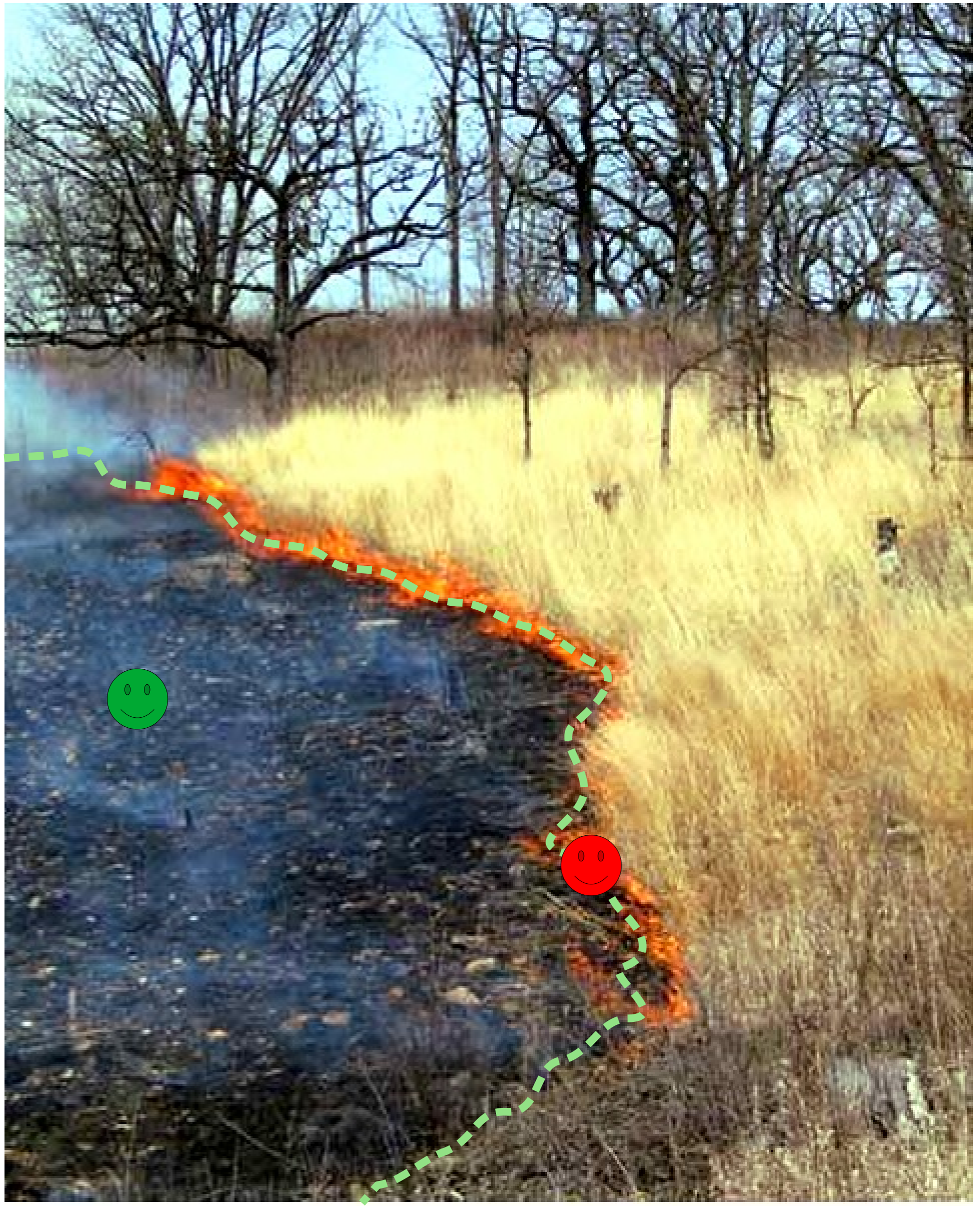
- **Resumos por aula:**
  - Redação de 250 a 500 palavras (apenas texto sem formato).
  - Envio pelo Tidia (prazo máx. ~~48h~~ 72h após cada aula).
  - Todos os resumos serão publicados na pág. da disciplina.

- **Prova de teoria (única):** 15/08
  - Serão abordados os conceitos vistos em aula.

- **Projeto (relatórios+apresentação):**
  - Mini-relatório 1 (1 página – 10%): ~~27/06~~ 04/07
  - Mini-relatório 2 (3 páginas – 20%): ~~25/07~~ 01/08
  - Mini-relatório 3 (5 páginas – 50%): 19/08 Não sera alterada
  - Apresentações orais (15min – 20%): 22, 26, 28 e 30/08

Sobre o projeto

Estado-da-arte →



(*) Figura - Fonte: http://ominho.pt/arcos-valdevez-estado-emergencia/

Sobre a entrega 2

- Preenchimento do formulário (apenas pelo aluno(a) representante do projeto).

- Deve dar maior ênfase:
  - À real proposta do artigo:

    O que propuseram os autores?
  - À parte que está sendo implementada

    O que e como está sendo implementado?

- Listar as limitações frente à ideia proposta pelos autores do artigo.