Modelos de Linguagem e Regex Aplicados

1 - Introdução a modelos de linguagem e Regex

1.1 - Modelos de linguagem

Modelo de linguagem

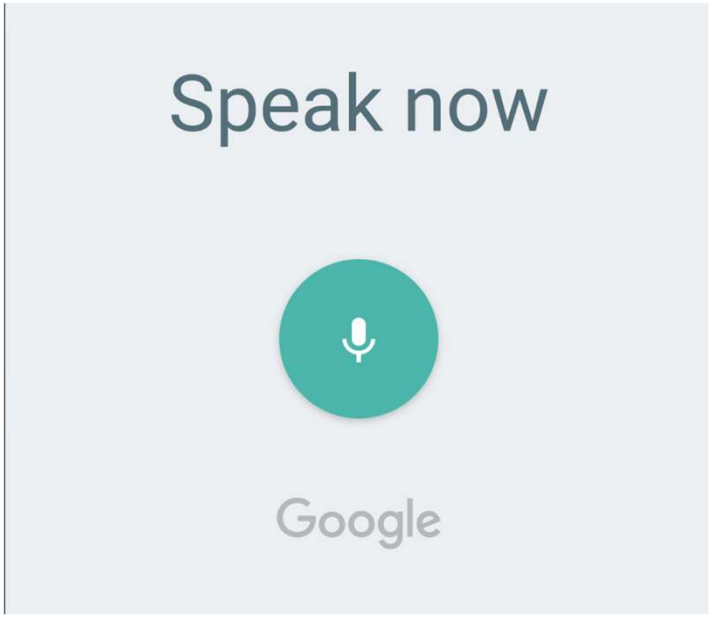
Formalmente são as regras da gramática paras as palavras e sentenças corretas em uma linguagem

- Ortografia
- Sintaxe
- etc

Em NLP podemos ter um modelo de probabilidade que avalia se uma sequência de caracteres ou palavras faz parte de uma linguagem

- cachrro (escrito errado mesmo) é mais provável que seja uma palavra de que linguagem?
- “feliz aniversário” é mais provável que seja uma frase de que linguagem? Tem outra linguagem que seja possível?

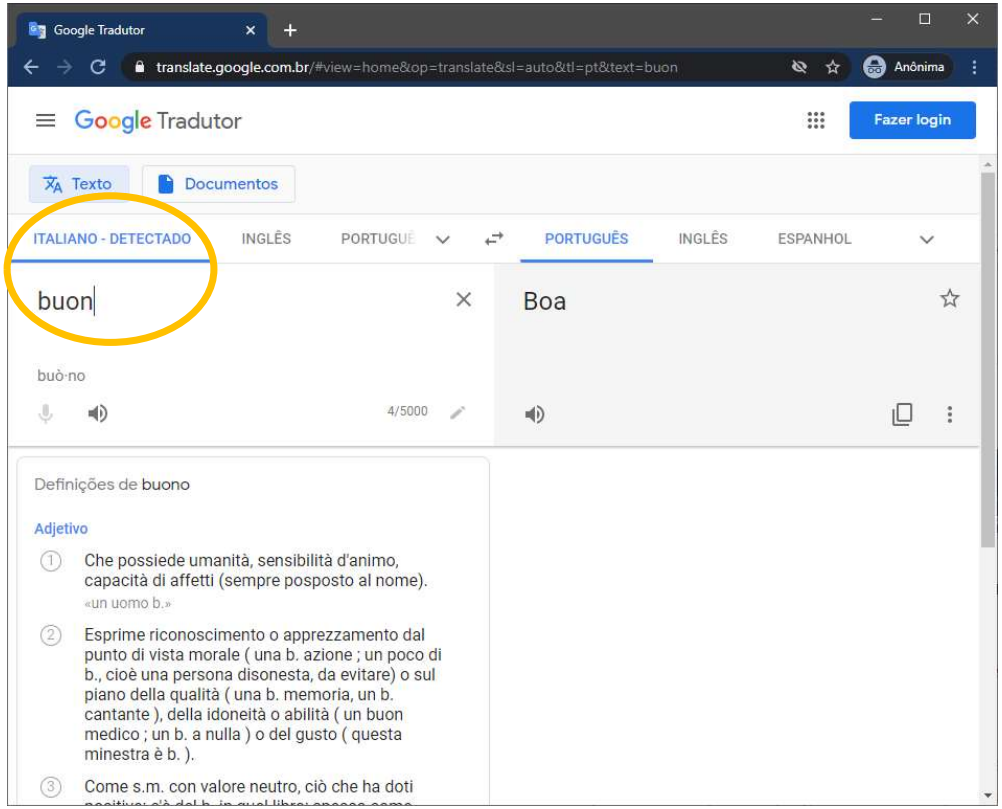
Reconhecimento de fala

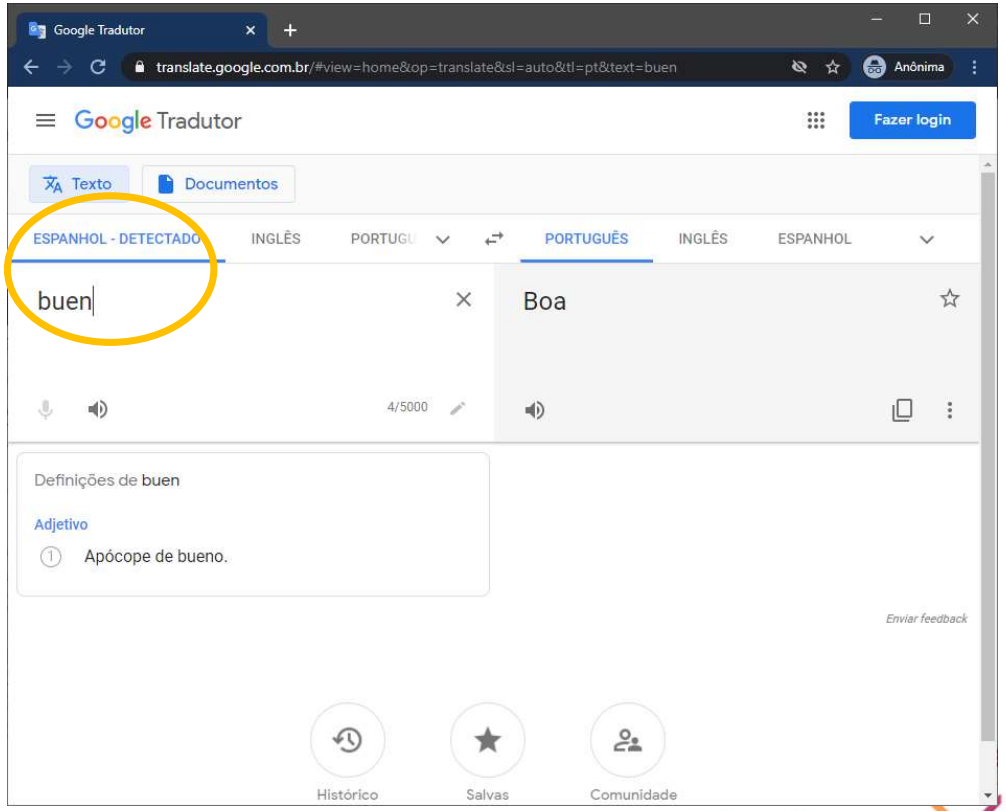


O som pode ser parecido para a máquina, mas qual é mais provável?

- Qual o comprimento do cavalheiro?
- Qual o cumprimento do cavaleiro?

Tradutor





Unyleya EDUCACIONAL

Autocompletar e autocorrigir

Boa tarde!

Espero que esteja tudo bem com você e sua família.
Estamos retomando aquisições para desenvolvimento de projetos.
Pode nos enviar uma cotação de um novo acoplamento?

Respeitosamente

Galdir Reges

```
states = df.state.unique()
statesClean = []
for s in states:
    if s != 'TOTAL':
        statesClean.ap
# df.drop
```

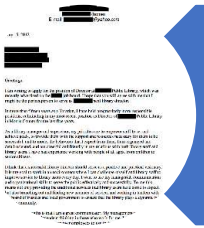
abc app
abc append
abc 3FnGE8fJT3GXwEOngsV7Zt27NXFoaoApmYm81iuXoPkFOJw…

Podem ser modelos de linguagem de grupos ou pessoas especificas, com suas manias!

Podem modelos de escrita de linguagens de programação!

Unyleya EDUCACIONAL

Amostras de linguagem



Para aprender modelos de linguagem precisamos de exemplos de linguagem.

Exemplos de livros podem representar a linguagem de seus escritores e suas épocas!

Exemplos de fóruns de discussão podem representar a linguagem comum na internet.

Unyleya EDUCACIONAL

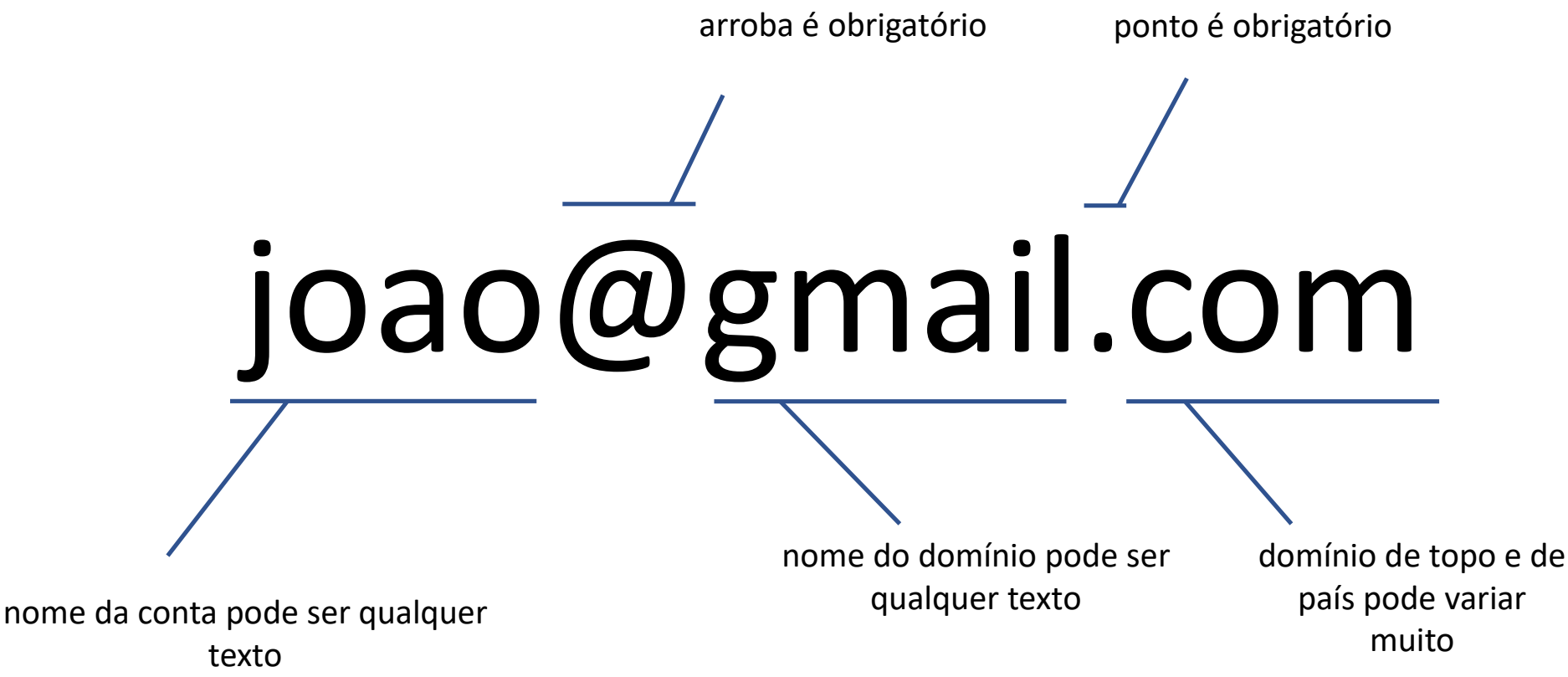
1.2 - Regex

Regex para tratar amostras

[^]*?@[^]*?\.[^]*

- Regex significa expressão regular.
- Pode ser considerada uma pequena linguagem de programação especializada
- Uma Regex é usada para usada para encontrar padrões em um texto.
  - HTML tem padrão: <h1>Seja Bem Vindo!</h1>
  - Emails tem padrão: joao@gmail.com

Unyleya EDUCACIONAL



Unyleya EDUCACIONAL

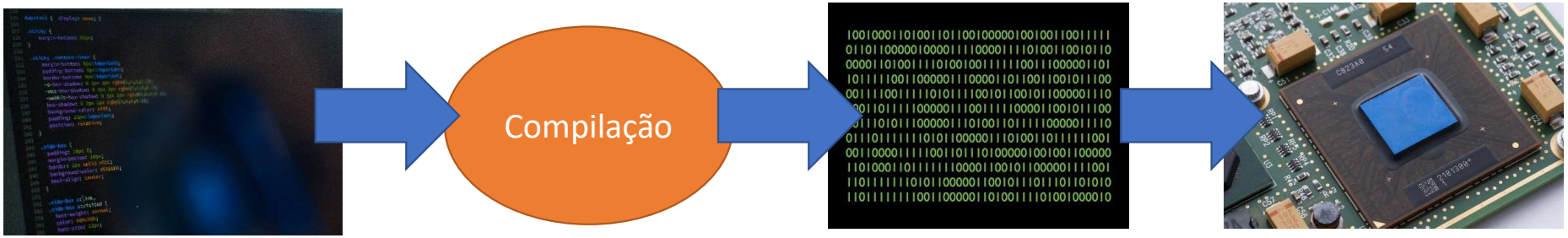
Metacaracteres em Regex

. ^ $ * + ? { } [ ] \ | ( )

2.1 – Acelerando Regex

Padrões Regex compilados

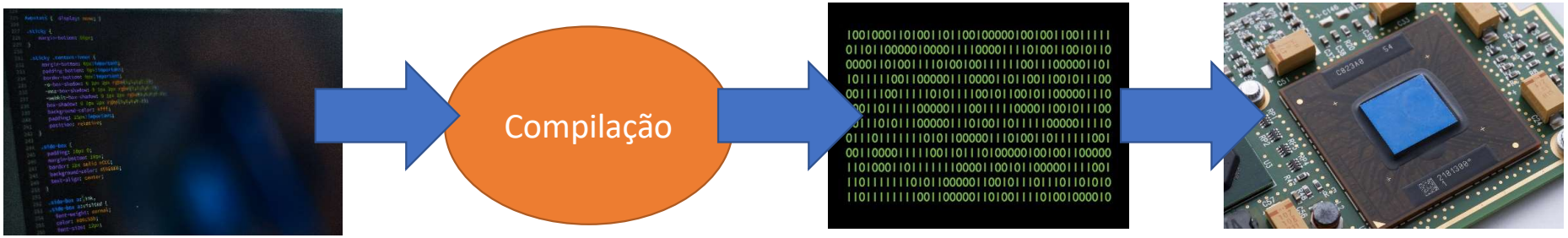
Padrões Regex no Python são compilados antes de serem executados.



Padrões Regex compilados

Para testes simples podemos usar as funções da biblioteca re diretamente, enviando um padrão Regex e uma string:

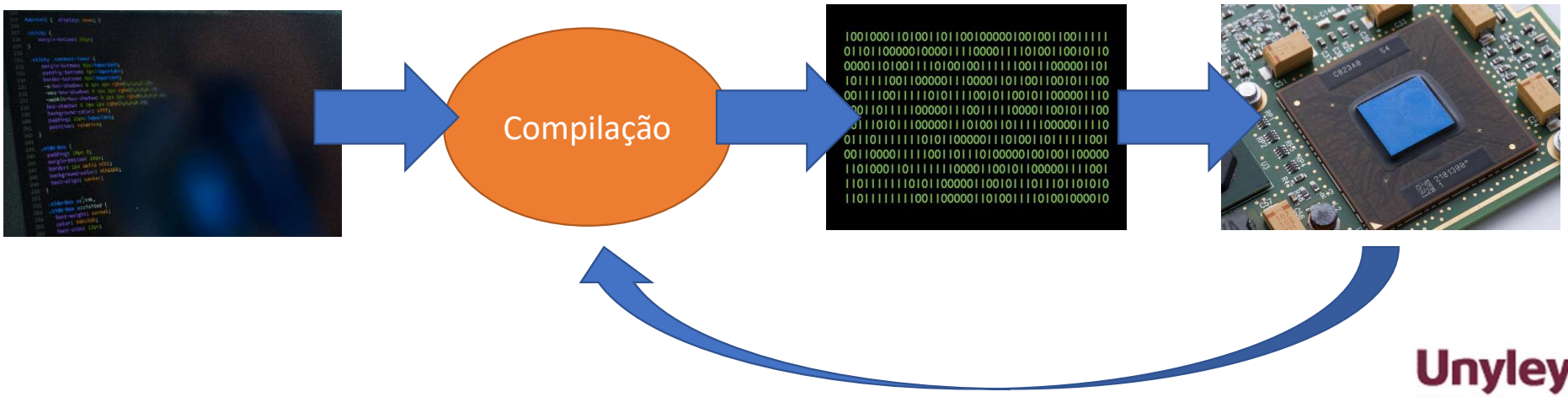
- re.findall(“<.*?>”, “<h1> Seja Bem Vindo </h1>”)
  - Resulta em [‘<h1>’,’</h1>’]



Padrões Regex compilados

Para processar grandes corpus isso não é eficiente:

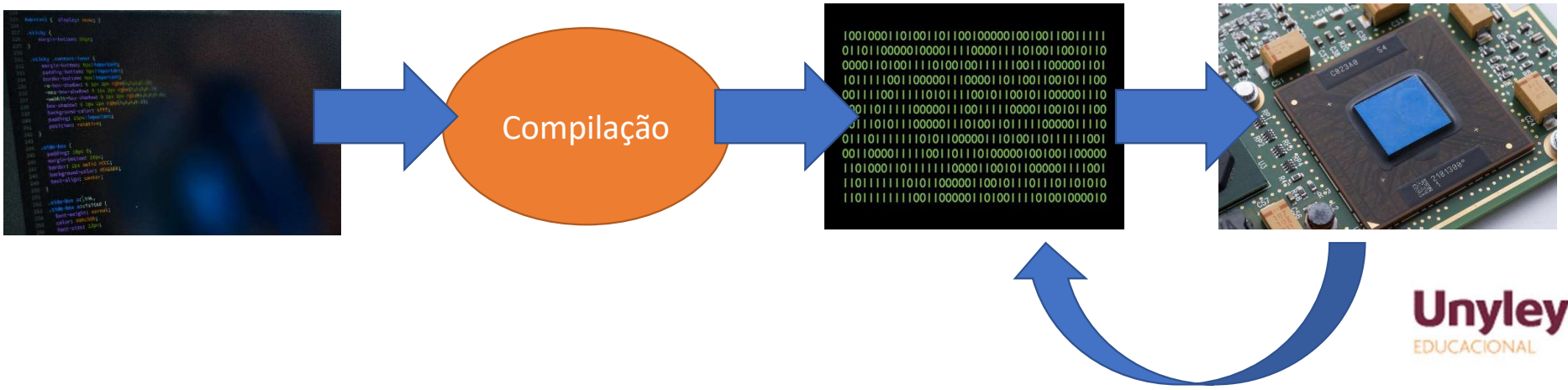
- for texto in textos:
  - trecho=re.findall(“<.*?>”, texto)



Regex compilada

A prática recomendada é fazer uma pequena pré-compilação do padrão Regex antes de usa-lo repetidamente:

- padrao=re.compile(“<.*?>”)
- for texto in textos:
  - trecho=padrao.findall(texto)



Unyleya
EDUCACIONAL

3.1 - Modelos de linguagem no NLTK

Modelo de linguagem

Qual a probabilidade da seguinte sequência de caracteres ser uma palavra em português?

- comédia

E escrito errado com abaixo?

- comedia
- P(comédia|português) > P(comedia|português)

E em inglês?

P(comedia|inglês) > P(comedia|português)

Modelo de linguagem

Qual a probabilidade da seguinte sequência bigramas ser uma palavra em português?

- P(me|co) =?
- P(ia|ed) =?

O modelo de linguagem vai ser ajustado com essas probabilidades a depender do corpus!

Lorem ipsum dolor si

Integer aliquam sem vitae ipsum vehicula e fringilla

Donec et nisi lorem, sed rhoncus odio. Pellentesque est nulla, commodo id accumsan ac, volutpat quis ligula. Nulla libero felis, venenatis id varius in, vehicula eu lectus. Suspendisse porttitor odio in massa luctus viverra interdum eros hendrerit. porttitor volutpat quis .

Mauris molestie consequat vulputate. Donec dignissim tempus suscipit. Sed tempor malesuada molestie. Etiam ullamcorper, orci vitae blandit malesuada, lacus orci consequat dolor, id adipiscing magna quam eu felis. Nunc sit amet turpis nisl. Cras nulla turpis, imperdiet non hendrerit vitae, ullamcorper varius ligula. Ut lacinia, risus sit amet sodales cursus,

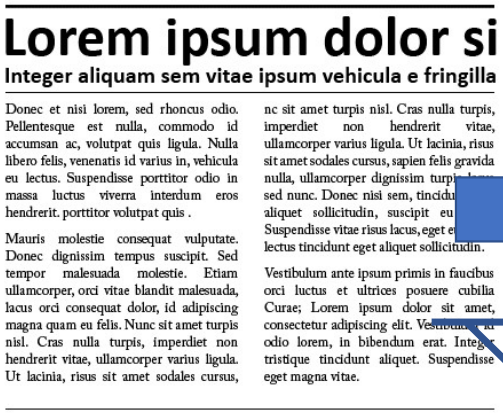
nc sit amet turpis nisl. Cras nulla turpis, imperdiet non hendrerit vitae, ullamcorper varius ligula. Ut lacinia, risus sit amet sodales cursus, sapien felis gravida nulla, ullamcorper dignissim turpis lacus sed nunc. Donec nisi sem, tincidunt eget aliquet sollicitudin, suscipit eu nulla. Suspendisse vitae risus lacus, eget euismod lectus tincidunt eget aliquet sollicitudin.

Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum id odio lorem, in bibendum erat. Integer tristique tincidunt aliquet. Suspendisse eget magna vitae.

Unyleya
EDUCACIONAL

4.1 – Fluxo de treinamento de modelo de linguagem

Pré-processamento padrão



Lorem ipsum dolor si

Integer aliquam sem vitae ipsum vehicula e fringilla

Donec et nisi lorem, sed rhoncus odio. Pellentesque est nulla, commodo id accumsan ac, volutpat quis ligula. Nulla libero felis, venenatis id varius in, vehicula eu lectus. Suspendisse porttitor odio in massa luctus viverra interdum eros hendrerit. porttitor volutpat quis .

Mauris molestie consequat vulputate. Donec dignissim tempus suscipit. Sed tempor malesuada molestie. Etiam ullamcorper, orci vitae blandit malesuada, lacus orci consequat dolor, id adipiscing magna quam eu felis. Nunc sit amet turpis nisl. Cras nulla turpis, imperdiet non hendrerit vitae, ullamcorper varius ligula. Ut lacinia, risus sit amet sodales cursus,

nc sit amet turpis nisl. Cras nulla turpis, imperdiet non hendrerit vitae, ullamcorper varius ligula. Ut lacinia, risus sit amet sodales cursus, sapien felis gravida nulla, ullamcorper dignissim turpis
sed nunc. Donec nisi sem, tincidu
aliquet sollicitudin, suscipit eu
Suspendisse vitae risus lacus, eget e
lectus tincidunt eget aliquet sollicitudin.

Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibu
odio lorem, in bibendum erat. Integ
tristique tincidunt aliquet. Suspendisse
eget magna vitae.

['<s>', 'a', 'b', 'c', '</s>']

pad_both_ends()

[('<s>', 'a'), ('a', 'b'), ('b', 'c'), ('c', '</s>')]

everygrams()

Unyleya
EDUCACIONAL

Pré-processamento padrão

Lorem ipsum dolor si

Integer aliquam sem vitae ipsum vehicula e fringilla

Donec et nisi lorem, sed rhoncus odio. Pellentesque est nulla, commodo id accumsan ac, volutpat quis ligula. Nulla libero felis, venenatis id varius in, vehicula eu lectus. Suspendisse porttitor odio in massa luctus viverra interdum eros hendrerit. porttitor volutpat quis .

Mauris molestie consequat vulputate. Donec dignissim tempus suscipit. Sed tempor malesuada molestie. Etiam ullamcorper, orci vitae blandit malesuada, lacus orci consequat dolor, id adipiscing magna quam eu felis. Nunc sit amet turpis nisl. Cras nulla turpis, imperdiet non hendrerit vitae, ullamcorper varius ligula. Ut lacinia, risus sit amet sodales cursus, nc sit amet turpis nisl. Cras nulla turpis, imperdiet non hendrerit vitae, ullamcorper varius ligula. Ut lacinia, risus sit amet sodales cursus, sapien felis gravida nulla, ullamcorper dignissim turpis lacus sed nunc. Donec nisi sem, tincidunt eget aliquet sollicitudin, suscipit eu nulla. Suspendisse vitae risus lacus, eget euismod lectus tincidunt eget aliquet sollicitudin.

Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum id odio lorem, in bibendum erat. Integer tristique tincidunt aliquet. Suspendisse eget magna vitae.

padded_everygram_pipeline()

[('<s>', 'a'), ('a', 'b'), ('b', 'c'), ('c', '</s>')]

Unyleya EDUCACIONAL

4.2 – Perplexidade

Probabilidade de um texto

Um bom modelo estima uma alta probabilidade para todos os ngramas de um texto *x* legitimo de sua linguagem.

Isso pode ser medido com entropia cruzada:

$$H(x) = -\frac{1}{N}\sum_{i=1}^{N} log_2\, p(x_i)$$

Ou com perplexidade:

$$perplexidade(L) = 2^{H(x)}$$

Em NLP usamos perplexidade, e quanto menor a probabilidade maior a perplexidade.

5.1 – Modelo de Laplace para lidar com perplexidade infinita

Perplexidade infinita

Caso a probabilidade uma sequencia, por exemplo yxwnvkm, seja 0:

$$p(x) = 0$$

A perplexidade vai ser infinita:

$$perplexidade(x) = \text{infinita}$$

E não podemos fazer computação com valores infinitos! O que fazer?

Modelo Laplace

Modelo Laplace usa uma suavização chamada de soma-um.

Para todo possível ngram será adicionada uma amostra de treino, o que faz a probabilidade de qualquer ngram ser diferente de zero.

- from nltk.lm import Laplace

Unyleya EDUCACIONAL

Obrigado!