Tokenization

John Gamboa

What is tokenization?

What is tokenization?

Word segmentation is the problem of dividing a string of written language into its component words.

What is tokenization?

Word segmentation is the problem of dividing a string of written language into its component words.

(Source: Wikipedia -- *if you search for "Tokenization", you will fall into "Word segmentation")*:

What is tokenization?

Word segmentation is the problem of dividing a string of written language into its component words.

(Source: Wikipedia -- *if you search for "Tokenization", you will fall into "Word segmentation")*:

For example

What is tokenization?

> *Word segmentation is the problem of dividing a string of written language into its component words.*
>
> (Source: Wikipedia -- *if you search for "Tokenization", you will fall into "Word segmentation")*:

For example

```
"The woman drank her coffee"
```

What is tokenization?

Word segmentation is the problem of dividing a string of written language into its component words.

(Source: Wikipedia -- *if you search for "Tokenization", you will fall into "Word segmentation")*:

For example

```
"The woman drank her coffee"

"woman", "coffee", "drank", "her", "The"
```

What is tokenization?

Word segmentation is the problem of dividing a string of written language into its component words.

(Source: Wikipedia -- *if you search for "Tokenization", you will fall into "Word segmentation")*:

For example

```
"The woman drank her coffee"
```

```
"woman", "coffee", "drank", "her", "The"
```

```
"The", "woman", "drank", "her", "coffee"
```

Tokens and Types

Tokens and Types

```
The woman saw another woman.
```

Tokens and Types

```
The woman saw another woman.
```

How many words?

Tokens and Types

```
The woman saw another woman.
```

How many words?

How many tokens?

Tokens and Types

```
The woman saw another woman.
```

How many words?

How many tokens?

How many types?

Tokens and Types

```
The woman saw the other woman.
```

How many words?

How many tokens?

How many types?

Tokens and Types

```
The woman saw the other woman.
```

How many words?

How many tokens?

How many types?

Hapax Legomenon: a *type* that appears *only once*

Why to learn about Tokenization?

Why to learn about Tokenization?

You know `nltk.word_tokenize()`

Why to learn about Tokenization?

You know `nltk.word_tokenize()`

Mostly, use it =)

Why to learn about Tokenization?

You know `nltk.word_tokenize()`

Mostly, use it =)

What if you want a different behavior?

Last words on Tokenization

Last words on Tokenization

English tokenization is “easy”

Last words on Tokenization

English tokenization is “easy” → use spaces

Last words on Tokenization

Last words on Tokenization

Other languages may be way harder…

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages (Latin)

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages

(Latin)

NEQVEPORROQVISQVAMESTQVIDOLOREMIPSVMQVIADOLORSITAMETCONSECTETVRADIPISCIVELIT

NEQVE•PORRO•QVISQVAM•EST•QVI•DOLOREM•IPSVM•QVIA•DOLOR•SIT•AMET•
CONSECTETVR•ADIPISCI•VELIT

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages

(Latin)

NEQVEPORROQVISQVAMESTQVIDOLOREMIPSVMQVIADOLORSITAMETCONSECTETVRADIPISCIVELIT

NEQVE•PORRO•QVISQVAM•EST•QVI•DOLOREM•IPSVM•QVIA•DOLOR•SIT•AMET•
CONSECTETVR•ADIPISCI•VELIT

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages

(Latin)

NEQVEPORROQVISQVAMESTQVIDOLOREMIPSVMQVIADOLORSITAMETCONSECTETVRADIPISCIVELIT

NEQVE•PORRO•QVISQVAM•EST•QVI•DOLOREM•IPSVM•QVIA•DOLOR•SIT•AMET•
CONSECTETVR•ADIPISCI•VELIT

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit

Nobody likes pain for its own sake, or looks for it and wants to have it, just because it is pain

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese

(example from Wikipedia)

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese

(example from Wikipedia)

北京在中国北方;广州在中国南方。

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese

(example from Wikipedia)

北京在中国北方;广州在中国南方。
北京　在　中国　北方；广州　在　中国　南方。

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese

(example from Wikipedia)

北京在中国北方;广州在中国南方。
北京　在　中国　北方；广州　在　中国　南方。
Běijīng zài Zhōngguó běifāng; Guǎngzhōu zài Zhōngguó nánfāng.

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese

(example from Wikipedia)

北京在中国北方;广州在中国南方。
北京　在　中国　北方；广州　在　中国　南方。
Běijīng zài Zhōngguó běifāng; Guǎngzhōu zài Zhōngguó nánfāng.
Beijing is in Northern China; Guangzhou is in Southern China.

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese
→ used nowadays in URLs

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese
→ used nowadays in URLs

gibtesheutepommes.de

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese
→ used nowadays in URLs
→ or also hashtags

Last words on Tokenization

Other languages may be way harder…

Scriptio continua: no spaces at all → used by older Indo-European languages
→ used, e.g., in Chinese
→ used nowadays in URLs
→ or also hashtags

#brainstorm → bra in storm