

# Course 2: Tokenization

# What is tokenization?

Turning text...

```
I love playing soccer!
```

...into *tokens*

```
['I', 'love', 'play', 'ing', 'soccer', '!']
```

# Historical Notions

# ***Tokenization Origins***

The concept comes from linguistics

“ *non-empty contiguous sequence of graphemes or phonemes in a document*

≈

split on blanks

”

# Tokenization Origins

```
old_tokenize("I love playing soccer!") = ['I', 'love', 'playing', 'soccer!']
```

- Different from *word-forms* ⚠
  - *damélo* → *da/mé/lo* (=give/me/it)

# ***Tokenization Origins***

Natural language is split into...

- Sentences, utterances, documents... (*macroscopical*)  
that are split into...
    - Tokens, word-forms... (*microscopical*)
- Used for linguistic tasks (POS tagging, syntax parsing,...)

# Tokenization & ML

Machine Learning relies on tokenization:

- Gives better performance
- **Fixed-size vocabulary** often required

# Tokenization & ML

Evolution of modeling complexity w.r.t. the sequence length

Model Type	Year	Complexity
Tf-Idf	1972	$O(1)$
RNNs	~1985	$O(n)$
Transformers	2017	$O(n^2)$

→ Need for a good downsampling mechanism



