Tokenization

School of Information Studies
Syracuse University

Text Representation/ Vectorization

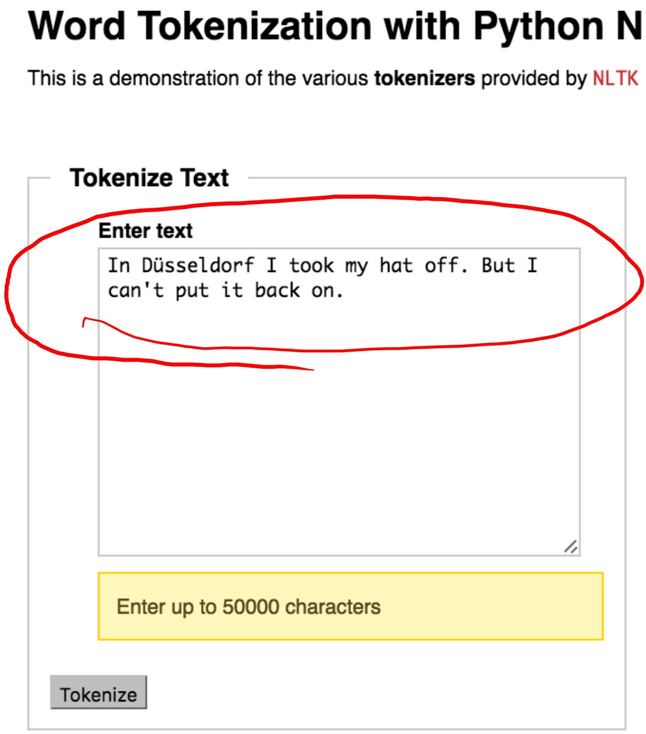
Computers can do only **one** thing, that is, **counting**!

First step toward text mining: convert text to numbers

- What to count?
- How to count?

What to Count? Tokens!

A tokenizer has a set of rules about grouping characters into tokens.



Word Tokenization with Python NLTK

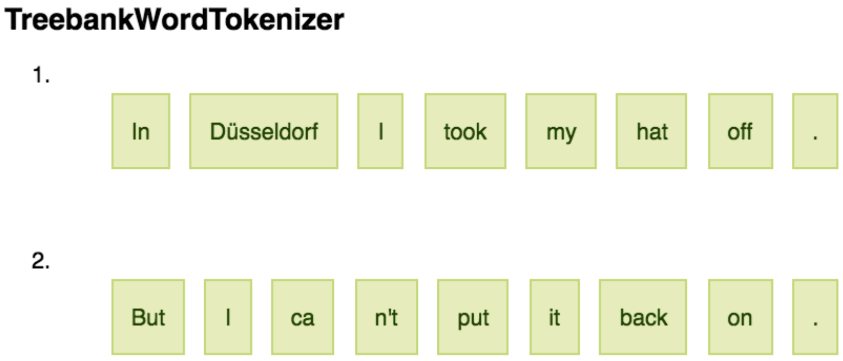
This is a demonstration of the various **tokenizers** provided by NLTK 2.0.4.

Tokenize Text

Enter text

```
In Düsseldorf I took my hat off. But I
can't put it back on.
```

Enter up to 50000 characters

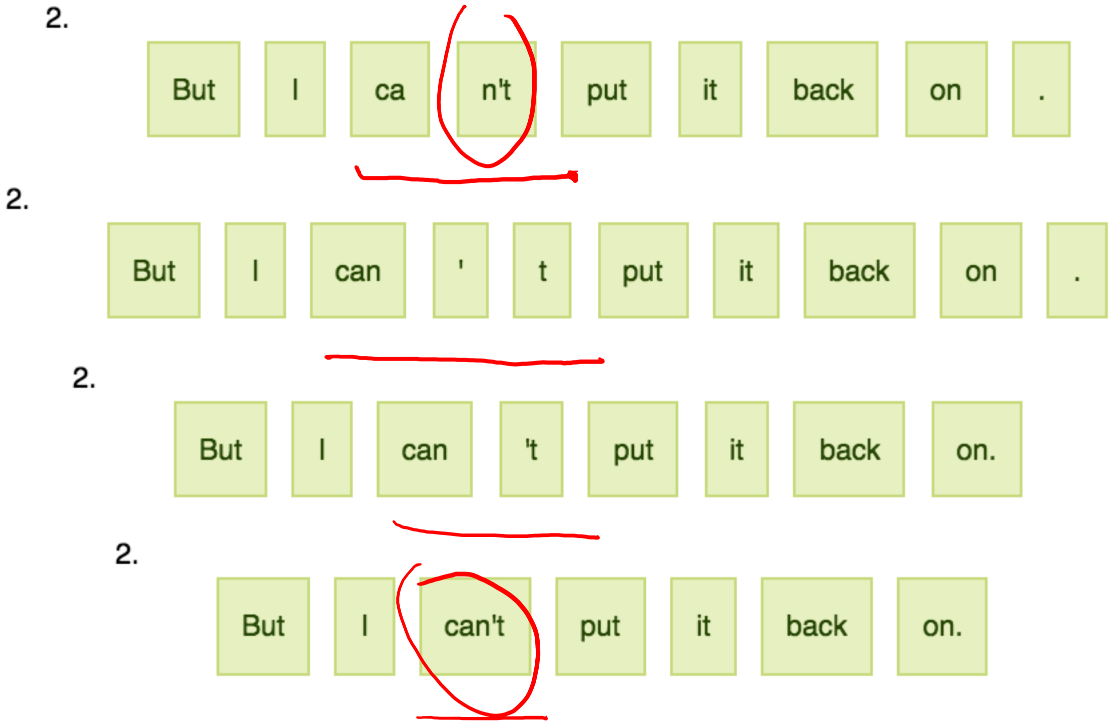


TreebankWordTokenizer

1. In | Düsseldorf | I | took | my | hat | off | .

2. But | I | ca | n't | put | it | back | on | .

Tokenization Rules Can Vary



2.	But	I	ca	n't	put	it	back	on	.	
2.	But	I	can	'	t	put	it	back	on	.
2.	But	I	can	't	put	it	back	on.		
2.	But	I	can't	put	it	back	on.			

N-Gram: Multi-Word Tokens

Bag-of-Word representation (BoW) ignores the context of words

Multi-word tokens (n-grams) can capture local context of words; e.g., "digital library"

Common n-grams:

- Uni-grams: tokens of individual words
- Bi-grams: tokens of two consecutive words
- Tri-grams: tokens of three consecutive words

Tokenization Is Not Easy

Tokenizing URLs

- Choosespain.com

Tokenization Is Not Easy

Tokenize text strings with no whitespace

Chinese (New Year couplets):

养猪大如山老鼠头头死

Raise|pigs|big|as|mountain|rats|all|die

养|猪|大|如|山|老鼠|头头|死

Raise|pigs|big|as|mountain rats, all|die

养|猪|大|如|山老鼠| 头头|死

Tokenization Is Not Easy

Lowercase vs. uppercase

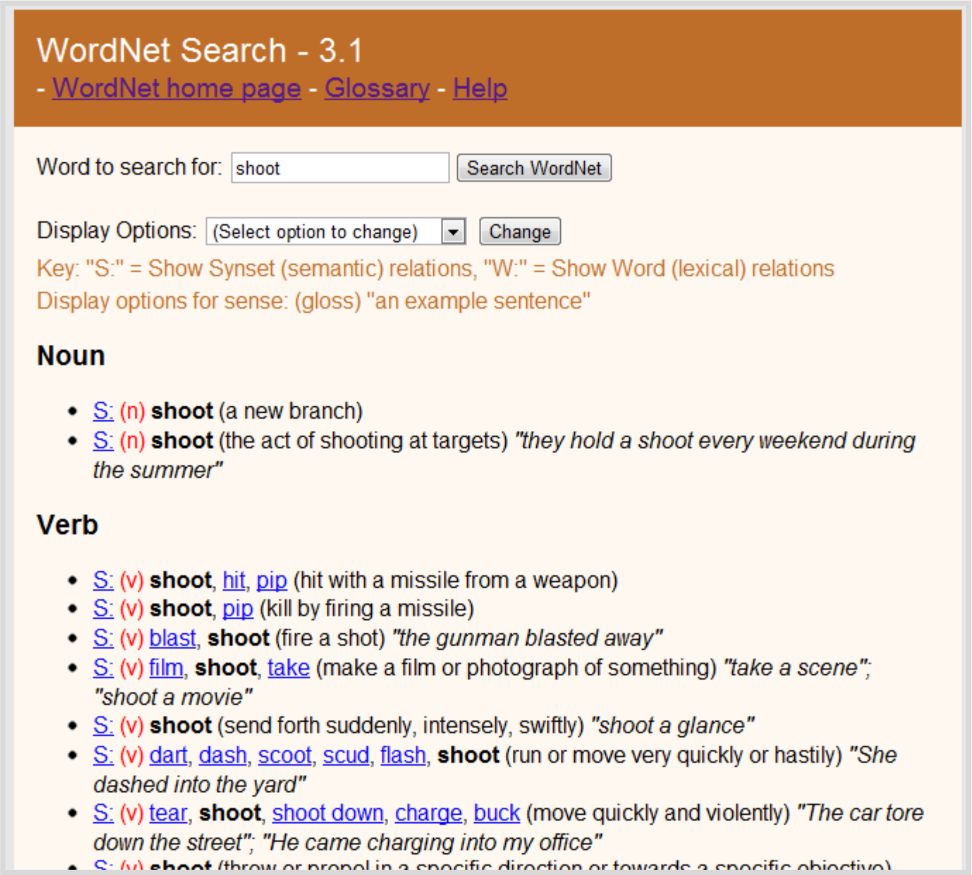
Words with inflected forms

- “dishwasher” vs. “dishwashers”

Words with multiple senses

- “There is a money **bank** near the river **bank**.”

WordNet



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: shoot

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **shoot** (a new branch)
- S: (n) **shoot** (the act of shooting at targets) *"they hold a shoot every weekend during the summer"*

Verb

- S: (v) **shoot**, hit, pip (hit with a missile from a weapon)
- S: (v) **shoot**, pip (kill by firing a missile)
- S: (v) blast, **shoot** (fire a shot) *"the gunman blasted away"*
- S: (v) film, **shoot**, take (make a film or photograph of something) *"take a scene"; "shoot a movie"*
- S: (v) **shoot** (send forth suddenly, intensely, swiftly) *"shoot a glance"*
- S: (v) dart, dash, scoot, scud, flash, **shoot** (run or move very quickly or hastily) *"She dashed into the yard"*
- S: (v) tear, **shoot**, shoot down, charge, buck (move quickly and violently) *"The car tore down the street"; "He came charging into my office"*
- S: (v) **shoot** (throw or propel in a specific direction or towards a specific objective)

Word Sense Disambiguation (WSD)

WSD techniques use word context to decide the word sense.

Could introduce more errors to next steps.

So far does not help search engines significantly.

Not widely used in text mining.