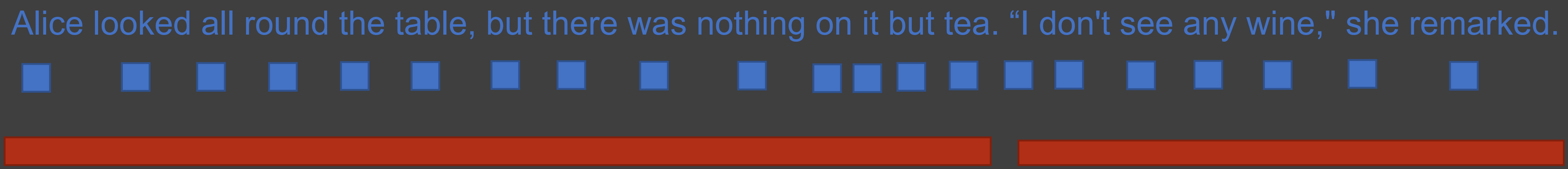
# Text Tokenization

Natalie Parde

UIC CS 421

# Text Tokenization: Step #1 in Most NLP Pipelines

- Text tokenization is critical to most NLP tasks
- A typical NLP pipeline begins by:
  - Separating words in running text
  - Normalizing word formats (e.g., favourite = favorite)
  - Segmenting sentences in running text



Alice looked all round the table, but there was nothing on it but tea. “I don't see any wine," she remarked.

# How many words?

- I do uh main- mainly business data processing
  - Fragments, filled pauses
- Seuss's **cat** in the hat is different from other **cats**!
  - **Lemma**: same stem, part of speech, rough word sense
    - cat and cats = same lemma
  - **Wordform**: the full inflected surface form
    - cat and cats = different wordforms

# How many words?

Alice looked all round the table, but there was nothing on it but tea.

- **Type**: an element of the vocabulary.
- **Token**: an instance of that type in running text.
- How many?
  - 14 tokens (or 16?)
  - 13 types (or 15?)

# How many words?

***N*** = number of tokens

***V*** = vocabulary = set of types

|*V*| is the size of the vocabulary

| **Dataset** | **Tokens = N** | **Types = \|V\|** |
|---|---|---|
| Switchboard phone conversations | 2.4 million | 20 thousand |
| Shakespeare | 884,000 | 31 thousand |
| Google N-grams | 1 trillion | 13 million |

# Issues in Tokenization

- Finland’s capital → Finland Finlands Finland’s ?
- what’re, I’m, isn’t → What are, I am, is not ?
- Hewlett-Packard → Hewlett Packard ?
- state-of-the-art → state of the art ?
- Lowercase → lower-case lowercase lower case ?
- San Francisco → one token or two?
- m.p.h., PhD. → ??

# Tokenization: Language Issues

## Contractions

- ***L'ensemble*** → one token or two?
  - ***L*** ? ***L'*** ? ***Le*** ?
  - Want ***l'ensemble*** to match with ***un ensemble***

## Tokens Not Delineated by Whitespace

- ***Lebensversicherungsgesellschaftsangestellter***
  - life insurance company employee
- 莎拉波娃现在居住在美国东南部的佛罗里达。
  - Sharapova now lives in Florida in the southeastern United States.

# Maximum Matching Word Segmentation Algorithm

Given a wordlist of Chinese and a string:

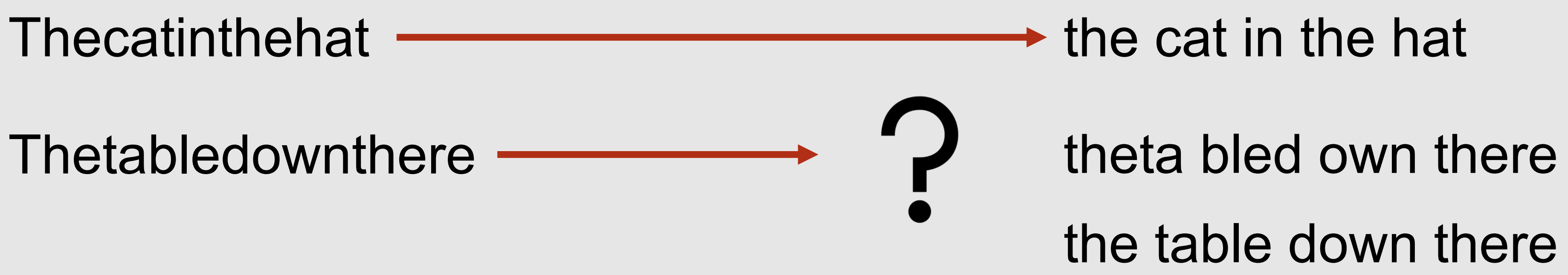
1) Start a pointer at the beginning of the string
2) Find the longest word in dictionary that matches the string starting at pointer
3) Move the pointer over the word in string
4) Go to 2

莎拉波娃现在居住在美国东南部的佛罗里达。

↓

莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达

# Doesn't generally transfer to English….



Thecatinthehat → the cat in the hat

Thetabledownthere → ?

theta bled own there

the table down there

- Nice Python tokenizers:
    - NLTK: http://www.nltk.org/api/nltk.tokenize.html
    - spaCy: https://spacy.io/api/tokenizer
    - StanfordNLP: https://stanfordnlp.github.io/stanfordnlp/