SMM694 — Week 2

Simone Santoni

Cass

May 27, 2020

Week 2

Main topics:

- Representing words and meanings
- Language modeling

Agenda

-
-
-

Models of Natural Language (NL)

What is an NL model?

How do we build an NL model?

Why should we care about NL models?

. . . let's focus on the **why** aspect first.

Why do we care about NL models?

Let's consider tokenization, a core task to any natural language processing analysis.

Now, let's apply different tokenizers to the below displayed sentence:

```
s = """Back in the golden age of hip-hop
       (the late '80s, youngsters), Rakim took
       lyricism to unfathomable heights,
       helping to usher in the wave of lethal
       MCs like Big Daddy Kane and Kool G Rap,
       who would go on to become icons. Two
       decades later, some of Ra's rhymes from
       '86 are still over people's heads: His
       wordplay remains a hip-hop measuring
       stick."""
```

Different tokenizers in action

:::::::::::::: {.columns} ::: {.column width="33%"} Naive tokenizer ::: ::: {.column width="33%"} NLTK ::: {.column width="33%"} spaCy ::::::::::::::

Naive tokenizer

Modeling natural language: Main challenges

Processing raw text intelligently is difficult:

- it's common for words that look completely different to mean almost the same thing
- the same words in a different order can mean something completely different
- most words are rare
- even splitting text into useful word-like units can be difficult in many languages (see Japanese)

Source is spaCy website

How do we build NL models?

- pre-DL models of the language
- post-DL models of the language

The architecture of Natural Language Models