Tokenization

- Whitespace tokenization
  - Segment on whitespace, compute vocabulary from top-k ranked words, add extra token for OOV words.
- Character tokenization
  - Segments on characters, very simple, but often limits performance on downstream tasks
- Subword tokenization methods
  - Byte-Pair Encoding (BPE)
    - Recursive algorithm to compute the common unicode character sequences in a dataset. Recursion stops based on frequency hyperparameter.
  - WordPiece
    - Similar to BPE, but instead of adding a sequence based on frequency alone, it normalizes frequency by the frequency of its constituent unicode character(s) / pairs.
  - Unigram Language Model
    - Starts with a complete vocabulary, and progressively shrinks it by removing tokens that result in the bottom percentile of log likelihood loss when removed.
  - SentencePiece
    - Completely agnostic to whitespace by including “\s” in the set of characters it recognizes, and is thus the only language agnostic tokenizer. It uses BPE+Unigram tokenization for subword regularization.

Introduction to Spacy

Explosion is a software company specializing in developer tools for Artificial Intelligence and Natural Language Processing. We're the makers of spaCy, one of the leading open-source libraries for advanced NLP and Prodigy, an annotation tool for radically efficient machine teaching.

Matthew Honnibal

Matthew is a leading expert in AI technology. He completed his PhD in 2009, and spent a further 5 years publishing research on state-of-the-art NLP systems. He left academia in 2014 to write spaCy and found Explosion.

Founder & Director
Berlin, Germany
matt@explosion.ai
@honnibal
honnibal

Ines Montani

Ines is a co-founder of Explosion and a core developer of the spaCy NLP library and the Prodigy annotation tool. She has helped set a new standard for user experience in developer tools for AI engineers and researchers.

Founder & Director
Berlin, Germany
ines@explosion.ai
@_inesmontani
ines

Spacy website

Spacy Courses

Spacy Github