Tokenization

Tokenization

- Process of splitting a text into smaller units(tokens)
- Smaller Units: words, numbers, symbols, characters
- Whitespace Tokenizer

Sentence: “I went to New-York to play football”
Tokens: “I”, “went”, “to”, “New-York”, “to”, “play”, “football”

- Regular Expression Tokenizer

Sentence: “Football,Cricket;Golf Tennis”
re.split(r’[;,\s]’,sentence)
Tokens: “Football”, “Cricket”, “Golf”, “Tennis”

Tokenization

Open the Jupyter Notebook

Analytics Vidhya

spaCy Tokenizer

Algorithm behind SpaCy Tokenizer:

1. Split the text on whitespaces and iterate over the whitespace-separated substrings.
2. Look for a token match. If a match, stop processing and keep it as a token.
3. Check if it is a special case like U.K., We're
4. Try to consume a prefix like $, (and go to step 2.
5. If didn't consume prefix, try to consume suffix like !, % and go to step 2.
6. If can't consume a prefix or a suffix, then check if it's an URL and keep it as a token.
7. If not an URL, then again check for the special cases.
8. Look for infixes like hyphen, slash, etc. and split substring on all infixes into tokens.
9. Can't do anything? Handle it as a token.

spaCy Tokenizer

I’m working as a Data Scientist in the U.S. and earning $140,000.

Thank You

Analytics
Vidhya