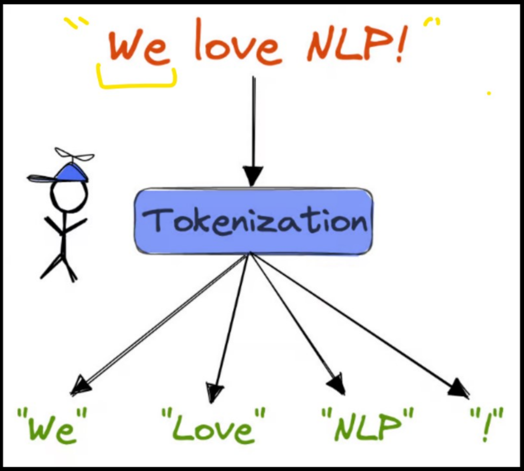
Tokenization

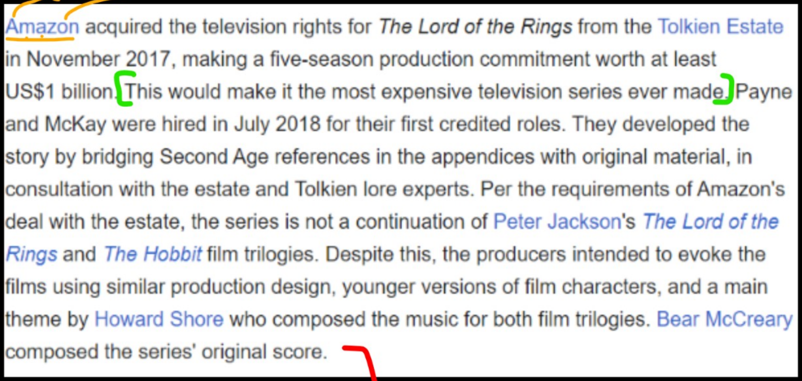


Tokenization is the process of creating tokens

Token :- builing unit of text sequence

1. characters
2. words
3. part of words
4. punctuations
5. sentences
6. regular expression

char [A-Z] — words → sentence → paragraph



Amazon acquired the television rights for *The Lord of the Rings* from the Tolkien Estate in November 2017, making a five-season production commitment worth at least US$1 billion. [This would make it the most expensive television series ever made.] Payne and McKay were hired in July 2018 for their first credited roles. They developed the story by bridging Second Age references in the appendices with original material, in consultation with the estate and Tolkien lore experts. Per the requirements of Amazon's deal with the estate, the series is not a continuation of Peter Jackson's *The Lord of the Rings* and *The Hobbit* film trilogies. Despite this, the producers intended to evoke the films using similar production design, younger versions of film characters, and a main theme by Howard Shore who composed the music for both film trilogies. Bear McCreary composed the series' original score.

Word Tokenization :- split the text data in words is called WT

sentence Tokenization :- corpus → sentences

Character Tokenization :- words → characters