Image from ChatGPT

PRE-PROCESSING

Luísa Coheur

TÉCNICO LISBOA

inesc id lisboa

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

LEARNING OBJECTIVES

LEARNING OBJECTIVES

- After this class, students should be able to:
  - Identify some pre-processing techniques
  - Be aware that they don't work in certain scenarios nowadays

TOPICS

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

TOKENIZATION

- Tokenization is the process of breaking down a stream of text into smaller, manageable units called tokens
- The goal is to create tokens that retain meaningful linguistic information while making the text more accessible for computational models

TOKENIZATION

- Consider the word “cats”. Which is the best input to a machine?
  - c + a + t + s
  - cats
  - cat + s
  - …

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

TOKENIZATION: CHARACTER-LEVEL

- Character-level tokenization tokenizes text by splitting it into individual characters
- Example:
  - Input: hello
  - Output: ['h', 'e', 'l', 'l', ‘o’]

- + Can manage out-of-vocabulary (OOV) elements
- - Each token (character) carries very little context

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

TOKENIZATION: WORD-LEVEL

- Word-level tokenization splits text into individual words based, for instance, on spaces and punctuation
- Example:
  - Mr. Smith finished his Ph.D on the 28th April.
  - ['Mr.', 'Smith', 'finished', 'his', 'Ph.D', 'on', 'the', '28th', 'April', '.']

- + It is intuitive, as it as words are natural linguistic units
- - Requires language-specific rules (ex: Ph.D)
- - Struggles with unknown words, typos, or words not present in the training vocabulary

EXERCISE

- Can you think of tokens you wouldn't want to split based on punctuation?

  - Examples:
    - Sr.
    - 55.5 or 55,5
    - www.google.com
    - FT-34-56
    - …
    - Rock 'n' roll
    - Toys'r us
    - U.S.A

ABOUT WORD-LEVEL TOKENIZATION

- It is possible that we also want to find sequences of words (compounds), that is, sequences of words that have some unified linguistic meaning
- Example:
  - Ice cream

ABOUT WORD-LEVEL TOKENIZATION

- Other scenarios:
  - Chinese:
    - don't have a white space between words
  - EN:
    - Nowherefast???? (best music ever! ☺)
  - Agglutinative languages:
    - Words are sequences of morphemes (such as Turkish)

    (we will talk about this in a next class)

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

TOKENIZATION: SUBWORD-LEVEL

- Subword-level tokenization splits words into smaller meaningful units, such as prefixes, suffixes, or frequent subword patterns
  - Example:
    - Input: unhappiness
    - Output: ['un', 'happi', 'ness']
- + Combines the benefits of character- and word-level tokenization by breaking down OOV words into known subwords
- + It is widely used nowadays
- - Requires more sophisticated algorithms to split text into subwords
- - May break down words in a way that loses meaningful context

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

SENTENCE/WORD'S MANIPULATION

- We call normalization to the preprocessing step that transforms raw text into "standardized" format to reduce noise and linguistic variability (and, thus, data sparseness)
- There are many manipulations we can do to "normalize" text

SENTENCE/WORD'S MANIPULATION

- Remove stop words
  - Stop words (mainly functional words)
    - Examples:
      - a, de, para, …
      - the, a, before, thus,
  - Problem: authorship identification!

There are lists of stopwords available.
Check them before you use them!!!!!!

SENTENCE/WORD'S MANIPULATION

- Remove punctuation
  - This can be ok, but it can also be a problem:
    - Os assassinos de D. Carlos, Afonso Costa e Buiça, foram...
    - Os assassinos de D. Carlos, Afonso Costa e Buiça foram...

    - The assassins of D. Carlos, Afonso Costa and Buiça, were...
    - The assassins of D. Carlos, Afonso Costa and Buiça were...

SENTENCE/WORD'S MANIPULATION

- Lowercasing
  - Avoid data sparseness
    - The dog is nice vs. I like the dog
  - Problem:
    - Us vs. us, Windows vs. windows, Figo vs. figo

SENTENCE/WORD'S MANIPULATION

- Normalization of dates, numbers, names, …
  - Examples:
    - 8th-Feb
    - 8-Feb-2013
    - 02/08/13
    - February 8th, 2013
    - Feb 8th
    - …

SENTENCE/WORD'S MANIPULATION

- Lemmatization: the process of reducing a word to its base or dictionary form (lemma) based on its meaning and context
- Example:
  - running → run
  - studies → study
  - went → go

SENTENCE/WORD'S MANIPULATION

- Stemming: the process of reducing a word to its root form by removing suffixes or prefixes, often without considering meaning
- Example:
  - running → run
  - studies → stud
  - went → went

Overview

- Learning objectives
- Topics
  - Tokenization
    - Concept
    - Character-level
    - Word-level
    - Subword-level
  - Sentence/word manipulation
  - Tips
- Key takeaways
- Suggested readings

TIP

- If you pre-process the training set, pre-process in the exact same manner the test set
- Pre-processing in probably not a good idea if you are using LLMs
- Pre-processing does not guarantee better results (so, sorry)

KEY TAKEWAYS

KEY TAKEWAYS

- Understand that corpora probably needs some pre-processing, although traditional operations, such as lemmatization or stop words removal make no sense nowadays in most cases

SUGGESTED READINGS

READINGS

- Sebenta: chapter 4 (4.1 and 4.2)
- Jurafsky: 2.5, 2.6, …