18-11-2020

NLP Cirriculum
Part 1

⇒ Regular Expressions

- Letters inside square brackets.

  [hH]ardik → Hardik, hardik.

  [1234567890] → Any digit.

- Ranges [A-Z]

  [A-Z] → Any upper case letter

  [a-z] → Any lower case letter

  [0-9] → Any single digit,

- Negation

  ^ means negation only when first in [ ]

  [^A-Z] → Not an upper case letter
  [^Ss] → Neither 'S' nor 's'
  [^e^] → Neithe 'e' nor '^'

- OR |

  [Ii]ndia|[Bb]harat

Before we can do any natural processing of a text, the text has to be normalized.

Normalization processes:

1. Tokenizing (segmenting) words
2. Normalizing word formats
3. Segmenting sentences.

2.4.1 Crude Tokenization and Normalization.

~~In~~ ~~text.~~ Naive approach for word tokenization and normalization

- Every sequence of non-alphabetic characters are changed to ~~be~~ newl

2.4.2 Word Tokenization

- By removing all the non-alphabetic characters, we are loosing a lot of information.

- Instead of neglecting them, we would also want to tokenize → these special characters.

- 'we're' should be converted to 2 tokens 'we' and 'are'

2.4.3 Byte Pair Encoding for Tokenization.

- In this we iteratively merge frequent pairs of characters.

2.4.4 Word Normalization, Lemmatization and Stemming.

- Case folding → Converting all words to lower case.

- lemmatization → Reduce inflections or varient forms to base form.
  - is, am, are → be

- **Morphemes**

The small meaningful units that make up words.

Word → stem + affixes

- **Stemming** :- Reduce terms to their stems.

  eg. Porter Stemmer