7. REGEX USES

RegExs in Natural Language Processing

Stages in Natural Language Processing

There are many stages in Natural Language Processing where RegExes can be used:

- ***Text Cleaning***: This is typically the first step in NLP, where a text file is taken into the NLP system, and unwanted characters are removed. These can include things such as special symbols, control characters, HTML tags, unwanted punctuation, and extra whitespaces. Regular Expressions can used extensively at this stage to remove such content, simply by only allowing content that matches "`[a-zA-Z0-9]*`" to proceed onto the next stage of the process.
- ***Stop Word Removal***: This stage involves the removal of "Stop words", which are words that are so widely used that they carry very little useful information, so, for example, "`a`", "`the`", "`with`", "`is`", "`are`", "`be`" can be put in a list for the RegEx as follows: `StopWords = (a|the|with|is|are|be).`
- ***Tokenization***: As the name suggests, this stage involves breaking the text into smaller units (called "Tokens"), this can include the following:
  - *Word Tokenization*, so for example, the sentence "`ChatGPT is a ChatBot`" becomes [`"ChatGPT","is","a","ChatBot"`].
  - *Character Tokenization*, so for example, the sentence "`ChatGPT is a ChatBot`" becomes [`"C", "h", "a", "t", "G", "P", "T", " ", "i", "s", " ", "a", " ", "C", "h", "a", "t", "B", "o", "t"`].
  - *Subword Tokenization*, so for example, the sentence "`ChatGPT is a ChatBot`" becomes [`"Chat", "GPT", "is", "a", "Chat", "Bot"`].
- ***Pattern Matching***: This stage involves locating specific patterns within the text, for example, identifying dates, times, names, addresses, and other predefined patterns. So, for example, if the sentence "`Jane Smith is a professor of Computer Science`" is scanned to search for names, it should return "`Jane Smith`".
- ***Entity Recognition***: This stage involves locating special patterns within the text, for example, identifying email addresses, phone numbers, bank account numbers, and social security numbers. We have already seen a RegEx for detecting emails: "`[A-Za-z+.]+@[A-Za-z.]+[A-Za-z]+`".
- ***Text Validation***: This stage involves validating input text against some predefined standard, so for example, validating user input like email addresses, passwords, or other structured data. And we have seen a RegEx for validating passwords: "`[a-zA-Z0-9]{9,15}`".
- ***Text Normalization:*** This stage involves standardizing the text by converting different representations of the same information into a single unified format. So, for example, dates may be in the same document in the following formats: "`23rd January 2021`", "`04/11/2019`", "`15-NOV-2002`".

These are just some of the stages that Regular Expressions can be used in NLP.