Text Cleaning

Analytics Vidhya

Learn everything about analytics

About the Module

❑ What is text noise

❑ Reasons for text noise

❑ Steps of text cleaning

What is text noise

- Unwanted or Useless information present in text data
- Examples of noise :
  - Common Entities : Stopwords, URLs, hashtags, punctuations, numbers
  - Slangs : words which are not present in the dictionaries
  - Spelling and Grammar errors
  - Keyword Variations

Reasons for text noise

- Human errors
- Data entry errors
- Less accurate digitization software
- Less accurate machine translation
- Web scraping

What is text cleaning

- Process of removal of text noise in a step by step manner
- Importance of text cleaning
  - Reduce the dimensions of data
  - Simple machine learning models
  - Getting rid of repetitive information
  - Focus on useful entities and information

Steps for text cleaning

- Fix the decoding of text
    - Text data obtained from web is often present in a different encoding
    - Can be converted into a standard encoding : utf8
- Escaping HTML Characters:
    - Examples : &nbsp; &lt; &gt; &amp;
    - Remove using Regular Expressions, List iteration, appropriate libraries

Steps for text cleaning

- Apostrophe Lookup:
    - Variations with the words
    - Example : it’s -> it is or it has
    - Solve using Regular expressions or lookup table
- Removal of stopwords:
    - Commonly used words : "a", "an", "the", "is", "am", "are", "of" etc.
    - Less informative about the context
    - Remove using appropriate libraries, list iteration

Steps for text cleaning

- Noisy Entities :
  - URLs, Hashtag words, Mention Words
  - Removal using regular expressions
- Punctuations:
  - All symbols : emoticons, regular punctuations, connectors
  - (I, went to New-York !!)
    I went to New York

Steps for text cleaning

- Keyword Normalization
  - Stemming or Lemmatization
  - Appropriate library
- Spelling Correction
- Grammar Correction