* Text Preprocessing

1. LowerCasing
   → Generates simplicity in data.
   → Avoids Confusion.

2. Remove HTML Tags [Unimportant data]
   → There is no need for html tags as ML model can't understand HTML Tags
   → Use Python Regular expression for removing unwanted data.

3. Remove URLs
   → Obtain mostly for Social media data.
   → Remove URL using RegX expressions
      i) https
      ii) http
      iii) www

4. Remove Punctuations

i) Punctuations work as a token during tokenization.

ii) ! " " # $ % & \ ' ( ) % + , - / . : ; < = > ? @ [\ ] ^ _ ` { | } ~'

→ Hence need to remove punctuations.

'Hello! How are you ?'

Hello, !, how, are, you, ? — Consider as Token

Hello!, How, are, you? — Consider Combine Token.

5. Chat Word Treatment

→ Using Shorthand. e.g lmao, Lol, GN, ASAP

6. Spelling Correction

Please read the notebook, and also like the ntebook

↳ Same word act as a different word after tokenization

7. Removing Stop Words

→ Sentence used Words used for sentence formation but has no meaning in itself.

→ Not useful in Sentiment analysis or Document classification.

→ Useful in Part of Speech

→ e.g at, the, a, an etc..

8. Handling emojis

1. Remove

2. Replace

9. Tokenization

Process of breaking text document into smaller parts known as tokens.
Tokens can be → word, Sentence, phrases

[I am an indian → [I, am, an, Indian] [based on words]
I love my country. Hi There → [I love my Country] [Hi There]

problem with Tokenization
- Prefix → Characters at the beginning
- Sufix → Characters at the end
- Infix → Characters in between
- Exception → Special-Case rule to split a string into several tokens or prevent token from being split when punctuation rules are applied.

10. Stemming:

inflection → Inflection is the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender and mood.

Stemming → The process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the language.
→ Used in information Retrieval Systems.
e.g Walking → Walk
Dancing → Dance
Stemmer is a algorithm with which performs stemming for given corpus.
Porter Stemmer → English language stemmer.
Snowball Stemmer → For other language stemming.

* 11. Lemmatization:

Lemmatization, unlike Stemming reduces the inflected words properly ensuring that the root word belongs to the language. In lemmatization root word is called Lemma. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form or citation form of a set of words.

Wordnet Lemma → Use for lemmatization
→ It is Comparatively slow since it searches for Dictionary